nature electronics

PERSPECTIVE

https://doi.org/10.1038/s41928-021-00646-1

Neuromorphic electronics based on copying and pasting the brain

Donhee Ham [1,2] ✉, Hongkun Park [3,4] ✉, Sungwoo Hwang [2,5] ✉ and Kinam Kim [5] ✉

Reverse engineering the brain by mimicking the structure and function of neuronal networks on a silicon integrated circuit was the original goal of neuromorphic engineering, but remains a distant prospect. The focus of neuromorphic engineering has thus been relaxed from rigorous brain mimicry to designs inspired by qualitative features of the brain, including event-driven signalling and in-memory information processing. Here we examine current approaches to neuromorphic engineering and provide a vision that returns neuromorphic electronics to its original goal of reverse engineering the brain. The essence of this vision is to 'copy' the functional synaptic connectivity map of a mammalian neuronal network using advanced neuroscience tools and then 'paste' this map onto a high-density three-dimensional network of solid-state memories. Our copy-and-paste approach could potentially lead to silicon integrated circuits that better approximate computing traits of the brain, including low power, facile learning, adaptation, and even autonomy and cognition.

Neuromorphic engineering began in the 1980s with the aim of using analogue integrated circuits to mimic the structure and function of neuronal networks in biological nervous systems[1,2]. The ultimate goal was to bring the remarkable computing abilities of the brain to a solid-state platform. However, rigorous mimicry of the brain's neuronal network has proved difficult, because we still do not know today how a large number of neurons wire inside the brain to create higher functions. As a result, the aims of neuromorphic engineering have been eased to include the development of designs inspired by qualitative features of the brain, including asynchronous, event-driven operation[3–6] and in-memory information processing[6–17]. The analogue design requirement has also been relaxed to mixed-signal design, leading to the creation of a range of sophisticated analogue and digital circuits[1–22].

In this Perspective, we explore the possibilities and limitations of these current approaches to neuromorphic engineering, and then provide a vision for neuromorphic electronics that returns the field to its original goal—reverse engineering the brain—through a combination of advanced neuroscience tools and state-of-the-art memory technology. The essence of our approach is to copy the functional synaptic connectivity map of a mammalian neuronal network using an intracellular neuro-electronic interface[23] and to paste this map to silicon integrated circuits including a high-density three-dimensional (3D) network of memories[24]. We also consider the key challenges involved in using this copy-and-paste strategy to develop silicon integrated circuits that can approximate the computing abilities—and ultimately the intelligence—of the brain.

Contemporary neuromorphic approaches

Current neuromorphic electronics generally fall into two categories: approaches motivated by artificial neural networks (ANNs) and approaches motivated by the brain's natural neuronal network (NNN) (Fig. 1). ANNs are the framework of machine learning and have led to a range of powerful artificial intelligence (AI) applications[25], and are of particular use in feature classification from big data. The ANN demands brute and precision calculations, and is thus best realized digitally, such as with central, graphics or neural processing units[26]. NNNs are the basis of natural intelligence and are powered by electrochemical reactions. They excel at different tasks to ANNs: they can learn easily from few or poorly conditioned data, can adapt to environments, are autonomous and are capable of cognition. These differences suggest that the organizing principles of the NNN, of which we still know little, are profoundly different from those of the ANN.

The NNN and neuromorphic engineering. Neuromorphic engineering was originally aimed at creating analogue integrated circuits based on the NNN organizing principle to emulate the brain's remarkable functions[1,2] (Fig. 1, Analogue mimicry). A family of silicon neuron circuits was developed to imitate the firing of action potentials (APs)—spikes—in biological neurons, with varying degrees of abstraction of the ion channel dynamics that underlie the AP firing[18,19]. These spiking silicon neurons were then connected through silicon synapse circuits, with synaptic integration also modelled in such connections[19]. Various AP firing patterns[20] and signal processing models[21] were obtained by networking silicon neurons and synapses.

Key advances were made in emulating the sensory peripherals of the brain, and in particular the retina. The structure and function of the biological retina's NNN were modelled, at least partially, on a silicon chip to approximate early visual processing[3,22]. Neuromorphic vision was further advanced to the event camera[3] that combines the silicon retina and spiking silicon neurons (in practice, there are different styles of realization[4,5]). Each pixel responds with a spike only when an event occurs: for example, when light intensity to the pixel changes due to motion. Each pixel thus outputs an asynchronous stream of spikes. Contrasting the standard frame-based, clock-driven complementary metal–oxide–semiconductor (CMOS) image sensor, this asynchronous event-driven vision can track motions with a time resolution of microseconds, and can be useful for applications in autopilot systems and robotics. It is now typically realized in mixed-signal mode, which

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [2]Samsung Advanced Institute of Technology, Samsung Electronics, Suwon, Republic of Korea. [3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [4]Department of Physics, Harvard University, Cambridge, MA, USA. [5]Samsung Electronics, Hwaseong, Republic of Korea. ✉e-mail: donhee@seas.harvard.edu; Hongkun_Park@harvard.edu; swnano.hwang@samsung.com; kn_kim@samsung.com

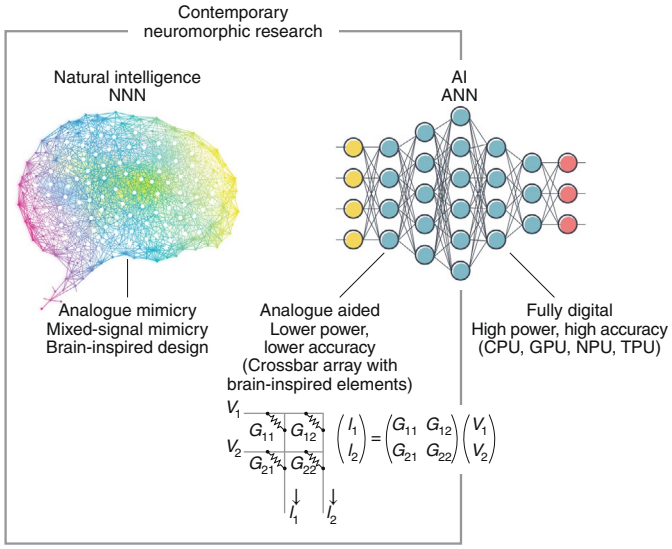


Fig. 1 | Contemporary neuromorphic research. Current neuromorphic electronics research may be categorized into the effort motivated by the NNN and that by the ANN. CPU, central processing unit; GPU, graphics processing unit; NPU, neural processing unit; TPU, tensor processing unit; *V*, voltage; *I*, current; *G*, memory conductance.

highlights the fact that neuromorphic electronics is no longer restricted to the analogue domain[3–5]. These advances in neuromorphic vision were possible due to the wealth of knowledge on the retinal NNN, while its embodiment of the asynchronous, event-driven operation is inspired by the NNN (that is, biological neurons respond only when necessary).

In contrast, because little is known about how neurons wire inside the brain (the cortex, for example) to create higher function, building a circuit that exhibits the unique computing abilities—and ultimately autonomy and cognition—of the NNN is fundamentally challenged, and a silicon chip that offers brain-like intelligence remains a distant prospect. Asynchronous, event-driven operation is an insightful clue to the NNN and can enable bio-inspired designs, but it cannot instruct us on how to wire a massive number of silicon neurons into a system that can compute similarly to the brain. Reproducing higher brain functions requires the NNN's functional wiring diagram, or functional connectivity map.

The ANN and neuromorphic engineering. The past decade has seen a sharp resurgence in neuromorphic engineering. This is driven by the AI boom, and represents another style of neuromorphic electronics trying to build an analogue-aided ANN processor (Fig. 1, Analogue aided) that consumes far less power in AI computing than fully digital ANN processors (Fig. 1, Fully digital)[7–17]. While this approach does not aim to reproduce the unique traits of natural intelligence via rigorous NNN mimicry, it is still regarded by many as neuromorphic electronics because its design is inspired by the in-memory computing attribute of the brain.

The backbone of this analogue-aided ANN is a crossbar array of conductive memories (resistive memories, for example) that performs multiply–accumulate operations—the most prevalent ANN algebra—in a physical manner[7–17]. Each memory stores an ANN optimization parameter (weight) as its conductance value. Input voltages fed to the rows of the array are multiplied by the weights via Ohm's law, and the resulting currents are accumulated in each column by Kirchhoff's law. This physical, and thus analogue, multiply–accumulate operation burns far less power than its digital counterpart. Here the colocalization of memory and computing in the crossbar array, which breaks away from the von Neumann paradigm, is inspired by the brain, where memory elements (biological synapses) are distributed across the network. The goal here, however, is to calculate the ANN algorithm, not to mimic the NNN to create the unique functions of the brain.

The considerable power reduction possible with the analogue-aided ANN has generated substantial interest[7–17]. This enthusiasm, however, must be balanced with the fact that the analogue-aided ANN borrows error-bound analogue methods to solve the precision ANN algorithm, trading accuracy for power savings. Consequently, the technology can be powerful for applications that demand low power but can afford reduced accuracy. A case in point is the always-on wake-up sensors in edge devices, which need not resolve fine features but must consume as little energy as possible. The technology thus has the potential to penetrate the large AI sensor market.

Other bio-inspired features beyond in-memory computing, such as asynchronous multiply–accumulate operations in the event- or data-driven manner[6] and unsupervised learning[27,28], have also been brought to the crossbar array. Nevertheless, the goal has remained the ANN calculation for AI applications, and not the rigorous mimicry of the NNN structure and function to emulate the unique traits of natural intelligence.

Copying the NNN

The original neuromorphic pursuit to mimic the NNN has been limited by the lack of the NNN's functional wiring diagram. As a result, the focus of the field has evolved from rigorous brain mimicry to design inspired by qualitative features of the brain such as asynchronous, event-driven operation and in-memory computing. Our aim is to turn back to the original idea of brain mimicry by leveraging recent advances in neuroscience tools, in particular, a silicon neuro-electronic interface[23] called the CMOS nanoelectrode array (CNEA). The CNEA can 'copy' the NNN's functional synaptic connectivity map (Fig. 2) through its massively parallel intracellular electrophysiological recording. Over the past decades, non-electrophysiological methods[29–31], such as optical and electron microscopy, genetically encoded indicators and a host of other experimental methods, have also made spectacular progress in deciphering anatomical and functional connections in the NNN. For instance, electron microscopy was used to obtain the anatomical map of the full nervous system of *Caenorhabditis elegans* in the mid-1980s[32] and has since been applied to various animal brains, culminating in the recent anatomical mapping of the complete *Drosophila* brain[33]. We, however, will focus mostly on the electrophysiological method, because it is a natural fit to the solid-state memory network to which the copied biological neuronal networks will be 'pasted'.

Parallelization of intracellular recording has been an important pursuit in neuroscience, because it would enable the functional synaptic connectivity mapping of the NNN[34–37]. The Nobel Prize-winning patch clamp electrode revolutionized neurobiology with its highly sensitive intracellular recording: it can measure not only APs but also subthreshold signals such as postsynaptic potentials (PSPs), and thus can find a synapse and measure its connection strength. However, because the bulky patch clamp cannot scale to a dense array, parallel patch recording has been limited to only about ten neurons[38], making it difficult to map a network-wide synaptic connectivity. Conversely, the microelectrode array records from many more neurons to monitor a network, but this extracellular method is not sensitive enough to record PSPs, making it difficult to study synaptic connections[39,40]. The CNEA—the latest version of which integrates 4,096 electronic channels in a CMOS chip with 4,096 vertical nanoelectrodes (Fig. 2a)—joins intracellular and parallel recording[23], and thus it can map the NNN's functional synaptic connectivity (Fig. 2b).

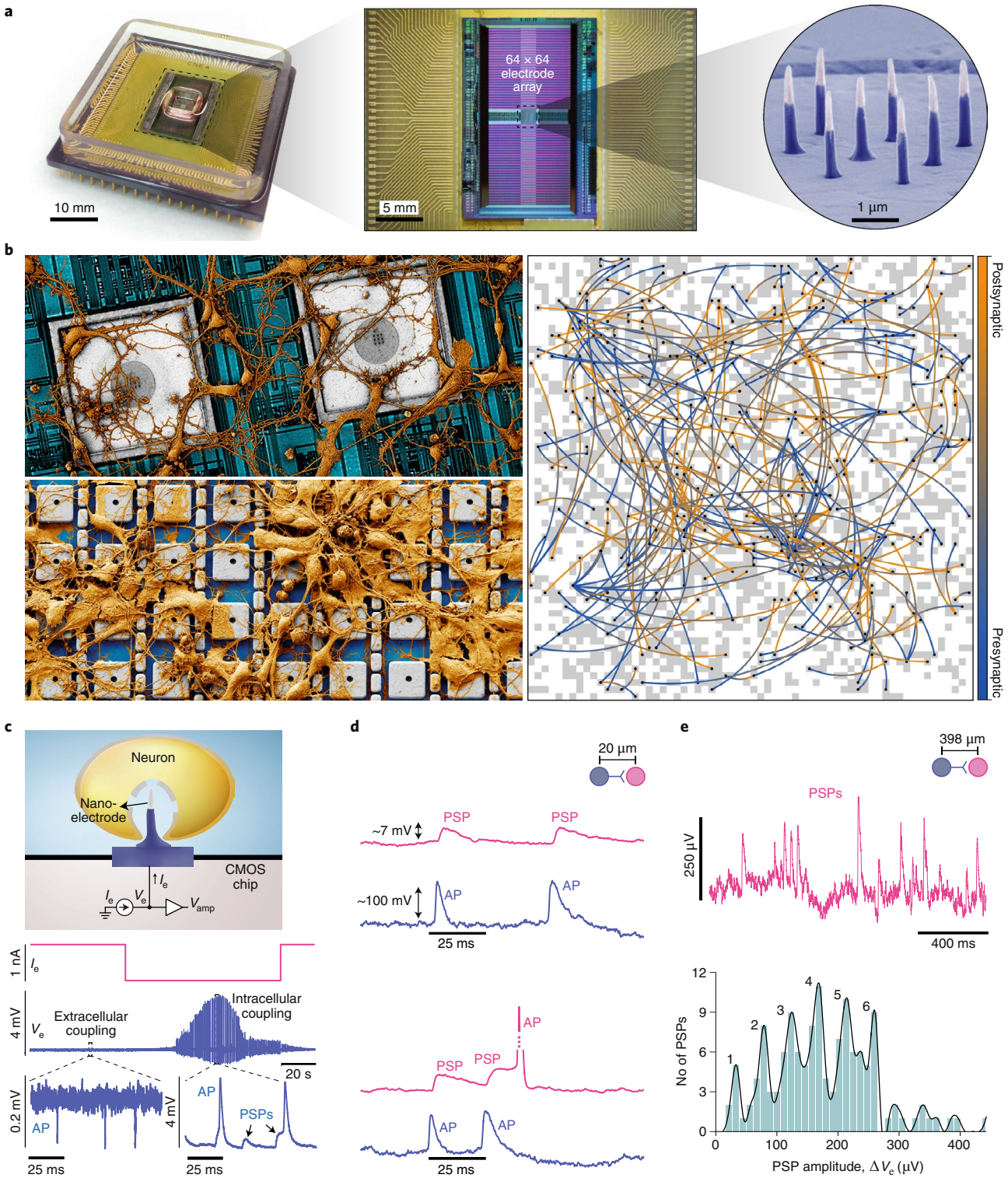


Fig. 2 | Copying the NNN. a, A CNEA[23]. **b**, Rat neurons on CNEAs (left) and synaptic connectivity map (right) extracted from network-wide intracellular recording data obtained by the CNEA. **c**, Intracellular recording at a CNEA channel is enabled by injecting a current I_e into the electrode and by concurrently recording the electrode voltage V_e, which is a scaled version of the membrane potential. Each recorded voltage shown here and in **d** and **e** is V_e. V_e serves as an input to the amplifier, whose output voltage is V_{amp}. **d**, Intracellular recording from a pair of connected neurons. APs of the presynaptic neuron and PSPs of the postsynaptic neuron are time correlated. When APs fire rapidly in the first neuron, synaptic integration occurs in the second neuron (bottom). **e**, In another pair, a histogram of 149 PSP amplitudes of the postsynaptic neuron shows quantization. Panels adapted with permission from: **a**, refs. [23,42], Springer Nature Ltd; **b**–**e**, ref. [23], Springer Nature Ltd.

Machinery of the CNEA. Each channel of the CNEA features a vertical nanoelectrode on the surface and can be configured to have a current injector and a voltage amplifier in the underlying CMOS chip (Fig. 2c)[23,41]. A neuron can wrap around the nanoelectrode. A current injection by the underlying electronics then permeabilizes the membrane around the electrode, giving the electrode intracellular

access (Fig. 2c). After this intracellular access is achieved, the current injection is sustained to compensate permeabilization-caused leakage from the neuron, stabilizing the cell's electrophysiology. Concurrently, the voltage amplifier measures the membrane potential. With this current injection with simultaneous voltage recording, or current clamp, the CNEA channel achieves a robust intracellular recording of not only APs but also PSPs on a routine basis.

Figure 2c shows an example recording of a rat cortical neuron at a CNEA channel. With no current injection, small and noisy extracellular signals register. With the current injection that causes intracellular access, the measured signal increases markedly, enabling definitive measurements of PSPs, as in the patch clamp (before this CNEA, intracellular access into a neuron using a nanoelectrode was difficult[42], and even when successful[43] PSPs could not be measured, due to the lack of integrated current-clamp electronics). What distinguishes the CNEA from the patch clamp is the scalability: the CNEA parallelizes the high-fidelity intracellular recording with its dense channels and can perform network-wide intracellular recording. For example, during a 19 min recording from a network of rat cortical neurons cultured on top in vitro, the 4,096-channel CNEA measured intracellular signals from 1,728 electrodes[23], a substantial leap from roughly ten patch recordings. This number can easily be further increased, as scalability—making a larger, denser CNEA—is the essence of CMOS technology.

Copying the functional synaptic connectivity map. In a pair of neurons connected by a chemical synapse, an AP from the presynaptic neuron elicits a time-correlated PSP in the postsynaptic neuron. Synaptic connections can thus be found by identifying time-correlated APs and PSPs in the network-wide intracellular recording data. Figure 2d shows an example neuronal pair thus found. This connection is further confirmed by the synaptic integration from the same data: when the presynaptic neuron fires a rapid sequence of APs, the resulting PSPs in the postsynaptic neuron summate to exceed the threshold, resulting in an AP in the postsynaptic neuron (Fig. 2d, bottom). By searching AP–PSP correlations in the 1,728 intracellular signals from the 19 min recording, we mapped 304 excitatory and inhibitory synaptic connections[23] (Fig. 2b), a scale and throughput unimagined with the patch clamp. The power of the intracellular recording to measure PSPs can be appreciated from the fact that the correlation analysis of presynaptic APs and postsynaptic APs (in lieu of PSPs) from the same dataset uncovers only 63 connections. This is because synaptic integration in a weak or inhibitory connection may not exceed the threshold, with the postsynaptic neuron not firing APs.

Not only can the CNEA's intracellular recording distinguish excitatory and inhibitory PSPs, but also it can resolve the PSP amplitude down to a unit quantum (Fig. 2e)[23]: the PSP quantization is a hallmark of the chemical synapse, where neurotransmitters are released in a discrete number of vesicles. This high-resolution PSP amplitude measurement, previously only possible using the patch clamp, can assign the strength to each synaptic connection, informing not only anatomy but also function.

In sum, we can extract the functional synaptic connectivity map from the network-wide intracellular recording data. The network intracellular recording can then be considered as the process of copying the functional synaptic connectivity map. The recording also provides other critical information, such as propagation delays in neuronal axons, feedback routings through neurons and ion channel properties in neuronal membranes. The last information on this list is obtained by engaging the CNEA recording in the voltage-clamp mode (voltage application with simultaneous current recording)[23].

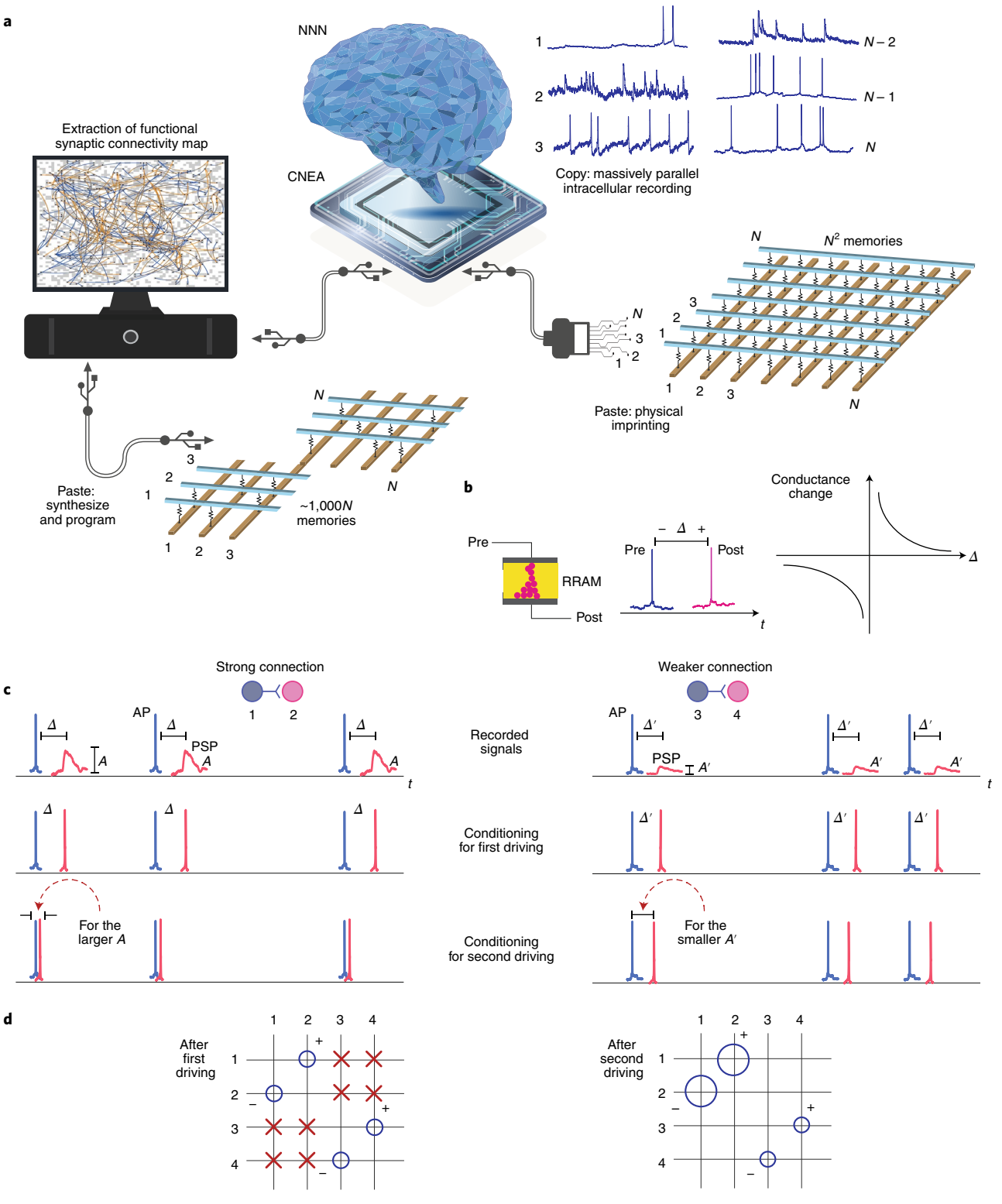
Working with a behavioural neuronal network. The current version of CNEA determined a synaptic connectivity map from in vitro cultured neurons. These neurons are live, connected and firing, but their random network does not give biological behaviour. The advance we have made thus far is hence a relatively small step towards the ultimate goal of copying neuronal networks from within the brain. Towards this goal, the nanoelectrodes should take different form factors to access 3D neuronal networks found in the brain, such as nanoelectrodes with long pillar geometry or placed at each pixel on a macroscopic shank[44]. To start, we are interrogating the neurons' functional connectivity in mouse retina and olfactory bulb/piriform cortex. These are attractive initial targets as they have clear functions, their inputs can be easily controlled and there are well-established protocols for preparing ex vivo samples. At the same time, they are different in terms of their neuronal organizations: in the retina, cells are arranged into regular lamina[45]; in the olfactory bulb/piriform cortex, such neat organization is absent[46]. As such, these studies will serve as a steppingstone to applying the CNEA to more complicated neuronal circuits in other brain regions and, ultimately, to probe cortex regions.

Fig. 3 | Pasting. a, A functional synaptic connectivity map extracted by the computer-aided analysis of N intracellular recordings can be used to synthesize a memory network (left) with ~1,000N memories. Alternatively, we can drive an $N \times N$ memory crossbar array with N^2 cross-point memories directly with the recorded signals for physical imprinting of the functional connectivity map (right). **b**, STDP. Δ, time delay between the two signals driving the two terminals of the RRAM; t, time. **c**,**d**, Illustration of STDP-based physical imprinting of the functional synaptic connectivity map, using an NNN toy model with four biological neurons and a 4 × 4 RRAM crossbar array. Neuron 1 (N1) synapses N2 strongly and N3 synapses N4 weakly, so in their recording (**c**, top) APs in N1 cause strong PSPs in N2, and APs in N3 induce weak PSPs in N4. We perform two rounds of driving of the crossbar array. The first round is to impress connections, but not their strengths. We precondition the recorded signals by converting PSPs into spikes at the same time positions (**c**, middle), which can be done by an analogue circuit, and feed these signals to both columns and rows of the crossbar array. As APs of N1 and PSPs-turned-spikes of N2 are time correlated, the RRAM at the intersection between the N1 row and N2 column, or at (1, 2), will increase its conductance (the amount of increase does not matter in this round) due to STDP (**d**, left). The same holds true for the RRAM at (3, 4) (**d**, left). Because STDP is sensitive to the sign of the time delay, RRAMs at (2, 1) and (4, 3) will decrease their conductance (**d**, left). At all other cross points driven by signals with no time correlations, STDP averages out the driving to cause no conductance change (**d**, left). Together, these conductance changes across the array (**d**, left) indicate that N1 synapses N2 and N3 synapses N4. The second round, performed after the memory reset, is to impress the strength on each connection identified in the first round. We precondition the recorded signals in such a way that the larger PSP of N2 is again converted to a spike but with a time shift to have a shorter time delay from the AP of N1 (we can do this because we know the N1–N2 connection from the first round; this amplitude-to-time conversion can be done with an analogue circuit), and the smaller PSP of N4 is converted to a spike with another time shift to have a longer time delay from the AP of N3 (**c**, bottom). As these signals drive the array, the conductance values of the RRAMs at (1, 2) and (3, 4) will increase, with the former being larger, due to STDP (**d**, right), thus correctly imprinting the connection strengths. RRAMs at (2, 1) and (4, 3) show symmetric behaviours. We ignore all the other cross-points with no synaptic connections. These two rounds conclude the physical imprinting of the functional synaptic connectivity map. Δ, time delay between each AP in N1 and the resulting PSP in N2; Δ', time delay between each AP in N3 and the resulting PSP in N4; A, voltage amplitude of the PSPs in N2; A', voltage amplitude of the PSPs in N4. Panel **a** adapted with permission from ref. [23], Springer Nature Ltd.

Pasting

The functional synaptic connectivity map extracted from the intracellular recording of N neurons can be pasted to a network of conductive memories (Fig. 3a), with each memory storing a conductance value that represents the strength of a corresponding biological synaptic connection. This memory network can then be weaved together with silicon neurons to reflect propagation delays, feedback routings and ion channels, which are also extracted from the recording data.

Memory candidates. While dynamic random access memory (DRAM) and flash memory are the pillars of memory technology[24], the industry has not stopped searching for 'new memories'—new in potential applications, not in concept—to complement fast but volatile DRAM and non-volatile but slow flash. Spin-transfer torque (STT) magnetic random access memory (MRAM)[47], phase change random access memory (PRAM)[48] and resistive random access memory (RRAM)[7,13,49–52] are promising examples. Of these commercial and new memories, flash, MRAM, PRAM and



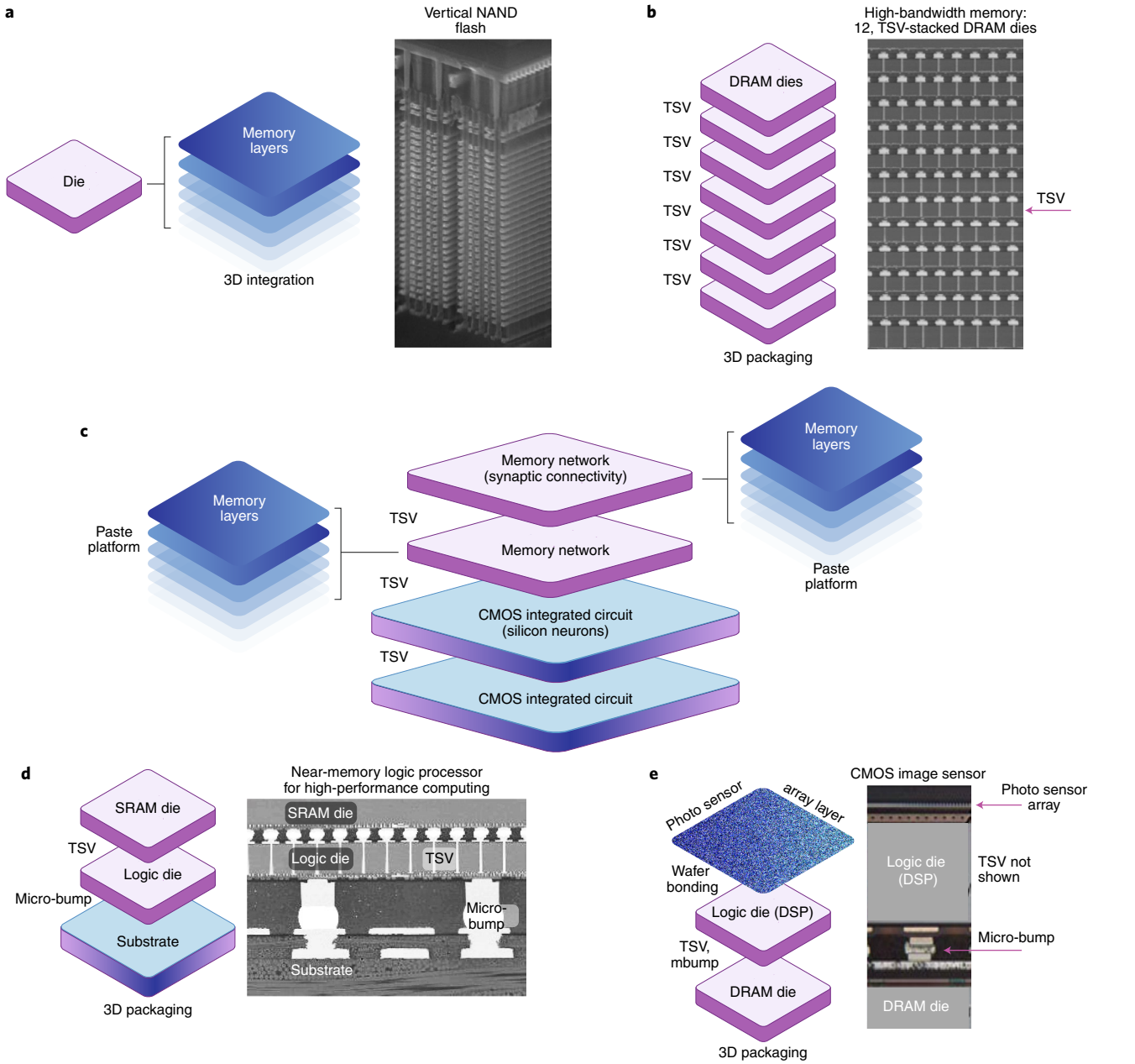


Fig. 4 | Neuromorphic scaling using 3D integration and packaging technology. a, 3D-integrated vertical NAND flash memory. **b**, High-bandwidth memory obtained by vertically stacking 12 DRAM dies using TSVs. **c**, A memory network representing a functional synaptic connectivity map can be fabricated into multiple layers within a given die using 3D integration, and multiple such dies can be vertically packaged using TSVs. These memory dies can be heterogeneously TSV-stacked with CMOS integrated circuits incorporating silicon neurons to model signal propagations, feedback routings and ion channels. **d**, Heterogeneous TSV stack of a CMOS logic and an SRAM for high-performance computing. **e**, Heterogeneous stacking in a CMOS image sensor: a photosensor array layer, a logic and a DRAM are stacked via wafer bonding, micro-bump bonding and TSV (not shown). DSP, digital signal processor. Credit: TechInsights (Ray Fontaine, http://rfontaine@techinsights.com; e (right))

RRAM—which are non-volatile, conductive memories and can be embedded in CMOS processes with read–write electronics—are well suited for the paste platform. Not all of them are yet ready for widespread adoption due to reliability issues, but the vision we present here is also for the future. Each of the four memory types has its own strengths.

Flash memory is a field-effect transistor with a floating gate that can trap different amounts of charge to give a range of channel conductance values. It is an attractive paste platform due to its ultrahigh-density 3D integration[53] and multi-level-cell (MLC) operation that can approximate the real-valued connection strength. It has a limited endurance and requires a high writing voltage due to moving electrons to and from the floating gate through a dielectric.

The STT-MRAM[47] consists of two ferromagnetic films separated by a thin oxide and is manipulated by a current to align or anti-align the magnetic moments of the two magnetic layers, creating low- and high-resistance states. It is thus a one-bit cell. Regarded as the possible alternative to DRAM in speed and endurance, this technology has advanced notably in recent years. For the paste work, several one-bit cells must combine to create an effective MLC, compromising the area density.

In the PRAM[48], a chalcogenide layer is electrically switched between a conductive crystalline phase and a resistive amorphous phase. It is also a mature technology: Samsung volume-produced 512 Mb PRAMs around a decade ago, supplanting NOR flash for mobile devices. Capable of attaining multiple intermediary resistance states, the PRAM can be an MLC, although resistance drift is a practical obstacle for MLC operation.

In the RRAM[7,13,49–52], a metal oxide or solid electrolytic medium lies between electrodes. Movement of ions (oxygen vacancies or metal ions, for example) through the medium causes a resistive switching, with MLC operation possible if properly designed. Due to its simple and small structure, it has attracted a great renewed interest, although device variability currently precludes widespread adoption.

Remarkably, RRAM or PRAM can learn the time correlation of the signals at its two terminals: when the two signals have a stronger correlation (a shorter time delay), its conductance changes more (Fig. 3b)[28,51,52,54–57]. This spike-timing-dependent plasticity (STDP), which is also sensitive to the sign of the time delay (Fig. 3b), can be exploited for the paste task, as seen later. Flash can also learn the time delay of two signals if both signals drive the same transistor gate through a system that converts their time delay into a gate charging time. STT-MRAM is less suitable for plasticity engineering due to its one-bit nature.

Computer-aided map extraction and memory programming. The pasting process can start with the extraction of a functional synaptic connectivity map via computer-aided analysis of AP–PSP correlations and PSP amplitudes from the intracellular recording dataset. This map can then be used to fabricate and program a network of memories where each memory represents the extracted strength of its corresponding biological synaptic connection: since one neuron has about ~1,000 synaptic connections in the brain, the map will have ~1,000N synaptic connections and the network will thus feature ~1,000N memories (Fig. 3a, left). From the semiconductor fabrication point of view, the key consideration will be the reduction of the production cost by designing a memory array platform that can be flexibly programmed to represent a variety of synaptic connectivity maps extracted from different NNNs. The memory programming itself can be done with the electronics integrated in the same chip that can rapidly write and verify the network of memories. In the 3D flash with three-bit cells, for example, write and verification is done at a rate exceeding 100 MB s^{-1}: hundreds of millions of connections can be programmed in 1 s. The main challenge of this paste approach is the computer-aided analysis due to the sheer volume and complexity of the recording data: even the 19 min recording by the 4,096-channel CNEA produces ~80 GB of data, which will rapidly rise with further scaling of the CNEA. The extraction of the functional connectivity map from the big data can benefit from new approaches such as machine-learning-based pattern recognition and crowd sourcing[58]. Despite this challenge, it offers an exciting opportunity, because such large-scale intracellular recording was previously unavailable.

Physical imprinting. Alternatively, we can bypass the computer analysis and directly imprint the connectivity map onto an RRAM or PRAM network by driving it with the N recorded signals (Fig. 3a, right). This exploits the STDP. Since we do not know the connectivity map a priori in this approach, we can use an $N \times N$ crossbar array as the memory network, where every silicon neuron (with one-to-one correspondence to a biological neuron) can be connected to every other silicon neuron with N^2 cross-point memories serving as artificial synaptic connections. If we drive both N rows and N columns of the crossbar array with the N neuronal recordings, the memory at a cross-point of a presynaptic neuron and a postsynaptic neuron will be strengthened due to the STDP, because of the time correlation of the presynaptic APs and the (preamplified) postsynaptic PSPs. In this way, the crossbar array will physically learn the connectivity map of the N neurons. A variant of this strategy can also make the crossbar array learn the strength of each identified connection, completing the imprinting of the functional connectivity map. Figure 3c,d details this STDP-based imprinting with an NNN toy model with four biological neurons and a 4×4 RRAM crossbar array. As mentioned earlier, the flash memory too can learn time correlations, so it can be used for physical imprinting in a varied form of crossbar array compatible with its three-terminal operation.

This physical imprinting is an elegant and powerful strategy to download the biological synaptic organization to the memory platform, but it also entails practical challenges. First, given only ~1,000N actual synaptic connections, most of the N^2 cross-point memories will be left unused. Second, RRAMs still suffer device-to-device variability due to the stochasticity of the ionic channel formation, and PRAMs the resistance drift, rendering their use as MLCs across a large network non-trivial. Also, their STDP studies have so far been largely focused on single devices rather than networks. It is encouraging, however, that a major advance has been made in using these memories in networks for AI computing[7–16]: although its goal differs from NNN mimicry, it shows improvement in engineering these memories for network usage. The mature flash memory may suffer less from the reliability issue when used for the physical imprinting. Third, even with an ideal memory, the imprinting process can be challenging due to the complex signal traffic caused by multi-input/multi-output connections and feedback routings among a large number of neurons. Tackling this problem alone could be a substantial research opportunity: one possibility is to mitigate the traffic during the recording via controlled excitation of the NNN[23] to prepare more straightforward data for the paste task.

Neuromorphic scaling with 3D technology. The network with ~1,000N memories pasted from the computer-extracted functional synaptic connectivity map (Fig. 3a, left) would occupy a prohibitively large area of ~30×30 cm^2 for $N \approx 100$ billion estimated for the human brain and a 30×30 nm^2 memory cell. This can be tackled using 3D integration, the technology that opened a new era for the memory industry[24]. For example, a 128-layer 3D integration will reduce the area to ~26×26 mm^2, a footprint feasible in the advanced technology node. The crossbar array used for the physical imprinting (Fig. 3a, right) presents a stiffer scaling challenge, since it has N^2 memories for ~1,000N actual synaptic connections; the chip area, even after a 128-layer 3D integration, will be ~260×260 m^2. The crossbar array can certainly implement a smaller network (for $N \approx 10$ million, a 128-layer integration yields a feasible ~26×26 mm^2), which still is a useful step. For a larger network, the physical imprinting platform must differ from the crossbar array. Designing such a platform would require some a priori knowledge of the map, hence mandating assistance from computer analysis of recording data to a certain degree.

To further appreciate the 3D technology, consider flash memory. In the face of the increasingly difficult scaling due to the lithographic limitation and the device physics at the reduced dimensions, advances have been made by using the third dimension: Samsung drove the 3D NAND flash production[53] (Fig. 4a), first vertically stringing 32 memory cells and now 128. This 3D integration has markedly increased the memory density and bulk data transfer speed, making the vertical NAND flash an enabler of the big data age. Another 3D revolution is found in DRAM in the form of 3D packaging using through-silicon vias (TSVs) that make shortest-path connections between silicon dies[24,59]. Twelve DRAM dies have been stacked (Fig. 4b), increasing the memory density to approximately gigabytes per square millimetre and the data transfer rate to hundreds of gigabytes per second, enabling high-end

graphics, computing and servers: TSV sizes are approaching the micrometre range, and thousands or more finely pitched TSVs are interconnected per die. If the 3D paste platform employs flash memory to represent the functional synaptic connectivity, the vertical NAND already approximates the solution. RRAMs and PRAMs can also be fabricated into multiple layers within a given die using 3D integration. Multiple such 3D-integrated dies can then be vertically stacked using TSV packaging to further boost the memory density (Fig. 4c). Due to its particular cell structure and fabrication rules, it might be more challenging to arrange the STT-MRAMs into a 3D stack, but this possibility merits further study.

The final neuromorphic system will combine the memory network representing the functional synaptic connectivity map with N spiking silicon neurons fabricated in the CMOS logic process: the latter is to model[18,19] individual neurons with their ion channels, signal propagations along axons and feedback routings, all extracted from the electrophysiological recording. As flash, MRAM, PRAM and RRAM can all be embedded in the CMOS logic process, integrating the two functional parts in the same chip is possible. However, the optimum process for CMOS circuits is generally different from that for memory. Thus, to optimize each function, we can implement the synaptic connectivity memory network in one chip (together with the read–write electronics), and the spiking silicon neurons in another chip. The two chips can then be vertically interconnected using TSVs to minimize footprint and latency (Fig. 4c). Such heterogeneous stacking was indeed a motivation to develop TSVs and has proven fruitful, as exemplified by the stacking of a CMOS logic and a static random access memory (SRAM) for high-performance computing (Fig. 4d)[59]. Another example of this heterogeneous stacking is the CMOS image sensor: it stacks a photodiode sensor array, a logic and a DRAM die via wafer bonding, micro-bump bonding and TSV packaging (Fig. 4e). These examples illustrate the potential for a variety of vertical combinations of multiple CMOS chips (spiking silicon neurons) and memory chips (synaptic connectivity).

From the circuit design viewpoint, the network with either 1,000N (Fig. 3a, left) or N^2 (Fig. 3a, right) memories may have to be broken down into multiple blocks. Otherwise, the voltage drop of interconnects and the fanout issue can be considerable, although analogue spiking silicon neurons may alleviate the fanout issue somewhat. In either network, the cointegrated read–write electronics can readily access an arbitrary memory cell by integrating a transistor switch selector for each cell.

Biological realism

The idea that the NNN stores information in its connection strength patterns and that this connection dictates the network dynamics for computing is a fundamental assumption of neuroscience[60] and provides the basis of both the original neuromorphic pursuit[1,2] and our vision to build an electronic brain by reverse engineering the functional connectivity of neurons. Connectomics[33,61–64], a path-breaking effort that seeks a dense reconstruction of the neuronal wiring of the brain, is also motivated by the same assumption[65,66]. A dominant method for connectomics is electron microscopy, which visualizes neuronal connections of serial brain slices at the synaptic resolution. The primary information obtained by the microscopy studies, at least so far, is an anatomical map, not a functional map, with no connection strengths quantified (some excitatory and inhibitory synapses in the mammalian brain may be distinguished due to differing morphologies). A given anatomical map can assume a variety of different connection strength patterns, thus producing different dynamics. Our copy-and-paste approach aims to reproduce the functional connectivity map, not just the anatomical map, specifying the connection strength patterns. Moreover, this download strategy also includes other neuronal attributes such as ion channels, feedback routings and delays. This functional connectivity map would better narrow down the scope of network dynamics, ideally with one-to-one correspondence, allowing for neuromorphic system development without having to uncover how dynamics are encoded in the map.

The pasted network is a snapshot of the functional connectivity during the time it is copied, just as the connectomes obtained from electron microscopy[32,33] are snapshots of the anatomical connectivity. At the same time, the NNN connectivity undergoes lifelong changes, even in its adult form, from, for example, learning and experience. It is important to note, however, that the NNN is a stable structure despite the plasticity, consisting of definite and elaborate baseline circuits to perform well-defined tasks, which are passed down through generations. Identifying such baseline circuits is a major theme of neuroscience[60]. Both connectomics and our Perspective build on the idea that the snapshots could not only identify baseline circuits but also track, if taken at different times, the slow changes in them from learning and experience, thus helping to understand plasticity[66]. Further, by analysing the changing behaviours and engineering them into the memories, we may, in principle, create a pasted network that exhibits similar long-term plasticity. The short-term plasticity influenced by, for example, sensory input, may be similarly analysed but with a higher time resolution, and may be engineered into the memories that exhibit short-term plasticity[56,67,68].

As discussed earlier, our approach has many challenges. An additional limitation is that the electronic paste platform greatly abstracts the immense chemical complexities and convoluted signal pathways of synaptic transmission into lumped, effective synaptic connection strengths represented by memory conductance values. The same line of abstraction goes for modelling many different types of neurons and their membrane proteins with silicon circuits. We, however, share the sense of optimism of the connectomics researchers that our pasted network may approximate some essential aspects of the brain's computing and would represent a first step towards brain reverse engineering.

Outlook

The lack of knowledge about the brain's functional wiring diagram makes reverse engineering the brain—the original goal of neuromorphic electronics—extremely challenging. Over the past few decades, the aims of neuromorphic engineering have thus been relaxed from rigorous brain reverse engineering to brain-inspired design that uses qualitative features of the brain (such as event-driven asynchronous signalling and in-memory information processing). These efforts have led to a number of exciting applications in dynamic vision sensing and low-power AI computing, but they are a far cry from creating a genuinely intelligent system. We have thus provided a vision to retarget the original goal of neuromorphic electronics. Advances in neuro-electronic interfaces have brought us closer to accessing the functional wiring diagram in the brain, and high-density memory technology—which has been made possible due to advances in 3D integration and packaging—offers a platform onto which to paste this wiring diagram. Our approach is ambitious, and there is no guarantee that all the challenges outlined here can be easily overcome. However, by working towards such a goal, we can, we believe, help push the boundaries of neuromorphic engineering, neuroscience and semiconductor technology.

Received: 7 September 2020; Accepted: 18 August 2021;
Published online: 23 September 2021

References

1. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
2. Mead, C. *Analog VLSI and Neural Systems* (Addison-Wesley, 1989).
3. Liu, S.-C. & Delbruck, T. Neuromorphic sensory systems. *Curr. Opin. Neurobiol.* **20**, 288–295 (2010).

4. Lichtsteiner, P., Posch, C. & Delbruck, T. A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid State Circuits* **43**, 566–576 (2008).
5. Son, B. et al. A 640×480 dynamic vision sensor with a 9 μm pixel and 300 Meps address-event representation. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)* 66–67 (IEEE, 2017).
6. Merolla, P. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
7. Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
8. Sheridan, P. M. et al. Sparse coding with memristor networks. *Nat. Nanotechnol.* **12**, 784–790 (2017).
9. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
10. Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nat. Commun.* **4**, 2072 (2013).
11. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60–67 (2018).
12. Choi, S., Shin, J. H., Lee, J., Sheridan, P. & Lu, W. D. Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. *Nano Lett.* **17**, 3113–3118 (2017).
13. Wang, Z. et al. Resistive switching materials for information processing. *Nat. Rev. Mater.* **5**, 173–195 (2020).
14. Lin, P. et al. Three-dimensional memristor circuits as complex neural networks. *Nat. Electron.* **3**, 225–232 (2020).
15. Wang, Z. et al. Reinforcement learning with analogue memristor arrays. *Nat. Electron.* **2**, 115–124 (2019).
16. Li, C. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).
17. Jang, H. et al. An atomically thin optoelectronic machine vision processor. *Adv. Mater.* **32**, 2002431 (2020).
18. Mahowald, M. & Douglas, R. A silicon neuron. *Nature* **354**, 515–518 (1991).
19. Indiveri, G. et al. Neuromorphic silicon neuron circuits. *Front. Neurosci.* **5**, 73 (2011).
20. Ryckebusch, S., Bower, J. M. & Mead, C. Modeling small oscillating biological networks in analog VLSI. *Adv. Neural Inf. Process. Syst.* **1**, 384–393 (1989).
21. Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M., Douglas, R. & Seung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **405**, 947–951 (2000).
22. Mahowald, M. & Mead, C. The silicon retina. *Sci. Am.* **264**, 76–82 (1991).
23. Abbott, J. et al. A nanoelectrode array for obtaining intracellular recordings from thousands of connected neurons. *Nat. Biomed. Eng.* **4**, 232–241 (2020).
24. Kim, K. Silicon technologies and solutions for the data-driven world. In *2015 IEEE International Solid-State Circuits Conference (ISSCC)* 8–14 (IEEE, 2015).
25. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
26. Song, J. et al. An 11.5TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8 nm flagship mobile SoC. In *2019 IEEE International Solid-State Circuits Conference (ISSCC)* 130–131 (IEEE, 2019).
27. Wang, Z. et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nat. Electron.* **1**, 137–145 (2018).
28. Ambrogio, S. et al. Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses. *Front. Neurosci.* **10**, 56 (2016).
29. Alivisatos, P. et al. Nanotools for neuroscience and brain activity mapping. *ACS Nano* **7**, 1850–1866 (2013).
30. Lin, M. Z. & Schnitzer, M. J. Genetically encoded indicators of neuronal activity. *Nat. Neurosci.* **19**, 1142–1153 (2016).
31. Cazemier, J. L., Clasca, F. & Tiesinga, P. H. E. Connectomic analysis of brain networks: novel techniques and future directions. *Front. Neuroanat.* **10**, 110 (2016).
32. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. B* **314**, 1–340 (1986).
33. Scheffer, L. K. et al. A connectome and analysis of the adult *Drosophila* central brain. *eLife* **9**, e57443 (2020).
34. Sasaki, T., Minamisawa, G., Takahashi, N., Matsuki, N. & Ikegaya, Y. Reverse optical trawling for synaptic connections in situ. *J. Neurophysiol.* **102**, 636–643 (2009).
35. Petreanu, L., Huber, D., Sobczyk, A. & Svoboda, K. Channelrhodopsin-2-assisted circuit mapping of long-range callosal projections. *Nat. Neurosci.* **10**, 663–668 (2007).
36. Shemesh, O. A. et al. Temporally precise single-cell-resolution optogenetics. *Nat. Neurosci.* **20**, 1796–1806 (2017).
37. Jäckel, D. et al. Combination of high-density microelectrode array and patch clamp recordings to enable studies of multisynaptic integration. *Sci. Rep.* **7**, 978 (2017).
38. Perin, R., Berger, T. K. & Markram, H. A synaptic organizing principle for cortical neuronal groups. *Proc. Natl Acad. Sci. USA* **108**, 5419–5424 (2011).
39. Frey, U. et al. Switch-matrix-based high-density microelectrode array in CMOS technology. *IEEE J. Solid State Circuits* **45**, 467–482 (2010).
40. Tsai, D., Sawyer, D., Bradd, A., Yuste, R. & Shepard, K. L. A very large-scale microelectrode array for cellular-resolution electrophysiology. *Nat. Commun.* **8**, 1802 (2017).
41. Abbott, J. et al. The design of a CMOS nanoelectrode array with 4096 current-clamp/voltage-clamp amplifiers for intracellular recording/stimulation of mammalian neurons. *IEEE J. Solid State Circuits* **55**, 2567–2582 (2020).
42. Abbott, J. et al. CMOS nanoelectrode array for all-electrical intracellular electrophysiological imaging. *Nat. Nanotechnol.* **12**, 460–466 (2017).
43. Robinson, J. T. et al. Vertical nanowire electrode arrays as a scalable platform for intracellular interfacing to neuronal circuits. *Nat. Nanotechnol.* **7**, 180–184 (2012).
44. Jun, J. J. et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).
45. Sanes, J. R. & Zipursky, S. L. Design principles of insect and vertebrate visual systems. *Neuron* **66**, 15–36 (2010).
46. Wilson, D. A. & Sullivan, R. M. Cortical processing of odor objects. *Neuron* **72**, 506–519 (2011).
47. Lee, Y. K. et al. Embedded STT-MRAM in 28-nm FDSOI logic process for industrial MCU/IoT application. In *2018 IEEE Symposium on VLSI Technology* 181–182 (IEEE, 2018).
48. Chung, H. et al. A 58 nm 1.8 V 1 Gb PRAM with 6.4 MB/s program BW. In *2011 IEEE International Solid-State Circuits Conference (ISSCC)* 500–501 (IEEE, 2011).
49. Fackenthal, R. et al. A 16 Gb ReRAM with 200 MB/s write and 1 GB/s read in 27 nm technology. In *2014 IEEE International Solid-State Circuits Conference (ISSCC)* 338–339 (IEEE, 2014).
50. Lee, M.-J. et al. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures. *Nat. Mater.* **10**, 625–630 (2011).
51. Xu, R. et al. Vertical MoS_2 double layer memristor with electrochemical metallization as an atomic-scale synapse with switching thresholds approaching 100 mV. *Nano Lett.* **19**, 2411–2417 (2019).
52. Jo, S. H. et al. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10**, 1297–1301 (2010).
53. Park, K.-T. et al. Three-dimensional 128 Gb MLC vertical NAND flash-memory with 24-WL stacked layers and 50 MB/s high-speed programming. In *2014 IEEE International Solid-State Circuits Conference (ISSCC)* 334–335 (IEEE, 2014).
54. Du, C., Ma, W., Chang, T., Sheridan, P. & Lu, W. D. Biorealistic implementation of synaptic functions with oxide memristors through internal ionic dynamics. *Adv. Funct. Mater.* **25**, 4290–4299 (2015).
55. Kim, S. et al. Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. *Nano Lett.* **15**, 2203–2211 (2015).
56. Wang, Z. et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **16**, 101–108 (2017).
57. Kuzum, D., Jeyasingh, R. G. D., Lee, B. & Wong, H.-S. P. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano Lett.* **12**, 2179–2186 (2012).
58. Marx, V. Neuroscience waves to the crowd. *Nat. Methods* **10**, 1069–1074 (2013).
59. Kim, D.-W. & Hwang, T. The future of advanced package solutions. In *2019 IEEE Symposium on VLSI Technology* 48–49 (IEEE, 2019).
60. Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A. & Hudspeth, A. J. *Principles of Neural Science* 5th edn (McGraw-Hill, 2012).
61. Bock, D. D. et al. Network anatomy and in vivo physiology of visual cortical neurons. *Nature* **471**, 177–182 (2011).
62. Briggman, K. L. & Bock, D. D. Volume electron microscopy for neuronal circuit reconstruction. *Curr. Opin. Neurobiol.* **22**, 154–161 (2012).
63. Briggman, K. L., Helmstaedter, M. & Denk, W. Wiring specificity in the direction-selectivity circuit of the retina. *Nature* **471**, 183–188 (2011).
64. Kleinfeld, D. et al. Large-scale automated histology in the pursuit of connectomes. *J. Neurosci.* **31**, 16125–16138 (2011).
65. Lichtman, J. W. & Sanes, J. R. Ome sweet ome: what can the genome tell us about the connectome? *Curr. Opin. Neurobiol.* **18**, 346–353 (2008).
66. Morgan, J. L. & Lichtman, J. W. Why not connectomics? *Nat. Methods* **10**, 494–500 (2013).
67. Berdan, R. et al. Emulating short-term synaptic dynamics with memristive devices. *Sci. Rep.* **6**, 18639 (2016).
68. Kim, M.-K. & Lee, J.-S. Short-term plasticity and long-term potentiation in artificial biosynapses with diffusive dynamics. *ACS Nano* **12**, 1680–1687 (2018).

Acknowledgements

We thank S. J. Kim, S. Jung, H. Lee and H. Kim of Samsung Advanced Institute of Technology, J. Abbott, T. Ye, K. Krenek, R. Gertner, S. Ban, Y. Kim, L. Qin, W. Wu, R. Xu, H. S. Jung and J. Wang of Harvard University and H. J. Baek (Executive Vice President),

J. Song (Executive Vice President), K. Choi (Senior Vice President), S. Yoon (Vice President), T. Hwang, J. Lim, D. Kwon, Y. Kim and J. Kim of Samsung Electronics for discussions, advice and/or contributions to materials used here from original research articles.

Author contributions

D.H., H.P., S.H. and K.K. conceived this Perspective. D.H., H.P., S.H. and K.K. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Donhee Ham, Hongkun Park, Sungwoo Hwang or Kinam Kim.

Peer review information *Nature Electronics* thanks Yoeri van de Burgt and Huaqiang Wu for their contribution to the peer review of this work.