PERSPECTIVE • OPEN ACCESS

Neurogrid simulates cortical cell-types, active dendrites, and top-down attention

To cite this article: Ben Varkey Benjamin *et al* 2021 *Neuromorph. Comput. Eng.* **1** 013001

NEUROMORPHIC
Computing and Engineering

PERSPECTIVE

Neurogrid simulates cortical cell-types, active dendrites, and top-down attention

Ben Varkey Benjamin[1,4], **Nicholas A Steinmetz**[2,5], **Nick N Oza**[3,6], **Jose J Aguayo**[3,7] and **Kwabena Boahen**[1,3,*]

[1] Department of Electrical Engineering, Stanford University, Stanford, CA 94305, United States of America
[2] Department of Neurobiology, Stanford University, Stanford, CA 94305, United States of America
[3] Department of Bioengineering, Stanford University, Stanford, CA 94305, United States of America
[*] Author to whom any correspondence should be addressed.
[4] Present address: Dexterity Inc., Redwood City, CA 94063.
[5] Present address: Department of Biological Structure, University of Washington, Seattle, WA.
[6] Present address: Myrrym Corporation, Palo Alto, CA.
[7] Present address: Amazon.com, Seattle, WA.

E-mail: boahen@stanford.edu

Keywords: brain simulation, hardware accelerator, mixed signal, active dendrites, cortical cell types, hybrid analog–digital

OPEN ACCESS

RECEIVED
12 April 2021

REVISED
17 May 2021

ACCEPTED FOR PUBLICATION
10 June 2021

PUBLISHED
15 July 2021

Abstract

A central challenge for systems neuroscience and artificial intelligence is to understand how cognitive behaviors arise from large, highly interconnected networks of neurons. Digital simulation is linking cognitive behavior to neural activity to bridge this gap in our understanding at great expense in time and electricity. A hybrid analog–digital approach, whereby slow analog circuits, operating in parallel, emulate graded integration of synaptic currents by dendrites while a fast digital bus, operating serially, emulates all-or-none transmission of action potentials by axons, may improve simulation efficacy. Due to the latter's serial operation, this approach has not scaled beyond millions of synaptic connections (per bus). This limit was broken by following design principles the neocortex uses to minimize its wiring. The resulting hybrid analog–digital platform, Neurogrid, scales to billions of synaptic connections, between up to a million neurons, and simulates cortical models in real-time using a few watts of electricity. Here, we demonstrate that Neurogrid simulates cortical models spanning five levels of experimental investigation: biophysical, dendritic, neuronal, columnar, and area. Bridging these five levels with Neurogrid revealed a novel way active dendrites could mediate top-down attention.

1. Multiscale neural simulations

A central challenge for systems neuroscience and artificial intelligence is to understand how cognitive behaviors arise from large, highly interconnected networks of neurons [1]. Bridging this gap in our understanding calls for large-scale yet biophysically realistic neural simulations to span levels of experimental investigation (biophysical, dendritic, neuronal, columnar, and area) [2]. Digital simulation bridges these levels, but the expense in time and electricity is great [3].

A Titan RTX GPU card consumes electricity at a rate of 250 W to simulate a cortex model with 4 million neurons connected by 24 billion synapses at a speed 510-fold slower than real-time (i.e. each biological second takes 510 s) [4]. These simulations are challenging because each neuron *distributes* its output to thousands of neurons and, in turn, *aggregates* inputs from thousands of neurons.

In the digital paradigm, replicating signals for distribution as well as summing signals for aggregation occurs entirely within computing cores (facilitated by programming constructs such as *MapReduce*). The communication network interconnecting these cores simply routes messages by relaying them across links that connect a core to its immediate neighbors. Sharing these core-to-core wires instead of routing dedicated point-to-point wires, as the cortex does, ameliorates the difficulty of mapping a three-dimensional brain onto a two-dimensional chip [5].

With shared wires communicating addresses that signal arrival of an action potential, or spike, at particular synapse, traffic scales as the total number of synaptic connections times the presynaptic population's average spike-rate [6]. This *address-event* bus has been successfully used to build networks with thousands of neurons with a few hundred synaptic connections each [7]. It has not scaled beyond millions of synaptic connections, the point at which traffic saturates the bus' signaling rate.

To break this communication bottleneck, Neurogrid adopts a hybrid analog–digital approach that follows design principles the neocortex uses to minimize its wiring [8]. Neurogrid uses fast digital routers, operating serially, to replicate signals for distribution and uses slow analog circuits, operating in parallel, to sum signals for aggregation. It follows the neocortex's design principles by performing both distribution and aggregation hierarchically. This neuromorphic architecture scales to billions of synaptic connections, between up to a million neurons.

Here, we demonstrate that Neurogrid simulates cortical models spanning five levels of experimental investigation: biophysical, dendritic, neuronal, columnar, and area (sections 5–7). Bridging these five levels with Neurogrid revealed a novel way active dendrites could mediate top-down attention. We begin with a brief review of Neurogrid's neuromorphic architecture (section 3) and simulation environment (section 4); detailed descriptions are available elsewhere [8].

2. Methods

2.1. Dimensionless neuron models

Neurogrid's models of soma, dendrite, gating variables and synapses are dimensionless. A leaky integrate-and-fire model of a neuron [9] is described by

$$C_{\mathrm{s}}\frac{\mathrm{d}V_{\mathrm{m}}}{\mathrm{d}t} = G_{\mathrm{leak}}\left(E_{\mathrm{leak}} - V_{\mathrm{m}}\right) + I_{\mathrm{sin}} + I_{\mathrm{Na}}, \tag{1}$$

where V_{m} is the soma membrane potential, C_{s} is the soma membrane capacitance, G_{leak} is the soma leak conductance and E_{leak} its reversal potential, I_{sin} is the injected current and I_{Na} is the voltage-dependent sodium current that generates spikes. Rewriting the above equation with E_{leak} as the reference voltage:

$$C_{\mathrm{s}}\frac{\mathrm{d}V_{\mathrm{s}}}{\mathrm{d}t} = -G_{\mathrm{leak}}V_{\mathrm{s}} + I_{\mathrm{sin}} + I_{\mathrm{Na}}, \tag{2}$$

where $V_{\mathrm{s}} = V_{\mathrm{m}} - E_{\mathrm{leak}}$. We modeled $I_{\mathrm{Na}}\left(V_{\mathrm{s}}\right)$ as

$$I_{\mathrm{Na}}\left(V_{\mathrm{s}}\right) = \frac{1}{2}\frac{V_{\mathrm{s}}^2}{V_{\mathrm{th}}}G_{\mathrm{leak}}, \tag{3}$$

where V_{th} is the threshold voltage. By definition, it is the voltage at which the sodium conductance $\left(\mathrm{d}I_{\mathrm{Na}}/\mathrm{d}V_{\mathrm{s}}\right)$ is equal to the leak conductance. At this voltage, the membrane potential stops decelerating and starts accelerating (inflection point), the onset of spiking. Substituting equation (3) into equation (2) and dividing by $G_{\mathrm{leak}}V_{\mathrm{th}}$ gives the dimensionless model for a quadratic leaky integrate-and-fire (QLIF) neuron [10, 11]:

$$\tau_{\mathrm{s}}\dot{v}_{\mathrm{s}} = -v_{\mathrm{s}} + i_{\mathrm{sin}} + \frac{v_{\mathrm{s}}^2}{2}, \tag{4}$$

where $\tau_{\mathrm{s}} = C_{\mathrm{s}}/g_{\mathrm{leak}}$ (membrane time constant), $v_{\mathrm{s}} = V_{\mathrm{s}}/V_{\mathrm{th}}$ and $i_{\mathrm{sin}} = I_{\mathrm{sin}}/\left(g_{\mathrm{leak}}V_{\mathrm{th}}\right)$. Thus, models of the form equation (1) can be expressed as dimensionless models of the form equation (4) by changing the reference voltage to E_{leak} and normalizing all voltages, conductances and currents by V_{th}, G_{leak} and $G_{\mathrm{leak}}V_{\mathrm{th}}$, respectively.

The soma compartment's dimensionless model is

$$\tau_{\mathrm{s}}\dot{v}_{\mathrm{s}} = -v_{\mathrm{s}} + i_{\mathrm{sin}} + \frac{v_{\mathrm{s}}^2}{2} - g_{\mathrm{K}}v_{\mathrm{s}} - g_{\mathrm{res}}v_{\mathrm{s}}p_{\mathrm{res}}(t) + v_{\mathrm{d}} + \sum g_{\mathrm{syn}_i}\left(e_{\mathrm{syn}_i} - v_{\mathrm{s}}\right), \tag{5}$$

where v_{s} is the membrane potential, τ_{s} the membrane time constant, i_{sin} the input current, and g_{syn_i} and e_{syn_i} the conductance and reversal potential, respectively, of synapse *i*. g_{res} is activated by $p_{\mathrm{res}}(t)$, a unit amplitude pulse that is active for a duration t_{res} after a spike (declared when $v_{\mathrm{s}} \geqslant 10$), to model the neuron's refractory period. g_{K} models a high-threshold potassium conductance and is only activated when a spike occurs, decaying afterward. We modeled its dynamics as

$$\tau_{\mathrm{K}}\dot{g}_{\mathrm{K}} = -g_{\mathrm{K}} + p_{\mathrm{res}}(t)g_{\mathrm{K}\infty}, \tag{6}$$

where τ_{K} is the decay time constant.

The dendrite compartment's dimensionless model is

$$\tau_{\mathrm{d}}\dot{v}_{\mathrm{d}} = -v_{\mathrm{d}} + i_{\mathrm{din}} + i_{\mathrm{bp}}p_{\mathrm{res}}(t) + \sum \zeta_i g_{\mathrm{syn}_i}\left(e_{\mathrm{syn}_i} - v_{\mathrm{d}}\right) + \sum g_{\mathrm{ch}_i}\left(e_{\mathrm{ch}_i} - v_{\mathrm{d}}\right), \tag{7}$$

where v_{d} is the membrane potential, τ_{d} the membrane time constant, i_{din} the input current, g_{syn_i}, e_{syn_i} and ζ_i the conductance, reversal potential and spatial decay factor, respectively, of synapse i, and g_{ch_i} and e_{ch_i} the conductance and reversal potential, respectively, of ion-channel population i. A current i_{bp} is injected for a duration t_{res} (same as in equation (5)) to model a backpropagating spike.

The decay factor, which models the spatial decay in dendritic trees, is given by [12]

$$\zeta(n) = \frac{1}{4\sqrt{\pi}}\left(1 + \left(\frac{1}{1-\gamma^2}\right)^{\frac{1}{4}}\right)^2 \frac{\gamma^n}{\sqrt{n}}, \tag{8}$$

where γ is the silicon dendritic tree's decay constant and n the distance traveled in number of neurons.

The synaptic population's dimensionless model [13] is

$$\tau_{\mathrm{syn}}\dot{g}_{\mathrm{syn}} = -g_{\mathrm{syn}} + p_{\mathrm{rise}}(t)g_{\mathrm{sat}}, \tag{9}$$

where g_{syn} is the instantaneous synaptic conductance, τ_{syn} the synaptic time constant and g_{sat} the maximum conductance. The unit amplitude pulse $p_{\mathrm{rise}}(t)$'s width t_{rise} models the duration for which neurotransmitter is available in the cleft following a pre-synaptic spike.

An ion-channel population's gating variable is modeled as:

$$\tau_{\mathrm{ch}}\dot{c} = -c + c_{\mathrm{ss}}, \tag{10}$$

where c_{ss} is its steady-state activation (or inactivation) and τ_{ch} its time constant. c_{ss} is given by

$$c_{\mathrm{ss}} = \frac{\alpha}{\alpha+\beta} \quad \text{or} \quad \frac{\beta}{\alpha+\beta}, \tag{11}$$

where α and β model a channel's voltage-dependent opening and closing rates and are described by

$$\begin{aligned} \alpha &= \frac{v_{\mathrm{d}} - v_{\mathrm{th}}}{2} + \frac{\sqrt{(v_{\mathrm{d}} - v_{\mathrm{th}})^2 + \frac{1}{4s^2}}}{2} \\ \beta &= -\frac{v_{\mathrm{d}} - v_{\mathrm{th}}}{2} + \frac{\sqrt{(v_{\mathrm{d}} - v_{\mathrm{th}})^2 + \frac{1}{4s^2}}}{2}. \end{aligned} \tag{12}$$

Here, v_{th} is the membrane potential at which $c_{\mathrm{s}} = 1/2$ and s is the slope at this point. α and β satisfy a difference relation, $\alpha - \beta = v_{\mathrm{d}} - v_{\mathrm{th}}$, and a reciprocal relation, $\alpha\beta = 1/\left(16s^2\right)$, resulting in a sigmoidal dependence of c_{ss} on v_{d}. The gating variable's time constant is given by

$$\begin{aligned} \tau_{\mathrm{ch}} &= \frac{\tau_{\mathrm{max}} - \tau_{\mathrm{min}}}{\tau_{\mathrm{max}}}\tilde{\tau}_{\mathrm{ch}} + \tau_{\mathrm{min}} \\ \tilde{\tau}_{\mathrm{ch}} &= \frac{1}{2s\left(\alpha+\beta\right)}\tau_{\mathrm{max}} \end{aligned} \tag{13}$$

τ_{ch} is bell-shaped with a maximum value of τ_{max} when $v_{\mathrm{d}} = v_{\mathrm{th}}$ and a minimum value of τ_{min} when $|v_{\mathrm{d}} - v_{\mathrm{th}}| \gg 1/(2s)$, to avoid unphysiologically short time constants.

An ion-channel population may have one or two gating variables. When using one gating variable, its associated maximum conductance g_{max} may be set by a synapse population's g_{syn} to model an ion-channel population that is voltage as well as ligand-gated (e.g., NMDA receptors). When using a pair of gating variables, the effective conductance is given by:

$$g_{\mathrm{ch}} = \frac{g_{\mathrm{max0}}c_0 g_{\mathrm{max1}}c_1}{g_{\mathrm{max0}}c_0 + g_{\mathrm{max1}}c_1}, \tag{14}$$

where g_{max0} and g_{max1} are the maximum conductances associated with gating variables c_0 and c_1, respectively. In this case, g_{max0} and g_{max1} may be set by two synapse populations' g_{syn} to model neuromodulation, and the gating variables' thresholds set low to eliminate voltage-dependence. Alternatively, g_{max0} and g_{max1} may be programmed to the same value, g_{max}, to model a channel that activates and inactivates. In this case, the above expression simplifies to $g_{\mathrm{ch}} = g_{\mathrm{max}}\frac{c_0 c_1}{c_0+c_1}$. The pair of gating variables always share a programmable reversal potential e_{ch}.

The parameter values used to obtain the simulation results presented in figures 2–4 are listed in tables 1–4. Reversal potentials and gating-variable thresholds were converted to dimensionless form by assuming that the

Table 1. Parameter values for models of cortical neuron types. Current injection for FS, RS and CH neuron types was modeled with i_{sin} and for IB neuron type with i_{din}. Parameter values were chosen to produce the best fit to the *in vitro* data. All time values are in ms.

									Channel							
	Soma					Dendrite			Conductance		Activation			Inactivation		
	τ_s	τ_K	$g_{K\infty}$	t_{res}	i_{sin}	τ_d	i_{bp}	i_{din}	g_{max}	e_{ch}	τ_{max}	v_{th}	s	τ_{max}	v_{th}	s
FS	3	200	0.005	0.8	3.7–9.8	—	—	_	—	—	—	—	—	—	—	—
RS	15	200	50	0.1	0.08–1.42	—	—	—	—	—	—	—	—	—	—	—
IB	18	200	50	1	0	54	0	1.09–2.5	1	7.5	1	0.5	1.25	50	0.2	−0.5
CH	13	50	250	2	30–39	12	100	0	—	—	—	—	—	—	—	—

Table 2. Parameter values used to model NMDA spikes in basal dendrites. NMDA and AMPA reversal potentials as well as NMDA threshold were chosen to match physiological values [14]. Other parameter values were chosen to produce the best fit to the *in vitro* data. All time values are in ms.

Soma					Dendrite			NMDA						AMPA			
τ_s	τ_K	$g_{K\infty}$	t_{res}	i_{sin}	τ_d	i_{din}	i_{bp}	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	v_{th}	s	τ_{syn}	t_{rise}	g_{sat}	e_{syn}
20	—	—	1	0	30	0	0	150	4	500	2.7	2.3	1	7.25	0.6	25–240	2.7

Table 3. Parameter values used to model coincidence detection in apical dendrites. Soma compartment was modeled with RS neuron parameter values (see table 1). Sodium channel parameters are similar to Mainen *et al* [15], adjusted together with other parameter values to produce the best fit to the *in vitro* data. All time values are in ms.

							Sodium channel							
Dendrite			Synapse				Conductance		Activation			Inactivation		
τ_d	i_{din}	i_{bp}	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	g_{max}	e_{ch}	τ_{max}	v_{th}	s	τ_{max}	v_{th}	s
3	0	100	8	4	80	2.7	10	5	0.4	0.9	1.2	35	0.5	−2

Table 4. Parameter values for attentional modulation simulation. Excitatory and inhibitory neurons' soma compartments were modeled with RS and FS neuron parameter values, respectively (see table 1). NMDA, AMPA and dendrite were modeled with the parameter values for NMDA spikes in basal dendrites (see table 2). All time values are in ms.

FEF excitatory														
Soma					AMPA					GABA				
τ_s	τ_K	$g_{K\infty}$	t_{res}	i_{sin}	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	γ	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	γ
15	200	50.0	0.1	0	7.25	0.6	8	2.7	0.65	15	1	0.01	0.1	0.97

FEF inhibitory									
Soma					AMPA				
τ_s	τ_K	$g_{K\infty}$	t_{res}	i_{sin}	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	γ
3	200	0.005	0.8	0.15	7.25	0.6	10	2.7	0.6

V4 excitatory																	
Soma					Dendrite		AMPA					NMDA					
τ_s	τ_K	$g_{K\infty}$	t_{res}	i_{sin}	τ_d	i_{din}	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	γ	τ_{syn}	t_{rise}	g_{sat}	e_{syn}	v_{th}	s
15	200	50	0.1	0	30	0.001	7.25	0.6	2.0	2.7	0.65	150	4	50	2.7	2.3	1

leak's reversal potential is −80 mV and the spike threshold is −50 mV (i.e., all voltages are normalized by $V_{th} = 30$ mV).

2.2. Programming Neurogrid

Neuronal and synaptic model parameters and connectivity are described in Python, similar to PyNN [16], and a C++ back-end uses this description to program Neurogrid's electronic circuit parameters and lookup

Table 5. Python constructs to program Neurogrid. The table lists example Python function calls for modeling a soma, dendrite, synapse and channel. All time values are in seconds.

Soma	Soma('quadratic_adaptive', 'tau_s': 0.015, 'i_sin': 0.6, 't_res': 0.0005, 'g_kinf': 200)
Dendrite	Dendrite('generic', 'tau_d': 0.02, 'i_din':0, 'i_bp': 100)
Synapse	Synapse('generic', 'tau_syn': 0.025, 'g_sat': 5, 'e_syn': 5, 't_rise': 0.001)
Channel	IonType('generic', 'e_ch': 5) IonChannel('first_order', 'tau_max': 0.02, 'g_max': 10, 'v_th': 0.7, 's': 7)

tables. Examples of Python descriptions of soma, dendrite, synapse and gating variables are listed in table 5. A calibration procedure is used to obtain a set of conversion factors to map the neuronal and synaptic model parameters to the circuit parameters [17]. In a population of silicon neurons, these parameters are lognormally distributed with a coefficient of variation (CV) of about 10% (e.g., $1/\tau_s$ has a CV of 7.2%). Lookup tables for intracolumn connections are stored in the Neuorocores' 256 × 16 bit RAM and for intercolumn connections in the daughterboard's 8M × 32 bit RAM.

3. Hierarchical distribution & aggregation

A neuron's axonal arbor distributes signals by hierarchically replicating them at branch points. Placing an axon's branch points as close as possible to its terminals replaces multiple axon segments with a single segment [18] (figure 1(a), top). This principle minimizes the axonal arbor's wiring.

Moreover, extending several dendritic branches to meet a terminal branch of an axon allows that branch to make multiple synapses (bouton cluster) [19] (figure 1(b), top). Synaptic signals from many axons sum in a dendritic branch and these branches' signals aggregate hierarchically. This principle minimizes the dendritic tree's wiring [18].

Hierarchical distribution is emulated by interconnecting silicon-neuron arrays with a tree-like (rather than a mesh-like) network [20, 21]. That allows address-events to be replicated close to their destination and reduces traffic on the array-to-array links (figure 1(a), bottom). Emulating an axon terminal's bouton cluster allows a single delivered address-event to evoke postsynaptic potentials in several silicon neurons (figure 1(b), bottom). This reduces traffic further.

Hierarchical aggregation is emulated efficiently by modeling multiple overlapping dendritic trees with a single two-dimensional resistive grid. This shared analog circuit replicates the exponential spatial decay that transforms postsynaptic potentials into current delivered to a dendritic tree's trunk [22].

4. Neurogrid

Emulating axonal arbors and dendritic trees enables Neurogrid to simulate cortical models scalably and efficiently. A cortical area is modeled by a group of *Neurocores* (figure 1(c)); Neurogrid has sixteen of these chips. Each cell layer (or cell type) is mapped onto a different Neurocore's two-dimensional silicon neuron array. Circular pools of neurons centered at the same (x, y) location on these Neurocores model a cortical column [24].

Intercolumn axonal projections are routed by using the presynaptic neuron's address to retrieve the target columns' centers (x, y) from an off-chip random-access memory (first distribution level). This memory is programmed to replicate the neocortex's function-specific intercolumn connectivity [25, 26].

Intracolumn axon collaterals are routed by copying a retrieved (x, y) address to all of a cortical area's Neurocores using the interchip tree network; unneeded copies are filtered using an on-chip memory (second distribution level). This memory is programmed to replicate the neocortex's stereotyped intracolumn connectivity [27].

Intralayer dendritic trees—arborizing over a circular disc centered on the cell body—are realized using the resistive grid mentioned earlier. Its space constant is adjusted electronically to match the arbor's radius; a transistor-based implementation makes this possible [28].

With intercolumn projections, intracolumn collaterals, and intralayer dendrites, five thousand synaptic connections may be made by: (1) routing an address-event to ten columns; (2) copying it to six layers in each column; and (3) evoking postsynaptic potentials in a hundred neighboring neurons in all but one layer (a 5.6 neuron arbor radius).

To demonstrate these routing mechanisms' functionality and scalability, we used Neurogrid to simulate a recurrent network of a million interneurons with billions of inhibitory synaptic connections (figure 1(d)). The interneurons were organized in sixteen 256 × 256-neuron arrays that formed a torus; the first and last layers

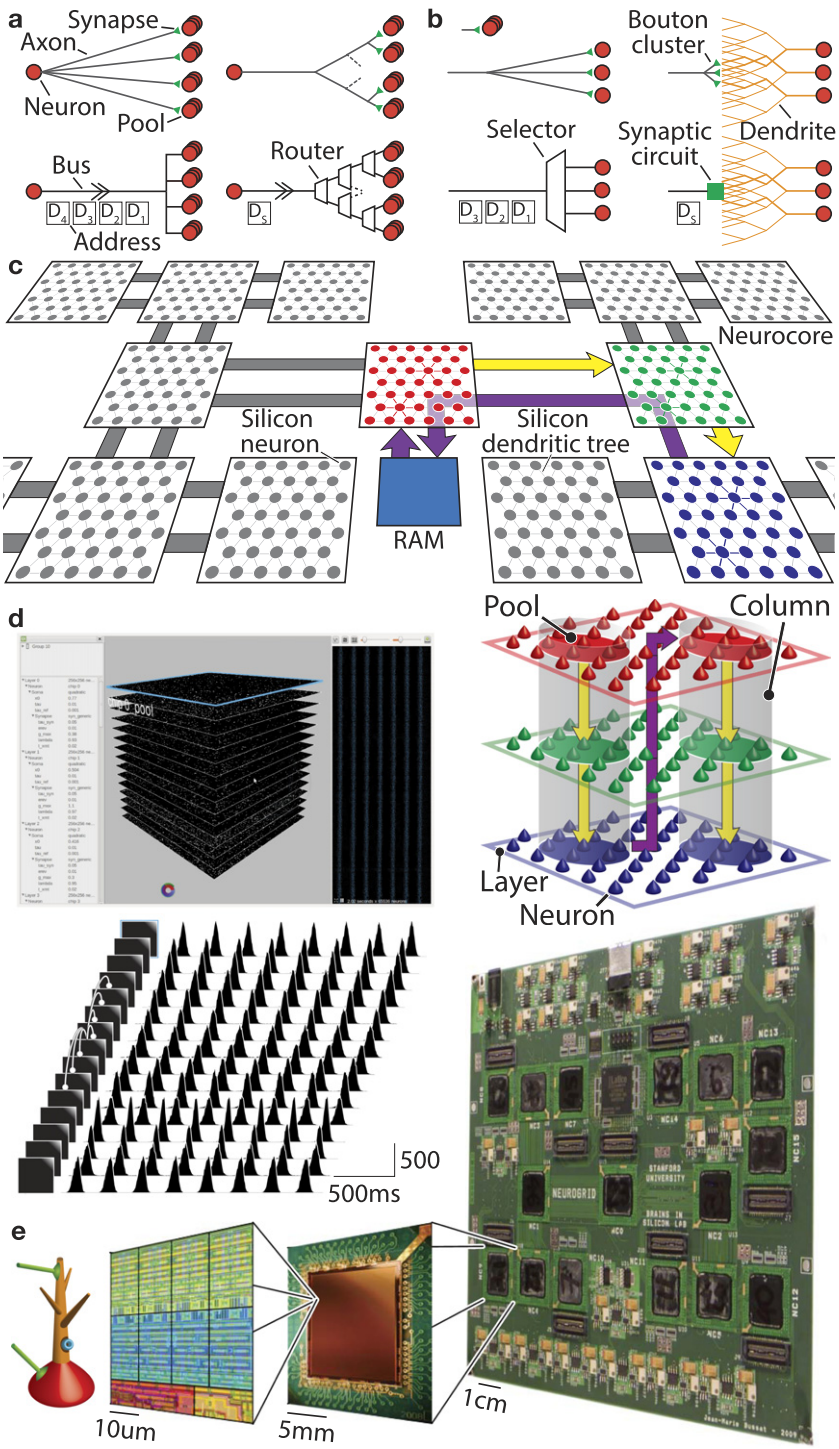

Figure 1. Modeling the cortex on Neurogrid. (a) Hierarchical distribution and (b) hierarchical aggregation in neural (top) and neuromorphic (bottom) networks: the traffic on a digital bus that emulates spike distribution by an axonal arbor is reduced by mimicking axonal and dendritic branching patterns. (c) Mapping cortical columns: cell layers (red, green and blue), intercolumn projections (purple), and intracolumn collaterals (yellow) are mapped onto different Neurocores, off-chip RAM (on a daughterboard), and on-chip RAM (in each Neurocore), respectively. (d) Simulating 1 M neurons with 8G synaptic connections in real-time. User-interface (top, left panel to right panel): displays model parameters, activity in all layers, and spike trains from a selected layer (the first one). Layer connectivity and spike histograms (bottom, left & right): each layer inhibits itself and its three immediate neighbors on either side (only the eighth layer's connectivity is shown). These local interactions synchronize activity globally. (e) Hardware (left to right). Silicon neuron schematic and layout: distinct soma and dendrite compartments express spike-generating, ligand-gated, and voltage-gated conductances. The last two occupy the most area (four types each). Neurocore chip: has 65 536 silicon neurons, tiled in a 256 × 256 array, and routing circuitry for interchip communication. Neurogrid board: has 16 Neurocores, connected in a binary-tree network (as in (c)).

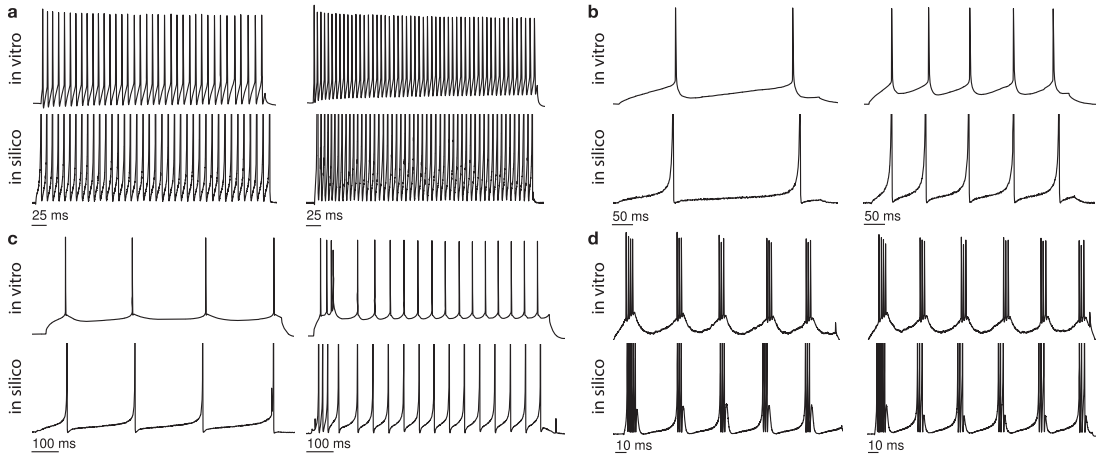

Figure 2. Different types of neocortical neurons and their Neurogrid simulations. Each quadrant shows membrane potentials recorded *in vitro* (upper traces) and *in silico* (lower traces) with two current injection levels (increasing from left to right). (a) FS neuron. (b) RS neuron. (c) IB neuron. (d) CH neuron. Over the ranges of input currents used (see section 2), the model's spike-times closely match the biological neuron's, even though the former are always reset to the same potential (not programmable). *In vitro* data adapted from Izhikevich [23].

were neighbors. Intracolumn collaterals routed an interneuron's spike to the three nearest layers on each side as well as back to its own layer. The inhibition evoked decreased exponentially with distance, due to spatial decay in the intralayer dendritic trees. Half of the inhibition a soma received came from 8000 neurons in a cylinder centered on it, 7 layers in height and 19 neurons in radius.

These local inhibitory interactions synchronized the interneurons globally, reproducing previous findings [29]. Spiking activity waxed and waned periodically (3.7 Hz), with individual interneurons skipping several cycles (0.42 spikes s^{-1} mean). Nevertheless, their spikes were entrained to the global rhythm across all 16 layers, demonstrating the functionality and scalability of Neurogrid's routing mechanisms.

Software infrastructure facilitates performing large-scale neural simulations on Neurogrid. Firmware and drivers initialize and configure the hardware. Calibration procedures establish correspondence between parameters of neuronal models and their silicon analogs [17, 30]. A mapping tool translates multiscale model descriptions (written in Python) into hardware configurations (programmed in memory). A user interface interacts with simulations in real-time. And visualization widgets render activity or plot spikes of up to a million neurons in real-time. For the details of this software infrastructure, see section 2.

5. Neocortical cell-types

Neurogrid's neuron model has separate soma and dendrite compartments (figure 1(e)). The soma compartment expresses conductances that generate spikes and adapt their rate. The dendrite compartment expresses ligand and voltage-gated conductances (up to four types each). The programmed maximum synaptic or membrane conductance is scaled by a 'gating variable' [31]. A pair of gating variables may be used to model a population of ion-channels that are activated *and* inactivated by voltage (e.g., T-type Ca channels), activated by ligand *and* voltage (e.g., NMDA receptors), or modulated by a second ligand (e.g., dopamine). For these components' equations and parameters, see section 2.

To demonstrate the expressiveness of this neuron model, we simulated four prominent neocortical cell-types classified by Nowak *et al* [32], namely fast spiking (FS), regular spiking (RS), intrinsic bursting (IB), and chattering (CH).

FS neurons, normally associated with inhibitory interneurons, respond to a depolarizing current input with a high-frequency spike-train, showing little or no adaptation. RS neurons, generally associated with excitatory stellate cells, respond to a depolarizing current input with a low-frequency spike-train, showing adaptation. For both types of neurons, increasing the injected current increases the spike rate.

We modeled FS and RS neurons with the soma compartment alone, including adaptation for the latter (figures 2(a) and (b)). These models matched the FS neuron's high spike-rate and relatively constant inter-spike intervals; the RS neuron's low spike-rate and increasing inter-spike intervals; and both cell types' overall spike-rate increase with injected current.

IB and CH neurons display more complex responses. IB neurons, generally associated with pyramidal cells, respond to a sufficiently large increase in injected current with a burst followed by single spikes, as observed

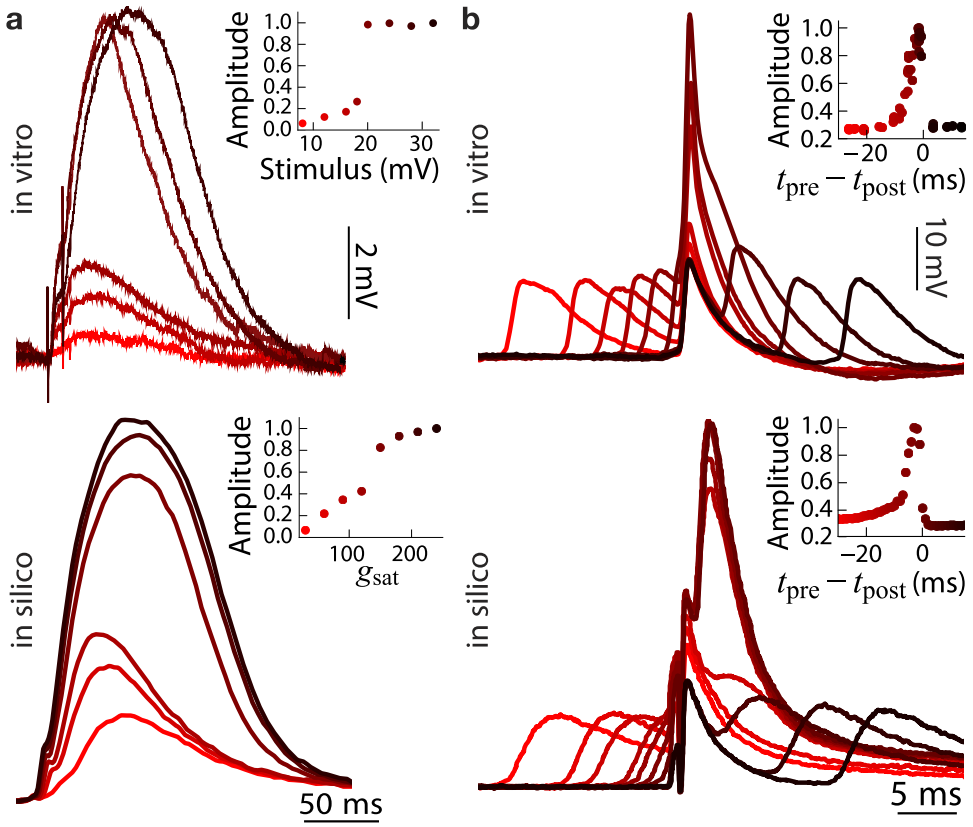

Figure 3. Active dendritic behaviors and their Neurogrid simulations. Each column compares *in vitro* (top) and *in silico* (bottom) membrane potential traces; insets plot traces' peak values versus stimulus conditions. (a) NMDA spikes in basal dendrites: the synaptic conductance was varied by stimulating afferent axons electrically *in vitro* or increasing the maximum synaptic conductance *in silico* (g_{sat}). (b) Coincidence detection in apical dendrites: the relative timing between the backpropagating spike (t_{pre}) and the EPSP (t_{post}) was varied. *In vitro* data in (a) and (b) adapted from Schiller *et al* [38] and Stuart and Häusser [39], respectively.

in guinea-pig cingulate neocortex [33]. Bursting does not occur if the current is injected when the membrane is depolarized. CH neurons, also associated with pyramidal cells, exhibit fast rhythmic bursting, as observed in cat striate cortex [34]. Their burst frequency does not increase much with increasing injected current.

We modeled IB and CH neurons with the soma and dendrite compartments. Slow calcium channels that are deinactivated near the resting potential contribute to bursting [35]. Hence, we equipped the dendrite compartment with an activating as well as an inactivating gating variable (figure 2(c)). This model matched the onset of bursting in an IB neuron and the increase in tonic spike-rate with increasing injected current.

Rhythmic bursting could arise from bidirectional interactions between the soma and the dendrite [36]. Neurogrid's neuron model supports this by allowing spikes to back-propagate from the soma compartment to the dendrite compartment (purely passive in this case). In turn, the dendrite depolarizes the soma. These reciprocal interactions produce repetitive spiking [37]. A build-up of K^+ conductance with each spike terminates the burst; this conductance's decay with time determines the inter-burst interval. This model matched the high spike-rate within a burst, the number of spikes per burst, and the moderate decrease in inter-burst interval with increasing injected current (figure 2(d)).

6. Active dendrites

Several lines of experimental evidence point to a critical role of active dendritic properties in sensorimotor processing. NMDA spikes in basal dendrites [38] have been shown to sharpen sensory tuning of neocortical neurons [40]. Such spikes double an excitatory post-synaptic potential's (EPSP) amplitude when they occur in basal dendrites of neocortical layer 5 pyramidal neurons. Coincidence of EPSPs (pre-spikes) and backpropagating action potentials (post-spikes) in apical dendrites [39] has been shown to amplify correlated sensory and motor signals during active sensation [41]. Thus, active dendrites enhance sensorimotor processing.

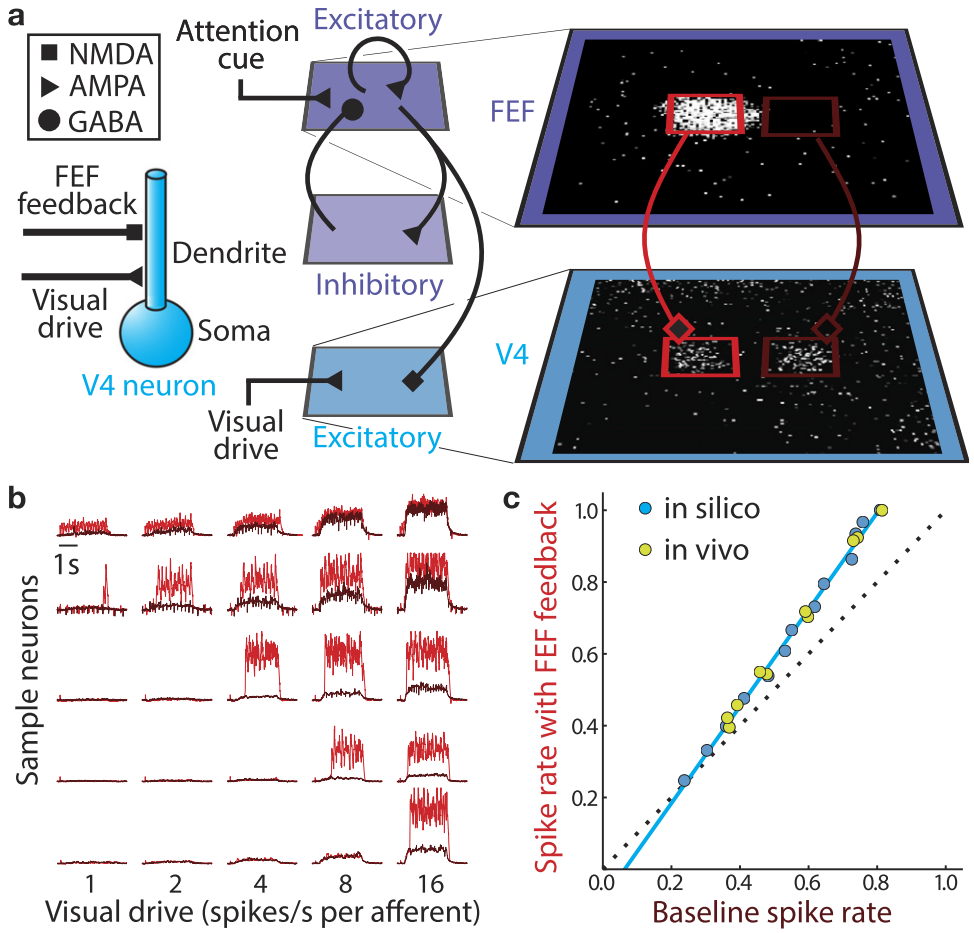

Figure 4. Modeling attention-related modulation of visual cortical responses. (a) FEF-V4 model: Cued activity in FEF (right, white pixels within top 19 × 19 red rectangle) drives NMDA receptors in a patch of excitatory neurons at the corresponding location in V4 (right, white pixels within bottom red rectangle); neurons there and elsewhere (right, bottom brown rectangle) receive identical external visual drive. (b) Heterogeneous responses in model V4: dendritic potentials with and without FEF drive (red and brown, respectively) for five sample neurons (rows) sorted according to the visual drive required to cross NMDA synapse's activation threshold. (c) Population activity in model V4: the average spike-rate scales multiplicatively with NMDA-mediated FEF feedback (calculated for 361 neurons), matching the behavior of V4 neurons recorded from awake, behaving macaques (adapted from McAdams and Maunsell [42]).

To demonstrate that Neurogrid's neuron model could express active dendritic behavior, we replicated the NMDA spike's all-or-none behavior and the backpropagating spike's amplifying action. An NMDA spike occurs when a small increment of synaptic strength, above some threshold, more than doubles the amplitude of the EPSP, which then remains unchanged for further increments [38].

To replicate the NMDA spike, we expressed an NMDA synaptic conductance in a dendrite compartment as well as an AMPA conductance. When the AMPA conductance drove the dendrite's potential above the NMDA conductance's voltage-activation threshold, the postsynaptic potential doubled in amplitude, replicating the all-or-none behavior of an NMDA spike (figure 3(a)).

To replicate the amplifying action of a backpropagating spike, we expressed sodium channels that activate and inactivate in a dendrite compartment as well as a synaptic conductance with time-constant between NMDA and AMPA conductances' to mimic a mix. When a backpropagating spike coincided with a post-synaptic potential, the sodium conductance activated and increased the dendrite's voltage further. The pre-spike's timing relative to the post-spikes' when voltage peaked as well as the voltage's asymmetric increase and decrease matched experimental observations in rat layer 5 pyramidal neurons [39] (figure 3(b)). Replicating these *in vitro* results did not require NMDA synapses' voltage-dependence, consistent with previous models [39].

7. Top-down attention

To demonstrate that Neurogrid can help relate cognitive phenomena to biophysical mechanisms, an important goal of neuroscience, we replicated the effects of spatially selective top-down attention on the responses of

visual cortical neurons recorded in awake, behaving macaques [42]. Amplification of postsynaptic potentials by NMDA or sodium conductances in apical dendrites of pyramidal cells could contribute to this multiplicative interaction. The former may underlie spike-rate changes associated with feedback-driven figure-ground modulations in macaque primary visual cortex [43]. This evidence led us to hypothesize that NMDA receptors could also account for the multiplicative changes in spike rate observed in visual cortical neurons during spatially selective attention [42].

Unlike most computational models, Neurogrid models are subjected to heterogeneity. Once calibrated, the median parameter value across a Neurocore's population of silicon neurons matches the specified model parameter value. Whereas the variance of parameter values is determined by the fabrication process, resulting in a spike-rate distribution that spans a decade. For comparison, firing rates vary by over three decades across a population of cortical neurons [44]. Hence, showing that a proposed mechanism is robust to heterogeneity increases its biological plausibility.

We tested robustness of NMDA-mediated multiplicative gain to heterogeneity by simulating top-down and bottom-up interactions between a visual cortical area (V4) and a frontal cortical area (the Frontal eye field, FEF) on Neurogrid (figure 4(a)). V4 was modeled with an excitatory neuronal population (128 × 128) while FEF was modeled with an excitatory and an inhibitory population (both 128 × 128). Besides V4's excitatory population, which had dendrite compartments, all others had just soma compartments.

FEF activity, temporally sustained by local recurrent excitation and spatially constrained by local recurrent inhibition, represented the locus of attention in a 2D map of visual space, inspired by Ardid *et al*'s (1D) attention model [45]. To explore NMDA receptors' role in attentional modulation, unlike in the previous model, columnar feedback projections from FEF to V4 modulated activity of V4 neurons exclusively through NMDA synapses.

We discovered that heterogeneity in AMPA conductances realizes gradual gain modulation from abrupt NMDA threshold crossings. This heterogeneity (CV of 26%) caused dendrite compartments to cross NMDA's voltage-activation threshold at different visual drives (figure 4(b)). This distributed threshold-crossing increased the V4 population's spike-rate gradually with visual drive; it would otherwise have increased abruptly. This increase was steeper with FEF feedback, matching multiplicative changes with attention observed in macaque V4 [42] (figure 4(c)).

8. Conclusions

Conductance-based synapses, active membrane conductances, multiple dendritic compartments, spike back-propagation, and cortical cell types have been emulated in neuromorphic chips [37, 46–53]. Our focus here was on the next step: deploying these *in silico* neuronal components in multiscale modeling. We simulated cortical models with up to 1 M neurons and 8G synaptic connections using 1800-fold less energy per synaptic activation than a GPU [4, 8] (120 pJ versus 210 nJ)[8].

A hybrid analog–digital architecture made this possible. It implements hierarchical distribution and aggregation and it maps cortical circuitry columnarly. Following these principles the neocortex uses to minimize its wiring reduced traffic between cores greatly. This neuromorphic approach enabled Neurogrid to simulate multiscale neural models that integrate findings across five levels of experimental investigation energy- and time-efficiently. Importantly, Neurogrid achieves scale and efficiency without sacrificing biophysical detail, and thus truly supports multiscale modeling.

Neurogrid's neuronal model is sufficiently detailed to simulate various cortical cell types as well as active dendritic behavior and neuromodulatory effects. This combination of scale and detail enabled us to discover a biologically plausible gain mechanism for attentional modulation. This advance in simulation capacity could not only lead to a better understanding of neurological disorders such as ADHD, it could also help reverse-engineer the brain's hybrid analog–digital computational paradigm, which could lead to artificial intelligence systems with billions of units and trillions of parameters that consume watts instead of kilowatts.

Author contributions

BVB developed the synchrony, NMDA-spike, coincidence detection, and cortical neuron models. NAS developed the attention model. NNO and JJA developed calibration, mapping, interaction and visualization software. KB directed the research effort. BVB, NAS and KB wrote the manuscript.

[8] $(250\ \mathrm{J\ s^{-1}} \times 510\ \mathrm{s})/(24\mathrm{G\ syn} \times 25\ \mathrm{spk\ s^{-1}} \times 1\ \mathrm{s}) = 212\ \mathrm{nJ\ syn^{-1}\ spk^{-1}}$.

Conflict of interest

KB and NNO are co-founders and equity owners of Femtosense Inc. The remaining authors declare that they have no competing financial interests.

Acknowledgments

We thank E Kauderer-Abrams for calibration algorithm development; B Softky and H S Seung for assistance in editing the manuscript; and K Chin for administrative support. This work was supported by an NIH Director's Pioneer Award (DPI-OD000965) and an NIH/NINDS Transformative Research Award (R01NS076460).

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. Correspondence and requests should be addressed to KB (email: boahen@stanford.edu).

ORCID iDs

Ben Varkey Benjamin https://orcid.org/0000-0002-6840-7136
Kwabena Boahen https://orcid.org/0000-0003-2301-3309

References

[1] Abbott L F *et al* 2020 The mind of a mouse *Cell* **182** 1372–6
[2] Urai A E, Doiron B, Leifer A M and Churchland A K 2021 Large-scale neural recordings call for new insights to link brain and behavior (arXiv:2103.14662)
[3] Markram H 2012 The human brain project *Sci. Am.* **306** 50–5
[4] Knight J C and Nowotny T 2021 Larger GPU-accelerated brain simulations with procedural connectivity *Nat. Comput. Sci.* **1** 136–42
[5] Boahen K A 2000 Point-to-point connectivity between neuromorphic chips using address events *IEEE Trans. Circuits Syst. II* **47** 416–34
[6] Mahowald M 1994 *An Analog VLSI System for Stereoscopic Vision* (New York: Springer)
[7] Vogelstein R J, Mallik U, Culurciello E, Cauwenberghs G and Etienne-Cummings R 2007 A multichip neuromorphic system for spike-based visual information processing *Neural Comput.* **19** 2281–300
[8] Benjamin B V *et al* 2014 Neurogrid: a mixed-analog–digital multichip system for large-scale neural simulations *Proc. IEEE* **102** 699–716
[9] Knight B W 1972 Dynamics of encoding in a population of neurons *J. Gen. Physiol.* **59** 734–66
[10] Ermentrout G B and Kopell N 1986 Parabolic bursting in an excitable system coupled with a slow oscillation *SIAM J. Appl. Math.* **46** 233–53
[11] Latham P E, Richmond B J, Nelson P G and Nirenberg S 2000 Intrinsic dynamics in neuronal networks: I. Theory *J. Neurophysiol.* **83** 808–27
[12] Feinstein D I 1988 The hexagonal resistive network and the circular approximation *Technical Report CaltechCSTR:1988.cs-tr-88-07*
[13] Destexhe A, Mainen Z F and Sejnowski T J 1994 An efficient method for computing synaptic conductances based on a kinetic model of receptor binding *Neural Comput.* **6** 14–8
[14] Jahr C and Stevens C 1990 Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics *J. Neurosci.* **10** 3178–82
[15] Mainen Z F, Joerges J, Huguenard J R and Sejnowski T J 1995 A model of spike initiation in neocortical pyramidal neurons *Neuron* **15** 1427–39
[16] Davison A P 2008 PyNN: a common interface for neuronal network simulators *Front. Neuroinf.* **2** 11
[17] Gao P, Benjamin B V and Boahen K 2012 Dynamical system guided mapping of quantitative neuronal models onto neuromorphic hardware *IEEE Trans. Circuits Syst.* I **59** 2383–94
[18] Chklovskii D *et al* 2004 Synaptic connectivity and neuronal morphology two sides of the same coin *Neuron* **43** 609–17
[19] Binzegger T, Douglas R J and Martin K A C 2007 Stereotypical bouton clustering of individual neurons in cat primary visual cortex *J. Neurosci.* **27** 12242–54
[20] Merolla P, Arthur J, Alvarez R, Bussat J-M and Boahen K 2013 A multicast tree router for multichip neuromorphic systems *IEEE Trans. Circuits Syst.* I **61** 820–33
[21] Park J, Yu T, Joshi S, Maier C and Cauwenberghs G 2016 Hierarchical address event routing for reconfigurable large-scale neuromorphic systems *IEEE Trans. Neural Netw. Learn. Syst.* **28** 2408–22
[22] Choi T Y W, Merolla P A, Arthur J V, Boahen K A and Shi B E 2005 Neuromorphic implementation of orientation hypercolumns *IEEE Trans. Circuits Syst.* I **52** 1049–60
[23] Izhikevich E M 2007 *Dynamical Systems in Neuroscience* (Cambridge, MA: MIT Press)
[24] Mountcastle V B 1957 Modality and topographic properties of single neurons of cat's somatic sensory cortex *J. Neurophysiol.* **20** 408–34

[25] Hubel D H and Wiesel T N 1977 Ferrier lecture: functional architecture of macaque monkey visual cortex *Proc. R. Soc.* B **198** 1–59
[26] Stettler D D, Das A, Bennett J and Gilbert C D 2002 Lateral connectivity and contextual interactions in macaque primary visual cortex *Neuron* **36** 739–50
[27] Binzegger T, Douglas R, Martin K and Binzegger T 2004 A quantitative map of the circuit of cat primary visual cortex *J. Neurosci.* **24** 8441–53
[28] Andreou A G and Boahen K A 1996 Translinear circuits in subthreshold MOS *Analog Integr. Circuits Signal Process.* **9** 141–66
[29] Arthur J V and Boahen K A 2007 Synchrony in silicon: the gamma rhythm *IEEE Trans. Neural Netw.* **18** 1815–25
[30] Kauderer-Abrams E and Boahen K 2017 Calibrating silicon-synapse dynamics using time-encoding and decoding machines in 2017 *IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp 1–4
[31] Hodgkin A L and Huxley A F 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve *J. Physiol.* **117** 500–44
[32] Nowak L G, Azouz R, Sanchez-Vives M V, Gray C M and McCormick D A 2003 Electrophysiological classes of cat primary visual cortical neurons *in vivo* as revealed by quantitative analyses *J. Neurophysiol.* **89** 1541–66
[33] Connors B W, Gutnick M J and Prince D A 1982 Electrophysiological properties of neocortical neurons *in vitro J. Neurophysiol.* **48** 1302–20
[34] Gray C M and McCormick D A 1996 Chattering cells: superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex *Science* **274** 109–13
[35] McCormick D A, Connors B W, Lighthall J W and Prince D A 1985 Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex *J. Neurophysiol.* **54** 782–806
[36] Mainen Z F and Sejnowski T J 1996 Influence of dendritic structure on firing pattern in model neocortical neurons *Nature* **382** 363–6
[37] Arthur J V and Boahen K A 2011 Silicon-neuron design: a dynamical systems approach *IEEE Trans. Circuits Syst.* I **58** 1034–43
[38] Schiller J, Major G, Koester H J and Schiller Y 2000 NMDA spikes in basal dendrites of cortical pyramidal neurons *Nature* **404** 285–9
[39] Stuart G J and Häusser M 2001 Dendritic coincidence detection of EPSPs and action potentials *Nat. Neurosci.* **4** 63–71
[40] Lavzin M, Rapoport S, Polsky A, Garion L and Schiller J 2012 Nonlinear dendritic processing determines angular tuning of barrel cortex neurons *in vivo Nature* **490** 397–401
[41] Xu N-l, Harnett M T, Williams S R, Huber D, O'Connor D H, Svoboda K and Magee J C 2012 Nonlinear dendritic integration of sensory and motor input during an active sensing task *Nature* **492** 247–51
[42] McAdams C J and Maunsell J H R 1999 Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4 *J. Neurosci.* **19** 431–41
[43] Self M W, Kooijmans R N, Super H, Lamme V A and Roelfsema P R 2012 Different glutamate receptors convey feedforward and recurrent processing in macaque V1 *Proc. Natl Acad. Sci.* **109** 11031–6
[44] Peyrache A *et al* 2012 Spatiotemporal dynamics of neocortical excitation and inhibition during human sleep *Proc. Natl Acad. Sci.* **109** 1731–6
[45] Ardid S, Wang X-J and Compte A 2007 An integrated microcircuit model of attentional processing in the neocortex *J. Neurosci.* **27** 8486–95
[46] Mahowald M and Douglas R 1991 A silicon neuron *Nature* **354** 515–8
[47] Van Schaik A 2001 Building blocks for electronic spiking neural networks *Neural Netw.* **14** 617–28
[48] Hynna K M and Boahen K 2007 Thermodynamically equivalent silicon models of voltage-dependent ion channels *Neural Comput.* **19** 327–50
[49] Yingxue Wang Y and Shih-Chii Liu S-C 2011 A two-dimensional configurable active silicon dendritic neuron array *IEEE Trans. Circuits Syst.* I **58** 2159–71
[50] Benjamin B V, Arthur J V, Gao P, Merolla P and Boahen K 2012 A superposable silicon synapse with programmable reversal potential *2012 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 771–4
[51] Ramakrishnan S, Wunderlich R, Hasler J and George S 2013 Neuron array with plastic synapses and programmable dendrites *IEEE Trans. Biomed. Circuits Syst.* **7** 631–42
[52] Chicca E, Stefanini F, Bartolozzi C and Indiveri G 2014 Neuromorphic electronic circuits for building autonomous cognitive systems *Proc. IEEE* **102** 1367–88
[53] Schemmel J, Kriener L, Müller P and Meier K 2017 An accelerated analog neuromorphic hardware system emulating NMDA-and calcium-based nonlinear dendrites *2017 Int. Joint Conf. on Neural Networks (IJCNN)* pp 2217–26