Neuron

CellPress

NeuroView

Nobel honors for John Hopfield, who ushered attractor dynamics into neuroscience

David Kleinfeld[1,2,*]
[1]Department of Physics, University of California, San Diego, San Diego, CA 92093, USA
[2]Department of Neurobiology, University of California, San Diego, San Diego, CA 92093, USA
*Correspondence: dkleinfeld@ucsd.edu
https://doi.org/10.1016/j.neuron.2024.11.002

John Hopfield's model on collective computation linked the recall of memories with interactions and dynamics associated with disordered magnetic systems. Insights from Hopfield's work catalyzed formulations that link the dynamics and emergent properties of recurrently connected generic neurons with the functional properties and signaling observed from brain circuits.

The Nobel Prize in Physics was awarded jointly to John Hopfield and Geoffrey Hinton. The Nobel committee notes the importance of Hopfield's contribution to " … inventions that enable machine learning with artificial neural networks …" Here, I provide a perspective on the influence of his work on neuroscience and on physicists who entered neuroscience because of him. I had the privilege to overlap with Hopfield for a decade at the former AT&T Bell Laboratories, Murray Hill, and to get to know him personally. Hopfield enjoys referring to himself as a dilettante, which correctly reflects his interest in many scientific areas yet belies his deep knowledge and significant contributions to those areas.

Hopfield began his scientific career in solid-state physics.[1] The subject of his 1958 thesis work, under the supervision of Albert Overhauser, was on electromagnetic excitation modes in solids. After completing his doctoral studies, Hopfield became a member of the theoretical physics department at Bell, only to leave after 2 years for the University of California, Berkeley (1961–1964), then onto Princeton University (1964–1980), the California Institute of Technology (1980–1997), and finally back to Princeton (1997). All the while, he remained a consultant at Bell. While Hopfield's scientific career was initially focused on the properties of materials and the interaction of light with matter, he showed a growing interest in the physics of life. Two of the papers that resulted from this shift, both published in 1974, are now classics. The first is on thermally assisted electron tunneling among reduction-oxidation centers in large biomolecules.[2] Electrons are always in equilibrium on the timescale of nuclear motion, and thus, thermal motion of large molecules provides the limit to electron transfer. This leads to quantum mechanical effects at room temperature in some large biomolecules, as seen in the initial step of photosynthesis. The second paper is on error correction in "highly specific biosynthetic reactions."[3] This led to a qualitative jump in the understanding of the replication and transcription of DNA. Hopfield showed how the precision of molecular reactions could be increased without bound through the consumption of energy. Thus, errors would be reduced to arbitrarily small levels. Hopfield's work in molecular biophysics was widely appreciated by biologists and physicists alike.

Despite the acclaim for his work in molecular biophysics, the appreciation of the neuroscience community to Hopfield's eponymous model was less than universal. In his 1982 paper titled "Neural networks and physical systems with emergent collective computational abilities,"[4] Hopfield proposed a deceptively simple and effective model network to address two predominant attributes of memory by brains. First, brains store memories and then recall those memories based only on partial information. For example, the seed "To … or not … be, that … " leads to the completed prose "To be or not to be, that is the question." Hopfield networks will complete a pattern in the sense that an input to the network of only part of a memory can seed the recall of the complete memory (Figure 1A). Second, the process of recall must occur as a parallel, robust process. This means that all model neurons continuously update their output so that, in some sense, the output "flows" to a memory state. This is the notion of an attractor. While precedents are found in the literature, e.g., by Shun-Ichi Amari in 1972, it was Hopfield who introduced a formulation that was tied to statistical mechanics and provided a ready path to general aspects of computation.

Hopfield's model contains only one type of neuron with either excitatory or inhibitory connections. This stands in opposition to the extensive diversity of neuronal cell types, each with different rules for connectivity and spiking activity, that are found in brains. Thus, Hopfield set up a dichotomy for the emergence of computational properties of nervous systems: can they arise solely from the collective dynamics of simple elements, as implied by his model, or is the vast breadth of biophysical properties found in neurons and synapses of fundamental importance?

Emergence of rich structure from underlying simplicity

The output of a Hopfield network is described in terms of states—that is, a vector that lists which neurons are active and which are quiet. This is not just an abstract notion. Concurrent recordings from

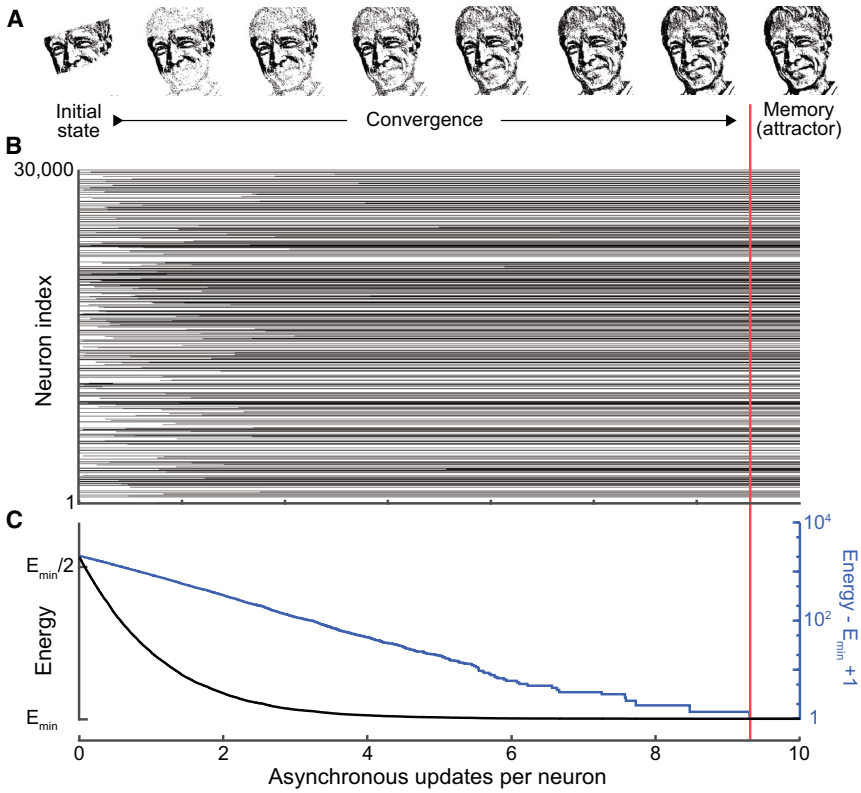

Figure 1. Pattern completion in a Hopfield network

Forty thousand threshold units were used as neurons, and 1,000 memories were stored. One neuron at a time was randomly selected, and its output was updated if it differed in sign from its input.

(A) States arranged as 200-by-200-pixel images. An incomplete memory forms the initial state. The final state matches the closest memory.

(B) The output of 30,000 of the 40,000 neurons shown as a raster plot, where black is $S_i = +1$ and white is $S_i = -1$. Each time step corresponds in N potential updates by Glauber dynamics.

(C) Energy as a function of update monotonically decreased as the network relaxed from the initial state to the nearest memory. The red line corresponds to convergence to the final stable state, a memory. The blue scale is logarithmic to highlight small changes in energy.

many sites in real nervous systems tend to exhibit persistent patterns of spiking activity that suggest the utility of a description in terms of states.

Hopfield showed that the sign and strength of synaptic weights, i.e., a switchboard of connections between neurons, could be chosen so that a network could store and recall a multiplicity of memories. Starting from some initial pattern, the dynamics of the network would drive the output to the closest memory. This process has limits; as more and more memories are summed together, the ability to recall memories is degraded because of incoherent interference between different stored memories and eventually lost.

A review of the mathematics of the Hopfield network reveals how stability of a memory is maintained, the role of quenched disorder in limiting the number of memories that can be stored, and the flow of network dynamics on an energy landscape.

Dynamics

The state of the network is given by the output of each neuron, which is taken to be +1 for active and −1 for quiescent. The state is represented by a vector, denoted by $\boldsymbol{S}$. This takes the form $\boldsymbol{S} = (+1\ +1\ -1\ -1\cdots)$, for example, when the first two neurons are spiking, the next two are quiescent, etc. The neurons interact through synaptic connections, a symmetric set of connections in this idealized case, which is represented by a switchboard or matrix, denoted $\boldsymbol{W}$. Element W_{ij} in the matrix connects the output (axon) of neuron "j" to the input (dendrite) of neuron "i." The recurrent dynamics of the network follows Glauber dynamics: neurons asynchronously update their output based on the summed input from all neighbors, i.e.,

$$S_i(t+\Delta t) = sign\left\{\sum_{j\neq i}^{N} W_{ij}S_j(t) + I_i^{ext}(t) - \theta_i\right\}$$

where $I_i^{ext}(t)$ is the external input, θ_i is the rheobase for spiking, sign$\{x\}$ imposes a threshold input/output relation, N is the number of neurons, and Δt is the time constant of a neuron. Absent an external input and taking $\theta_i = 0$ for all neurons as befits the case of random patterns, we have

$$S_i(t+\Delta t) = \text{sign}\left\{\sum_{j\neq i}^{N} W_{ij}S_j(t)\right\}$$

where the term in brackets is solely the recurrent input to the neuron.

Recall with one memory

Consider the case of just one pattern, $\boldsymbol{S} = \xi$, that we want to memorize to motivate the rule for forming the connection weights. The condition for this pattern to correspond to a stable state is just

$$\xi_i = \text{sign}\left\{\sum_{j\neq i}^{N} W_{ij}\xi_j\right\}$$

since the update rule produces no changes. It is easy to verify that the outer product rule

$$W_{ij} \propto \xi_i\xi_j$$

satisfies stability since ξ_j^2 is always one. This storage rule corresponds to "neurons that are co-active strengthen their excitatory synaptic connections," while "neurons that are anti-correlated strengthen their inhibitory synaptic connections." If fewer than half of the elements of the initial state are incorrect, i.e., $\xi_i = -\xi_i$, they will be overwhelmed in the sum by the majority that are correct. Then sign$\{\sum_{j\neq i}^{N} W_{ij}\xi_j\}$ will still yield ξ_i. Thus, an initial state near memory ξ will flow to the memory state ξ, and the network will have successfully performed pattern completion.

Recall with many memories

How does the Hopfield network perform pattern completion to the closest memory state when many memories are stored in the network? The simplest approach is to form the synaptic weights by summing together the outer products of each of the memories, denoted ξ^{ν}, where "ν" indexes

the memory and there is a total of P memories. This corresponds to

$$W_{ij} = \frac{1}{N}\sum_{\nu=1}^{P}\xi_i^\nu\xi_j^\nu$$

and may be viewed as a formalization of the Hebb rule. The connection matrix is symmetric, i.e., $W_{ji} = W_{ij}$. To the extent that the memory states remain the ground states of the network, and for the case of low levels of random (thermal) noise, the state of the network will relax from an initial state to the nearest memory (Figure 1B).

Limitations to storage

How many memories can be embedded in a network before the overlap among memories leads to faulty pattern completion? The answer leads to the notion of quenched disorder, i.e., noise that is frozen into the network because of the storage of multiple memories. Let $\boldsymbol{S} = \xi^1$, one of the stored memory states. After plugging in terms, the input to neuron i is

$$\sum_{j\neq i}^{N} W_{ij}S_j = \left(1-\frac{1}{N}\right)\xi_i^1 + \frac{1}{N}\sum_{\nu\neq 1}^{P}\xi_i^\nu\sum_{j\neq i}^{N}\xi_j^\nu\xi_j^1.$$

In the limit of a large network, the first term on the righthand side leads to stability, while the second term corresponds to disorder and potentially leads to instability. The mean input is

$$\text{mean}\left\{\sum_{j\neq i}^{N} W_{ij}S_j\right\} = \xi_i^1.$$

The memory ξ^1 is stable if the magnitude of the second term is smaller than 1. As in the example of a single memory, a small fraction of neuronal outputs that are different from a memory will be corrected, and an initial state near to ξ^1 thus flows to ξ^1. Yet, the memory ξ^1 can become unstable if the magnitude of the second term changes the sign of the output from that of ξ_i^1 to $-\xi_i^1$. The variance of this term, in the limit of large P and N, is just the fraction of memories to neurons, i.e.,

$$\text{variance}\left\{\sum_{j\neq i}^{N} W_{ij}S_j\right\} = \frac{P}{N}.$$

This variance is constant in time and defines the quenched disorder. It limits the number of memories that can be stored. The severe constraint that the network will produce, at most, one bit of error, i.e., one neuron's output in only one of the memory states that can be incorrect, leads to the statistical bound $P/N < 0.25/lnN$. This bound is relaxed for the case of near but imperfect recall. As discussed later, the network performs pattern completion as long as $P/N < 0.14$. There is a phase transition from a region of memory retrieval with $P/N < 0.14$ to one with catastrophic forgetting with $P/N > 0.14$.

Attractors and the energy landscape

When the state of the network is initialized near a memory, the neuronal activity flows through many of the 2^N possible patterns until the memory is reached (Figures 1A and 1B). A lasting contribution of Hopfield was to introduce an energy landscape into neural network theory to conceptualize this flow.[4] The landscape spans all N dimensions, depends on the synaptic weights, and defines a hilly surface with deep pits in the valleys. The central property of an energy function is that it either decreases or remains constant as the output of the network evolves according to the update rule. The flow ends at an attractor, which is defined by the deep pits and corresponds to a memory so long as the network is in phase ($P/N < 0.14$) with retrieval. An energy function exists only if the connection strengths are symmetric. While symmetry is an unreasonable assumption for brain circuits, experimental data show that symmetric synapses occur more than expected by chance and that attractor dynamics can hold close to minima even for the case of weak symmetric interactions.

By analogy with the interaction energy in lattices of magnetic spins, the energy of each state $\boldsymbol{S}$ is defined by

$$\text{Energy} = -\sum_{i}^{N}\sum_{j\neq i}^{N} S_i W_{ij} S_j$$

The change in energy for a change in output at neuron i is

$$\Delta\text{Energy} = -\Delta S_i \sum_{j\neq i}^{N} W_{ij}S_j$$

where $\Delta S_i = S_i(t+\Delta t) - S_i(t)$. The energy is constant when the state remains unchanged and decreases for any change in state, since the signs of ΔS_i and the input $\sum_{j\neq i}^{N} W_{ij}S_j(t)$ are the same. Changes in the state of the network continue until a local minimum in energy, or a pit in the landscape, is reached (Figure 1C), for which $\Delta S_i = 0$ for all values of i. Neuronal dynamics in real brains is, of course, more complex than a flow through many patterns until a memory state is reached. Yet, Hopfield's abstraction provides a starting point to characterize dynamics for any neuronal computation, from memory recall to motor control. Further, it had a direct impact on the formulation of the "Boltzmann Machine" by Geoffrey Hinton and Terrence Sejnowski, as reported in 1983.

Embrace by physicists

Hopfield's eponymous model[4] was published at a time of intense interest in the magnetic properties of disordered systems, both theoretically and experimentally, in terms of dilute magnetic alloys. In fact, William Little had pointed out the potential connection of recurrent neural network to spin systems in 1970s. Thus, Hopfield's model was rapidly absorbed and analyzed by the physics community. Daniel Amit, Hanoch Gutfreund, and Haim Sompolinsky saw an opportunity to solve the thermodynamics of a system that exhibited quenched disorder, i.e., variability in synaptic strengths caused by the interference of stored patterns, as well as fast, random noise. They used the replica method of Samuel Edwards to derive the different phases of the output of the Hopfield network. The phase diagram, published in 1985, is rich.[5] It exhibits regions of perfect recall, as expected from the statistical analysis, as well as regions with recall close to the memory states, regions where the memories are no longer the most stable patterns, and an ergodic region where the output of the network no longer has a relation to any of the memories. In 1988, Elizabeth Gardner showed how the learning rule could be altered to store as many memories as there are neurons. As Hopfield hypothesized, many of the assumptions that violated biological reality, such as all-to-all connectivity, could be softened, and the performance of the network to complete patterns is merely degraded.

As time progressed, many of the theoretical efforts became driven by specific experimental observations, particularly those that involved invariance to stimulus parameters. Hopfield shifted his attention toward concentration-invariant odor

recognition.[6] Sompolinsky and I formed a connection between attractor dynamics and central pattern generators for locomotion.[7] In 1995, Sompolinsky conceived an extension of the Hopfield model to a continuous "ring" of attractor states.[8] This model highlights the competition between feedforward and recurrent connections in determining the dynamics of a network and was later found to capture the representation of heading relative to a landmark. Two decades later, observations by Vivek Jayaraman and Johannes Seelig at HHMI Janelia demonstrated that the ring model captured the neural computation of heading within the central complex in flies. In 1996, Sebastian Seung proposed the "line" attractor to understand the stability of eye movement and neural integrators.[9] These and related successes suggest a close correspondence between the minimalist approach of Hopfield and biological reality, yet a critique is that the matches only apply for behaviors and underlying circuits with few degrees of freedom.

Hopfield and neuroscience at Murray Hill and beyond

In 1981, Hopfield convinced then-Bell president Arno Penzias (a 1978 Nobel laureate in Physics) to take a position in neuroscience given the potential impact on computing and algorithms. Hopfield further argued that one needed a thriving experimental, as well as theoretical, effort to make progress, even if the goal for AT&T was advancement in computation. John Connor and Alan Gelperin joined Bell in 1982. David Tank joined a year later. Follow-up work by Hopfield in 1984 relaxed the form of the nonlinear input/output relation from a threshold to a smoothly saturating function, and Hopfield and Tank used this as a starting point to extend recurrent networks to problems in computation and optimization.[10]

Hopfield's insistence on an experimental program in neuroscience led to new technologies and findings. A particularly fruitful circle of discovery that was catalyzed by Hopfield's scientific and institutional roles concerns intracellular ionic calcium (Ca^{2+}) imaging and the evidence for a ring attractor. Connor published the first paper on digital imaging of space-time patterns of intracellular Ca^{2+} in neurons in 1986, starting with cells in culture. He utilized Roger Tsien's then newly developed fluorescence-based ratiometric Ca^{2+} indicators (Tsien shared the 2008 Nobel Prize in Chemistry), which incorporated a prodrug method to trap the indicator in a cell, and a cooled charge-coupled device (CCD) as a low-noise imager (Bell scientists Willard Boyle and George Smith shared half of the 2009 Nobel Prize in Physics for inventing the CCD). Tank then teamed with Connor and Rodolfo Llinas (NYU) and extended Ca^{2+} imaging to the cerebellar brain slice in 1988. However, single-cell *in vivo* measurements in the mammalian brain were all but impossible because of light scattering by overlying brain tissue. The answer to this conundrum was the introduction of two-photon laser-scanning microscopy (TPLSM) to neuroscience. Winfried Denk, whose graduate work with Watt Webb included the invention of TPLSM, joined Bell in 1993 and pioneered this application. This technique permits excitation of fluorophores deep in tissue when the scattering of light is predominantly in the forward direction, as occurs in neocortex.

Denk first used TPLSM to make intracellular measurements of Ca^{2+} from spines in brain slices, working separately with Llinas and with then-postdoctoral fellow Rafael Yuste in 1995; the latter work demonstrated the signature of pre- and postsynaptic coincident activity. With functional imaging now feasible and following Denk's and my 1994 observation that anatomically labeled neurons in rat vibrissa cortex could be imaged *in vivo*, a team of Denk, fellow Karel Svoboda, Tank, and I measured Ca^{2+} dynamics within individual neurons in rat vibrissa cortex in 1997. Just a few years later, Jing Wang, a fellow with Gelperin, brought *in vivo* TPLSM imaging to the fly brain in a second fellowship with Richard Axel (a 2004 Nobel laureate in Physiology or Medicine) at Columbia. The year 2003 marked the introduction of TPLSM Ca^{2+} imaging to study networks of neurons: Axel, Wang, and their colleagues introduced genetic expression of a Ca^{2+}-sensitive fluorescent protein in the fly brain, while Arthur Konnerth's Munich group introduced multicellular labeling in the mouse brain. Finally, three decades after Connor's initial measurements, the circle was closed by Jayaraman's adoption of TPLSM Ca^{2+} imaging and his 2015 report of evidence for a ring attractor in the fly brain.

Epilogue

Hopfield gifted us with insights that pulled physics into neuroscience. His model laid bare an unsettled dichotomy between minimalist models and complexity in nervous systems. Time will tell how each of these views will inform us about how intelligence can emerge from the interactions among neurons.

ACKNOWLEDGMENTS

I am grateful to Pantong Yao for preparing Figure 1 and to Johnatan Aljadeff, Winfried Denk, Beth Friedman, Andreas Herz, Sebastian Seung, and Haim Sompolinsky for discussions and/or their critique of early versions of the text. The author is funded by the NIH BRAIN Initiative (grants U19 NS123717 and U19 NS137920), NIBIB (grant U24 EB028942), and NINDS (grant R35 NS097265).

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. Hopfield, J.J. (2014). Whatever happened to solid state physics? Annu. Rev. Condens. Matter Phys. *5*, 1–13.
2. Hopfield, J.J. (1974). Electron transfer between biological molecules by thermally activated tunneling. Proc. Natl. Acad. Sci. USA *71*, 3640–3644.
3. Hopfield, J.J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc. Natl. Acad. Sci. USA *71*, 4135–4139.
4. Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558.
5. Amit, D.J., Gutfreund, H., and Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. Phys. Rev. Lett. *55*, 1530–1533.
6. Hopfield, J.J. (1991). Olfactory computation and object perception. Proc. Natl. Acad. Sci. USA *88*, 6462–6466.
7. Kleinfeld, D., and Sompolinsky, H. (1988). Associative neural network model for the generation of temporal patterns: Theory and application to central pattern generators. Biophys. J. *54*, 1039–1051.
8. Ben-Yishai, R., Bar-Or, R.L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. Proc. Natl. Acad. Sci. USA *92*, 3844–3848.
9. Seung, H.S. (1996). How the brain keeps the eyes still. Proc. Natl. Acad. Sci. USA *93*, 13339–13344.
10. Hopfield, J.J., and Tank, D.W. (1985). "Neural" computation of decisions in optimization problems. Biol. Cybern. *52*, 141–152.