Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Thalamo-koniocortical simulations reveal hardware-oriented principles for Transformer-like architectures

**Francisco Javier Peláez**

francisco.pelaez@ufabc.edu.br

Universidade Federal do ABC

**Carlos Fellip Rabadan Braga**

Universidade Federal do ABC

**Ricardo Tiosso Panassiol**

Universidade de São Paulo

**Diego Andina**

Universidad Politécnica de Madrid

**Research Article**

**Keywords:** thalamus, koniocortex, Hebbian crosstalk, attention, winner-take-all, softmax, Transformer architectures

**Posted Date:** November 17th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11751492/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Thalamo-koniocortical simulations reveal hardware-oriented principles for Transformer-like architectures.

*Francisco Javier Ropero Peláez[1,2], Carlos Fellip Rabadan Braga[1], Ricardo Tiosso Panassiol[1,2] & Diego Andina[3]
francisco.pelaez@ufabc.edu.br; carlosfrbraga@gmail.com; ricardotiosso@gmail.com; d.andina@upm.es

[1] Center for Mathematics, Computing, and Cognition, Federal University of ABC, São Paulo, Brazil; [2] Department of Neuroscience and Behavior, Institute of Psychology, University of São Paulo, Brazil
[3] Grupo de Automatización en Señal y Comunicaciones, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain

## Abstract

Transformer architectures process distributed representations through explicit matrix operations, whereas biological circuits compute through local electrical and synaptic dynamics. We ask whether Transformer-like functions can emerge in three separately simulated phenomenological models of a thalamo-koniocortical pathway. The thalamic model converts overlapping sensory activity into less-overlapping components and reconstructs incomplete inputs; the koniocortical model transforms these components into sparse, stable, pattern-selective events; and the cortical model binds ordered events into frame-like representations and temporal sequences. Mathematical analysis links Hebbian retrieval and Transformer attention through similarity-dependent weighting. These correspondences suggest that algorithmic operations may arise from the physical dynamics of interacting units, motivating hardware-oriented and neuromorphic alternatives. Although simulated separately, the models define a testable pathway from sensory recoding to sparse representation, temporal binding and competitive selection.

**Keywords:** thalamus; koniocortex; Hebbian crosstalk; attention; winner-take-all; softmax; Transformer architectures

## 1. Introduction

Transformer architectures achieve context-sensitive computation through vector representations, normalization, attention, iterative recoding and competitive weighting [1–3], implemented as explicit high-dimensional matrix algebra. Biological systems face related demands through local synaptic interactions, inhibition, recurrence, oscillations, threshold nonlinearities, intrinsic excitability and population codes [4–7]. We therefore ask whether biologically plausible circuits can realize functionally comparable operations through local dynamics and suggest hardware-oriented alternatives [8–10]. We examine three phenomenological models of thalamic, koniocortical and cortical processing. The comparison is functional, not anatomical: we do not propose that the brain is a Transformer, but test whether normalization, recoding, interference reduction, competitive selection, binding and iterative refinement can emerge through mechanisms distinct from artificial matrix procedures [1–7].

Figure 1 maps a distributed sensory pattern (the letters D–N–A) through thalamic normalization and component extraction, sparse koniocortical representation and pyramidal binding. A Transformer similarly converts vectorized inputs into context-dependent representations before weighting or selection [1–3]. Despite different substrates, both must stabilize representations, reduce interference and select alternatives; we examine these functions through relay–reticular dynamics, koniocortical competition and pyramidal integration [11–13].

Our framework starts from a limitation of local Hebbian learning [14]: overlap among stored patterns introduces destructive crosstalk during retrieval [15,16], so representational overlap must be reduced before reliable binding and retrieval.

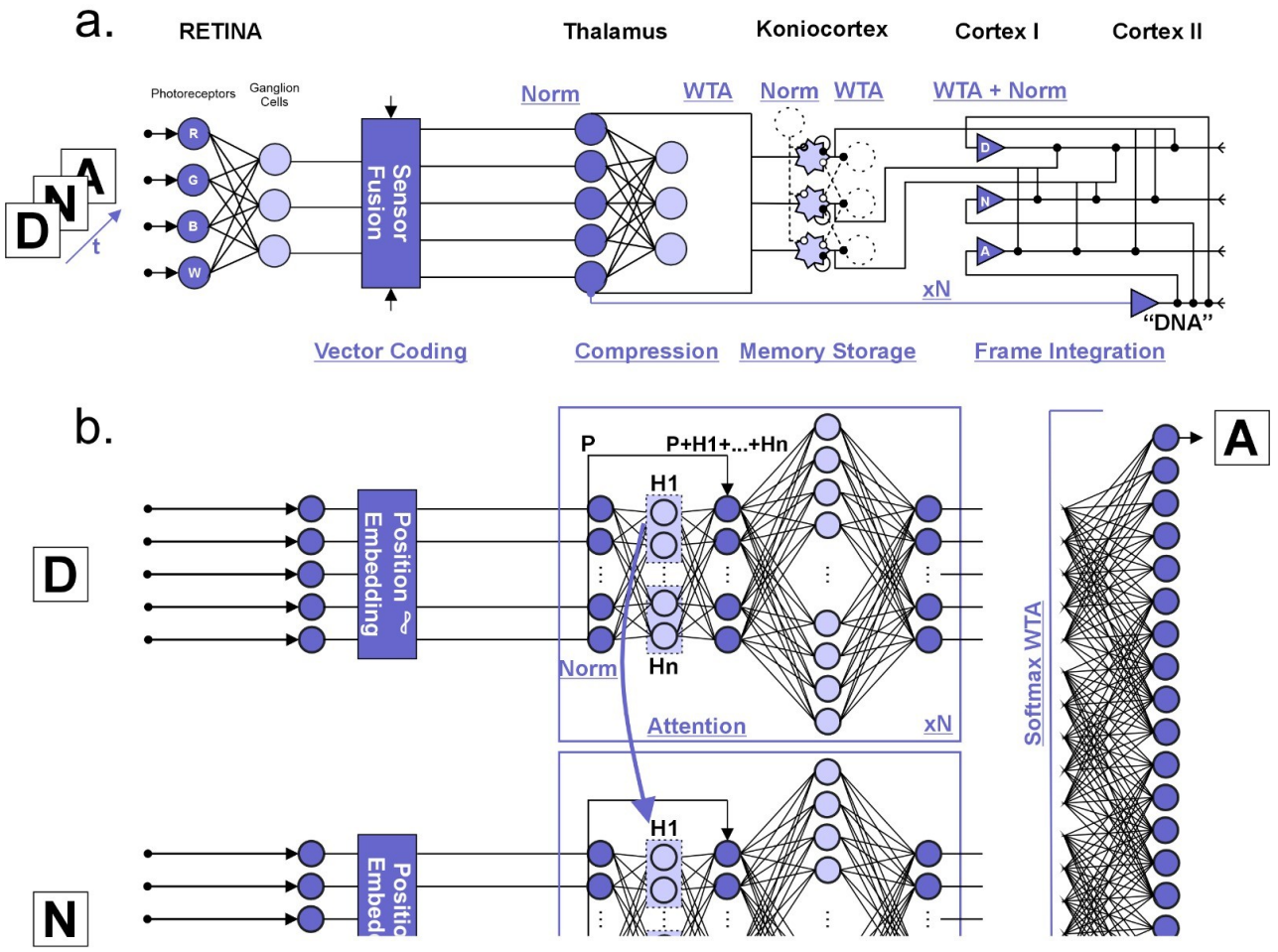



Figure 1. Conceptual roadmap of the biological and Transformer pathways. (a) A sequence of letter-like sensory patterns (D–N–A) enters the biological pathway. Retinal population coding, ganglion-cell compression and sensor fusion form a distributed representation processed by thalamic normalization and WTA-like component selection, koniocortical normalization, competition and memory storage, and cortical integration into D, N and A frame-like representations and temporal sequences. Cortex I and II denote successive competitive and frame-integration stages; ×N indicates recurrent thalamocortical refinement. (b) In the Transformer pathway, D and N tokens receive positional embeddings, normalization, multi-head attention (H1–Hn) and feedforward recoding. Repetition (×N) progressively updates representations before softmax-based selection completes D–N–A with A. Labels identify the main computational landmarks: vector coding, normalization, compression, attention, recurrence, memory storage, frame integration and output selection. The alignment is functional rather than substrate-level.

The thalamic model addresses this by interpreting thalamic circuitry as a folded autoassociative architecture [11,12]. Classical autoassociation reconstructs inputs through a hidden representation supporting compression and completion [17,18]; folding this architecture assigns relay neurons the combined input/output/comparison role and reticular neurons a competitive bottleneck [11,12]. The model thereby normalizes input, extracts component-like representations and reconstructs patterns, with a principal-component-like interpretation that generates more independent coordinates for subsequent Hebbian learning [11–13,19].

Two cortical stages follow. Koniocortical circuits transform distributed thalamic components into sparse, pattern-selective spiny-stellate (S)-neuron responses via inhibition, synaptic plasticity and intrinsic plasticity [13]; a pyramidal-unit-like module then binds temporally ordered S-neuron events into frame-like representations. Because the modules were developed and simulated independently, their interconnection remains a physiologically grounded integration hypothesis rather than an integrated thalamocortical simulation.

This progression also reframes crosstalk. In Hebbian associative models, overlap destabilizes pattern formation and retrieval [15,16], whereas after separation and binding, similarity-dependent interactions among stable frames may support context. The mathematical Appendix formalizes a structural relation between Hebbian retrieval and Transformer attention: stored representations contribute according to similarity with an input or query [1,15,16]. In associative memory these unregulated contributions appear as crosstalk; in attention, learned projections, scaling and softmax convert them into controlled contextual weighting [1]. Competitive selection provides another correspondence. Transformer softmax emphasizes dominant alternatives within a normalized distribution [1–3]. In the thalamic reticular model, a related selective effect emerges from synchronized oscillation, nonlinear amplification, threshold crossing, feedback inhibition and refractoriness [11]; in the koniocortex-like network it emerges from inhibition, steep activation functions and plasticity [13]. These mechanisms do not reproduce softmax, but show that contrast enhancement and winner selection can arise from local circuit dynamics rather than external maximum search.

This is relevant to hardware-oriented NeuroAI because biological computation is embodied in the physical evolution of neuronal and synaptic states [8–10]. Divisive inhibition, shunting or circuit-level gain control can implement normalization [5,20], while coupled oscillatory and threshold dynamics can implement competitive selection. Such mechanisms suggest architectures in which substrate dynamics perform some operations now executed as sequential numerical instructions. We identify these principles but do not quantify electronic latency, energy or scalability.

Finally, Transformers refine representations by stacking blocks [1], whereas thalamocortical systems may revisit, correct and stabilize them through recurrence [4,6,7]. Artificial depth and biological recurrence may thus offer alternative routes to progressive refinement.

Accordingly, we place the demonstrated operations of the three models in a common NeuroAI framework: thalamic normalization, folded autoassociative recoding, orthogonalization, competitive component extraction and pattern completion; koniocortical sparse learning; cortical binding and sequencing; the formal relation between Hebbian crosstalk and attention; and hardware implications of oscillatory winner-take-all dynamics. We explicitly separate simulated results from integration hypotheses and do not claim either that the brain implements a Transformer or that the present models form a complete integrated architecture.

## 2. Methods

### 2.1 General dynamical-systems framework

Simulations used phenomenological neurocomputational models in discrete time to test whether biologically inspired thalamic, koniocortical and cortical circuits could implement normalization, recoding, orthogonalization, competitive selection and binding, rather than reproduce detailed neuronal biophysics.

At iteration t, the activity state was:

$$O(t)=\left[O_1(t),O_2(t),\cdots,O_n(t)\right]^T, \quad (1)$$

where $O_i(t)\in[0,1]$ is the normalized activity, firing-rate proxy or firing probability of neuron i.

Transitions used the synaptic matrix

$$W(t), \quad (2)$$

yielding the net-input vector

$$N(t)=W(t)O(t-1). \quad (3)$$

or, for neuron i,

$$N_i(t)=\sum_{j=1}^{n} W_{ij}(t)O_j(t-1). \quad (4)$$

Positive and negative weights represented excitatory and inhibitory projections, including feedback and lateral inhibition.

Sensory patterns were imposed by clamping predefined input neurons, not by adding an external transition term. For sensory set S and pattern value $I_i^p$,

$$O_i(t)=I_i^p,\ i\in S. \quad (5)$$

All other neurons followed:

$$O_i(t)=f_i\left(N_i(t),s_i(t)\right),\ i\notin S. \quad (6)$$

### 2.2 Neuronal activation function

Unless otherwise specified, neurons used a sigmoidal activation function:

$$O_i(t)=\frac{1}{1+e^{-k_i(N_i+0.5-2s_i)}}, \quad (7)$$

where $N_i(t)$ is net input, $k_i$ the sigmoid slope, and $s_i(t)$ an excitability shift. The effective firing threshold was

$$\theta_i(t)=2s_i(t)-0.5. \quad (8)$$

Increasing $s_i(t)$ reduces excitability; decreasing it increases excitability. Competitive neurons used steeper sigmoids to amplify small input differences into winner-take-all-like responses.

Intrinsic plasticity updated si according to

$$s_i(t+1)=\frac{\upsilon O_i(t)+s_i(t)}{\upsilon+1}, \quad (9)$$

where υ sets the adaptation rate. Repeated activity therefore reduced excitability, preventing persistent dominance and allowing other units to compete.

A binary mask $P_i$ restricted intrinsic plasticity to selected neuronal populations:

$$s_i(t+1)=\begin{cases}\frac{\upsilon O_i(t)+s_i(t)}{\upsilon+1}, P_i=1\\ s_i(t), P_i=0.\end{cases} \quad (10)$$

Thus, non-plastic populations, including inhibitory units in selected simulations, retained fixed thresholds.

### 2.3 Synaptic plasticity and structural constraints

Synaptic plasticity used an incremental probabilistic presynaptic rule:

$$\Delta W_{ij}(t)=\xi O_j(t-1)\left[O_i(t)-W_{ij}(t)\right], \quad (11)$$

where ξ is the learning rate; active presynaptic connections strengthened toward the postsynaptic response and weakened when that response was low.

This derives from

$$W_{ij}\approx P(o_i/o_j). \quad (12)$$

where lowercase $o_i$ and $o_j$ are binary postsynaptic and presynaptic activity realizations, respectively [21]. Incremental updating allowed gradual adaptation.

To prevent plasticity from creating anatomically absent connections, a structural mask C preserved circuit architecture:

$$C_{ij}=\begin{cases}1, \textit{if the connection } j\to i \textit{ is allowed}\\ 0, \textit{otherwise}\end{cases} \quad (13)$$

After each update,

$$W(t+1)\leftarrow W(t+1)\odot C, \quad (14)$$

where ⊙ denotes element-wise multiplication. Additional masks specified excitatory/inhibitory projections, plastic/fixed synapses, and neurons with/without intrinsic plasticity.

### 2.4 Divisive normalization

Normalization made responses depend on relative input structure rather than absolute magnitude. Most inhibition was represented by negative weights in W, but GABA-A-like divisive normalization [5,20] required an explicit operation:

$$x_{norm}=\frac{x}{\epsilon+\|x\|_1}, \quad (15)$$

with

$$\|x\|_1=\sum_j x_j \ \vee, \quad (16)$$

where ϵ>0 prevents division by zero. Explicit $L_1$-like normalization was applied only when required; subtractive, feedback and lateral inhibition remained embedded in W.

### 2.5 Parameter initialization and plasticity settings

Initial plastic weights were sampled as $w_{ij}(0)=0.01+0.1$ rand, with rand uniform on [0,1]; fixed connections were typically set near 0.85. Fixed weights included successive apical-dendritic compartments, non-plastic inhibitory projections and autapses. Among reticular inhibitory projections, only synapses onto thalamic relay neurons were plastic, representing reticulothalamic facilitation.

For neurons with intrinsic excitability, $s_i(0)=0.5$. Divisive inhibitory units had no shift parameter and integrated inputs approximately linearly. Their sigmoid slope, and that of dendritic compartments, was k=15; WTA neurons—including principal competitors and basket- and chandelier-like interneurons—used steep sigmoids (k≈70).

Plastic synapses used ξ=0.001, and adaptive-excitability neurons used υ=0.025; these values supported WTA emergence. Shunting basket units lacked intrinsic adaptation because they had no shift parameter. Intralaminar projections remained fixed, with their plasticity reserved for future, more complex architectures.

### 2.6 Units, iterations and simulation-specific architectural parameters

Activities, weights, net inputs, shift parameters and normalized variables were dimensionless; neuronal outputs represented normalized firing-rate proxies or firing probabilities, and weights dimensionless connection strengths or probabilistic estimates.

Time was measured in iterations comprising net-input calculation, activation, sensory clamping, enabled synaptic/intrinsic plasticity, and structural-mask reapplication. Patterns remained active through all sequential processing stages plus one additional iteration. Simulations were developed in MATLAB R2014a.

## 3. Results

Three complementary phenomenological models test whether biologically grounded local dynamics can implement computations consistent with thalamic, koniocortical and cortical organization. The thalamic model performs normalization, component extraction and pattern completion; the koniocortex-like network yields sparse pattern-selective responses; the cortical module binds them into frame-like representations and temporal sequences.

The koniocortex-like network has been quantitatively tested on biomedical, financial and signal-classification datasets, and the thalamic model generated a clinically assessed feature-based therapy for arachnophobia. The cortical module poses a further hypothesis: after destructive overlap is reduced, interactions among stabilized representations may support context-sensitive processing formally related to attention.

### 3.1 Thalamic computation implements normalization, orthogonal recoding, winner-take-all selection and pattern completion

#### *3.1.1 From folded autoassociation to triadic thalamic normalization*

The thalamic model is interpreted as a biologically inspired folded autoassociative system. Classical autoassociative networks reconstruct inputs through a hidden layer that compresses distributed patterns into reduced representations [17,18,22]. Because non-orthogonal Hebbian memories generate crosstalk, preprocessing that reduces overlap before cortical association is advantageous [15], and principal-component-like decomposition provides more independent coordinates [19,23].

Because input and target output coincide, the architecture can be folded around the hidden reticular layer (Figure 2). Thalamocortical (TC) relay neurons then form a combined input/output/comparison stage, while reticular neurons provide the reduced competitive bottleneck. TC activity is thereby recoded into sparse, component-like and progressively less-overlapping states. This is a functional computational analogy, not a claim of exact anatomical equivalence.

Earlier thalamic models implemented component extraction and reconstruction of incomplete sensory patterns [24,12], while a later formulation added synaptic-triad normalization and reticular inhibitory facilitation [11]. This conceptual folding therefore leads to the biologically detailed relay–reticular architecture shown in Figure 3, which provides the first stage of the present thalamo-koniocortical framework linking normalization, orthogonalization, competitive selection, pattern completion and cortical binding. Figure 3 shows the biologically detailed thalamic architecture, comprising corresponding 9 × 9 relay and reticular layers in the computational model that will be introduced in the following section. Sensory patterns are imposed on TC neurons, which excite the reticular population and receive inhibitory feedback from it. This recurrent organization provides the substrate for component extraction, orthogonalization and pattern completion [11,24,12]. Normalization is proposed to arise from thalamic synaptic triads. Each sensory afferent excites a relay dendrite, while a thalamic interneuron integrates overall afferent activity. Its axoniform dendrites distribute an approximation of this summed activity to synaptic triads, where dendritic GABA release inhibits corresponding TC inputs [11,4,25,26]. GABA-A-mediated shunting can therefore scale each local input relative to global activity [11,27,28], consistent with normalization as a canonical neural computation [5].

By reducing sensitivity to absolute intensity, triadic normalization makes subsequent reticular competition depend mainly on relative pattern structure and precedes recoding, WTA selection and progressive orthogonalization [11].

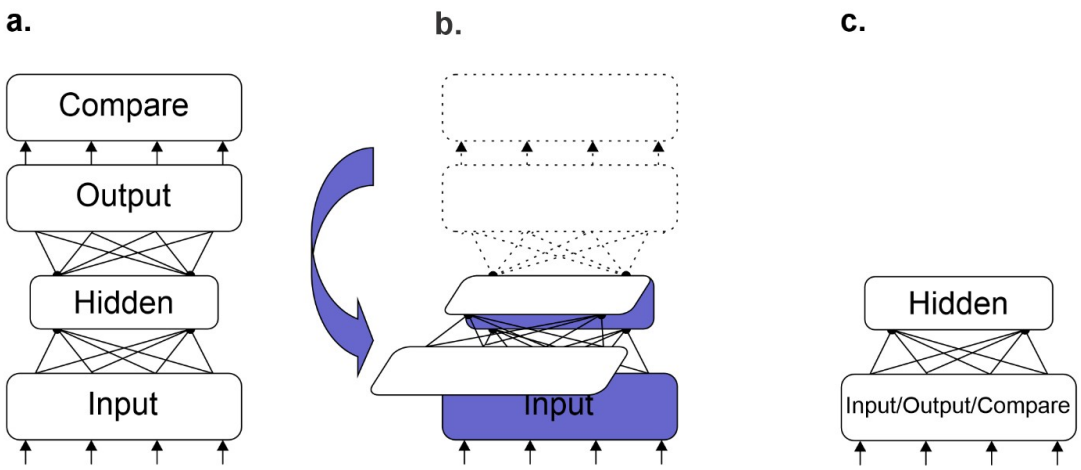



Figure 2. Folding an autoassociative network into a thalamic-like architecture. (a) A classical autoassociative network reconstructs its input through a hidden layer. (b) Because input and target coincide, the network can be folded around that layer. (c) TC neurons then form the combined input/output/comparison stage and reticular neurons the hidden bottleneck, preserving compression and reconstruction with fewer explicit layers.

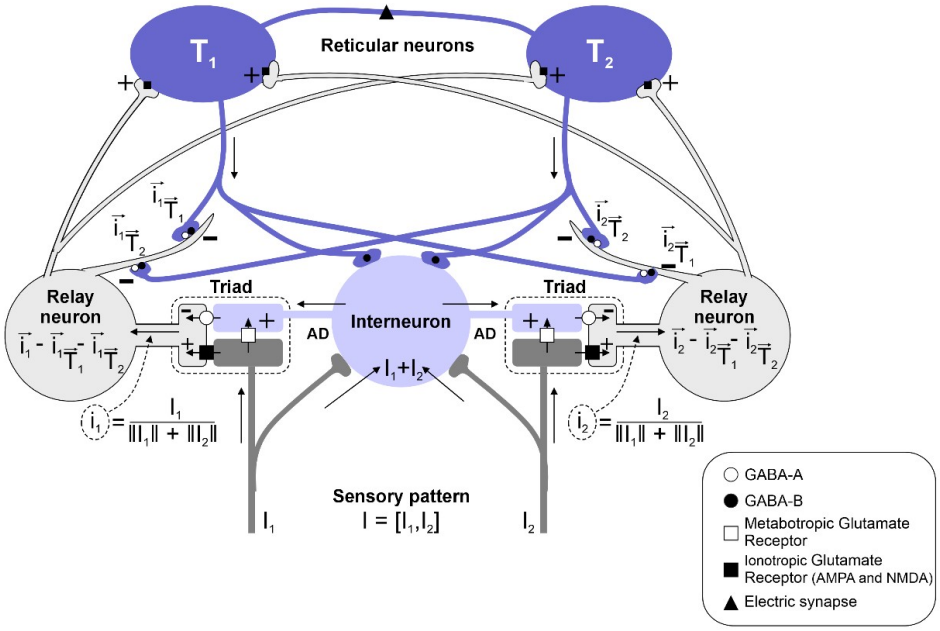



Figure 3. A simplified representation of the 9×9 relay and 9×9 reticular layers is shown. Thalamocortical relay neurons, hereafter referred to as TC neurons, receive sensory input and form the folded network's combined input/output/comparison layer; reticular neurons receive TC excitation and return inhibitory feedback, acting as a competitive bottleneck. Synaptic triads provide GABA-A-mediated divisive normalization. TC–reticular dynamics generate sparse orthogonal winner states in reticular space, WTA component selection and reconstruction of incomplete patterns.

### *3.1.2 Oscillatory reticular competition generates sparse winner states and sequential component extraction*

After triadic normalization, distributed TC activity is projected to the reticular layer. Although both layers contain 9×9 units, their computational roles differ: the relay layer represents the current sensory state, whereas reticular competition selects candidate component-like directions [11,24,12]. This recoding addresses the crosstalk produced

by overlapping, non-orthogonal Hebbian representations [15,16]. Decomposition into more independent coordinates is functionally related to principal-component methods [19,23], although the reticular layer is not proposed as an exact implementation of PCA.

At each WTA episode, one reticular neuron becomes a transient winner, idealized as a one-hot vector:

$$r^{(k)}=[0,0,\ldots,1,\ldots,0]^T, \tag{17}$$

followed by another winner

$$r^{(l)}=[0,0,\ldots,1,\ldots,0]^T, \quad k\neq l \tag{18}$$

so that

$$r^{(k)}\cdot r^{(l)}=0, \quad k\neq l. \tag{19}$$

Exact orthogonality therefore applies to reticular winner vectors; their inhibitory projections onto the relay layer are interpreted more cautiously as sparse, component-like and approximately orthogonal. Across oscillatory phases, different winners form a sparse sequential code that reduces overlap before downstream Hebbian binding. Figure 4 shows how synchronized activity becomes a sequence of transient winners. Its pseudo-spikes are threshold-crossing events, not sodium/potassium action potentials, generated by TC input, shared reticular oscillation, nonlinear amplification and recurrent inhibition [11].

A synchronized oscillatory state is biologically motivated by electrical coupling among thalamic reticular neurons by means of dendro-dendritic electrical synapses [6,29-31]; intrinsic membrane properties and reciprocal TC interactions also support oscillatory and spindle-like activity [29-32]. In the model, this oscillation provides a shared time-varying excitability signal [11].

Reticular neurons rise and fall together but receive different relay inputs, so the rising phase progressively magnifies small differences until the neuron best matched to the current relay pattern crosses threshold first. The oscillation therefore provides both synchronization and a temporal window for converting graded differences into a discrete winner. Selection emerges from the relay–reticular loop rather than an externally imposed maximum. The winner inhibits the relay layer sustaining the common reticular drive, collapsing the wave, while refractoriness prevents immediate reselection [11].

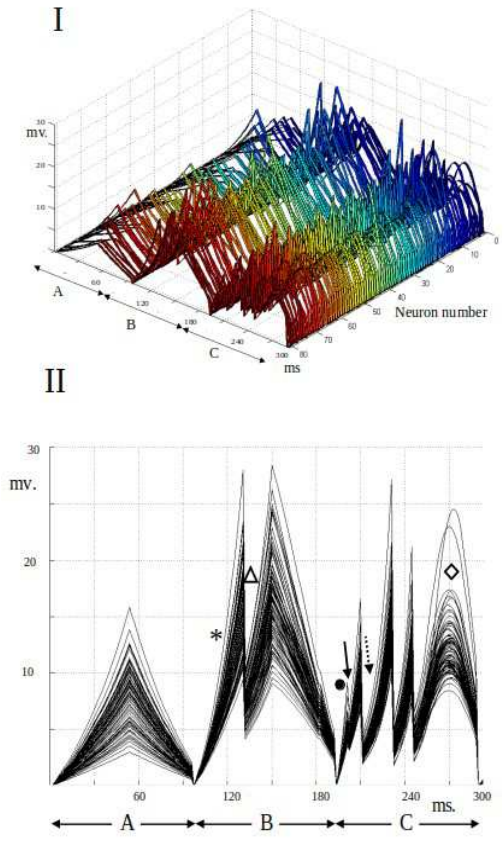



Figure 4. Synchronized reticular oscillations generate softmax-like WTA selection and sparse orthogonal reticular states. Activity of 81 reticular neurons arranged in a 9×9 matrix is shown as a 3D surface (I) and superimposed traces (II). During the rising shared oscillation, stronger TC inputs are amplified more steeply until one neuron crosses threshold and becomes a transient one-hot winner (*). The winner inhibits the relay layer, suppressing common drive and collapsing the wave. Recovery while that neuron remains refractory permits the next winner (Δ). Repeated oscillation, amplification, thresholding, feedback inhibition and refractoriness generate sequential WTA selection.

As activity recovers, the next-best non-refractory neuron can win, yielding an ordered sequence of component-like fragments. This is consistent with WTA models in which recurrence, inhibition, delayed feedback, threshold nonlinearities and recovery generate selection [33-37]. Such local dynamics also motivate future neuromorphic implementations that avoid a centralized maximum-search operation.

The nonlinear transformation coupling each neuron's TC input to the shared oscillation is

$$R_i(t)=\left(N_i(t)+\epsilon\right)^{\alpha B_i(t)}-\epsilon \tag{20}$$

where $R_i(t)$ is reticular activation, $N_i(t)$ relay input, $B_i(t)$ the basal oscillatory state, $\alpha$ the amplification factor and $\epsilon$ a

small term preventing null responses. The implemented simulation used

$$R_i(t) = \left(N_i(t) + 1\right)^{25\,B_i(t)} - 1 \quad (21)$$

When $B_i(t)$ is low, responses to different inputs remain similar; as it rises, the increasing exponent magnifies those differences, allowing the best-matched neuron to separate and cross threshold first. The equation is phenomenological rather than a membrane-conductance model, but captures oscillation as both timing and contrast amplification, consistent with synchronized reticular activity and strong reticulothalamic regulation [11,30,6,29].

The effect is operationally, not mathematically, related to softmax. Conventional softmax exponentiates scores and normalizes their sum [38,39,3], and Transformers use it to convert scaled query–key similarities into attention weights [1]. The thalamic model computes neither exact softmax nor global normalization. Instead, nonlinear amplification determines the first threshold crossing; reticulothalamic inhibition suppresses shared drive; and refractoriness permits subsequent components, consistent with recurrent competitive neural models [33-35,37]. Thus, amplification contributes to ordered orthogonalizing component extraction and identifies a circuit principle for selected WTA-like functions rather than an exact softmax replacement.

Each pseudo-spike identifies the current winner and its component-specific inhibitory projection [11], but successive events also reshape the shared oscillation. Because reticulothalamic inhibition facilitates with repeated activation, persistent winners produce progressively stronger inhibitory effects. The oscillatory envelope is consequently sculpted by the sequence of pseudo-spikes: its deflections deepen, its amplitude progressively decreases and, near the end of the simulated extraction episode (interval C in Figure 4), only a modest hump remains identifying the end of the oscillation. The waveform thus retains a dynamic trace of the inhibitory history of the circuit rather than merely displaying independent selection events. Earlier simulations reproduced electrophysiological findings in which high-frequency reticular stimulation progressively increased inhibitory postsynaptic responses in TC neurons [11,40]. The waveform is also qualitatively consistent with intrinsic oscillations, low-threshold responses, nonlinear membrane dynamics, ramp-like depolarization and rapid pre-threshold upswings reported in reticular neurons [32,29,41], although ionic conductances are not explicitly modeled.

Each threshold event therefore produces an approximately one-hot reticular state and a component-like relay projection. By suppressing one selected contribution at a time, successive winners modify the remaining relay representation and allow new components to emerge. Progressive disappearance of the oscillation marks cumulative component extraction, linking oscillatory dynamics, inhibitory facilitation and sequential orthogonal recoding.

### 3.1.3 Sequential component extraction reconstructs incomplete patterns

Figure 5 illustrates how the preceding mechanisms reconstruct an incomplete pattern. A corrupted, noisy B is imposed on the relay layer, normalized and projected to the reticular population, where synchronized WTA dynamics sequentially select component-like fragments [11]. During each rising oscillatory phase, learned weights produce different compatibilities with the current relay state; nonlinear amplification magnifies these differences until one neuron crosses threshold. The winner sends component-specific inhibition to the relay layer and becomes refractory. In the simulation, the first inhibitory projection was an I-like vertical component (Figure 5.b) ; subsequent winner produced an intermediate E-like pattern (Figure 5.d) through a superposed inhibition and, at the end, a third inhibitory component, reconstructed the complete B [11] (Figure 5.f).

Reconstruction therefore emerges from successive WTA events that modify the relay representation rather than from single-step template matching [11,24,12]. Through intrinsic plasticity, relay neurons gradually estimate the training-set mean [13]. Deviations are projected to the reticular population, where the strongest defines the first principal-component-like direction; inhibitory feedback suppresses that contribution, allowing remaining deviations to be processed in descending relevance [11]. The model is thus related to, but not identical with, formal PCA.

Exact sparse orthogonality applies to distinct one-hot-like reticular winners, whereas their relay projections are component-like and approximately orthogonal.

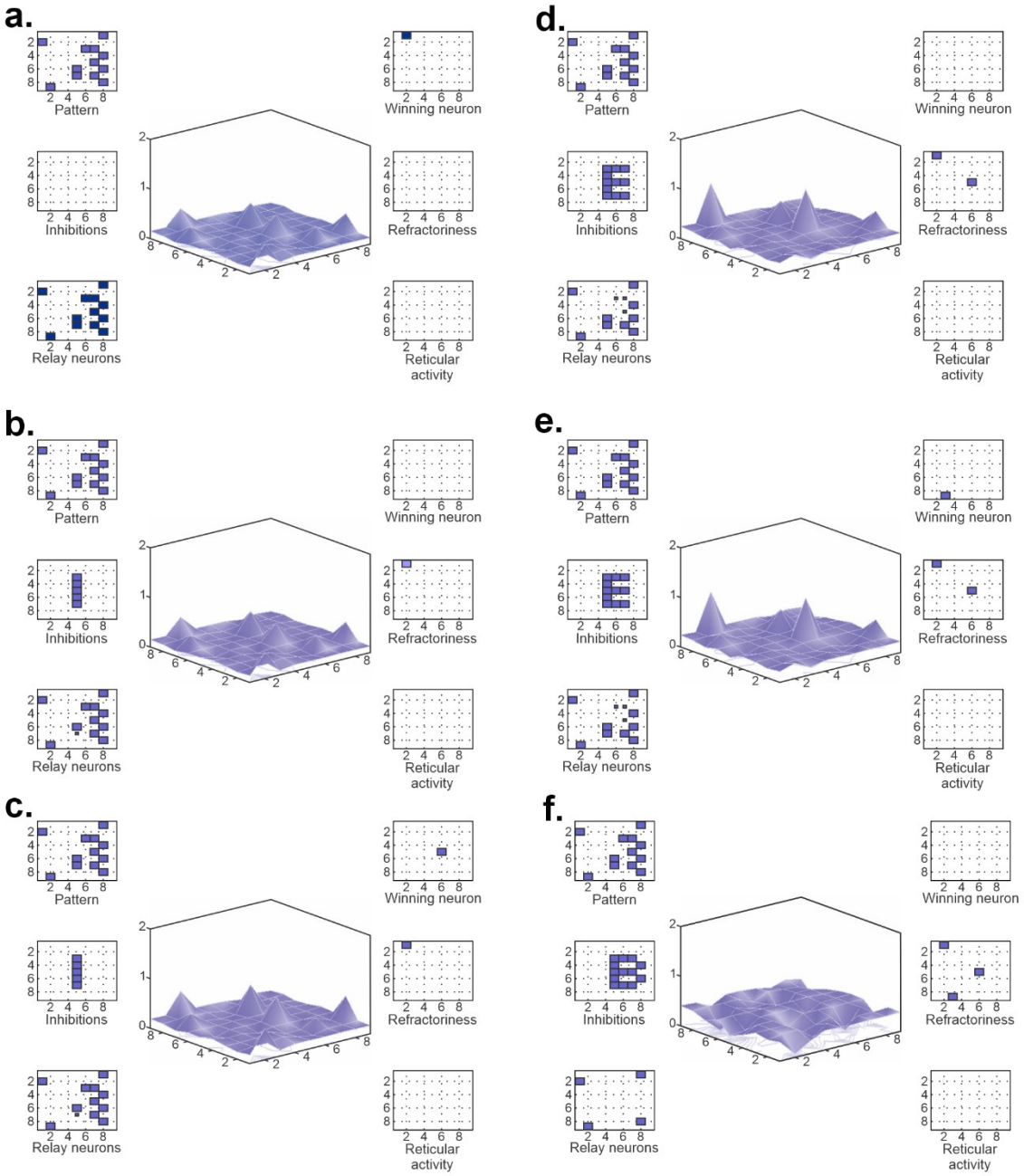



Figure 5. Sequential completion of a corrupted letter B by orthogonal reticular component extraction. The network receives a corrupted, noisy B. Left maps show input, reticular feedback and relay activity; right maps show the current winner, refractory neurons and activation history; the central surface shows synchronized reticular activity. During each rising phase, nonlinear amplification selects the best-matched neuron, whose component-specific inhibition collapses the oscillation. Recovery during refractoriness permits the next component. Repeated cycles progressively reconstruct the B. Distinct winners form sparse orthogonal reticular states, whereas their relay projections are component-like and approximately orthogonal, reducing overlap before downstream Hebbian learning.

Normalization, oscillatory amplification, WTA selection, refractoriness and recurrent subtraction thereby transform partial noisy input into a completed representation, extending the model beyond classification and supporting its folded-autoassociative interpretation [11]. Figures 2–5 therefore describe one continuous recurrent computation: folding assigns relay and reticular layers complementary representational and bottleneck roles; triadic normalization prepares afferent activity; oscillatory amplification selects the best-matched reticular neuron; inhibitory feedback and refractoriness reshape relay activity and permit subsequent components. This produces a softmax-like competitive function without implementing the softmax equation or explicit global comparison. Repeated winners progressively reconstruct the input while converting an overlapping pattern into an ordered sequence of less-overlapping components, preparing it for downstream Hebbian learning and suggesting a biological principle for neuromorphic sequential selection.

#### *3.1.4 Translational impact of the thalamic model: spiderless therapy and neuropsychiatric hypotheses*

The thalamic model, developed to explain biological computation rather than classify, received its main external test translationally. It predicted that arachnophobia might be treated without real, virtual or imagined spiders [42]. Spiderless arachnophobia therapy (SLAT) followed from the hypothesis that the thalamus generates a preliminary feature-based sensory code reaching limbic fear circuits [42,24,12]. If pathological thalamo-amygdala associations depend on combinations of spider-related features, activating subsets below the full phobic-response threshold might weaken them. SLAT used non-spider images containing arachniform color, texture or symmetry while deliberately avoiding the complete object [42], thus deriving directly from the neurocomputational model [24,12].

Clinical outcomes were assessed with the Behavioral Avoidance Test, Subjective Units of Discomfort Scale and Fear of Spiders Questionnaire, plus repeated-measures multivariate ANOVA and k-means classification [42]. Relative to placebo, treatment showed a significant group-by-time interaction: 41.7% of treated participants shifted to the non-arachnophobic cluster after four weeks and 91.7% after six months. Behavioral avoidance toward a real spider improved by approximately 90%, and several participants opened the tarantula cage or touched the animal [42].

These results do not validate every architectural detail, but support the premise that incomplete feature-based codes can modify emotional learning through limbic circuits [42,24,12], yielding a measurable clinical intervention from a mechanistic prediction.

The pattern-completion framework was later extended to hallucinations, delusions, transference and repetition compulsion [43,44], proposing that ambiguous, incomplete or poorly gated inputs may be completed into maladaptive perceptual, cognitive or affective patterns [43,44]. Although lacking SLAT's quantitative clinical evidence, these extensions broaden its relevance to neuropsychiatric phenomena involving ambiguity, sensory gating and pattern completion [42-44].

### 3.2 Koniocortex as a competitive sensory learning stage

The koniocortex, or granular cortex, contains a prominent layer IV rich in spiny stellate neurons, especially in primary somatosensory, visual and auditory areas, but less so in higher association cortex [13] . This suggests a specialized role in transforming thalamic input into sparse cortical representations. TC neurons project both to reticular neurons and to spiny stellate (S) neurons. Intrathalamic relay activity is proposed to remain predominantly tonic, whereas selected components reach layer IV through rebound bursts, as detailed in Section 3.4.

Layer IV does more than relay thalamic signals. Afferent–local interactions can generate sensory selectivity and map reorganization within the thalamocortical–layer IV system [45,46]. Competition is central: strongly tuned spiny stellate neurons become active while most neighbors remain weak or silent, yielding sparse WTA-like responses [47,48].

Barrel cortex studies in rats illustrates experience-dependent granular processing. Whisker inputs form discrete layer-IV barrels, and sensory manipulation alters responses, receptive fields and synaptic organization [49,50]. Thalamocortical synapses also show short-term depression during repeated stimulation and long-term depression during nervous system development [51, 52]. These observations fit the model’s presynaptic rule: active inputs are potentiated when they strongly activate the

winning S neuron and depressed otherwise, producing selective potentiation across a sparse winner population.

The koniocortex-like network formalizes this principle through inhibition, synaptic plasticity and intrinsic plasticity, converting normalized thalamic activity into sparse, pattern-selective responses [13]. It complements thalamic preprocessing: the thalamus reduces overlap and extracts component-like patterns, while the koniocortex compresses each distributed component into a stable cortical event for later binding and association. Once stabilized, controlled overlap may support the context-sensitive interactions discussed later.

#### *3.2.1 Koniocortical architecture transforms thalamic components into sparse competitive cortical representations*

Figure 6 shows how the koniocortex-like network transforms a distributed thalamocortical pattern into sparse activity within a small competitive cortical population. In the example, a 5×3 numerical pattern activates 15 thalamocortical units (TC1–TC15). Digits provide a transparent test case; biologically, these inputs represent sensory components previously normalized and partially orthogonalized by the thalamus. Because the digit patterns still overlap, their separation tests the network's ability to form sparse representations, with more orthogonal thalamic components predicted to be at least as tractable [13].

Each TC unit projects to ten spiny-stellate-like neurons (S1–S10), the principal excitatory competitors. Spiny stellate neurons are abundant in cortical layer IV and receive thalamocortical input [45,46,48]. Initially, every S neuron receives the complete distributed pattern, but learning makes different units selective for different inputs. Basket neurons B1–B10 mediate competition, allowing the winner to emerge through local excitation, inhibition and plasticity rather than an external maximum operation [13,47,35,36]. Distributed thalamocortical activity is thus compressed into one dominant S-neuron response [13].

A shunting basket unit (SB) controls competition by estimating total TC activity and divisively scaling synaptic drive before S-neuron activation. The model therefore normalizes input drive rather than assuming exact division of final firing rate, consistent with evidence that shunting inhibition can scale synaptic input without perfectly divisive output responses [20,27,28]. This prevents magnitude from dominating pattern structure and parallels thalamic normalization: the thalamus normalizes sensory activity before component extraction, while the koniocortex normalizes thalamocortical drive before cortical selection [20,5].

Autaptic S-neuron connections provide brief persistence and subsequent decay of winning responses, a biologically plausible mechanism because autapses do exceptionally occur in spiny stellate neurons [13,53]. Intrinsic plasticity prevents persistent dominance: repeatedly active S neurons become less excitable, allowing less active units to compete and distributing learning across the population [13,54,55].

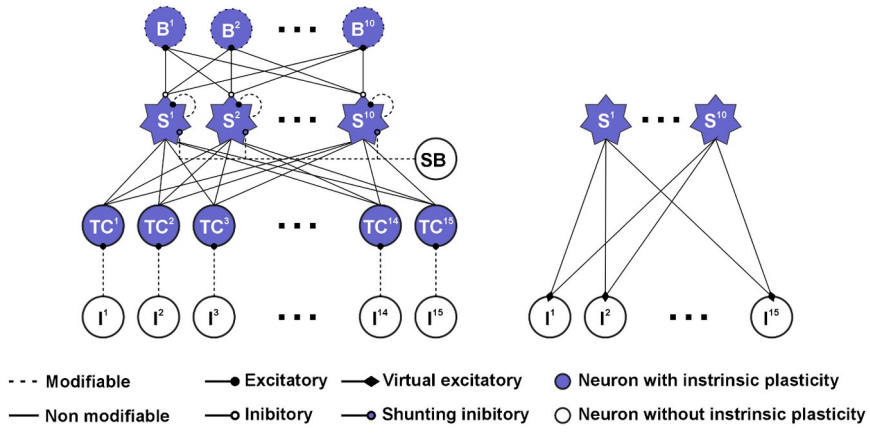



Figure 6. Koniocortex-like network and auxiliary representation readout. (a) Fifteen TC units encoding a 5×3 pattern project to ten competing S neurons. Basket neurons mediate competition, a shunting basket unit (SB) normalizes thalamocortical drive, and S-neuron autapses stabilize responses. Synaptic and intrinsic plasticity support sparse, pattern-specific WTA-like activity. (b) An auxiliary, non-biological readout maps S-neuron activity back to input space to reveal learned specialization; in the example, S neurons specialize for digits 0–9, corresponding biologically to thalamically preprocessed sensory components.

Synaptic plasticity generates selectivity: active TC inputs are potentiated when they contribute to a winning S neuron and depressed when the postsynaptic neuron does not win. Potentiation is therefore concentrated on winners, while depression is distributed across non-winners, progressively separating S-neuron weight vectors and providing a possible computational explanation for prominent depression in granular cortical learning [13,56,57] .

After training, each pattern ideally activates one S neuron strongly while leaving others nearly silent, producing one-hot-like, approximately orthogonal population states better suited to downstream Hebbian association and less vulnerable to crosstalk [1,17]. The auxiliary readout in Figure 6 simply visualizes this specialization and is not part of the biological model [13].

The koniocortex therefore complements thalamic processing: the thalamus separates a sensory frame into less-overlapping distributed components, whereas the koniocortex compresses each component into a stable cortical event through normalization, competition and adaptive specialization. This prepares the sequence for later pyramidal binding and illustrates how normalization and selection may emerge from local circuitry rather than explicit artificial operations.

#### *3.2.2 Quantitative validation of KLN performance across real-world datasets*

The Koniocortex-Like Network (KLN) was introduced as a biologically plausible model of natural competition [56] and later extended to associative learning [13]. Real-world applications tested whether this architecture could provide measurable performance beyond its theoretical formulation (Table 1).

On the Wisconsin Breast Cancer Database, 10-fold cross-validation yielded 99.94% sensitivity and 96.91% accuracy, with a 0.06% false-negative rate [58], demonstrating discriminative unsupervised representations in a clinically important task.

On the Australian Credit Approval Database (ACAD), the best configuration reached 87.66% sensitivity and 83.10% accuracy under cross-validation, confusion-matrix, learning-curve, runtime and benchmark analyses [59]. Although optimized supervised methods performed better, the non-optimized unsupervised KLN remained competitive and converged within few epochs. A Business Intelligence application likewise stabilized rapidly without target labels [60].

Three parallel binary KLNs applied to MIT-BIH-derived ECG patterns reached 96.98% specificity, 97.02% sensitivity and 97.01% accuracy after 20 epochs [61], extending the architecture from tabular data to biomedical signals.

These results do not imply universal superiority over optimized classifiers, but show that the same biologically motivated competitive and plastic mechanisms generate useful sparse representations across markedly different domains.

### 3.3 Cortical binding, synaptic specialization and temporal sequencing

We next examined the cortical stage as an independent module that recombines component-derived activity into frame-like representations and temporal sequences. It was not simulated within a closed thalamo-koniocortical-cortical loop; instead, it received S-neuron activation patterns compatible with sparse events expected after thalamic processing and koniocortical integration [13,11].

The simulations tested two operations: temporally gated dendritic binding of component-derived S-neuron activity by pyramidal-unit-like modules, and assignment of successive frame-like representations to different pyramidal units. Figure 7 illustrates both using the D-N-A sequence introduced in Figure 1.

#### *3.3.1 Cortical module architecture*

The cortical module comprised S-neuron-like inputs, compartmentalized pyramidal units and chandelier-like inhibitory units, the latter omitted from Figure 7b for clarity. S-neuron-like units (Figure 7a) served as controlled proxies for expected koniocortical output and, biologically, for the granular cortical interface of sensory cortex receiving thalamically related component-derived inputs; these activations were bound by pyramidal-unit-like structures [11,13].

For the D-N-A sequence, D was encoded by S1, S3 and S14; N by S2, S6 and S4; and A by S2 and S5. Rather than indivisible symbols, these patterns were represented as ordered S-neuron event sequences, whose labels identify the associated alphanumeric components.

| Application and dataset | Validation | Best reported KLN performance | Main result |
|---|---|---|---|
| **Breast cancer diagnosis, WBCD** [Fombellida et al. 2017] | 10-fold CV; sensitivity; accuracy; confusion matrix | Sensitivity 99.94%; Accuracy 96.91% | High sensitivity to malignant cases |
| **Credit scoring, ACAD** [Fombellida & Andina 2018; Fombellida et al. 2018] | 10-fold CV; sensitivity; accuracy; confusion matrix; runtime; benchmark comparison | Sensitivity 87.66%; Accuracy 83.10% | Competitive unsupervised classification of heterogeneous financial data |
| **Business intelligence, ACAD** [Fombellida & Andina 2018] | Sensitivity; accuracy; learning across epochs; benchmark comparison | Sensitivity 87.66%; Accuracy 83.10% | Stable performance after few iterations |
| **Cardiac arrhythmia classification, MIT-BIH-derived ECG** [Torres-Alegre et al. 2019] | 10-fold CV; specificity; sensitivity; accuracy; benchmark comparison | Specificity 96.98%; Sensitivity 97.02%; Accuracy 97.01% | Extends KLN performance to biomedical signals |

Table 1. Quantitative validation of KLN performance across real-world datasets. CV, cross-validation; WBCD, Wisconsin Breast Cancer Database; ACAD, Australian Credit Approval Database.

Each pyramidal unit was compartmental rather than point-like. PNtuft represented the distal apical compartment receiving the intralaminar-like signal that initiated the carrier wave; PNden1–PNden4 were sequential dendritic integration sites; and PNsoma was the somatic readout of successful integration. This abstraction reflects nonlinear pyramidal compartmental computation and earlier pyramidal-unit-like models [62-67].

An intralaminar-like unit initiated the carrier wave at PNtuft, which propagated through PNden compartments toward the soma. Without component-derived S-neuron input, the wave attenuated and did not trigger PNsoma, thus acting as a temporal gate for binding rather than a substitute for sensory input [24,67].

*During sequencing, chandelier-like units suppressed coincident PNsoma activations when multiple pyramidal units responded. The model does not reproduce the full diversity of cortical interneurons; inhibition here only resolves competition among candidate pyramidal responses [24,12,67].*

#### *3.3.2 Temporally gated binding of component-derived inputs*

The first cortical simulation tested whether component-derived inputs could be bound into single frame-like representations. Letter-like patterns D, N and A served as minimal examples, each represented by an ordered set of S-neuron activations delivered to the modular pyramidal unit described above.

The cortical module did not receive complete letter templates. S-neuron inputs were available to all dendritic compartments, but effective binding required temporal coincidence with the carrier wave propagating along the pyramidal dendritic tree. A component contributed only when its activation matched the local dendritic state, so binding depended jointly on component identity, dendritic position and temporal gating [68,69,67].

Each episode began when intralaminar input activated PNtuft and initiated a carrier wave toward the soma. As it reached successive PNden compartments, active S-neuron events were bound at temporally matching sites. In the D-N-A simulation this produced specific component-to-dendrite maps: D through PNden2-S1, PNden3-S3 and PNden4-S14; N through PNden2-S2, PNden3-S6 and PNden4-S4; and A through PNden3-S2 and PNden4-S5 [67]. Thus, each frame emerged from temporally ordered component–dendrite associations rather than presentation of a complete letter template.

Binding was therefore not a simple feedforward sum. Progressive coincidence between the carrier wave and S-neuron signals activated successive dendritic segments until

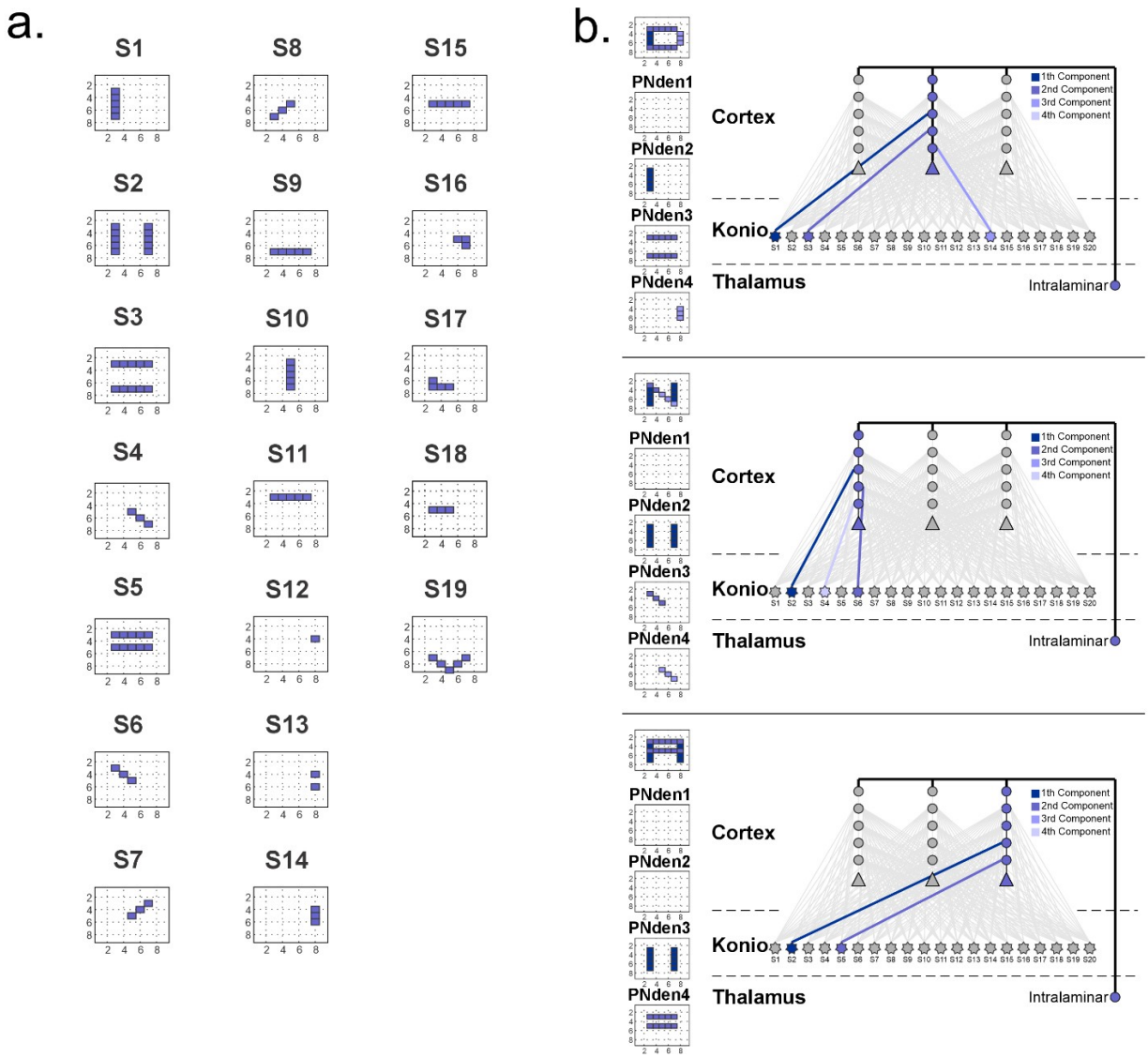



Figure 7. Component-derived inputs and D-N-A sequencing in the independent cortical module. (a) Learned component dictionary represented by S neurons after training with alphabet-like patterns; each inset shows a subletter-like spatial component that can be recombined into a larger frame. (b) During D-N-A presentation, selected S-neuron outputs are delivered in component order to pyramidal dendritic compartments, beginning at PNden4. Intralaminar input initiates a PNtuft carrier wave that enables temporally gated binding and competitive frame assignment: D→PN2, N→PN1 and A→PN3. S-neuron outputs were controlled to emulate expected koniocortical activity rather than generated online by a fully integrated loop.

*combined activity triggered PNsoma. The somatic response marks, at the model level, completion of a composite frame-like representation [70,67].*

In this restricted sense, the D-N-A simulation provides a minimal model of cortical frame formation: previously separated components are bound into structured units rather than full perceptual scenes. In the broader framework, thalamic and koniocortical stages first generate sparse component-derived S-neuron activations, which the cortical module then binds through temporally gated dendritic integration. The result is deliberately limited: the independent pyramidal-unit-like module can assemble letter-like frames from S-neuron components, but does not demonstrate autonomous reconstruction by a closed thalamocortical loop, context-dependent completion or self-attention.

### *3.3.3 Synaptic specialization after frame binding*

A transient PNsoma response alone would not demonstrate stable component-to-dendrite learning, so the S-neuron-to-PNden weight matrix was compared before and after training [67, Figure 8]. Initially, no organized mapping matched the final D-N-A sequence. After training, pathways involved in temporally gated binding were selectively strengthened: D → PN2 through PNden2-S1, PNden3-S3 and PNden4-S14; N → PN1 through PNden2-S2, PNden3-S6 and PNden4-S4; and A → PN3 through PNden3-S2 and PNden4-S5 [67, Figure 8b].

This post-training structure provided convergent evidence that frame-like representations were encoded not only by PNsoma activation but also by the synaptic pathways supporting each bound state. Synaptic change followed a presynaptic Hebbian rule constrained by postsynaptic activity and fixed cortical architecture: S-neuron inputs coincident with the carrier wave and contributing to somatic activation were selectively reinforced, whereas nonparticipating inputs showed weaker or no comparable strengthening [57,70,21,67].

Thus, ordered S-neuron events were integrated through the PNden sequence until PNsoma activation, while participating S-neuron-to-PNden synapses were selectively strengthened. Within this independent module, binding was therefore not imposed by a pre-existing map but emerged from compartmental pyramidal architecture, carrier-wave timing and presynaptic Hebbian learning. The module assembled and stabilized frame-like representations from component-derived inputs through its own architecture, dynamics and learning rule.

### *3.3.4 Sequencing of D-N-A frame-like representations*

The final cortical simulation tested whether binding could extend from individual frame-like representations to an ordered sequence. D, N and A were presented successively using the S-neuron component sets defined above: D = S1/S3/S14, N = S2/S6/S4 and A = S2/S5 [67, Figure 7a].

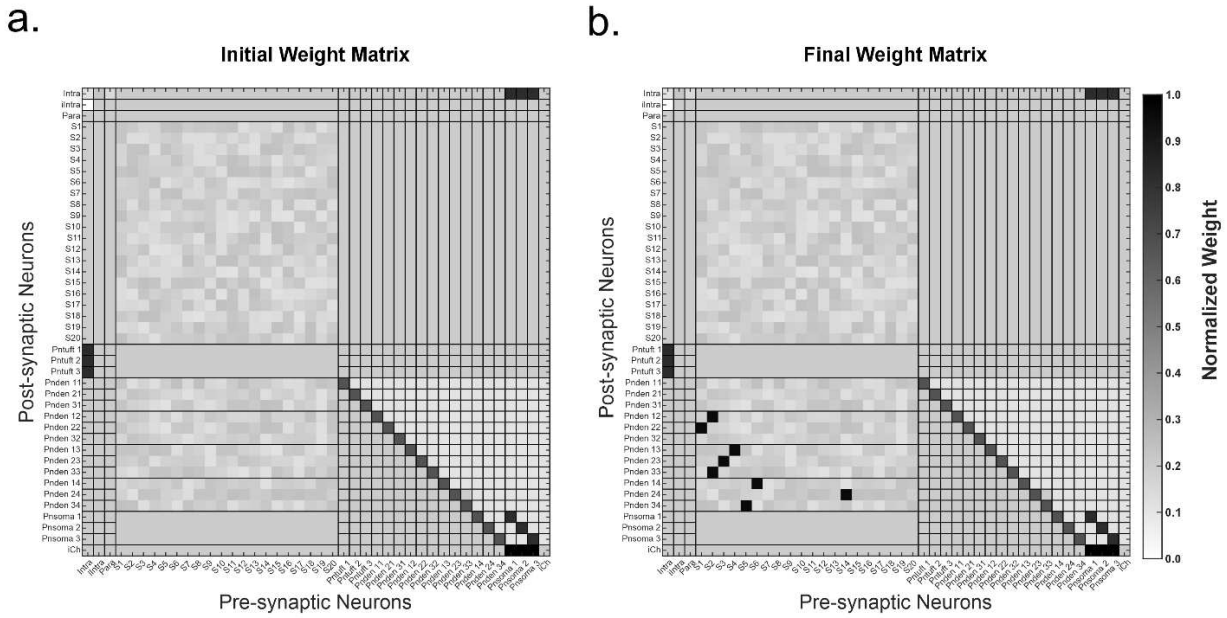



Figure 8. Synaptic specialization after frame binding. (a) The initial S-neuron-to-PNden matrix lacks an organized dendritic map. The long dark diagonal denotes strong fixed connections between compartments within each pyramidal unit; the shorter diagonal denotes fixed phenomenological self-recurrent connections producing gradual activity decay and temporal persistence, not necessarily biological autapses. Chandelier and intralaminar connections are also strong and fixed. (b) After training, sparse high-weight S-neuron-to-PNden pathways selectively support D-N-A binding rather than reflecting nonspecific potentiation. Each pyramidal unit contains one tuft, four dendritic compartments and one soma: PN1 = PNtuft1 + PNden11–PNden14 + PNsoma1; PN2 = PNtuft2 + PNden21–PNden24 + PNsoma2; PN3 = PNtuft3 + PNden31–PNden34 + PNsoma3. The strengthened associations correspond to the D/N/A mappings described above.

Three competing pyramidal units were used. Before each letter, an induced Intra pulse initiated a carrier wave in their PNtufts, enabling binding of incoming S-neuron-derived components. Each episode therefore combined carrier-wave initiation, component presentation, dendritic binding, somatic activation and competitive selection. Chandelier-like inhibition imposed WTA-like control over competing PNsoma responses, allowing a single pyramidal unit to stabilize for the current frame [67].

Across repeated presentations, synaptic adjustment, intrinsic plasticity and refractory dynamics biased different PNs toward different sequence positions. Letter “D” was assigned to PN2, N to PN1 and letter “A” to PN3, producing the ordered sequence PN2-PN1-PN3 [24,12,67]. The model therefore operates hierarchically: S-neuron activity represents subletter components, each pyramidal unit binds a component combination into a frame, and distinct pyramidal units encode successive frames.

Because intralaminar activation was induced before each letter, this result represents sequencing under controlled stimulation rather than autonomous generation by a complete intralaminar-pyramidal loop that will be computationally modeled in a future implementation. Nevertheless, the module succesfully allocated successive component-derived frames to distinct cortical units while preserving temporal order.

Overall, the cortical simulations demonstrate binding of component-derived S-neuron inputs into frame-like representations, training-dependent stabilization of S-neuron-to-PNden pathways, and sequencing of multiple frames. They do not demonstrate autonomous corticothalamic recurrence, context-dependent completion or self-attention.

### 3.4 Intermodular transmission from thalamic components to cortical frames

This section specifies the information exchanged among the three modules and its coding format. The proposed pathway converts a distributed sensory frame into an ordered sequence of compact cortical events and then into a frame-specific pyramidal representation.

#### *3.4.1 What is transmitted between modules*

Let us recall the whole process from a sensory frame like “D” initially activating a distributed population of thalamocortical relay neurons. After normalization, recurrent relay–reticular processing sequentially selects component-like patterns; approximately seven to nine selections can occur during one ascending reticular oscillatory episode, depending on input, threshold and refractory dynamics. Each transient reticular winner projects a specific inhibitory component onto the relay layer, progressively reshaping the distributed representation and contributing to pattern completion.

Reticular winners are one-hot-like and mutually orthogonal, whereas their relay projections are more cautiously described as sparse, component-like and approximately orthogonal. Thus, thalamic output is an ordered sequence of distributed components rather than a single compressed label [24,12,11].

The koniocortex receives each component in distributed form and, through local inhibition, synaptic plasticity and intrinsic plasticity, transforms convergent relay input into one dominant S-neuron response [13]. This many-to-one transformation preserves component identity while converting each distributed thalamic pattern into a compact cortical event.

These sparse S-neuron events reach the pyramidal module in temporal order. As the carrier wave traverses apical dendritic compartments, successive events encounter different integration windows; their ordered combination activates one pyramidal-unit-like representation corresponding to the complete frame [67].

The coding progression is therefore distributed component → sparse cortical event → bound frame. It reduces destructive crosstalk during initial frame formation, while stabilized frames may later exploit regulated overlap for the context-sensitive interactions developed in the Discussion and mathematical Appendix.

#### *3.4.2 Why the thalamic output must be temporally gated*

Thalamic reconstruction passes through incomplete relay states that should not recruit layer IV as strongly as completed components, or S neurons could learn unstable mixtures. The proposed interface therefore separates thalamic processing from cortical delivery.

During relay–reticular computation, TC neurons are proposed to fire tonically at low rates, whereas selected components are delivered more effectively through high-

frequency bursts. Tonic spikes can still reach cortex, but should recruit recurrent layer-IV circuitry less strongly; the koniocortex thus acts as a frequency-sensitive gate, responding preferentially when component-specific relay populations enter burst mode.

Most modeled processing occurs within an oscillatory interval of ~120 ms. Reticular competition both selects/subtracts components and prolongs inhibition of their relay populations, providing a plausible code transition: intermediate states remain mainly tonic, whereas selected components emerge as an ordered burst sequence more effectively detected by layer IV.

#### *3.4.3 How reticulothalamic inhibition can convert components into bursts*

The proposed tonic-to-burst switch relies on TC relay-cell voltage dependence. In relatively depolarized states, T-type calcium channels remain largely inactivated and firing is tonic. Reticulothalamic inhibition hyperpolarizes component-specific TC populations; if sufficiently strong and sustained for ~100 ms, T-type channels deinactivate. Release from inhibition can then evoke a low-threshold calcium potential followed by a high-frequency sodium-spike burst [71,72].

This timescale is comparable to the modeled oscillatory episode, suggesting that the same inhibition that marks and subtracts a component may also prepare its TC population for rebound transmission. Sequential inhibition of component-specific populations should therefore preserve component order in the resulting cortical burst sequence.

Bursting can recruit cortex efficiently: the preceding silent interval permits recovery from short-term synaptic depression, strengthening the first spike, while closely spaced spikes summate postsynaptically [73]. Brief thalamocortical spike bursts around 40–100 Hz can also recruit recurrent layer-IV activity [74]. Thus, each distributed component can be delivered in a form readily compressed by the competitive S-neuron population into one dominant response.

The interface therefore changes representational format across stages: distributed thalamic components → rebound bursts → sparse S-neuron events → pyramidal frame binding.

Because the three modules were simulated independently, tonic-to-burst integration remains a physiologically grounded hypothesis rather than a demonstrated result. It nevertheless makes a specific prediction: component-specific reticulothalamic inhibition should produce an ordered rebound-burst sequence whose S-neuron activations preserve the order required for pyramidal frame binding.

## 4. Discussion

We use the Transformer as a functional reference to ask whether biologically inspired circuits can achieve related information-processing goals through different physical mechanisms, without implying that the brain implements a Transformer. We examine whether representation formation, normalization, context-sensitive interaction, iterative refinement and competitive selection can emerge from excitation, inhibition, plasticity, recurrence, dendritic integration and oscillatory dynamics.

The three models form a processing sequence: relay–reticular dynamics normalize input, reduce overlap and extract component-like patterns; koniocortical competition converts them into sparse, stable S-neuron events; and temporally gated pyramidal integration binds ordered events into frame-like representations and assigns successive frames to different cortical units. Once stabilized, interactions among these representations may support context-sensitive processing related to attention, while local dynamics may implement selected softmax-like or winner-take-all functions.

### 4.1 Embeddings and positional encoding: from raw inputs to useful internal representations

Transformer computation converts raw tokens or multimodal inputs into vector representations for subsequent processing [1,2,17]. Autoencoder latent spaces illustrate the broader principle that distributed, overlapping inputs can be reorganized into internal representations preserving relevant structure [17]. Although mathematically distinct, both approaches transform raw input into a more useful representational basis.

The thalamo-koniocortical pathway offers a biological analogue of this principle. TC neurons encode the current distributed sensory pattern; reticular competition separates it into sequential component-like states, while recurrent

inhibition supports reconstruction of incomplete inputs [24,12,11]. The koniocortex then compresses each distributed component into a sparse S-neuron event suitable for binding, sequencing and contextual interaction.

A related functional progression appears in distributed language networks. Posterior temporal–parietal regions associated with Wernicke's area contribute strongly to lexical-semantic representation and access, whereas frontal networks associated with Broca's area contribute to sequencing, integration and context-dependent control [75]. Rather than isolated modules, these are landmarks within distributed, reciprocally connected systems and illustrate a progression from representation formation and stabilization to ordered, context-dependent transformation.

Positional encoding adds sequence information to parallel token representations:

$$x_i = e_i + p_i \tag{22}$$

where $e_i$ is token content, $p_i$ position and $x_i$ their combination. More generally, ordinal variables can be transformed into distributed internal coordinates. Biological population coding provides an analogy: wavelength is represented not simply as a scalar but by distributed activity across photoreceptor classes and downstream populations [76]. Wavelength and sequence position differ, but both illustrate how simple variables can become richer coordinates able to interact with other information.

### 4.2 Recoding, binding and iterative refinement

A Transformer block updates representations through context-dependent interaction followed by feedforward transformation [1]. Functionally, these stages recode the current vector and apply learned nonlinear mappings.

The thalamic model provides a biological recoding stage through folded autoassociation [24,12,11]. TC neurons form a combined input/output/comparison layer, while the reticular population acts as a competitive bottleneck that extracts sparse component-like states whose feedback reshapes TC activity.

The koniocortex converts each distributed component into a sparse, stable S-neuron event; pyramidal-unit-like structures then bind ordered events into composite frame-like representations. These are model-level units, not claims about conscious percepts or complete scenes. The progression is:

thalamic recoding → koniocortical sparse stabilization → cortical frame formation

This is not anatomically equivalent to a Transformer block, but both frameworks transform unstable inputs through successive recoding and memory-like nonlinear processing.

Transformers create computational depth by stacking blocks [1], whereas thalamocortical systems may achieve a related function through temporal recurrence. Cortical feedback can influence subsequent component extraction and integration [68,69]; descending corticothalamic axons outnumber ascending thalamocortical axons by ~10:1, supporting a substantial feedback role [77,78].

One recurrent cycle would transform sensory activity into thalamic components, sparse koniocortical events and a bound cortical frame; feedback could then influence the next cycle. Repetition would convert artificial depth into temporal depth by reusing the same circuitry. Because the present simulations establish these operations separately, an integrated recurrent model is still required to test whether successive cycles improve completion, stability and context sensitivity.

### 4.3 Normalization and the transition from crosstalk to attention

Normalization provides a clear functional correspondence between Transformers and biological neural systems. Transformer LayerNorm stabilizes representations across repeated transformations [1,2], whereas the biological models use different mechanisms. The thalamic model applies $L_1$-like divisive normalization motivated by synaptic triads, with GABA-A-mediated shunting regulating sensory input before reticular competition [4,25–28]; the koniocortex-like network uses a shunting basket-like unit to limit global thalamocortical drive before sparse competition [13,5]. These mechanisms are not implementations of LayerNorm or necessarily exact divisions of firing rate; functionally, they constrain activity before nonlinear transformation or selection. The cortical module contained no separate normalization operation because its simulated scope was dendritic binding, synaptic specialization and temporal sequencing, not because biological cortex lacks gain control.

Normalization and component separation also frame the proposed relation between Hebbian crosstalk and Transformer attention. In a heteroassociative Hebbian network, test input produces a similarity-weighted combination of stored outputs [14–16]:

$$O_t = \xi \sum_{p=1}^{n} O_p \left( I_p \cdot I_t \right) \tag{23}$$

where $I_t$ is the test input, $I_p$ and $O_p$ a stored input–output pair, and $\xi$ the learning factor. For unit-normalized vectors,

$$I_p \cdot I_t = \cos \theta_p \tag{24}$$

so that

$$O_t = \xi \sum_{p=1}^{n} O_p \cos \theta_p \tag{25}$$

Associative generalization and interference arise from the same mechanism: when several stored inputs overlap with the test pattern, their outputs contribute simultaneously, producing unregulated similarity-dependent crosstalk [15,16].

The thalamic and koniocortical stages reduce this problem during early representation formation: thalamic normalization and component extraction decrease overlap, and koniocortical competition converts distributed components into sparse cortical events before binding.

Once components are separated and stabilized, similarity can instead support context-sensitive interaction. As formalized in the Mathematical Appendix, Hebbian retrieval and scaled dot-product attention share a basic structure: outputs combine stored values according to similarity between inputs/queries and stored inputs/keys. Scaling and softmax convert raw similarities into positive, competitive, normalized attention weights [1]. Attention can therefore be viewed as controlled crosstalk: similarity-dependent interaction becomes useful when regulated for contextual integration.

### 4.4 Oscillatory winner selection and physical computation

Softmax converts candidate scores into a normalized distribution:

$$\mathrm{softmax}\left( z_i \right) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{26}$$

Exponentiation amplifies score differences and normalization sets relative contributions [1]. The thalamic reticular model does not compute this equation; it reaches a related selective outcome through local dynamics.

Reticular neurons share a synchronized oscillatory state but receive different TC inputs. The rising oscillation amplifies these differences until the strongest neuron crosses threshold and becomes a transient winner. Reticulothalamic inhibition suppresses the TC drive sustaining the oscillation, refractoriness prevents immediate reselection, and recovery permits the next component to win [6,29–31,11]. Repeated cycles generate sparse sequential states rather than a full graded distribution.

This suggests a hardware opportunity. In software, oscillation, amplification and thresholding are numerical operations; in neuromorphic hardware they could be embodied in currents, voltages, conductances and device nonlinearities. Candidate scores could modulate coupled elements, shared circuitry provide oscillatory drive, and first threshold crossing signal the winner, making competition a consequence of substrate dynamics rather than explicit maximum search.

Because candidates evolve concurrently, selection could emerge within one competitive cycle. Larger candidate sets increase area, connectivity and synchronization demands, but not necessarily sequential search time. Latency, energy, noise tolerance, coherent coupling and communication load become measurable implementation properties.

The proposal targets softmax's selective function, not its full graded distribution, and is therefore most relevant to argmax, routing, gating, top-k-like operations or sparse attention; conventional normalization remains appropriate when the full distribution is required.

The other modules support the same substrate-level principle: local inhibition and plasticity produce S-neuron competition [13], while inhibitory, adaptive and refractory dynamics assign successive frames to different pyramidal units. The thalamic model adds an explicit oscillatory mechanism for parallel comparison and winner emergence.

Softmax and LayerNorm can become important latency and energy costs once dense matrix multiplication is accelerated. Existing strategies include hardware–software co-design, approximate softmax, specialized exponential

instructions, in-memory top-k selection and softmax-free spiking Transformers [79–84]. Oscillatory WTA circuits could complement GPUs, TPUs or in-memory accelerators by offloading routing, gating and sparse-selection operations to coupled physical circuits.

### 4.5 Scope, limitations and testable implications

The simulations support functions assigned to the thalamic, koniocortical and cortical modules but do not demonstrate autonomous operation of an integrated system. Sequential thalamic transmission through rebound bursting therefore remains a physiologically grounded hypothesis: component-specific reticulothalamic inhibition should generate ordered rebound bursts whose koniocortical activations preserve the timing required for pyramidal frame binding.

The models are phenomenological rather than conductance-based, and hardware advantages remain unquantified. An integrated recurrent implementation should test whether repeated cycles improve reconstruction, stability and contextual sensitivity. Electronic implementations should compare physical WTA with digital argmax, top-k and approximate-softmax baselines while measuring latency, energy, scalability, robustness and synchronization.

The central hypothesis is that shared oscillation, nonlinear amplification and threshold crossing can turn competitive selection from algorithmic search into an emergent physical event. More broadly, these operations may arise from interacting physical states, suggesting a neuroscience-derived route to adaptive, recurrent and hardware-oriented Transformer-like architectures.

## 5. Conclusion

This work presents a simulation-grounded framework in which thalamic, koniocortical and cortical models provide biological principles for Transformer-like computation. The comparison is functional: both systems must form useful representations, stabilize activity, reduce interference, select alternatives, bind components and organize information over time.

The framework starts from a limitation of local Hebbian learning: overlapping patterns generate crosstalk. Thalamic relay–reticular dynamics normalize sensory input, reduce overlap and extract component-like representations; the koniocortex converts each distributed component into a sparse, stable cortical event; and pyramidal dendritic integration binds ordered events into frame-like representations and temporal sequences. Representational overlap therefore plays two roles: it is reduced during early learning to limit interference, but after stabilization similarity-dependent interactions can support contextual computation. Transformer attention can thus be viewed formally as a regulated form of the same vector interactions that produce crosstalk in simpler associative memories.

The models also suggest hardware-oriented circuit motifs. Divisive normalization, recurrent recoding, sparse competition, dendritic binding and temporal refinement distribute computation across substrate dynamics. In the thalamic reticular model, synchronized oscillation, nonlinear amplification, threshold crossing, inhibitory feedback and refractoriness allow winners to emerge as circuit events. For routing, gating, argmax and top-k-like operations, this could shift selected computations from explicit exponentiation and global normalization toward parallel physical dynamics, complementing efforts to reduce softmax and LayerNorm latency and energy costs [79–84].

These modular phenomenological simulations define a testable path toward an integrated recurrent thalamo-koniocortical-cortical system linking sensory input, adaptive learning, contextual refinement and competitive selection. Electronic implementations can quantify latency, energy, scalability, robustness and synchronization, while embodied platforms can test real-time sensorimotor operation.

The resulting design trajectory is: normalize and separate sensory activity, sparsify and bind components, allow stabilized representations to interact contextually, and let selection emerge from local dynamics. Translating these biological operations into computational and hardware hypotheses may support adaptive, hardware-oriented and embodied artificial intelligence.

# 6. Mathematical Appendix: From Hebbian crosstalk to attention

This Appendix shows that heteroassociative Hebbian retrieval already contains the central algebraic operation

underlying attention in Transformers: a query combines stored values according to similarity with their associated keys. In a classical Hebbian network, partially matching associations appear as crosstalk; scaling and softmax normalization transform these raw contributions into the controlled contextual weighting used in Transformer attention.

Consider n key–value associations,

$$(k_1, v_1), (k_2, v_2), \ldots, (k_n, v_n), \tag{A1}$$

$k_p \in \mathbb{R}^{1 \times d_k}$ represents an input pattern and each value $v_p \in \mathbb{R}^{1 \times d_v}$ where each key kp $\in \mathbb{R}^{1xdk}$ represents an input pattern and each value vp $\in \mathbb{R}^{1xdv}$ its associated output. Arranging keys and values as rows gives

$$K = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}, V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}. \tag{A2}$$

A linear heteroassociative network stores these associations through the Hebbian outer-product rule [15,16]:

$$W = \xi \sum_{p=1}^{n} k_p^T v_p = \xi K^T V, \tag{A3}$$

$\xi$ where ξ is the learning factor.

$q \in \mathbb{R}^{1 \times d_k}$ When a query q $\in \mathbb{R}^{1xdk}$ is presented, the network produces

$$O_t = q W. \tag{A4}$$

Substituting Eq. A3 into Eq. A4 gives

$$O_t = \xi q K^T V. \tag{A5}$$

and expanding over stored associations yields

$$O_t = \xi \sum_{p=1}^{n} \left( q k_p^T \right) v_p. \tag{A6}$$

Equation A5 is the key result: each stored value contributes to the output with a coefficient determined by query–key similarity. Hebbian retrieval is therefore a similarity-weighted aggregation of values.

For unit-normalized queries and keys,

$$\|q\| = \|k_p\| = 1, \tag{A7}$$

the similarity coefficient becomes

$$q k_p^T = \cos\theta_p, \tag{A8}$$

$\theta_p$ where θp is the angle between q and kp, so Eq. A5 becomes

$$O_t = \xi \sum_{p=1}^{n} v_p \cos\theta_p. \tag{A9}$$

This formulation makes the origin of crosstalk explicit. If the query coincides with a stored key $k_r$, then

$$O_t = \xi v_r + \xi \sum_{\substack{p=1 \\ p \neq r}}^{n} \left( k_r k_p^T \right) v_p. \tag{A10}$$

The first term retrieves the matching value, whereas other stored values contribute whenever their keys overlap with the query. These additional similarity-dependent contributions constitute Hebbian crosstalk.

The Hebbian expression can now be written in a form closely related to attention. Define the vector of query–key similarities as

$$s = q K^T, \tag{A11}$$

so that Eq. A4 becomes

$$O_t = \xi s V. \tag{A12}$$

The computation therefore follows:

query → key similarities → weighted combination of values

This is also the central sequence implemented by attention, but the Transformer regulates how the similarity coefficients are used. Raw inner products are first scaled according to key dimensionality:

$$\tilde{s} = \frac{q K^T}{\sqrt{d_k}}. \tag{A13}$$

This scaling controls their magnitude as $d_k$ increases. Softmax then transforms the scaled similarities into positive, competitive and normalized weights:

$$\alpha_p = \frac{\exp\left( \frac{q k_p^T}{\sqrt{d_k}} \right)}{\sum_{j=1}^{n} \exp\left( \frac{q k_j^T}{\sqrt{d_k}} \right)}. \tag{A14}$$

These coefficients satisfy

$$\sum_{p=1}^{n} \alpha_p = 1. \quad \text{(A15)}$$

Replacing the raw Hebbian similarity coefficients by these regulated weights gives

$$O_t = \sum_{p=1}^{n} \alpha_p v_p, \quad \text{(A16)}$$

or, in matrix form,

$$O_t = \text{softmax}\left(\frac{q K^T}{\sqrt{d_k}}\right) V, \quad \text{(A17)}$$

which is the scaled dot-product attention equation for a single query [1].

Thus, the transition from Eq. A10 to Eq. A17 preserves the central Hebbian computation: query–key similarities determine the contributions of associated values. Softmax adds competition and normalization, increasing the influence of strongly compatible keys while reducing weakly compatible contributions.

In a Transformer, learned projections generate queries, keys and values:

$$Q = X W_Q, K = X W_K, V = X W_V, \quad \text{(A18)}$$

and multiple queries are processed simultaneously:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) V. \quad \text{(A19)}$$

These projections define the spaces in which compatibility and value aggregation occur while preserving the same sequence of comparison and weighted retrieval.

Hebbian crosstalk can thus be viewed as unregulated similarity-weighted aggregation: stored values contribute whenever their keys resemble the query. Attention preserves this associative principle but regulates it through dimensional scaling, competitive normalization and learned projections, converting interference into a context-sensitive mixture.

This supports the proposed progression: thalamic and koniocortical processing first reduce overlap to form stable components and cortical frames; at higher cortical levels, similarity-dependent interactions may be reintroduced in regulated form. Attention can therefore be interpreted as a controlled, normalized form of the mechanism that generates Hebbian crosstalk.

## Data availability

No new datasets were generated in this study. Data underlying the previously published results discussed here are available through the cited original publications and their associated repositories.


## Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.


## Code availability

The MATLAB code used to generate the simulations reported in this study is available at https://osf.io/hkx9w.

## Author contributions

All authors contributed equally to this work, participated in its conception, analysis, interpretation and manuscript preparation, and approved the final version.

## Competing interests

The authors declare no competing interests.

## 7. References


[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).

[2] Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. arXiv:1607.06450 (2016).

[3] Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. MIT Press (2016).

[4] Sherman, S. M. & Guillery, R. W. Exploring the Thalamus and Its Role in Cortical Function. MIT Press (2006).

[5] Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. Nature Reviews Neuroscience 13, 51–62 (2012).

[6] Steriade, M., McCormick, D. A. & Sejnowski, T. J. Thalamocortical oscillations in the sleeping and aroused brain. Science 262, 679–685 (1993).

[7] McCormick, D. A. & Bal, T. Sleep and arousal: thalamocortical mechanisms. Annual Review of Neuroscience 20, 185–215 (1997).

[8] Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. Neuron 95, 245–258 (2017).

[9] Richards, B. A. et al. A deep learning framework for neuroscience. Nature Neuroscience 22, 1761–1770 (2019).

[10] Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. Nature Communications 10, 3770 (2019).

[11] Panassiol, R. T. & Peláez, F. J. R. Evaluating the cognitive and electrophysiological plausibility of a thalamic computational model. Advanced Neurology 3, 3188 (2024). https://doi.org/10.36922/an.3188

[12] Peláez, J. R. Towards a neural network based therapy for hallucinatory disorders. Neural Networks 13, 1047–1061 (2000).

[13] Peláez, F. J. R., Aguiar-Furucho, M. A. & Andina, D. Intrinsic plasticity for natural competition in koniocortex-like neural networks. International Journal of Neural Systems 26, 1650040 (2016). https://doi.org/10.1142/S0129065716500404

[14] Hebb, D. O. The Organization of Behavior. Wiley (1949).

[15] McClelland, J. L. & Rumelhart, D. E. Explorations in Parallel Distributed Processing. MIT Press (1988).

[16] Hertz, J., Krogh, A. & Palmer, R. G. Introduction to the Theory of Neural Computation. Addison-Wesley (1991).

[17] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, 318–362. MIT Press (1986).

[18] Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal 37, 233–243 (1991).

[19] Oja, E. Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology 15, 267–273 (1982).

[20] Carandini, M. & Heeger, D. J. Summation and division by neurons in primate visual cortex. Science 264, 1333–1336 (1994).

[21] Peláez, F. J. R. & Andina, D. Do biological synapses perform probabilistic computations? Neurocomputing 114, 24–31 (2013).

[22] Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. Science 313, 504–507 (2006).

[23] Diamantaras, K. I. & Kung, S. Y. Principal Component Neural Networks: Theory and Applications. Wiley (1996).

[24] Peláez, J. R. Plato's theory of ideas revisited. Neural Networks 10, 1269–1288 (1997).

[25] Sherman, S. M. & Guillery, R. W. Functional Connections of Cortical Areas: A New View from the Thalamus. MIT Press (2013).

[26] Famiglietti, E. V. & Peters, A. The synaptic glomerulus and the intrinsic neuron in the dorsal lateral geniculate nucleus of the cat. Journal of Comparative Neurology 144, 285–334 (1972).

[27] Borg-Graham, L. J., Monier, C. & Frégnac, Y. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. Nature 393, 369–373 (1998).

[28] Holt, G. R. & Koch, C. Shunting inhibition does not have a divisive effect on firing rates. Neural Computation 9, 1001–1013 (1997).

[29] Destexhe, A., Contreras, D., Sejnowski, T. J. & Steriade, M. A model of spindle rhythmicity in the isolated thalamic reticular nucleus. Journal of Neurophysiology 72, 803–818 (1994).

[30] Landisman, C. E. et al. Electrical synapses in the thalamic reticular nucleus. Journal of Neuroscience 22, 1002–1009 (2002).

[31] Fuentealba, P. & Steriade, M. The reticular nucleus revisited: intrinsic and network properties of a thalamic pacemaker. Progress in Neurobiology 75, 125–141 (2005).

[32] Mulle, C., Madariaga, A. & Deschênes, M. Morphology and electrophysiological properties of reticularis thalami neurons in cat: in vivo study of a thalamic pacemaker. Journal of Neuroscience 6, 2134–2145 (1986).

[33] Yuille, A. L. & Grzywacz, N. M. A winner-take-all mechanism based on presynaptic inhibition feedback. Neural Computation 1, 334–347 (1989).

[34] Ermentrout, B. Complex dynamics in winner-take-all neural nets with slow inhibition. Neural Networks 5, 415–431 (1992).

[35] Coultrip, R., Granger, R. & Lynch, G. A cortical model of winner-take-all competition via lateral inhibition. Neural Networks 5, 47–54 (1992).

[36] Mao, Z. H. & Massaquoi, S. G. Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition. IEEE Transactions on Neural Networks 18, 55–69 (2007).

[37] Brandt, S. F. & Wessel, R. Winner-take-all selection in a neural system with delayed feedback. Biological Cybernetics 97, 221–228 (2007).

[38] Bridle, J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Neurocomputing: Algorithms, Architectures and Applications, 227–236. Springer (1990).

[39] Bishop, C. M. Pattern Recognition and Machine Learning. Springer (2006).

[40] von Krosigk, M., Monckton, J. E., Reiner, P. B. & McCormick, D. A. Dynamic properties of corticothalamic excitatory postsynaptic potentials and thalamic reticular inhibitory postsynaptic potentials in thalamocortical neurons of the guinea-pig dorsal lateral geniculate nucleus. Neuroscience 91, 7–20 (1999).

[41] Steriade, M., Jones, E. G. & McCormick, D. A. Thalamus: Organization and Function. Elsevier (1997).

[42] Granado, L. C., Ranvaud, R. & Ropero Peláez, J. A spiderless arachnophobia therapy: comparison between placebo and treatment groups and six-month follow-up study. Neural Plasticity 2007, 10241 (2007).

[43] Javanbakht, A. Sensory gating deficits, pattern completion, and disturbed fronto-limbic balance: a model for description of hallucinations and delusions in schizophrenia. Medical Hypotheses 67, 1173–1184 (2006).

[44] Javanbakht, A. & Ragan, C. L. A neural network model for transference and repetition compulsion based on pattern completion. Journal of the American Academy of Psychoanalysis and Dynamic Psychiatry 36, 255–278 (2008).

[45] Miller, K. D., Pinto, D. J. & Simons, D. J. Processing in layer IV of neocortical circuits: new insights from visual and somatosensory cortex. Current Opinion in Neurobiology 11, 488–497 (2001).

[46] Ferster, D., Chung, S. & Wheat, H. Orientation selectivity of thalamic input to simple cells of cat visual cortex. Nature 380, 249–252 (1996).

[47] Miller, K. D. Synaptic economics: competition and cooperation in synaptic plasticity. Neuron 17, 371–374 (1996).

[48] da Costa, N. M. & Martin, K. A. C. How thalamus connects to spiny stellate cells in the cat's visual cortex. Journal of Neuroscience 31, 2925–2937 (2011).

[49] Diamond, M. E., Armstrong-James, M. & Ebner, F. F. Experience-dependent plasticity in adult rat barrel cortex. Proceedings of the National Academy of Sciences USA 90, 2082–2086 (1993).

[50] Fleidervish, I. A., Binshtok, A. M. & Gutnick, M. J. Functionally distinct NMDA receptors mediate horizontal connectivity within layer IV of mouse barrel cortex. Neuron 21, 1055–1065 (1998).

[51] Chung, S., Li, X. & Nelson, S. B. Short-term depression at thalamocortical synapses contributes to rapid adaptation of cortical sensory responses in vivo. Neuron 34, 437–446 (2002). https://doi.org/10.1016/S0896-6273(02)00659-1

[52] Feldman, D. E., Nicoll, R. A., Malenka, R. C. & Isaac, J. T. R. Long-term depression at thalamocortical synapses in developing rat somatosensory cortex. Neuron 21, 347–357 (1998). https://doi.org/10.1016/S0896-6273(00)80544-9

[53] Okhotin, V. E. Cytophysiology of spiny stellate cells in the striate cortex and their role in the excitatory mechanisms of intracortical synaptic circulation. Neuroscience and Behavioral Physiology 36, 825–836 (2006).

[54] Desai, N. S. Homeostatic plasticity in the CNS: synaptic and intrinsic forms. Journal of Physiology Paris 97, 391–402 (2003).

[55] Desai, N. S., Rutherford, L. C. & Turrigiano, G. G. Plasticity in the intrinsic excitability of cortical pyramidal neurons. Nature Neuroscience 2, 515–520 (1999).

[56] Ropero Peláez, F. J. & Godoi, A. C. From forced to natural competition in a biologically plausible neural network. In Estévez, P. A., Príncipe, J. C. & Zegers, P. (eds.) Advances in Self-Organizing Maps, Advances in Intelligent Systems and Computing, vol. 198, 95–104. Springer (2013).

[57] Artola, A., Bröcher, S. & Singer, W. Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. Nature 347, 69–72 (1990).

[58] Fombellida, J., Ropero Peláez, F. J. & Andina, D. Koniocortex-like network unsupervised learning surpasses supervised results on WBCD breast cancer database. In IWINAC 2017, LNCS 10338, 32–41. Springer (2017).

[59] Fombellida, J., Martín-Rubio, I., Romera-Zarza, A. & Andina, D. KLN, a new biological koniocortex based unsupervised neural network: competitive results on credit scoring. Natural Computing (2018). https://doi.org/10.1007/s11047-018-9698-6

[60] Fombellida, J. & Andina, D. Koniocortex-like network application to business intelligence. 2018 World Automation Congress, 1–5 (2018). https://doi.org/10.23919/WAC.2018.8430298

[61] Torres-Alegre, S., Benchaib, Y., Ferrández Vicente, J. M. & Andina, D. Application of koniocortex-like networks to cardiac arrhythmias classification. In IWINAC 2019, LNCS 11487, 264–273. Springer (2019). https://doi.org/10.1007/978-3-030-19651-6_26

[62] DeFelipe, J., & Fariñas, I. The pyramidal neuron of the cerebral cortex: morphological and chemical characteristics of the synaptic inputs. Progress in Neurobiology 39, 563–607 (1992).

[63] Poirazi, P., Brannon, T., & Mel, B. W. Pyramidal neuron as two-layer neural network. Neuron 37, 989–999 (2003).

[64] Papoutsi, A., Kastellakis, G., Psarrou, M., Anastasakis, S., & Poirazi, P. Coding and decoding with dendrites. Journal of Physiology Paris 108, 18–27 (2014).

[65] Eyal, G., Verhoog, M. B., Testa-Silva, G., Deitcher, Y., Benavides-Piccione, R., DeFelipe, J., de Kock, C. P. J., Mansvelder, H. D., Segev, I. Human cortical pyramidal neurons: from spines to spikes via models. Frontiers in Cellular Neuroscience 12, 181 (2018).

[66] Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., Holtkamp, M., Vida, I., & Larkum, M. E. Dendritic action potentials and computation in human layer 2/3 cortical neurons. Science 367, 83–87 (2020).

[67] Panassiol, R. T. Em Direção a um Sistema Computacional Tálamo-Cortical: Uma Rede Neural Bioinspirada pela Citoarquitetura e Fisiologia Cerebral. Doctoral thesis, Institute of Psychology, University of São Paulo (2024).

[68] Llinás, R., Ribary, U., Joliot, M., & Wang, X.-J. Content and context in temporal thalamocortical binding. In Temporal Coding in the Brain (1994).

[69] Llinás, R., Leznik, E., & Urbano, F. J. Temporal binding via cortical coincidence detection of specific and nonspecific thalamocortical inputs: a voltage-dependent dye-imaging study in mouse brain slices. Proceedings of the National Academy of Sciences USA 99, 449–454 (2002).

[70] Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science, 275, 213–215 (1997).

[71] Jahnsen, H. & Llinás, R. Electrophysiological properties of guinea-pig thalamic neurones: an in vitro study. Journal of Physiology 349, 205–226 (1984).

[72] Sherman, S. M. Tonic and burst firing: dual modes of thalamocortical relay. Trends in Neurosciences 24, 122–126 (2001).

[73] Swadlow, H. A. & Gusev, A. G. The impact of 'bursting' thalamic impulses at a neocortical synapse. Nature Neuroscience 4, 402–408 (2001).

[74] Beierlein, M., Fall, C. P., Rinzel, J. & Yuste, R. Thalamocortical bursts trigger recurrent activity in neocortical networks: layer 4 as a frequency-dependent gate. Journal of Neuroscience 22, 9885–9894 (2002).

[75] Friederici, A. D. The brain basis of language processing: from structure to function. Physiological Reviews 91, 1357–1392 (2011).

[76] Kandel, E. R. et al. Principles of Neural Science. 5th ed. McGraw-Hill (2013).

[77] Steriade, M. The GABAergic reticular nucleus: a preferential target of corticothalamic projections. Proceedings of the National Academy of Sciences USA 98, 3625–3627 (2001).

[78] Crandall, S. R., Cruikshank, S. J. & Connors, B. W. A corticothalamic switch: controlling the thalamus with dynamic synapses. Neuron 86, 768–782 (2015).

[79] Stevens, J. R., Venkatesan, R., Dai, S., Khailany, B. & Raghunathan, A. Softermax: Hardware/software co-design of an efficient softmax for Transformers. Proceedings of the 58th ACM/IEEE Design Automation Conference (2021).

[80] Dong, S. et al. Topkima-Former: Low-energy, low-latency inference for Transformers using top-k in-memory ADC. arXiv:2411.13050 (2024).

[81] Wang, R. et al. VEXP: A low-cost RISC-V ISA extension for accelerated softmax computation in Transformers. arXiv:2504.11227 (2025).

[82] Kim, R., Lee, D., Kim, J., Park, J. & Lee, S. E. Hardware accelerator for approximation-based softmax and layer normalization in Transformers. Electronics 14, 2337 (2025).

[83] Wang, W., Zhou, S., Sun, W., Sun, P. & Liu, Y. SOLE: Hardware-software co-design of Softmax and LayerNorm for efficient Transformer inference. arXiv:2510.17189 (2025).

[84] Zhou, C. et al. Winner-take-all spiking Transformer for language modeling. arXiv:2604.11321 (2026).

[85] Maass, W. On the computational power of winner-take-all. Neural Computation 12, 2519–2535 (2000).