Multihead self-attention in cortico-thalamic circuits

Arno Granier[a,b] and Walter Senn[a,c,*]

[a]Department of Physiology, University of Bern, Switzerland
[b]Graduate School for Cellular and Biomedical Sciences, University of Bern, Switzerland
[c]Center for Artificial Intelligence in Medicine, University of Bern, Switzerland
[*]correspondence: walter.senn@unibe.ch

May 2, 2025

Keywords: cortical circuits | transformers | thalamus | gradient-based learning

Abstract

Both biological cortico-thalamic networks and artificial transformer networks use canonical computations to perform a wide range of cognitive tasks. In this work, we propose that the structure of cortico-thalamic circuits is well suited to realize a computation analogous to multihead self-attention, the main algorithmic innovation of transformers. We start with the concept of a cortical unit module or microcolumn, and propose that superficial and deep pyramidal cells carry distinct computational roles. Specifically, superficial pyramidal cells encode an attention mask applied onto deep pyramidal cells to compute attention-modulated values. We show how to wire such microcolumns into a circuit equivalent to a single head of self-attention. We then suggest the parallel between one head of attention and a cortical area. On this basis, we show how to wire cortico-thalamic circuits to perform multihead self-attention. Along these constructions, we refer back to existing experimental data, and find noticeable correspondence. Finally, as a first step towards a mechanistic theory of synaptic learning in this framework, we derive formal gradients of a tokenwise mean squared error loss for a multihead linear self-attention block.

Significance statement

While artificial intelligence has been inspired by neuronal processing in the brain, the success of artificial intelligence, in turn, inspires neuroscience. The notion of self-attention is the central algorithmic innovation underlying recent progress in artificial intelligence, including Large Language Models. In essence, self-attention allows the representation of each element in a sequence to be influenced by representations of the other elements (a 'river bank' and an 'investment bank' are different 'banks'). We show that self-attention can be realized by neuronal circuits involving the thalamus and the cerebral cortex, their known structured connectivity, as well as the known properties of cortical pyramidal neurons and their organization in functionally specialized types.

1 Introduction

In the mammalian neocortex, basic elements and rules of connectivity are similar across functionally specialized areas and mammalian species (Harris and Shepherd 2015). Recent evidence supports this hypothesis of a canonical cortical structure (Powell et al. 2024; Meyer et al. 2025). Furthermore, every cortical area sends and receives projections to and from the thalamus, and the classical focus on purely cortical computation gives way to a view where the cortex and the thalamus are tightly intertwined (Suzuki, Pennartz, and Aru 2023; Sherman and Usrey 2024). On the functional side, transformer networks (Vaswani et al. 2017; Phuong and Hutter 2022) have achieved impressive feats in a variety of cognitive tasks including, but not limited to, natural language processing (e.g.,

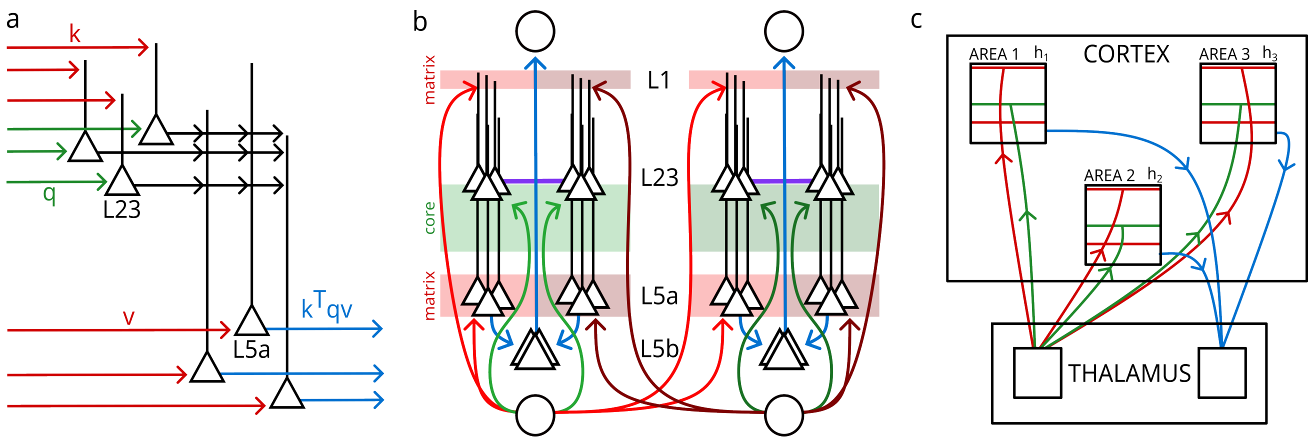

Figure 1: Neuronal circuits for multihead self-attention. (a) A cortical attention microcolumn. (b) One head of self-attention with cortical microcolumns. Bottom circles are two initial sequence elements, while top circles are context-aware representations for the same elements. Blue projections have weights $\boldsymbol{W}_O^h$, green projections $\boldsymbol{W}_Q^h$, red projections to the apex of superficial pyramidal cells $\boldsymbol{W}_K^h$, and red projections to the base of deep pyramidal cells $\boldsymbol{W}_V^h$. Divisive normalization used in the attention kernel is represented in purple. One vertical set of superficial and deep IT pyramidal cells (L5a) acts as in panel a; connections from superficial to deep pyramidal cells are omitted in this drawing. Attention modulated values are pooled in deep PT pyramidal cells (L5b). (c) Multihead self-attention. One square within the bottom structure (thalamus) represents a full sequence, while one square in the top structure (cortex) represents a head of self-attention, analogous to a cortical area.

Dosovitskiy et al. 2021; OpenAI 2022; Alayrac et al. 2022). Given that both transformers and cortico-thalamic circuits excel in a variety of cognitive tasks using canonical architectures, we ask whether these architectures have commonalities. Recent work shows that transformer networks develop brain-like representations of auditory (Li et al. 2023) and language processing (Schrimpf et al. 2021; Caucheteux and King 2022). The implementations of transformer-like computation in a circuit of neurons and astrocytes (Kozachkov, Kastanenka, and Krotov 2023) and a model of the hippocampal formation (Whittington, Warren, and Behrens 2022; see also Gershman, Fiete, and Irie 2025) have been suggested. In this work, we propose that the structure of cortico-thalamic circuits, cell types, pathways and interactions, are well suited to implement multihead self-attention, the main algorithmic innovation of transformers.

2 Cortical attention microcolumns

We adopt the concept of a cortical unit module or microcolumn, and interpret its operation in terms of the elementary computation of self-attention in transformer networks, modulating values by a similarity matching of keys and queries (Vaswani et al. 2017; Phuong and Hutter 2022). Within a cortical microcolumn, multiple cell-type-specific populations interact, see fig. 1a. We propose that superficial (layer 2/3) and deep (layer 5) pyramidal cells fulfill fundamentally distinct computational roles. Superficial pyramidal cells compute an attention signal by comparing keys and queries formed by the thalamic inputs at their apical and basal dendrites, respectively. Intratelencephalic (IT) deep pyramidal cells in turn receive inputs from local superficial pyramidal cells at their apical dendrites and combine it with values computed in their basal dendrites, to represent the attention-modulated values in the soma. Consequently, within a microcolumn, the number of superficial pyramidal cells is the dimension of the key and query vectors d_k, and the number of deep pyramidal cells is the dimension of the value vector d_v. Both of these computations are realized by computing the product of basal and apical inputs in the soma, interpreted as gain-modulated somatic activity (Larkum, Senn, and Lüscher 2004), see fig. 2a. In accordance with this motif, Quiquempoix et al. (2018) show that superficial pyramidal cells act as controllers of the gain of deep pyramidal cells, see fig. 2b. With all-to-all unitary lateral weights $\boldsymbol{A}$ from superficial to deep IT pyramidal cells, this motif implements the core operation of multiplying a vector ($\boldsymbol{v}$) by the dot product of two other vectors ($\boldsymbol{k}$ and $\boldsymbol{q}$), namely $(\boldsymbol{A}(\boldsymbol{k} \odot \boldsymbol{q})) \odot \boldsymbol{v} = (\boldsymbol{k}^T \boldsymbol{q})\, \boldsymbol{v}$, with $\boldsymbol{A}$ the matrix full of ones of dimension $d_v \times d_k$ (learning $\boldsymbol{A}$ would add another degree of freedom in the computation of the attention signal for layer 5 neuron).

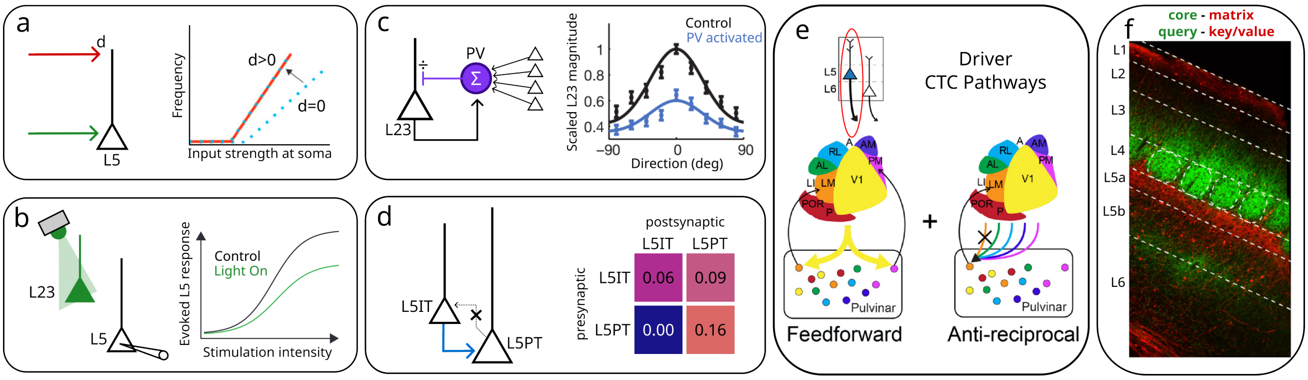

Figure 2: Experimental observation supporting cortical self-attention. (a) Apical input has a multiplicative impact on the somatic firing of deep pyramidal cells. Adapted from Larkum, Senn, and Lüscher (2004). (b) Superficial pyramidal cells modulate the gain of deep pyramidal cells. In this experiment superficial pyramidal cells are photo-inhibited (green line). Adapted from Quiquempoix et al. (2018). (c) Divisive normalization in superficial cortical layers. Circuit diagram inspired by Carandini and Heeger (1994). Experimental data adapted from Wilson et al. (2012). (d) The connectivity between layer 5 IT and layer 5 PT pyramidal cells is unidirectional (from IT to PT). Data from Campagnola et al. (2022). Numbers are connection probabilities. (e) Layer 5 PT pyramidal cells form feedforward cortico-thalamic-cortical (CTC) pathways ('anti-reciprocal') and avoid reciprocal loops. Reproduced from Cassidy et al. (2025). (f) Core thalamo-cortical projections (green) are dense, modular, and mainly target layer 4 and lower layer 3. Matrix thalamo-cortical projections (red) are more diffuse and target mainly layers 1 and 5a. Fluorescence imaging data reproduced from Sermet et al. (2019).

3 Wiring cortical microcolumns for self-attention

We continue by showing how to wire microcolumns into a circuit equivalent to a single head h of self-attention, see fig. 1b. To achieve this goal, we use n^2 microcolumns, organized into n macrocolumns composed each of n microcolumns, where n represents the number of tokens ('words') processed in parallel. This incurs a quadratic scaling of the number of neurons with n and, at least formally, duplications of synaptic weights.Macrocolumn i will compute the context-aware value $\widetilde{\boldsymbol{v}}_i^h = \sum_j \kappa(\boldsymbol{k}_j, \boldsymbol{q}_i)\, \boldsymbol{v}_j$ of its associated sequence element $\boldsymbol{x}_i$, with $\boldsymbol{q}_i = \boldsymbol{W}_Q^h \boldsymbol{x}_i$, $\boldsymbol{k}_j = \boldsymbol{W}_K^h \boldsymbol{x}_j$, and $\boldsymbol{v}_j = \boldsymbol{W}_V^h \boldsymbol{x}_j$. Vectors $\boldsymbol{x}_i$ are of embedding dimension d_e, $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ of dimension d_k, and $\boldsymbol{v}_j$ of dimension d_v (typically, $d_k = d_v = d_e/H$ with H the number of heads). To compute $\widetilde{\boldsymbol{v}}_i^h$, each microcolumn j in the i-th macrocolumn calculates the same query $\boldsymbol{q}_i$ in the basal dendrites of superficial pyramidal cells, but its own key $\boldsymbol{k}_j$ in the apical dendrites of superficial pyramidal cells and value $\boldsymbol{v}_j$ in the basal dendrites of deep pyramidal cells, see fig. 1b. As for the attention kernel $\kappa : \mathbb{R}^{d_k} \times \mathbb{R}^{d_k} \to \mathbb{R}$, the classical choice of $\mathrm{softmax}(\boldsymbol{k}_j^T \boldsymbol{q}_i / \sqrt{d_k})$ appears hard to implement neuronally, notably because it would use the pooled activity of apical dendrites of deep pyramidal cells. The kernel $\kappa(\boldsymbol{k}_j, \boldsymbol{q}_i) = \phi(\boldsymbol{k}_j)^T \phi(\boldsymbol{q}_i)/Z_i$ with $Z_i = \sum_{j'} \phi(\boldsymbol{k}_{j'})^T \phi(\boldsymbol{q}_i)$, called linear self-attention (Katharopoulos et al. 2020; Peng et al. 2021; Schlag, Irie, and Schmidhuber 2021; Choromanski et al. 2022), is more straightforward. It only implies the introduction of a (dendritic) positive nonlinearity ϕ applied elementwise to keys (apical dendrites) and queries (basal dendrites), and a division by the pooled somatic activity of superficial pyramidal cells realized by divisive lateral and recurrent inhibition within each macrocolumn, see fig. 2c. Finally, the sum of attention-modulated values $\kappa(\boldsymbol{k}_j, \boldsymbol{q}_i)\, \boldsymbol{v}_j$ is performed by pooling the activity of deep IT pyramidal cells (L5a) into a set of deep pyramidal tract (PT) pyramidal cells (L5b), yielding the output of a macrocolumn $\widetilde{\boldsymbol{v}}_i^h$, see fig. 1b. This motif is supported by the reported unidirectional local connectivity from IT to PT neurons, see fig. 2d.

4 Multihead self-attention in cortico-thalamic pathways

We suggest the parallel between one head of attention, as just defined, and a cortical area. The number of macrocolumns in an area and the number of microcolumns in a macrocolumn should then be equivalent. At least for the well-studied primary somatosensory cortex of the rodent responsible for whisker perception (barrel cortex), this is consistent with the reported dimensions of these structures. Specifically, there are $n = 33$ macrocolumns ('barrels') of diameter $\sim 200\mu\text{m}$ (Petersen 2019). Within one macrocolumn, ~ 33 microcolumns of diameter $\sim 30\mu\text{m}$ (Buxhoeveden and Casanova 2002, see also Maruoka et al. 2017) can be packed. The ~ 100 neurons in each microcolumn (Buxhoeveden and Casanova 2002) would allow for keys, queries, and values of dimension $d_k = d_v \simeq 33$, considering a third of all neurons in a microcolumn as layer 5 pyramidal cells, and another third as layer 2/3 pyramidal cells. Generalizing the numbers reported for mouse barrel cortex, each single head (cortical area) would contain on the order of magnitude of $\sim 10^5$ neurons. The whole mouse cortex, with on the order of magnitude of $\sim 10^7$ neurons (Herculano-Houzel, Mota, and Lent 2006), could then host a total of ~ 100 heads.

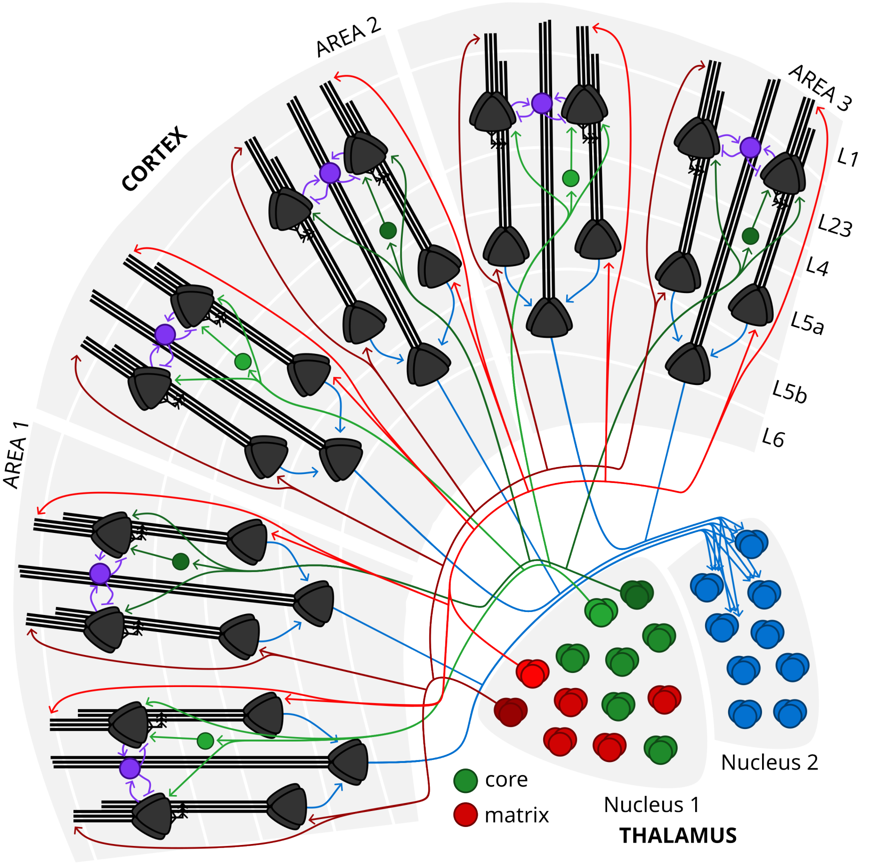

Figure 3: Multihead self-attention in cortico-thalamic pathways. The initial sequence is encoded in the thalamic nucleus 1, corresponding to the input of a multihead self-attention block. It is then distributed to H cortical areas through core (green) and matrix (red) thalamo-cortical projections. Local cortical computation results in the computation of context-aware values $\widetilde{\boldsymbol{v}}_i^h$ in deep PT pyramidal cells (L5b), as described in the main text. The purple interneurons depict the divisive normalization used in the attention kernel. PT pyramidal cells then project to another (higher-order) thalamic nucleus 2 (blue), where context-aware values coming from multiple areas are summed. The final context-aware representation $\widetilde{\boldsymbol{x}}_i$ is thus represented in the thalamic nucleus 2, corresponding to the output of a multihead self-attention block.

In a multihead self-attention block, attention heads act independently and additively (Elhage et al. 2021), processing different aspects of the input in parallel (for example in the visual system: shape, color, motion, etc.). The output of the block, the context-aware representations, can then be written as $\widetilde{\boldsymbol{x}}_i = \sum_h \boldsymbol{W}_O^h \widetilde{\boldsymbol{v}}_i^h$, with $\boldsymbol{W}_O^h$ of dimension $d_e \times d_v$. In a thalamo-cortico-thalamic pathway, the sequence elements $\boldsymbol{x}_i$, originating from a thalamic nucleus, are processed in parallel by a subset of H cortical areas, and their outputs are integrated in another thalamic nucleus to yield $\widetilde{\boldsymbol{x}}_i$, see fig. 1c. In such a pathway, the thalamo-cortical projections are assigned the synaptic weights $\boldsymbol{W}_Q^h$, $\boldsymbol{W}_K^h$, and $\boldsymbol{W}_V^h$, while the cortico-thalamic projections are assigned the weights $\boldsymbol{W}_O^h$. Algorithm 1 sums up the necessary computation with an exact implementation of multihead linear self-attention annotated with the mapping to cortico-thalamic circuits. This general organization into 'transcortical' pathways connecting two thalamic nuclei through a subset of cortical areas is consistent with the observation that single thalamic cells integrate information from different cortical areas (Sampathkumar et al. 2021), and that deep PT pyramidal cells avoid driving thalamic nuclei projecting to their cortical area (Cassidy et al. 2025), see fig. 2e. Moreover, in our interpretation, query projections need to be dense, focused, and target cortical layers 3 and 4, matching the observed structure of core thalamo-cortical projections. Key and value projections need to be sparse, diffuse, and target cortical layers 1 and 5a, matching the observed structure of matrix thalamo-cortical projections (Jones 1998; Sermet et al. 2019), see fig. 2f and fig. 3.

Algorithm 1 MultiHead linear Self-Attention (MHSA) in cortico-thalamic circuits.

procedure MHSA($\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) ▷ MHSA input ($n$ tokens) — Thalamus (e.g. via L4)
  $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n = \text{zeros}(d_e), \ldots, \text{zeros}(d_e)$ ▷ $d_e =$ token embedding dim — #neurons per token
  $\boldsymbol{A} = \text{ones}(d_v, d_k)$ ▷ Local weights from L23 to Apical L5
  for $h = 1 : H$ **do** ▷ For all heads / cortical areas
    for $i = 1 : n$ **do** ▷ For all query / macrocolumn indices
      $Z = 0\,,\ \widetilde{\boldsymbol{v}} = \text{zeros}(d_v)$ ▷ $d_v =$ value dimension, typically, $d_v = d_e/H$
      for $j = 1 : n$ **do** ▷ For all key / microcolumn indices
        $Z = Z + \mathbf{1}^T(\phi(\boldsymbol{W}_K^h \boldsymbol{x}_j) \odot \phi(\boldsymbol{W}_Q^h \boldsymbol{x}_i))$ ▷ Divisive normalization — Interneurons L23
      end for ▷ End summing all L23 activities within macrocolumn
      for $j = 1 : n$ **do** ▷ For all key / microcolumn indices
        $\boldsymbol{q} = \boldsymbol{W}_Q^h \boldsymbol{x}_i$ ▷ Query — Basal L23 (in general: $\boldsymbol{q}^{h,ij} = \boldsymbol{W}_Q^{h,ij} \boldsymbol{x}_i$)
        $\boldsymbol{k} = \boldsymbol{W}_K^h \boldsymbol{x}_j$ ▷ Key — Apical L23 (in general: $\boldsymbol{k}^{h,ij} = \boldsymbol{W}_K^{h,ij} \boldsymbol{x}_j$)
        $\boldsymbol{s} = \phi(\boldsymbol{k}) \odot \phi(\boldsymbol{q})/Z$ ▷ Salience — Soma L23 ($\boldsymbol{s} = \boldsymbol{s}^{h,ij}$)
        $\boldsymbol{v} = \boldsymbol{W}_V^h \boldsymbol{x}_j$ ▷ Value – Basal L5IT (in general: $\boldsymbol{v}^{h,ij} = \boldsymbol{W}_V^{h,ij} \boldsymbol{x}_j$)
        $\boldsymbol{\alpha} = \boldsymbol{A}\boldsymbol{s}$ ▷ Attention / summed salience — Apical L5IT
        $\boldsymbol{y} = \boldsymbol{\alpha} \odot \boldsymbol{v}$ ▷ Attention-modulated value — Soma L5IT (L5a)
        $\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{v}} + \boldsymbol{y}$ ▷ Sum $\boldsymbol{y}$ across keys/values (j) = macrocolumn output — L5PT (L5b)
      end for ▷ End summing attention-modulated values within macrocolumn
      $\widetilde{\boldsymbol{x}}_i = \widetilde{\boldsymbol{x}}_i + \boldsymbol{W}_O^h \widetilde{\boldsymbol{v}}$ ▷ Sum across queries (i) and heads (h) — Higher-order Thalamus
    end for ▷ End summing context-aware values across macrocolumns / queries
  end for ▷ End summing context-aware values across areas / heads
  return $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n$ ▷ Output of MHSA — Higher-order Thalamus
end procedure ▷ End MHSA for the $\#h$ of cortical areas

5 Formal gradients of multihead self-attention parameters

After showing how neurons could be wired to realize the computation of a multihead self-attention block, a pressing question is the one of learning the synaptic weights $\boldsymbol{W}^h_{\{K,Q,V,O\}}$. Energy-based approaches offer a formalism to derive local learning rules in neuronal circuits by jointly minimizing a layerwise loss and a cost on the outputs (Scellier and Bengio 2017; Senn et al. 2024; Song et al. 2024; Ellenberger et al. 2024; see also Hoover et al. 2023 for inference as energy minimization). The energy is most often a sum of the square norm of the local errors $\boldsymbol{e}_i$, in our case the error of a multihead linear self-attention block with targets $\boldsymbol{y}_i$,

$$E = \sum_i \tfrac{1}{2}\|\boldsymbol{e}_i\|^2 = \sum_i \tfrac{1}{2}\|\boldsymbol{y}_i - \widetilde{\boldsymbol{x}}_i\|^2 \,, \tag{1}$$

with again $\widetilde{\boldsymbol{x}}_i = \sum_h \boldsymbol{W}_O^h \widetilde{\boldsymbol{v}}_i^h$, $\widetilde{\boldsymbol{v}}_i^h = \sum_j \kappa(\boldsymbol{k}_j, \boldsymbol{q}_i)\boldsymbol{v}_j$, $\kappa(\boldsymbol{k}_j, \boldsymbol{q}_i) = \phi(\boldsymbol{k}_j)^T\phi(\boldsymbol{q}_i)/Z_i$, and $Z_i = \sum_{j'} \phi(\boldsymbol{k}_{j'})^T\phi(\boldsymbol{q}_i)$.

Quantities of interest in the derivation of cost-minimizing synaptic learning rules are the negative energy gradients, along which the synaptic weights are changed,

$$\dot{\boldsymbol{W}}_O \propto -\partial_{\boldsymbol{W}_O} E = \sum_i \boldsymbol{e}_i \widetilde{\boldsymbol{v}}_i^T \,, \tag{2}$$

$$\dot{\boldsymbol{W}}_V \propto -\partial_{\boldsymbol{W}_V} E = \sum_{i,j} \kappa(\boldsymbol{k}_j, \boldsymbol{q}_i)\boldsymbol{\Delta}_{ij} \,, \tag{3}$$

$$\dot{\boldsymbol{W}}_Q \propto -\partial_{\boldsymbol{W}_Q} E = \sum_{i,j} Z_i^{-1}[\boldsymbol{\Gamma}^q_{ijj} - \kappa(\boldsymbol{k}_j, \boldsymbol{q}_i)\sum_{j'}\boldsymbol{\Gamma}^q_{ijj'}] \,, \tag{4}$$

$$\dot{\boldsymbol{W}}_K \propto -\partial_{\boldsymbol{W}_K} E = \sum_{i,j} Z_i^{-1}[\boldsymbol{\Gamma}^k_{ijj} - \kappa(\boldsymbol{k}_j, \boldsymbol{q}_i)\sum_{j'}\boldsymbol{\Gamma}^k_{ijj'}] \,. \tag{5}$$

Here, $\boldsymbol{\Delta}_{ij} = \boldsymbol{W}_O^T \boldsymbol{e}_i \boldsymbol{x}_j^T$, $\boldsymbol{\Gamma}^q_{ijj'} = [\phi'(\boldsymbol{q}_i) \odot \phi(\boldsymbol{k}_{j'})]\boldsymbol{v}_j^T \boldsymbol{\Delta}_{ii}$, $\boldsymbol{\Gamma}^k_{ijj'} = [\phi(\boldsymbol{q}_i) \odot \phi'(\boldsymbol{k}_{j'})]\boldsymbol{v}_j^T \boldsymbol{\Delta}_{ij'}$, and we omit the head index which is always h. These are the formal gradient of a tokenwise mean squared error loss backpropagated through one multihead linear self-attention block. We check these results against numerical approximations based on finite differences and forward automatic differentiation (`https://github.com/arnogranier/Formal-MHSA-gradient-numerics`).

Details of neural mechanisms and circuits allowing synaptic learning rules to follow these gradients are left pending. Yet, as a first step, we might consider unnormalized attention with $Z_i = 1$. In that case, eq. (4) reduces to $\dot{\boldsymbol{W}}_Q \propto \sum_{i,j} \boldsymbol{\Gamma}^q_{ijj}$ and eq. (5) to $\dot{\boldsymbol{W}}_K \propto \sum_{i,j} \boldsymbol{\Gamma}^k_{ijj}$. As a second step, we may define microcolumn-specific weights $\boldsymbol{W}^{ij}_Q, \boldsymbol{W}^{ij}_K$, $\boldsymbol{W}^{ij}_V$, and $\boldsymbol{W}^{i}_O$ (see Algorithm 1). In that case, eqs. (2) to (5) reduce to local learning rules. For instance, the learning rule for keys, eq. (5), becomes $\dot{\boldsymbol{W}}^{ij}_K \propto -\partial_{\boldsymbol{W}^{ij}_K} E = \boldsymbol{\Gamma}^k_{ijj} = (\boldsymbol{v}_j^T \boldsymbol{b}_i)\ \widetilde{\boldsymbol{q}}_i^{(j)} \boldsymbol{x}_j^T$, with $\widetilde{\boldsymbol{q}}_i^{(j)} = \phi(\boldsymbol{q}_i) \odot \phi'(\boldsymbol{k}_j)$ and $\boldsymbol{b}_i = \boldsymbol{W}^{iT}_O \boldsymbol{e}_i$ the 'backpropagated error'. For each entry of the synaptic weight matrix $\boldsymbol{W}^{ij}_K$, the plasticity rule takes the form of a product of a modulated postsynaptic activity (the corresponding component of $\widetilde{\boldsymbol{q}}_i^{(j)}$ modulated by the common factor $\boldsymbol{v}_j^T \boldsymbol{b}_i$) times the presynaptic activity (the corresponding component of $\boldsymbol{x}_j^T$). It is crucial to realize that these expressions do not require full matrix multiplications, but instead only scalar products scaling outer products, and this can be read as correlation-based synaptic learning rules.

Making the learning rules local in space alleviates the problem of weight sharing, and allows for a potential implementation in terms of dedicated dendritic compartments that calculate the postsynaptic components of the plasticity rule (see e.g. Sacramento et al. 2018; Senn et al. 2024). The microcolumn-specific weights for keys and queries further allows for personalizing the query for each key and vice verse. Whether the biological realism retains the performance of classical transformer networks (or even improves it) remains to be investigated.

6 Discussion

Summary. In this work, we draw parallels between multihead linear self-attention and the structure of cortico-thalamic circuits. We first adopt the concept of a cortical unit module or microcolumn, in which superficial pyramidal cells compute an attention signal while deep pyramidal cells compute attention-modulated values. We show how to wire microcolumns into a circuit equivalent to a single head of self-attention. We then suggest the parallel between one head of attention and a cortical area, and show how to wire cortico-thalamic circuits to perform multihead self-attention. Finally, we derive formal gradients of a tokenwise mean squared error loss for a linear multihead self-attention block.

Fully-connected blocks, residual stream, normalization. Reproducing the full computation of a transformer block entails the sequential computation of multihead self-attention and a tokenwise 2-layers fully connected block, their integration into a residual stream, and their normalization (Elhage et al. 2021, see Shen, Wang, and Navlakha 2021 for an account of biologically plausible normalization). A future direction is to analyze whether these additional computations could be realized by direct cortico-cortical projections (see fig. 4b), reciprocal thalamo-cortical loops through pyramidal cells of cortical layer 6 (see fig. 4c and Cassidy et al. 2025), local circuits, or further relays of projections (e.g., through cortical layer 4). For instance, the subcomponent of core thalamocortical projections ending in deep cortical layers and directly contacting layer 5 PT pyramidal cells (Constantinople and Bruno 2013; see fig. 2f) might implement a skip connection, summed up to the context-aware values computed in layer 5 IT pyramidal cells.

Cross-attention. A natural extension of our model is to consider that keys and values on the one hand and queries on the other hand are computed as functions of two distinct sequences $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_i\}$. That is, going back to the definition of section 3, $\boldsymbol{q}_i = \boldsymbol{W}^h_Q \boldsymbol{x}_i$, $\boldsymbol{k}_j = \boldsymbol{W}^h_K \boldsymbol{y}_j$, and $\boldsymbol{v}_j = \boldsymbol{W}^h_V \boldsymbol{y}_j$. This would realize a form of cross- (instead of self-) attention, also present in the initial encoder-decoder architecture of Vaswani et al. (2017). For example, the two sequences might encode information from two different modalities, e.g., $\{\boldsymbol{x}_i\}$ encodes visual information while $\{\boldsymbol{y}_i\}$ encodes auditory information, see fig. 4a. In that case, cross-attention allows interpreting elements of the visual sequence $\{\boldsymbol{x}_i\}$ as weighted combinations of elements of the auditory sequence $\{\boldsymbol{y}_i\}$. This multimodal representation would then be integrated in a higher-order thalamic nucleus.

Beyond quadratic scaling. The n 'duplications' of each key, query, and value, alongside the weights producing them, resulting from the 'naive' quadratic scaling of transformers, remains a strong limitation. In biological

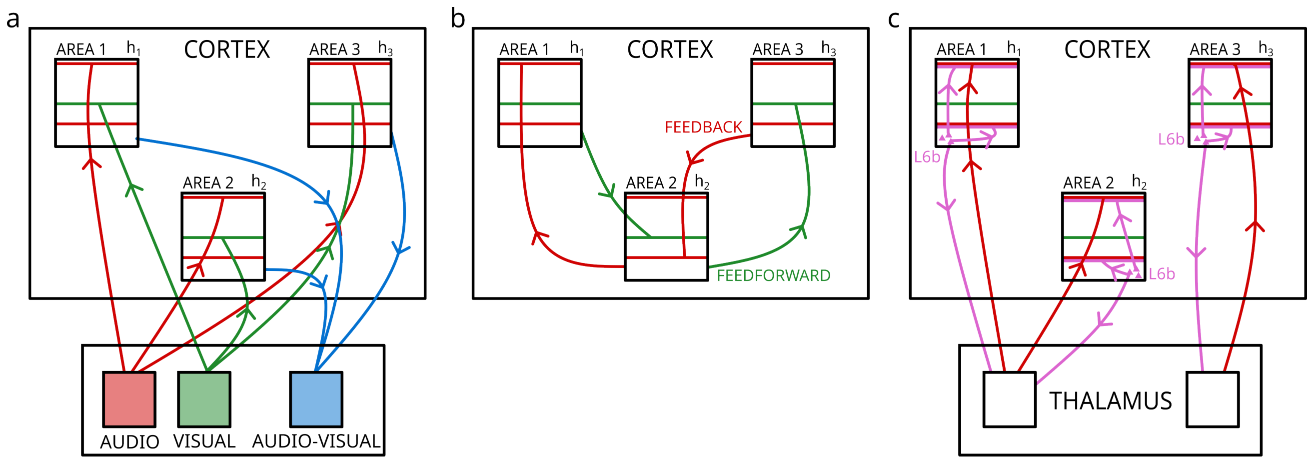

Figure 4: Additional motifs of connectivity in cortex and thalamus. (a) A thalamo-cortical wiring for cross-attention. Here the red an green thalamic nuclei encode two different modalities (audio, visual). The (putatively associative) cortical areas are queried by the visual input, while the keys and values are computed based on the auditory input. The resulting audiovisual representations are integrated in yet another blue thalamic nucleus. (b) Laminar target patterns of cortico-cortical feedforward (green) and feedback (red) connections are similar to the ones of thalamo-cortical core and matrix projections, respectively. One area is querying another area (via feedforward projections, green). This other area is returning context-aware values (via feedback projections, red), used to construct keys and values in layers 1 and 5 in the original area. Layer 6b (pink) sends cortico-thalamic projections, while also sending local projections targeting the same cortical layers as thalamo-cortical projections from matrix thalamus (Zolnik et al. 2024). Cortico-thalamic projections from layer 6 preferentially form reciprocal loops rather than transthalamic pathways (Cassidy et al. 2025).

circuits, a combination of innate structures in connectivity and synaptic learning might lead to this repetition of similar weights along the surface of a cortical area (similar to a convolution filter), whilst needing to be robust to small differences between those weights. More straightforwardly, the linear self-attention mechanism that we adopt here in theory allows for a reorganization of the computation that scales linearly rather than quadratically with the length of the sequence (hence the name; Katharopoulos et al. 2020; see also Yang et al. 2024 their table 2 for a recent overview). The trick is to compute only once the outer product of keys and values, store it in memory, and reuse it for every query. Other machine learning propositions also aim to circumnavigate the original quadratic scaling from Vaswani et al. (2017), seeking to formulate again inference as a recurrent integration of context (Sun et al. 2023) or using subquadratic operations such as gating by element-wise multiplication (Poli et al. 2023). Future work might evaluate the plausibility of these propositions with respect to cortico-thalamic circuits.

Encoding temporal context. Another conundrum lies in the way the cortico-thalamic complex encodes the past, to be taken as context when dealing with temporal sequences (note that not all sequences necessarily unfold through time, e.g. as when transformer networks are applied to image recognition; Dosovitskiy et al. 2021). An elegant solution is based on cortical traveling waves (Muller, Churchland, and Sejnowski 2024). A similar solution might be envisioned here, although it would most straightforwardly make use of thalamic rather than cortical traveling waves (Bhattacharya et al. 2021). Finally, memories of the more distant past might also be taken as contexts. This would certainly involve the hippocampal formation and its targeting of cortical layer 1 (Doron et al. 2020; thus involved in the computation of keys, see also Gershman, Fiete, and Irie 2025).

Cortico-cortical projections. The structure of cortico-cortical feedforward and feedback connections respectively resembles the structure of core and matrix projections in their laminar targets (see fig. 4b and Markov et al. 2014; Harris et al. 2019). The computation of these projections might then be analyzed in our framework as participating in the formation of keys and values (matrix, feedback) and queries (core, feedforward). Cortico-cortical feedback targeting layer 1 would be involved in the computation of the attention kernel, providing top-down keys. In support, recent evidence from direct inactivation of cortico-cortical feedback projections demonstrates their role in attentional gain modulation (Debes and Dragoi 2023). Based on our observations, we can speculate on the computational roles of the observed reciprocal asymmetric connectivity between cortical areas, see fig. 4b.

An area h_1 queries another area h_2 through a feedforward projection. Area h_2 sends back to area h_1 an interpretation of h_1's query in terms of h_2's current keys and values through a feedback projection. Formally, a feedback projection from h_2 to h_1 would then transport $\widetilde{\boldsymbol{v}}_i^{h_1,h_2} = \sum_j \kappa(\boldsymbol{k}_j^{h_2}, \boldsymbol{W}_Q^{h_2,h_1} \hat{\boldsymbol{x}}_i^{h_1})\, \boldsymbol{v}_j^{h_2}$, with $\hat{\boldsymbol{x}}_i^{h_1}$ the representation in the neurons sending the feedforward connection from macrocolumn i in h_1. From these feedback projections, the apical dendrites of layer 2/3 pyramidal cells and the basal dendrites of layer 5 pyramidal cells in h_1 themselves construct keys and values, respectively. Note that this cross-talk between heads does not exist in the original transformer algorithm. Anatomically, feedback projections to layer 1 and deep layers are carried by L5PT (Harris et al. 2019) and contact L5IT in the target area (Bodor et al. 2023); feedforward projections to layer 2/3 are carried by L5IT (Harris et al. 2019; although these projections also target layer 1).

Cortico-thalamic enhancement via layer 6b. The structure of local projections by cortical layer 6b has also been linked to matrix projections, targeting layers 1 and 5a (see fig. 4c and Zolnik et al. 2024). Functionally, cortical layer 6b seems to be implicated as a 'volume knob' on cortico-thalamic loops (Zolnik et al. 2024), and could potentially replace or supplement superficial pyramidal cells as an attention mask.

Self-attention in the hippocampal formation. So far, hippocampal processing has been linked to transformers via Hopfield-type networks that bind abstract locations with sensory observations (Whittington, Warren, and Behrens 2022), or retrieving values with keys in the recurrent memory (Gershman, Fiete, and Irie 2025). Based on considerations of connectivity (reviewed in Kesner and Rolls 2015, see their fig. 1), we suggest differentiated roles for the CA3 and CA1 hippocampal subregions, computing respectively a self-attention signal and an attention-modulated representation. More specifically, our suggestion is as follows. CA3 receives inputs from layer 2 of the entorhinal allocortex both through a direct pathway and a pathway relaying through the dentate gyrus, and computes an attention signal as a similarity measure between inputs from these two pathways. CA1 receives inputs both from CA3 and directly from layer 3 of the entorhinal allocortex, with the representation built from the direct entorhinal input modulated by the CA3 input. A division of labor between CA3 and CA1 is observed in human memory recall, CA3 activity differentiates memories within a context (episode) through task-relevant attention, while CA1 activity adds a differentiation of the context itself (Aly and Turk-Browne 2016; Dimsdale-Zucker et al. 2018).

Self-attention in cortical evolution. Among mammalian species, a large neocortex with more cortical areas is a hallmark of primates and in particular humans (Kaas 2009). The architecture of the cortex seems to be scalable, both in terms of parallelization of computation and in high 'performance' gain with scale (Herculano-Houzel 2012; Meyer et al. 2025); this is also the case for the architecture of multihead self-attention. Moreover, both the proportion of cortical thickness allocated to cortical layer 2/3 and the complexity of layer 2/3 pyramidal cells increase significantly from rodents to primates (Galakhova et al. 2022). Within our framework, we can interpret this expansion of layer 2/3 as the addition or enhancement of an attention mechanism on top of a pre-existing forward processing streams supported by layer 5 pyramidal cells.

Outlook. Despite our report of analogies between the structure of a subset of cortico-thalamic circuits and multihead self-attention, we do not claim that the brain implements transformers *per se*. However, we suggest that the resulting general concepts might be illuminating for a mechanistic understanding of cortical computation. First, locally computed attention signals multiplicatively gate the processing of representations in cortical circuits. Second, the cortico-thalamic complex forms a shallow hierarchy where cortical areas process different aspects of the input in parallel before integrating the results in the thalamus. It would be interesting to investigate whether the additional structure in cortical and thalamic circuits ignored here imply an improvement compared to classical transformer networks in terms of performance.

7 Acknowledgments

We thank Timothée Proix for discussions on human speech processing and intracortical recordings, Nicolas Deperrois for implementing a previous version of our cortical transformers, Katharina Wilmes and Ausra Saudargiene for working out a time-continuous version within a different project, and Michael Marmaduke Woodman for discussing how to implement it in The Virtual Brain. This work was funded by the European Union's Horizon

Europe Programme under the Specific Grant Agreement No. 101147319 (EBRAINS 2.0 Project).

References

Alayrac, Jean-Baptiste et al. (2022). “Flamingo: a visual language model for few-shot learning”. In: *Advances in neural information processing systems* 35, pp. 23716–23736.

Aly, Mariam and Nicholas B. Turk-Browne (2016). “Attention promotes episodic encoding by stabilizing hippocampal representations”. In: *Proceedings of the National Academy of Sciences* 113.4, E420–E429.

Bhattacharya, Sayak et al. (2021). “The impact of a closed-loop thalamocortical model on the spatiotemporal dynamics of cortical and thalamic traveling waves”. In: *Scientific Reports* 11.1, pp. 1–19.

Bodor, Agnes L. et al. (2023). “The Synaptic Architecture of Layer 5 Thick Tufted Excitatory Neurons in the Visual Cortex of Mice”. In: *bioRxiv*.

Buxhoeveden, Daniel P and Manuel F Casanova (2002). “The minicolumn hypothesis in neuroscience”. In: *Brain* 125.5, pp. 935–951.

Campagnola, Luke et al. (2022). “Local connectivity and synaptic dynamics in mouse and human neocortex”. In: *Science* 375.6585, eabj5861.

Carandini, Matteo and David J Heeger (1994). “Summation and division by neurons in primate visual cortex”. In: *Science* 264.5163, pp. 1333–1336.

Cassidy, Rachel M et al. (2025). “Complementary Organization of Mouse Driver and Modulator Cortico-Thalamo-Cortical Circuits”. In: *Journal of Neuroscience*.

Caucheteux, Charlotte and Jean-Rémi King (2022). “Brains and algorithms partially converge in natural language processing”. In: *Communications biology* 5.1, p. 134.

Choromanski, Krzysztof et al. (2022). “Rethinking Attention with Performers”. In: *arXiv*.

Constantinople, Christine M and Randy M Bruno (2013). “Deep cortical layers are activated directly by thalamus”. In: *Science* 340.6140, pp. 1591–1594.

Debes, Samantha R and Valentin Dragoi (2023). “Suppressing feedback signals to visual cortex abolishes attentional modulation”. In: *Science* 379.6631, pp. 468–473.

Dimsdale-Zucker, Halle R. et al. (2018). “CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields”. In: *Nature Communications* 9.1.

Doron, Guy et al. (2020). “Perirhinal input to neocortical layer 1 controls learning”. In: *Science* 370.6523, eaaz3136.

Dosovitskiy, Alexey et al. (2021). “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*.

Elhage, Nelson et al. (2021). “A Mathematical Framework for Transformer Circuits”. In: *Transformer Circuits Thread*.

Ellenberger, Benjamin et al. (2024). “Backpropagation through space, time, and the brain”. In: *arXiv*.

Galakhova, A. A. et al. (2022). “Evolution of cortical neurons supporting human cognition”. In: *Trends in Cognitive Sciences* 26.11, pp. 909–922.

Gershman, Samuel J, Ila Fiete, and Kazuki Irie (2025). “Key-value memory in the brain”. In: *Neuron*.

Harris, Julie A et al. (2019). “Hierarchical organization of cortical and thalamic connectivity”. In: *Nature* 575.7781, pp. 195–202.

Harris, Kenneth D and Gordon MG Shepherd (2015). “The neocortical circuit: themes and variations”. In: *Nature neuroscience* 18.2, pp. 170–181.

Herculano-Houzel, Suzana (2012). “The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost”. In: *Proceedings of the National Academy of Sciences* 109.supplement_1, pp. 10661–10668.

Herculano-Houzel, Suzana, Bruno Mota, and Roberto Lent (2006). “Cellular scaling rules for rodent brains”. In: *Proceedings of the National Academy of Sciences* 103.32, pp. 12138–12143.

Hoover, Benjamin et al. (2023). “Energy transformer”. In: *Advances in neural information processing systems* 36, pp. 27532–27559.

Jones, EG (1998). “The core and matrix of thalamic organization”. In: *Neuroscience* 85.2, pp. 331–345.

Kaas, Jon H. (2009). “Evolution of the Brain in Mammals”. In: *Encyclopedia of Neuroscience*. Ed. by Marc D. Binder, Nobutaka Hirokawa, and Uwe Windhorst. Springer Berlin Heidelberg, pp. 1292–1295.

Katharopoulos, Angelos et al. (2020). "Transformers are rnns: Fast autoregressive transformers with linear attention". In: *International Conference on Machine Learning*. PMLR, pp. 5156–5165.
Kesner, Raymond P and Edmund T Rolls (2015). "A computational theory of hippocampal function, and tests of the theory: new developments". In: *Neuroscience & Biobehavioral Reviews* 48, pp. 92–147.
Kozachkov, Leo, Ksenia V Kastanenka, and Dmitry Krotov (2023). "Building transformers from neurons and astrocytes". In: *Proceedings of the National Academy of Sciences* 120.34, e2219150120.
Larkum, Matthew E, Walter Senn, and Hans-R Lüscher (2004). "Top-down dendritic input increases the gain of layer 5 pyramidal neurons". In: *Cerebral cortex* 14.10, pp. 1059–1070.
Li, Yuanning et al. (2023). "Dissecting neural computations in the human auditory pathway using deep neural networks for speech". In: *Nature Neuroscience* 26.12, pp. 2213–2225.
Markov, Nikola T et al. (2014). "Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex". In: *Journal of comparative neurology* 522.1, pp. 225–259.
Maruoka, Hisato et al. (2017). "Lattice system of functionally distinct cell types in the neocortex". In: *Science* 358.6363, pp. 610–615.
Meyer, Emily E et al. (2025). "Expansion of a conserved architecture drives the evolution of the primate visual cortex". In: *Proceedings of the National Academy of Sciences* 122.3, e2421585122.
Muller, Lyle, Patricia S Churchland, and Terrence J Sejnowski (2024). "Transformers and cortical waves: encoders for pulling in context across time". In: *Trends in neurosciences.*
OpenAI (2022). *ChatGPT.*
Peng, Hao et al. (2021). "Random Feature Attention". In: *arXiv.*
Petersen, Carl CH (2019). "Sensorimotor processing in the rodent barrel cortex". In: *Nature Reviews Neuroscience* 20.9, pp. 533–546.
Phuong, Mary and Marcus Hutter (2022). "Formal algorithms for transformers". In: *arXiv.*
Poli, Michael et al. (2023). "Hyena hierarchy: Towards larger convolutional language models". In: *Machine Learning*. PMLR, pp. 28043–28078.
Powell, Nathaniel J et al. (2024). "Common modular architecture across diverse cortical areas in early development". In: *Proceedings of the National Academy of Sciences* 121.11, e2313743121.
Quiquempoix, Michael et al. (2018). "Layer 2/3 pyramidal neurons control the gain of cortical output". In: *Cell reports* 24.11, pp. 2799–2807.
Sacramento, João et al. (2018). "Dendritic cortical microcircuits approximate the backpropagation algorithm". In: *Advances in Neural Information Processing Systems (NeurIPS)* 2018, pp. 8721–8732. URL: https://intranet.physio.unibe.ch/Publikationen/Dokumente/Sacramento2018Dendriticc.pdf.
Sampathkumar, Vandana et al. (2021). "Integration of signals from different cortical areas in higher order thalamic neurons". In: *Proceedings of the National Academy of Sciences* 118.30, e2104137118.
Scellier, Benjamin and Yoshua Bengio (2017). "Equilibrium propagation: Bridging the gap between energy-based models and backpropagation". In: *Frontiers in computational neuroscience* 11, p. 24.
Schlag, Imanol, Kazuki Irie, and Jürgen Schmidhuber (2021). "Linear transformers are secretly fast weight programmers". In: *International Conference on Machine Learning*. PMLR, pp. 9355–9366.
Schrimpf, Martin et al. (2021). "The neural architecture of language: Integrative modeling converges on predictive processing". In: *Proceedings of the National Academy of Sciences* 118.45, e2105646118.
Senn, Walter et al. (2024). "A neuronal least-action principle for real-time learning in cortical circuits". In: *ELife* 12, RP89674.
Sermet, B Semihcan et al. (2019). "Pathway-, layer-and cell-type-specific thalamic input to mouse barrel cortex". In: *Elife* 8, e52665.
Shen, Yang, Julia Wang, and Saket Navlakha (2021). "A correspondence between normalization strategies in artificial and biological neural networks". In: *Neural computation* 33.12, pp. 3179–3203.
Sherman, S Murray and W Martin Usrey (2024). "Transthalamic Pathways for Cortical Function". In: *Journal of Neuroscience* 44.35.
Song, Yuhang et al. (2024). "Inferring neural activity before plasticity as a foundation for learning beyond backpropagation". In: *Nature Neuroscience* 27.2, pp. 348–358.
Sun, Yutao et al. (2023). "Retentive Network: A Successor to Transformer for Large Language Models". In: *arXiv.*
Suzuki, Mototaka, Cyriel MA Pennartz, and Jaan Aru (2023). "How deep is the brain? The shallow brain hypothesis". In: *Nature Reviews Neuroscience*, pp. 1–14.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Whittington, James C. R., Joseph Warren, and Tim E.J. Behrens (2022). "Relating transformers to models and neural representations of the hippocampal formation". In: *International Conference on Learning Representations*.

Wilson, Nathan R et al. (2012). "Division and subtraction by distinct cortical inhibitory networks in vivo". In: *Nature* 488.7411, pp. 343–348.

Yang, Songlin et al. (2024). "Parallelizing linear transformers with the delta rule over sequence length". In: *arXiv*.

Zolnik, Timothy Adam et al. (2024). "Layer 6b controls brain state via apical dendrites and the higher-order thalamocortical system". In: *Neuron* 112.5, pp. 805–820.