



MASTER'S THESIS

Modelling nAChR Dynamics and Working Memory in Alzheimer's Disease:

A Ring-Attractor Extension of a Local Cortical Circuit

Valentin Jacquin

under the supervision of Dr. Boris Gutkin

Group for Neural Theory — ENS Paris, France

May 20, 2026

This thesis investigates a circuit-level mechanism for working-memory impairment in early Alzheimer's disease. Building on a four-population local model of prefrontal cortex in which amyloid- β disrupts nicotinic acetylcholine receptors on specific interneuron classes, we extend the framework to a ring-attractor architecture and ask how cell-type-specific nAChR dysfunction degrades the stability and precision of persistent activity supporting spatial working memory.

1 Introduction

1.1 Alzheimer’s disease

Alzheimer’s disease (AD) is a widespread neurodegenerative disorder with a high prevalence in the elderly population, currently affecting over 50 million individuals worldwide (Breijyeh and Karaman, 2020). In its early stages, the disease is primarily mediated by the medial temporal lobe and manifests as a decline in episodic memory, while in later stages executive function and attention become increasingly disrupted (Almkvist, 1996). This underlines the involvement of the prefrontal cortex (PFC) in AD-related cognitive decline (Flores-Martínez and Peña Ortega, 2017).

Working memory tests have been considered as one of the earliest and most reliable cognitive markers of AD (Kirova et al., 2015). More recent work by Perpetuini et al. (2020) has shown that electrophysiological recordings during a working memory task can themselves serve as a cognitive marker, underlining the link between the observed cognitive impairment and the underlying circuit-level dysfunction.

While the symptoms and impairments in subjects’ lives are well documented, the underlying circuit-level mechanisms remain poorly understood. As the review by Corriveau-Lecavalier et al. (2024) states, the relationship between AD pathology and cortical activity is complex and non-monotonic, depending on disease stage, brain region, and severity, which makes establishing a clear mechanistic link challenging. However, during the first stage of the disease, commonly referred to as mild cognitive impairment (MCI), the consensus is that the prefrontal cortex (PFC) exhibits a hyperactivation of neural activity. While our work does not aim to give a stage- and region-specific account, we argue that focusing on the underlying mechanisms at the neuronal level could provide a mechanistic explanation that would contribute to understanding the global interactions with amyloid- β ($A\beta$).

1.2 Amyloid- β and nicotinic acetylcholine receptors

Among the molecular drivers of early AD pathology, $A\beta$ peptides are the most studied and the most therapeutically relevant. $A\beta$ accumulation precedes overt neurodegeneration by years, and soluble oligomeric forms are now thought to cause the earliest synaptic and circuit-level abnormalities (Milà-Alomà et al., 2020). The link between $A\beta$ and those abnormalities is tightly related to nicotinic acetylcholine receptors (nAChRs). In their review, Lombardo and Maskos (2015) present evidence for this link, showing that $A\beta$ can bind to nAChRs and disrupt their function. They further underline the differences in binding affinity across nAChR subunit compositions, and discuss the therapeutic potential of ligands targeting specific subunit compositions.

The nAChRs are multimeric ligand-gated ion channels. In the cortex, two compositions dominate: the homomeric $\alpha 7$ receptor, highly calcium-permeable and rapidly desensitising, and the heteromeric $\beta 2$ -containing receptors (notably $\alpha 4\beta 2$ and $\alpha 5\alpha 4\beta 2$), which show slower kinetics and lower calcium permeability (Albuquerque et al., 2009). These structural differences translate into distinct interactions with ligands: $\alpha 7$ receptors have low agonist affinity and fast gating kinetics, whereas $\alpha 4\beta 2$ receptors bind nicotine with high affinity and display slower channel dynamics (Albuquerque et al., 2009). $A\beta$ differentially affect both subtypes, with sensitivity depending on its concentration and the receptor subunit composition (Albuquerque et al., 2009).

These subtypes are not uniformly distributed across interneuron classes: in layer II/III PFC, $\alpha 5$ receptors are enriched in VIP (vasoactive intestinal peptide-positive) interneurons, $\alpha 7$ -containing receptors dominate in PV (parvalbumin-positive) interneurons, and SOM (somatostatin-positive) cells express both $\alpha 7$ and $\beta 2$ subunits (Koukoulis et al., 2025). This cell-type-specific distribution provides a route through which $A\beta$ -induced nAChR disruption translates into interneuron-class-specific effects on cortical circuit dynamics.

A key anatomical fact is important to highlight: in layer II/III of mouse PFC, functional

nAChRs are effectively restricted to interneurons, not pyramidal cells (Poorthuis et al., 2013). This implies that $A\beta$ cannot disrupt pyramidal firing directly through nAChR binding at those layers, so its effects on excitatory activity should be mainly mediated by altered interneuron drive. Because each interneuron population expresses a different subunit composition, the effects of $A\beta$ on each of them are distinct, as previously discussed. Therefore, the impact of $A\beta$ on pyramidal cells is complex. Consequently, any model of $A\beta$ -induced WM impairment that aims to give a precise mechanistic interpretation must explicitly represent the different interneuron populations and their associated nAChRs.

Koukoulis et al. (2025) take this stance directly: they build a four-population local rate model of PFC (one excitatory pyramidal population (PYR) and three inhibitory classes (PV, SOM, VIP)) whose connectivity and nAChR expression follow the biological patterns described above, keeping it grounded in known physiology. Fitted to firing rates recorded in both wild-type and APP-mouse (simulating Alzheimer’s disease) conditions, the model shows that WT-versus-APP differences can be explained by local nAChR modulation alone, without invoking network-level pathology.

1.3 Cortical interneurons: PV, SOM, and VIP

Explicitly representing these populations requires a clear characterisation of each class. Here we follow the now-standard taxonomy of three main interneuron classes in mouse PFC: parvalbumin-positive (PV), somatostatin-positive (SOM), and vasoactive intestinal peptide-positive (VIP) cells (Tremblay et al., 2016). Each class has distinct anatomical and functional properties that shape its role in cortical computation.

PV interneurons are fast-spiking cells that target the perisomatic region of pyramidal neurons (either the soma and proximal dendrites (basket cells) or the axon initial segment (chandelier cells)) exerting strong control over pyramidal spike generation (DRUGA et al., 2023). Functionally, PV activity is often described as providing gain control or divisive-like inhibition over local excitation. Beyond their local reach, PV cells send horizontal axons across adjacent columns, positioning them as a form of global inhibitor, regulating the overall excitatory/inhibitory ratio participating of stabilizing behavioral responses (Ferguson and Gao, 2018).

SOM interneurons are essentially Martinotti cells in the layer 2/3 of the PFC. They are slower and target the distal dendrites of pyramidal cells, filtering incoming excitatory drive in what is considered subtractive inhibition (Tremblay et al., 2016; Yavorska and Wehr, 2016). Their horizontal projections showcase their involvement in genuine lateral inhibition between cortical columns: local pyramidal cells excite local SOM interneurons, which project horizontally to inhibit pyramidal cells in neighbouring columns (Karnani et al., 2016). This PYR→SOM→neighbour-PYR motif is thought to be central to the selective sharpening of cortical activity patterns and, by extension, may contribute to the sharpness and stability of localised activity bumps in attractor networks.

VIP interneurons are also implicated in lateral inhibition (Karnani et al., 2016), and their role, as mentioned before, is disinhibitory. Activation of VIP transiently silences SOM, releasing pyramidal cells from dendritic filtering. VIP cells are also described as targets of top-down processes supporting various cognitive functions (reinforcement signals, attention, ...) (Pi et al., 2013), placing them at the interface between cortical state modulation and circuit-level excitability.

Two additional points matter for the modelling work that follows. First, the nAChR drive that $A\beta$ disrupts is not delivered uniformly: each interneuron class receives a distinct combination depending on the corresponding expression of $\alpha 7$ and $\beta 2$ -containing receptors. Second, the three interneuron classes are not independent. PV and SOM cells inhibit each other, while VIP cells inhibit both PV and SOM, creating multiple recursive interactions that shape the circuit’s response to nAChR disruption and, by extension, the observed WM impairment.

1.4 Working memory and ring attractors

At the mechanistic level, WM is typically modelled as persistent delay-period activity mimicking real working-memory tasks such as the oculomotor delayed response task. A population of neurons maintains an elevated firing rate in the absence of any external cue, until that activity is either read out or extinguished. The foundational computational accounts of this phenomenon come from (Compte et al., 2000), who showed that networks with excitatory and inhibitory populations can give rise to self-sustaining activity states encoding continuous variables such as spatial location.

The canonical architecture for spatial WM is the ring attractor Compte et al. (2000). It consists of a set of local neural populations, each selective for a preferred spatial direction, linked by distance-dependent recurrent connectivity, with short-range excitation and broader inhibition. When properly tuned, such a network supports a localised “bump” of activity that can be induced by a transient input.

Two features of ring attractors are critical for what follows. First, bump persistence relies on a form of bistability: in the absence of external drive, the network must be able to sit in either a quiescent state (no bump anywhere) or an active state (a bump at some location). Without this, cues either fail to initiate activity or fail to be retained. Second, the architecture is a network-level property: while local properties manipulate the dynamics of individual columns, the spatial structure and stability of the bump are emergent properties of the coupled system.

1.5 Bistability and its mechanistic origins

Bistability is the coexistence of two stable fixed points at a given set of parameters. It is considered to be the mathematical substrate of persistent memory Zylberberg and Strowbridge (2017). A transient cue flips the system from a “low” resting state to a “high” active state, and the network then self-sustains in the active state until a reset signal (or intrinsic relaxation) returns it to rest. The empirical signature in cortex is the observation of distinct low- and high-activity regimes in delay-task recordings (Sreenivasan and D’Esposito, 2019).

In a model of neural population where firing rates evolve according to dynamical equations, bistability can be obtained by a positive-feedback loop whose gain is a saturating function of activity. The saturation can enter either at the output, as a sigmoid ceiling on the transfer function, or at the synapse, as a gating variable whose saturation is bounded between zero and one. Each route has precedent in the literature, with Papasavvas et al. (2015) taking the former and Wong and Wang (2006) the latter. The sigmoid ceiling is phenomenological: saturation parameters are tuned to fit data, with no principled link to underlying receptor kinetics. The NMDA-gating approach is biologically grounded: the saturation arises from the fraction of NMDA channels open at steady state, a quantity directly computable from synaptic physiology. For PFC WM specifically, thought to depend on NMDA receptors (Wang et al., 2013), the gating-variable approach offers a more direct mechanistic interpretation.

Two distinctions can be made in our context of ring-attractor bistability. Local bistability refers to the ability of a single column, in isolation, to support both low and high stable firing rates. Network bistability refers to the ability of the coupled ring to support a stable bump of activity, which may or may not be underpinned by local bistability at the single-node level (see (Camperi and Wang, 1998)).

Which of the two scenarios, local or network bistability, actually applies in cortex remains an open empirical question (Zylberberg and Strowbridge, 2017). Empirically, single-cell recordings showing intrinsic stimulus-locked persistent firing in the absence of synaptic input demonstrate that local mechanisms can in principle sustain elevated activity (Egorov et al., 2002), though whether such mechanisms operate in vivo in PFC remains unresolved (Zylberberg and Strowbridge, 2017).

1.6 Gap and research question

The aim of this thesis is to extend the Koukouli et al. (2025) model introduced above. This model provides the biological detail, with three interneuron classes and cell-type-specific nAChR drive, but no spatial architecture. Standard ring attractors provide the spatial architecture for WM but collapse inhibition into a single population, losing the interneuron resolution that makes $A\beta$ effects interpretable. We aim to bridge this gap by embedding the local circuit into a ring that preserves the interneuron structure, and to ask whether local $A\beta$ -induced nAChR disruption translates into WM deficits at the network level.

Research question. Can a biologically grounded local cortical circuit embedded in a ring-attractor architecture produce stable working-memory dynamics? Does $A\beta$ -simulated nAChR disruption selectively impair them?

Objectives.

- Build a ring-attractor model with Gaussian PYR→PYR lateral excitation and interneuron-mediated inhibition that sustains a self-maintained activity bump during a delay period, with population resting-state rates matching biological data.
- Characterise how $A\beta$ -induced nAChR disruption affects the network’s ability to form and sustain a bump of activity, and to interpret the results in terms of the underlying circuit mechanisms.

2 Methods

The original plan, as outlined in our pre-registration (Appendix A), was to take the fitted Koukoulouli model, embed it directly in a ring architecture, and probe WM performance under nAChR disruption. An attempt to extend the original code, however, revealed several discrepancies between the implementation and the published equations, and restarting from the paper’s specification alone did not reproduce the reported fits. The model described below was therefore built from first principles, following the previously described biological constraints and the general architecture of the original model, but with a revised transfer function to make bistability achievable and using available firing-rate data as fitting targets. We also needed to develop an optimisation procedure to fit the model parameters to the data, which is not trivial given the need to enforce bistability while fitting both low and high states. The following sections describe the model and the fitting procedure in detail.

2.1 Overview of the population model

Each cortical column is represented by four rate-based populations: excitatory pyramidal neurons (PYR) and three inhibitory interneuron classes (PV, SOM, VIP). Each population i is characterised by a single time-varying firing rate $r_i(t)$, updated according to a Wilson-Cowan-like equation:

$$\tau_i \dot{r}_i(t) = -r_i(t) + \Phi_i(I_i^{\text{net}}(t)), \quad (1)$$

where τ_i is the membrane time constant of population i , Φ_i is its input-output transfer function, and $I_i^{\text{net}}(t)$ is the net synaptic current integrating excitatory and inhibitory inputs together with external drive and, for PYR, a slow adaptation term.

The local connectivity graph follows the Koukoulouli taxonomy. PYR receives recurrent self-excitation and inhibition from PV (perisomatic) and SOM (dendritic). PV receives excitation from PYR and cross-inhibition from VIP. SOM receives excitation from PYR and strong disinhibition from VIP. VIP receives excitation from PYR and modest cross-inhibition. All four populations additionally receive an external current I_i^0 and, for interneurons, a cell-type-specific nAChR drive I_i^{ACh} .

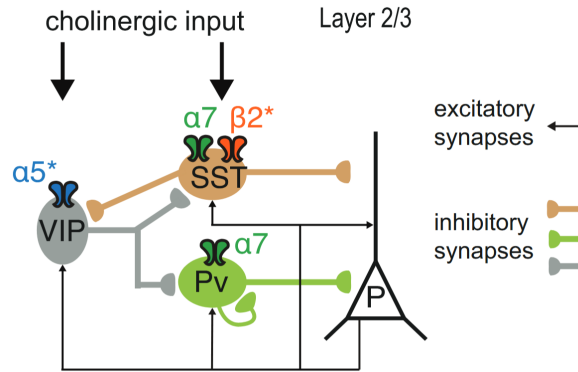


Figure 1: Schematic of the local circuit architecture from Koukoulouli (2025).

2.2 From the Koukoulouli (2025) equation to ours

The starting point is the original Koukoulouli et al. (2025) formulation, which uses a Wong-Wang-type transfer function for each population and treats GABAergic inhibition to PYR in two distinct ways depending on the source.

Divisive versus subtractive inhibition. Indeed, PV-mediated inhibition is modelled as divisive, acting on the gain of the pyramidal input-output curve rather than shifting its threshold. This is motivated by the perisomatic targeting of PV synapses, which controls the output gain of pyramidal cells without preferentially cancelling specific dendritic inputs (El-Boustani and Sur, 2014). SOM-mediated inhibition, by contrast, is modelled as subtractive: dendritic-targeting SOM synapses subtract from the net excitatory input integrated at the soma.

The pyramidal net input takes the form

$$I_{\text{PYR}}^{\text{net}} = \frac{w_{ee} r_{\text{PYR}} - w_{se} r_{\text{SOM}} + I_{\text{PYR}}^0 + I_{\text{PYR}}^{\text{adapt}}}{1 + g_{\text{GABA}} w_{pe} r_{\text{PV}}}, \quad (2)$$

where the denominator $1 + g_{\text{GABA}} w_{pe} r_{\text{PV}}$ implements the divisive effect of PV activity (with w_{pe} the PYR→PV weight and g_{GABA} a global GABAergic gain), and $I_{\text{PYR}}^{\text{adapt}}$ is a slow hyperpolarising adaptation current with its own dynamics (see Section 2.2). Interneuron inputs follow the same additive form but without the divisive denominator (see the full equations in Appendix F). Throughout the rest of the thesis, weights are denoted w_{ij} , where i is the source and j the target population.

The Wong-Wang transfer function. Each population uses the standard Wong-Wang input-output curve (Wong and Wang, 2006),

$$\Phi_i(I) = \frac{u_i}{1 - \exp(-g_i u_i)}, \quad u_i = \alpha_i (I - \theta_i), \quad (3)$$

with population-specific parameters $(\alpha_i, \theta_i, g_i)$. This function is smooth, approximately linear for large I , and has a well-defined zero at $I = \theta_i$. Crucially, it does not saturate: as $I \rightarrow \infty$, $\Phi_i(I) \rightarrow \alpha_i (I - \theta_i)$, which grows without bound. The parameters are fitted to a leaky integrate-and-fire model with biologically realistic synaptic dynamics, and therefore admit a mechanistic interpretation in terms of underlying biophysics (Wong and Wang, 2006). This is the main motivation for using this transfer function. In the original Koukoulis formulation, however, the fast-synapse approximation $S \approx \tau r$ linearises the NMDA gating variable, removing the fold from the PYR nullcline and making bistability structurally unachievable regardless of parameter tuning. We address this directly in Section 2.3 by reintroducing the full NMDA gating variable on the PYR→PYR recurrence.

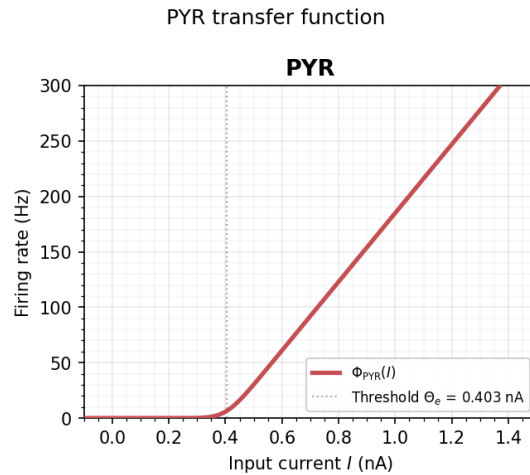


Figure 2: PYR transfer function $\Phi_{\text{PYR}}(I)$.

Adaptation on PYR. A slow hyperpolarising current is added to the pyramidal population, with dynamics:

$$\tau_{\text{PYR}} \dot{I}_{\text{PYR}}^{\text{adapt}} = -I_{\text{PYR}}^{\text{adapt}} - g_{\text{adapt}} r_{\text{PYR}}, \quad (4)$$

where τ_{PYR} is the adaptation time constant. While the time-constant value is fixed based on known characteristics of the population, the adaptation strength g_{adapt} is a free parameter. Here we use $\tau_{\text{PYR}} = 600$ ms for PYR, the value used in previous work (Koukoulis et al., 2025) and originally taken from (Destexhe, 2009). This models the activity-dependent hyperpolarisation observed in pyramidal neurons (Sanchez-Vives et al., 2000).

External and nicotinic drive. Each population receives a constant external current I_i^0 . Interneurons additionally receive a nAChR drive I_i^{ACh} that represents the cholinergic tone filtered through the subunit composition specific to that cell class. This is the channel through which $A\beta$ is modelled to exert its effect (Koukoulis et al., 2025).

2.3 Bistability and saturation: fix on transfer function

The formulation in the previous subsection cannot support bistability at the single-node level. This section explains why, reviews the two candidate fixes, and justifies the choice made.

The problem: a missing saturation. The PYR nullcline, defined by $r^* = \Phi_{\text{PYR}}(I^{\text{net}}(r^*))$, can support two stable fixed points only if the function $F(r) \equiv \Phi_{\text{PYR}}(I^{\text{net}}(r)) - r$ crosses zero at least three times, producing the sign pattern $+, -, +, -$ along the sweep. This requires the right-hand side $\Phi_{\text{PYR}}(I^{\text{net}}(r))$ to fold: to increase, then flatten (or decrease), then increase again. In the Wong-Wang formulation above, $I^{\text{net}}(r)$ grows roughly linearly with r via the $w_{ee}r_{\text{PYR}}$ term, Φ_{PYR} is monotonic, and no intrinsic saturation bounds the recurrent excitation. The result $\Phi_{\text{PYR}}(I^{\text{net}}(r))$ is therefore monotonically increasing and cannot fold. Bistability is therefore structurally impossible in this formulation: it is the shape of the equations, not the parameter values, that rules it out. We only consider local connections here. In principle, the ring coupling could create the fold at the network level, but this is a distinct mechanism (introduced in Section 1.5), and our parameter searches failed to produce such a fold, suggesting that the ring coupling alone is not sufficient.

Option 1 (Papassavas lineage): bounded sigmoid on every population. One resolution, used in previous work by Papasavvas et al. (2015), is to replace the Wong-Wang transfer function with an explicit sigmoid that saturates at a maximum rate r_i^{max} . This produces bistability robustly, as shown in (Rooy et al., 2021). The difficulty is that the saturation parameter is phenomenological: it is not derived from any biophysical quantity, and the choice of ceiling, slope, and half-activation is effectively arbitrary. While this approach works mechanically, it sacrifices the biophysical interpretability that motivated the Koukoulis formulation in the first place.

Option 2 (adopted): NMDA gating on PYR→PYR. The alternative, following Wong and Wang (2006), is to reintroduce the slow NMDA gating variable $S(t)$ on the pyramidal recurrent connection. The full equation reads

$$\tau_{\text{NMDA}} \dot{S}(t) = -S(t) + \gamma[1 - S(t)]r_{\text{PYR}}(t), \quad (5)$$

whose steady-state solution is the saturating function

$$S^*(r) = \frac{\gamma\tau_{\text{NMDA}}r}{1 + \gamma\tau_{\text{NMDA}}r}. \quad (6)$$

The pyramidal net input then becomes

$$I_{\text{PYR}}^{\text{net}} = \frac{J_{\text{NMDA}} S^*(r_{\text{PYR}}) - w_{se} r_{\text{SOM}} + I_{\text{PYR}}^0 + I_{\text{PYR}}^{\text{adapt}}}{1 + g_{\text{GABA}} w_{pe} r_{\text{PV}}}, \quad (7)$$

where J_{NMDA} replaces w_{ee} as the recurrent excitatory strength. With $\tau_{\text{NMDA}} = 100$ ms and $\gamma = 0.641$ fixed at canonical values (Wong and Wang, 2006), J_{NMDA} becomes the free recurrent-coupling parameter.

This change is both minimal and biologically grounded. It adds one dynamical variable and replaces one parameter, while introducing a biologically documented mechanism. The resulting $S^*(r)$ introduces exactly the saturation needed to create a fold in the PYR nullcline, and enables genuine single-node bistability in the fitted circuit. Note that this change effectively sets a cap on the input received by the pyramidal population: this cap is given by J_{NMDA} , which sets the maximum contribution of the recurrent excitation (when $S^* \rightarrow 1$). This value will be fitted during the optimisation.

Interneuron saturation: hyperbolic soft ceiling. The NMDA gating mechanism is specific to PYR→PYR recurrence mediated by NMDA receptors and cannot be applied to the interneuron populations, whose outputs are GABAergic and whose inputs are dominated by fast AMPA receptors. To prevent pathological runaway firing in PV, SOM, and VIP during optimisation a hyperbolic soft ceiling is applied post-hoc to the Wong-Wang transfer function for each interneuron population X :

$$\Phi_{\text{capped}}^X(I) = \frac{r_{\text{max}}^X \cdot \Phi^X(I)}{r_{\text{max}}^X + \Phi^X(I)}. \quad (8)$$

This form leaves the low-rate regime unchanged ($\Phi_{\text{capped}} \approx \Phi$ when $\Phi \ll r_{\text{max}}^X$) and asymptotes smoothly to r_{max}^X at high inputs. Ceiling values are set to $2\times$ the Rooy high-state target for each population, yielding $r_{\text{max}}^{\text{PV}} = r_{\text{max}}^{\text{SOM}} = 70$ Hz and $r_{\text{max}}^{\text{VIP}} = 137$ Hz. This choice is internally consistent: the same dataset that defines the optimisation targets also defines the ceilings, and the $2\times$ factor guarantees that the full physiologically observed operating range remains unconstrained. This ceiling approach is admittedly phenomenological. Unlike the NMDA saturation, it is nonetheless derived from biological data in order to stay as close as possible to biological reality. Possible mechanistic adaptations of the model are discussed later (see the Discussion).

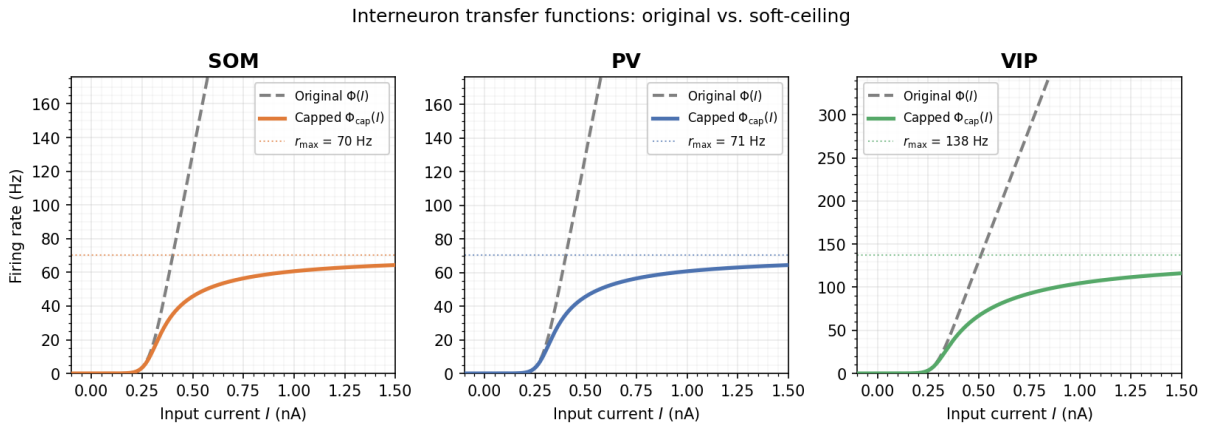


Figure 3: Interneuron transfer functions with (solid) and without (dashed) the hyperbolic soft ceiling, for SOM, PV, and VIP.

2.4 Ring architecture

To address the research question, we embed the local circuit in a ring of N columns arranged on a circle with angular spacing $\Delta\theta = 2\pi/N$. Each node carries a full four-population local circuit. The ring connectivity is derived from the single-node fitted weights via a row-sum principle: at the homogeneous fixed point (all nodes firing identically), the total synaptic drive onto any population must equal the value produced by the single-node circuit. Any spatially distributed weight matrix that is row-sum normalised to the fitted scalar satisfies this constraint exactly, so resting-state firing rates, KO conditions, and single-node bistability are all preserved by construction. The spatial distribution of that total drive is then determined by the kernel shape specific to the known connectivity of each interneuron class.

PYR→PYR: unified Gaussian kernel including self-weight. All pyramidal recurrent excitation, both local and inter-columnar, is carried by a single Gaussian weight matrix with a non-zero diagonal:

$$W_{k\ell}^{\text{PYR}} \propto \exp\left(-\frac{d(\theta_k, \theta_\ell)^2}{2\sigma_{\text{PYR}}^2}\right), \quad \sum_{\ell} W_{k\ell}^{\text{PYR}} = J_{\text{NMDA}}^{\text{fitted}}, \quad (9)$$

where $d(\theta_k, \theta_\ell)$ is the angular distance and $\sigma_{\text{PYR}} = 15^\circ$. Each row is normalised to $J_{\text{NMDA}}^{\text{fitted}}$, and the diagonal entry W_{kk}^{PYR} is the Gaussian evaluated at distance zero, replacing the separate local recurrence term. The NMDA drive to column k becomes a weighted sum over all columns:

$$I_k^{\text{PYR,NMDA}} = \frac{\sum_{\ell} W_{k\ell}^{\text{PYR}} S_{\ell}^{\text{NMDA}}}{1 + g_{\text{GABA}} \sum_{\ell} W_{k\ell}^{\text{PV}} r_{\ell}^{\text{PV}}}, \quad (10)$$

where S_{ℓ}^{NMDA} is the NMDA gating variable of node ℓ (one scalar per node, since $S_{j\ell} \equiv S_{\ell}$). This short-range excitatory kernel is the standard mechanism by which bump locality is produced in ring attractors (Compte et al., 2000).

PV→PYR: uniform fully divisive inhibition. PV axons project broadly across cortex and produce perisomatic shunting inhibition, which is divisive regardless of whether the PV soma is local or distal. All PV→PYR inhibition (including the local column's own PV population) therefore enters the divisive denominator via a flat uniform kernel over all N columns:

$$W_{k\ell}^{\text{PV}} = \frac{w_{pe}^{\text{fitted}}}{N} \quad \text{for all } \ell, \quad (11)$$

row-sum normalised to w_{pe}^{fitted} .

SOM→PYR: annular lateral inhibition. SOM interneurons receive input from the local pyramidal population but project their inhibitory output to neighbouring columns, realising the motif $\text{PYR}_k \rightarrow \text{SOM}_k \rightarrow \text{PYR}_{\ell \neq k}$. Crucially, this lateral inhibition should peak outside the PYR excitatory zone rather than at the source column itself. We therefore use an annular (offset) Gaussian kernel whose peak is displaced from zero by μ_{SOM} :

$$W_{k\ell}^{\text{SOM}} \propto \exp\left(-\frac{(d(\theta_k, \theta_\ell) - \mu_{\text{SOM}})^2}{2\sigma_{\text{SOM}}^2}\right) \cdot \mathbf{1}[\ell \neq k], \quad \sum_{\ell \neq k} W_{k\ell}^{\text{SOM}} = w_{se}^{\text{fitted}}, \quad (12)$$

with $\mu_{\text{SOM}} = 3\sigma_{\text{PYR}} = 45^\circ$ and $\sigma_{\text{SOM}} = 15^\circ$. Because $d(\theta_k, \theta_\ell)$ is always non-negative, the kernel creates two symmetric inhibitory lobes on the two sides of the active column, both peaking at

angular distance μ_{SOM} (Figure 11). Setting $\mu_{\text{SOM}} = 3\sigma_{\text{PYR}}$ places the lobes just beyond the edge of the PYR excitation zone (which concentrates most of its weight within $\sim 2\sigma_{\text{PYR}} = 30^\circ$). The diagonal is explicitly zeroed (no self-inhibition), so each column’s SOM population suppresses its neighbours rather than itself. This is not a standard kernel in the ring-attractor literature, but it is motivated by the known connectivity of SOM interneurons and provides a lateral-inhibition mechanism that is distinct from the global divisive inhibition of PV.

It is important to note here that the fitted weight w_{se}^{fitted} comes from the single-node fit, which has different connectivity (local inhibition) than the ring model (lateral inhibition). This connectivity should therefore be regarded as exploratory rather than fully validated. We compare results with and without this ring connectivity in the Results section.

Weight normalisation. All inter-columnar kernels are row-sum normalised to the corresponding single-node fitted scalar, ensuring that homogeneous fixed-point firing rates are preserved exactly for any ring size N and kernel width.

2.5 Data and fitting targets

Two data sources are used.

Koukouli receptor desensitisation data. We take from Koukouli et al. (2025) the percentage reduction in nAChR-mediated current in the APP mouse compared to wild type (given in Appendix G). These deactivation percentages are applied as multiplicative scaling factors on the corresponding nAChR drive parameters ($I_{\alpha 7, \text{PV}}$, $I_{\alpha 7, \text{SOM}}$, $I_{\beta 2, \text{SOM}}$, $I_{\alpha 5, \text{VIP}}$) to simulate the effect of A β accumulation on the ring model (Section 3.6). The firing-rate data from the same experiments are not used as fitting targets: all rate targets are drawn from the Rooy dataset described below.

Rooy high-vs-low state rates. Because the Koukouli averages cannot serve as valid low-state targets in a bistable fit without a mixture model (Obstacle 3 above), the low- and high-state targets are drawn from Rooy et al. (2021), which explicitly separates spontaneous and active regimes for each population:

Population	Low state (Hz)	High state (Hz)
PYR	1.75	60.2
SOM	1.12	35.2
PV	1.04	35.3
VIP	1.33	68.8

Tension between the two data sources. The Koukouli data quantify condition-specific differences (WT vs APP vs KO) that are essential for the AD question, while the Rooy data resolve the low/high-state structure needed for bistability. There is no single dataset providing both. The practical consequence is that our bistable fits use Rooy targets, and the bridge back to AD is made post-hoc through nAChR-drive perturbations applied to the Rooy-fitted model. We flag this as a central limitation, discussed in the Discussion.

2.6 Loss functions

To fit the model parameters such that they both allow for bistability and match the target rates, we use a machine-learning approach. We define a loss function that shapes the loss landscape to enforce the bistable sign pattern while fitting both low and high states.

L_{bistab} : adaptive sign-pattern enforcement (weight $w_{\text{bistab}} = 5.0$). The loss operates on $F(r) = \Phi_{\text{PYR}}(I^{\text{net}}(r)) - r$ along a sweep $r \in [0, 80]$ Hz, and enforces the four-segment pattern $+, -, +, -$ characteristic of bistability through four additive sub-terms:

- **Adaptive low-basin positivity.** The first down-crossing of F locates the actual low fixed point r_{low}^* , which need to match the target. F is required to remain positive throughout $[0, r_{\text{low}}^*]$, with the penalty scaled by $(1 + |r_{\text{low}}^* - r_{\text{low}}^{\text{target}}|/r_{\text{low}}^{\text{target}})$ so that a misplaced FP incurs both a shape penalty and an implicit rate-matching penalty. This adaptive scaling couples the shape term and the rate term in a way that a fixed-probe penalty cannot.
- **Valley.** F must go negative somewhere after the low FP. Penalty $\text{relu}(\min_{r > r_{\text{low}}^*} F)$, zero as soon as any valley exists.
- **Windowed high-basin positivity.** F must rise back above a margin (1 Hz by default) somewhere in the window $[0.7, 1.0] \times r_{\text{PYR}}^{\text{high, target}} \approx [42, 60]$ Hz. The upper bound is $1.0 \times$ (not above) because the high basin is strictly the ascending region before the high FP: above the FP, F is negative by definition and is covered by the tail term. Using a windowed rather than global maximum prevents spurious low-rate crossings (e.g. a second zero at 17 Hz) from trivially satisfying the bistability check.
- **Tail.** F must be negative in the top 15% of the sweep, ensuring the high FP is stable and doesn't drift upwards indefinitely. We set the threshold at 15% rather than depending to the high target to have a loss even without a valid high FP.

The full loss is $L_{\text{bistab}} = L_{\text{low-basin}} + L_{\text{valley}} + L_{\text{high-basin}} + L_{\text{tail}}$, which is zero if the complete $+, -, +, -$ pattern is satisfied with the low FP close to target.

L_{rate} : low-FP rate matching (weight $w_{\text{rate}} = 1.0$). At the lowest stable fixed point, the squared relative error across all four populations:

$$L_{\text{rate}} = \sum_{i \in \{\text{PYR}, \text{SOM}, \text{PV}, \text{VIP}\}} \left(\frac{r_i^{\text{low}} - r_i^{\text{low, target}}}{r_i^{\text{low, target}}} \right)^2. \quad (13)$$

This mean-squared percentage error (MSPE) penalises relative rather than absolute deviations, which is biologically appropriate: a 1 Hz error on a 1.75 Hz target matters more than a 1 Hz error on a 60 Hz target.

$L_{\text{rate, high}}$: high-FP rate matching (weight $w_{\text{rate, high}} = 1.5$). Symmetric to L_{rate} but evaluated at the highest stable fixed point, with Rooy high-state targets. When the model is monostable, there is no second stable FP and the term is set to zero. The bistability penalty L_{bistab} already exerts strong pressure to recover the high state, and a phantom high-FP penalty would produce a misleading gradient.

L_{margin} : fixed-point separation (weight $w_{\text{margin}} = 2.0$). A minimum 15 Hz separation between the two stable fixed points is enforced via $L_{\text{margin}} = \text{relu}(\Delta r_{\text{min}} - (r^{\text{high}} - r^{\text{low}}))$. When monostable, the term takes a fixed penalty of 30. Without this term the optimiser can satisfy the sign-pattern check with marginally separated fixed points that are functionally useless for WM.

L_{jac} : Jacobian regulariser (weight $w_{\text{jac}} = 0.1$). The 4×4 Jacobian of the local circuit is evaluated at the low fixed point, and entries exceeding 5 in absolute value are penalised: $L_{\text{jac}} = \text{relu}(\max_{ij} |J_{ij}| - 5)^2$. This prevents the optimiser from finding bistability via a single implausibly dominant pathway rather than the intended distributed disinhibitory circuit. The computation of this Jacobian is detailed in Appendix B.

L_{peak} : **nullcline peak penalty (weight $w_{\text{peak}} = 0$, off by default)**. Caps the maximum of $\Phi(I^{\text{net}}(r))$ below a threshold $r_{\text{peak,max}}$ to limit the overshoot during cue-driven transitions. Not used in the reported runs, but available as an optional constraint.

Total loss.

$$L_{\text{total}} = w_{\text{bistab}}L_{\text{bistab}} + w_{\text{rate}}L_{\text{rate}} + w_{\text{rate,high}}L_{\text{rate,high}} + w_{\text{margin}}L_{\text{margin}} + w_{\text{jac}}L_{\text{jac}} + w_{\text{peak}}L_{\text{peak}}. \quad (14)$$

2.7 Ring-level metrics for working-memory performance

Ring simulations are evaluated against a set of WM-relevant metrics computed over a task consisting of a pre-cue period, a cue of duration T_{cue} at a chosen angular location, and a delay period.

Bump formation and cue response. A bump is defined by thresholding the peak PYR activity across the ring. The fractions `delay_bump_frac` represent the fraction of the delay period during which a bump is present.

Distractor robustness. A secondary cue at a different angular location is presented during the delay, and the bump’s response (shift, split, or retention) is quantified.

2.8 Simulation and optimisation pipeline

Numerical integration. Dynamics are integrated with a forward Euler scheme at $\Delta t = 0.1$ ms, implemented in Python with Numba just-in-time compilation Lam et al. (2015).

Optimisation. Parameter fitting proceeds in two stages. In the first stage, the single-node loss L_{total} is minimised over all local weight parameters using the non-gradient-based “TwoPointsDE” (Two-Points Differential Evolution) optimiser. The output of this stage is a set of fitted scalars (all the currents and conductances for each population and nicotinic receptor, together with $J_{\text{NMDA}}^{\text{fitted}}$, w_{pe}^{fitted} , w_{se}^{fitted} , and the remaining local weights) that define the target row-sums for the ring kernels. In the second stage, these scalars are transferred to the ring via the row-sum normalisation described in Section 2.4. No additional optimisation is performed. The only structural free parameters introduced at the ring level are the Gaussian widths $\sigma_{\text{PYR}} = \sigma_{\text{SOM}} = 15^\circ$, which are fixed rather than fitted.

Noise. Stochastic input is injected into all four populations $X \in \{\text{PYR}, \text{SOM}, \text{PV}, \text{VIP}\}$ at each node i as an additive current term $\sigma_{\text{noise}} \cdot I_{\text{ext}}^X \cdot \xi_i(t)$, where σ_{noise} is a dimensionless amplitude (default 0.3) and I_{ext}^X is the population’s baseline external drive. All four populations at a given node share the same amplitude of $\xi_i(t)$ at the same timestep, so variability is correlated across populations but the total amplitude scales with the population’s own drive. Two processes are supported: (i) white Gaussian noise, $\xi_i(t) \sim \mathcal{N}(0, 1)$ i.i.d. across nodes and integration steps, and (ii) Ornstein-Uhlenbeck noise, $d\xi_i = -\xi_i/\tau_{\text{noise}} dt + \sqrt{2/\tau_{\text{noise}}} dW_i$, which introduces a temporal correlation τ_{noise} and is the natural diffusion limit of Poisson spiking input (Smith, 2010).

2.9 Parameter bounds and weight units

Units. Synaptic weights w_{ij} have units of nA/Hz: multiplied by a presynaptic rate (Hz) they give a current in nA that enters Φ_i . External drives I_i^0 and nAChR currents I_i^{ACh} are in nA directly. Adaptation strengths J_{adapt} are also nA/Hz. The exception is J_{NMDA} , in nA: it multiplies the dimensionless gating variable $S^*(r) \in [0, 1]$, which saturates to ≈ 1 at any

physiological firing rate, so J_{NMDA} effectively sets a constant ceiling on recurrent excitatory current.

Role of the bounds. The search bounds listed in Table 1 serve primarily to keep the optimisation in the right region, not to encode strong biological priors. Without bounds, optimization could reach unrealistic values. Bounds prevent this by keeping every pathway in a range where it can plausibly interact with the others.

As the units are consistent, we can estimate the order of magnitude of the weights from the target firing rates and the transfer-function shape. For example, with a PYR high-state target of 60.2 Hz and the Wong-Wang slope $\alpha_e = 310$ Hz/nA, the total net input needed is $\Theta_e + r/\alpha_e \approx 0.40 + 0.19 \approx 0.60$ nA. Similarly, for PV at its target of 35.3 Hz: $\Theta_i + r/\alpha_i \approx 0.29 + 0.06 \approx 0.35$ nA. This gives insight into the order of magnitude of the currents, and consequently of the weights, by relating them to the firing rates of the presynaptic populations.

Group	Parameter	Lo	Init	Hi	Mode	Unit
Adaptation strength	$J_{\text{adapt,PYR}}$	0.001	0.002	0.2	log	nA/Hz
GABA gain	$g_{\text{GABA,base}}$	0.1	1.0	5.0	lin	–
	$g_{\alpha 7}$	0.1	0.1	5.0	lin	–
Synaptic weights	J_{NMDA}	0.05	0.3	2.0	log	nA
	w_{ep}, w_{es}, w_{ev}	0.001	0.002	0.05	log	nA/Hz
	w_{pe}	0.001	0.002	0.05	log	nA/Hz
	w_{pp}	0.001	0.002	0.05	log	nA/Hz
	w_{se}, w_{sp}	0.001	0.002	0.05	log	nA/Hz
	w_{vp}, w_{vs}	0.001	0.002	0.05	log	nA/Hz
External drives	I_{PYR}^0	0.01	0.44	1.5	lin	nA
	I_{PV}^0	0.01	0.35	0.6	lin	nA
	I_{SOM}^0	0.01	0.35	0.6	lin	nA
	I_{VIP}^0	0.01	0.35	0.6	lin	nA
nAChR currents	$I_{\alpha 7, \text{PV}}$	0.01	0.20	0.5	lin	nA
	$I_{\alpha 7, \text{SOM}}$	0.01	0.20	0.5	lin	nA
	$I_{\beta 2, \text{SOM}}$	0.01	0.20	0.5	lin	nA
	$I_{\alpha 5, \text{VIP}}$	0.01	0.20	0.5	lin	nA

Table 1: Optimisation search bounds.

3 Results

3.1 Fitting result: single-node bistable optimisation

The optimiser minimises the composite loss L_{total} (Section 2.6) over the full parameter space. As shown in Figure 4, there is a rapid early decrease, reflecting the optimiser’s discovery of a bistable state ($L_{\text{bistab}} \rightarrow 0$), followed by a slower tail corresponding to the joint minimisation of the rate-matching terms L_{rate} and $L_{\text{rate,high}}$.

The optimiser converges to a solution early (around 3000 iterations), with a non-negligible total loss that we analyse in the next section.

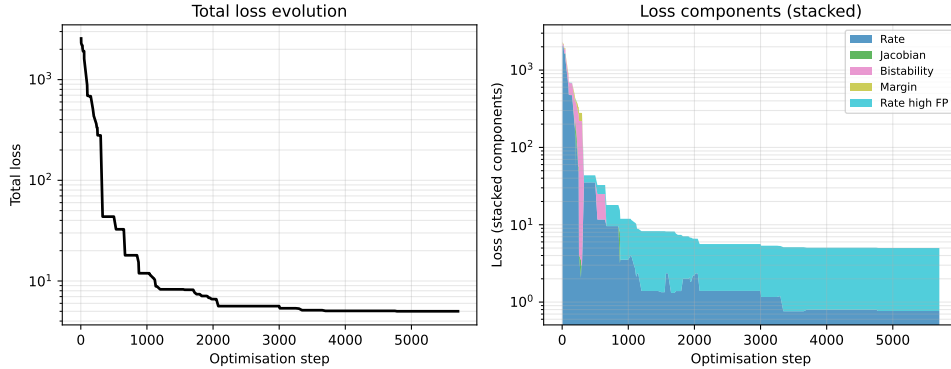


Figure 4: Loss evolution during optimisation.

The evolution of each individual loss component, and their relative contributions to the total loss, confirms the interpretation above: L_{bistab} dominates the early phase, while L_{rate} and $L_{\text{rate,high}}$ dominate the later phase. The margin loss L_{margin} is zero after bistability is achieved, which confirms that the fixed points are well separated.

The final fitted parameters are given in Appendix E. Notably, none of the fitted values reach either the lower or the upper bound, which indicates that the bounds are in a reasonable range and that the optimiser is not constrained by them.

3.2 SOM extinction at the high fixed point

As previously mentioned, the optimiser converges to a solution with a non-zero loss, which indicates that not all conditions are satisfied. The most striking deviation from the targets is the complete silencing of SOM in the high state: the fitted firing rate is 0 Hz compared to a target of 35.2 Hz (Table 2).

Population	Low-FP (Hz)	Target low (Hz)	High-FP (Hz)	Target high (Hz)
PYR	0.37	1.75	50.12	60.20
SOM	1.44	1.12	0.00	35.20
PV	0.98	1.04	29.42	35.30
VIP	1.65	1.33	73.07	68.80

Table 2: Fixed-point firing rates at the fitted optimum compared with Rooy (2021) targets.

Looking at the mechanisms underlying the fitted solution, we find that the VIP→SOM disinhibitory pathway is strong compared to the VIP→PV pathway. This means that, in the high state, VIP strongly suppresses SOM, which in turn releases PYR from inhibition and allows it to reach the high firing rate, while PV inhibits PYR to regulate its high state. There is strong pressure here to achieve bistability, and the optimiser finds a solution that uses VIP to silence

SOM in the high state. Whether this solution is biologically plausible is an open question. The VIP→SOM pathway is indeed a key disinhibitory motif in cortex, as previously mentioned in the Introduction (see (Karnani et al., 2016)). However, the complete silencing of SOM in the high state is not consistent with the Rooy data, which show a significant increase from low to high state.

Two non-mutually-exclusive explanations can be proposed. First, the local circuit may be structurally incomplete: in vivo, additional interneuron classes or inter-layer connections might prevent total VIP-mediated SOM silencing in the high state. Second, the columnar topology provides a geometric explanation: we know from the literature that SOM cells receive input from the local pyramidal population and project their output to neighbouring columns. A column in the high state therefore silences its neighbours’ SOM, not its own. The SOM rate observed locally at the high fixed point may thus reflect the SOM of the columns projecting into the active column, which are necessarily in the low state. Both hypotheses are developed further in the Discussion.

An independent random parameter search over the same bounds (Appendix D) partially rules out a loss-landscape artefact: SOM silencing is the dominant outcome across random bistable hits, and no random sample reaches a high state with both SOM and PV at biologically plausible rates. The optimiser should therefore be operating in a landscape where the only way to achieve bistability is through SOM silencing.

3.3 The optimised nullcline exhibits genuine bistability

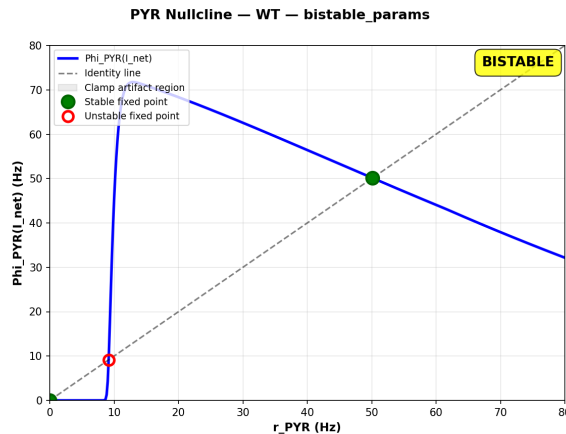


Figure 5: PYR nullcline at the fitted optimum.

While the network does not fully match the experimental observations, we can still verify that the fitted parameters produce a genuine bistable configuration. Figure 5 shows the PYR nullcline $F(r) = \Phi_{\text{PYR}}(I^{\text{net}}(r)) - r$ at the fitted optimum. The function exhibits the required sign pattern (+, −, +, −): two stable fixed points at approximately 0.4 Hz and 50 Hz are separated by an unstable branch, confirming that our machine-learning approach successfully found a bistable configuration. The low FP lies below the target of 1.75 Hz, while the high FP is close to the target of 60.2 Hz. This nearly silent low point is interesting, and could be explained by the absence of noise in our optimisation procedure. Indeed, noise would likely push the system transiently out of the low state, so the average low-state rate observed in simulation would be higher than the deterministic low FP. We return to this point in the Discussion: it could be addressed by including noise in the optimisation procedure, which is currently not done.

The result is notable: the loss function, designed purely from abstract geometric requirements on $F(r)$, recovers a biophysically grounded bistable circuit without requiring further fine-tuning. However, such an approach is likely to produce a degenerate solution space, with

multiple parameter sets producing the same nullcline shape. We discuss this point further in the Discussion.

3.4 Single-node stochastic dynamics

To verify that the nullcline result above carries over to a stochastically driven bistable network, we run simulations of the single-node model under stochastic input and a transient cue, sweeping over multiple noise levels and cue amplitudes. We observe that the fitted parameters can indeed support a cue-triggered switch from the low to the high state, and we characterise the set of noise and cue conditions for which this occurs.

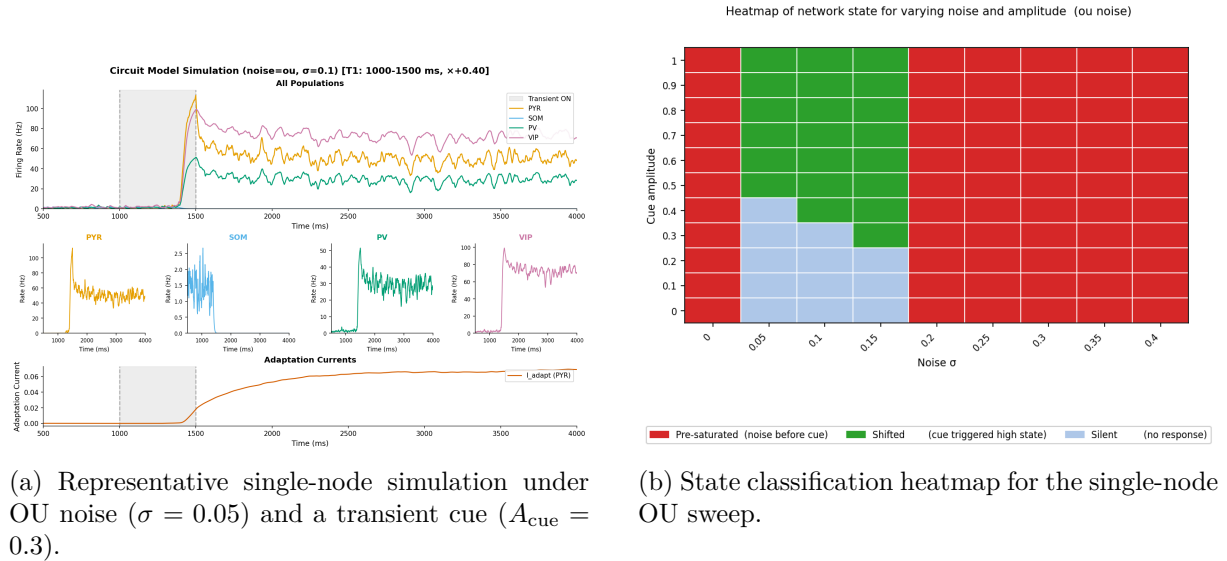


Figure 6: Single-node transient-response results.

Figure 6a shows a representative run, while Figure 6b summarises the sweep. While noise mainly saturates the circuit when it is too high, a moderate amount is in fact necessary to stabilise the network in the low state: with $\sigma = 0.0$, the circuit is saturated (see the heatmap in Figure 6b and the corresponding example in Appendix 12). We can also note that the noise level changes the cue amplitude required to trigger the switch, with higher noise requiring a higher cue amplitude. We discuss this point further in the Discussion: noise would help inhibit the network overall, and would therefore participate in stabilising the low state.

3.5 Ring attractor dynamics

Having established that the fitted parameters can reproduce a sustained high state at the single-node level and in a noisy environment, we now move on to the multi-columnar architecture defined earlier in order to determine whether it supports a spatially localised bump. We compare the no-SOM control with a Gaussian SOM kernel (see the SOM connectivity diagram in Appendix 11).

First, we focus on the structure without SOM lateral inhibition (Figure 7): we observe the same behaviour as in the single-node case, with a cue-triggered transition to a high-activity state. The influence of each parameter (noise level and cue amplitude) is also very close to the single-node case, with higher noise and lower cue amplitude leading to more silent runs. However, we do not observe the $\sigma = 0.0$ saturation effect seen in the single-node case, which is interesting and could be explained by the fact that, since each column's activity inhibits the whole network, the ring architecture itself stabilises the activity.

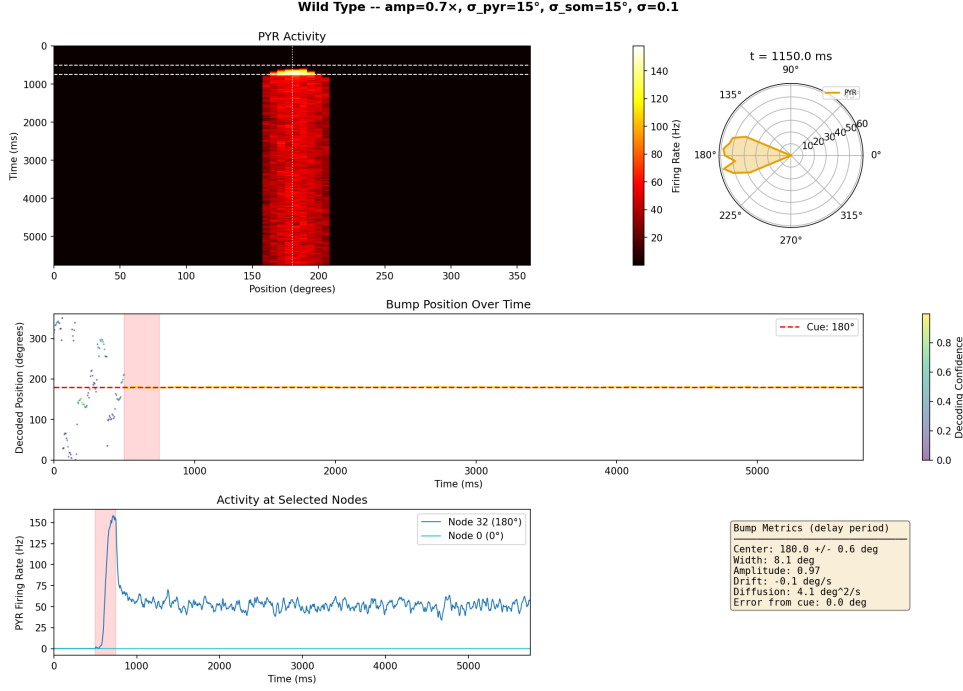


Figure 7: Ring simulation without lateral SOM inhibition.

The bump we obtain is very clear and stable despite the noise, and the heatmap shows the same pattern as in the single-node case (see Appendix 13). Spatially, the bump does not spread at all, which is expected from the architecture's discrete distinction between the active and the low state. However, this does not align with the behavioural data, which show a more spread-out bump. We discuss this point further in the Discussion.

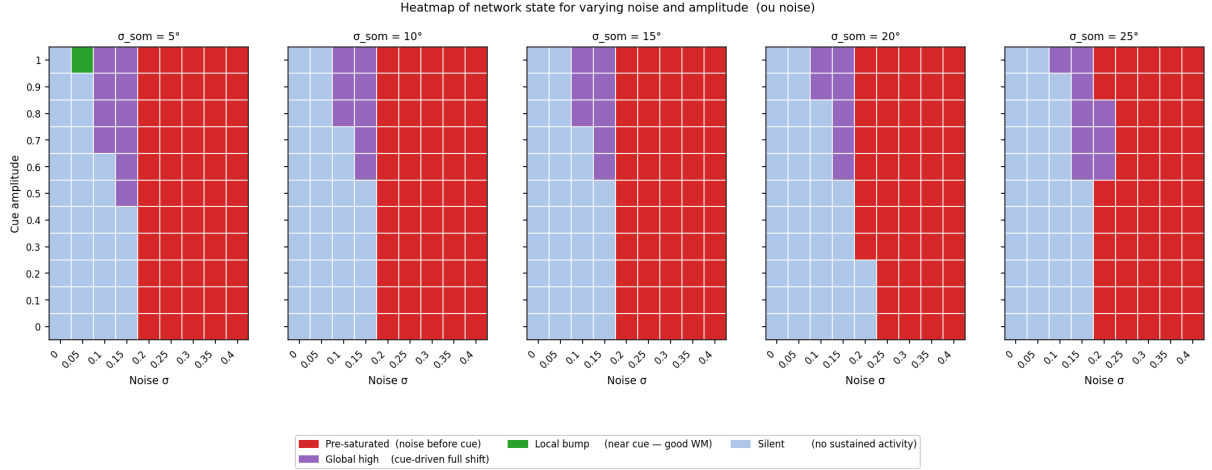


Figure 8: State classification heatmap for the ring simulation with Gaussian SOM inhibition.

We now turn to our previously defined SOM lateral-inhibition architecture. As shown in the heatmap (Figure 8), almost no setting allows the network to form a genuine bump: it is either silent, saturated, or the bump fails to localise, drifts and spreads across the ring, eventually leading to a fully saturated network (see Appendix 14 for an example of this behaviour). This is a disappointing result, but one point in the heatmap does show a bump. Figure 9 examines this case more closely and confirms the presence of a bump. While this is a promising sign that the architecture can support a genuinely localised ring response, the fact that it only occurs in a

very narrow parameter range is concerning and suggests that the solution is not robust. Recall that the single-node fit produces a high state in which SOM is fully silenced (Section 3.2). On the ring, this means that an active column delivers very little SOM-mediated inhibition to its neighbours, which could explain why the lateral inhibition does not appear strong enough to restrain the bump. We discuss this point further in the Discussion.

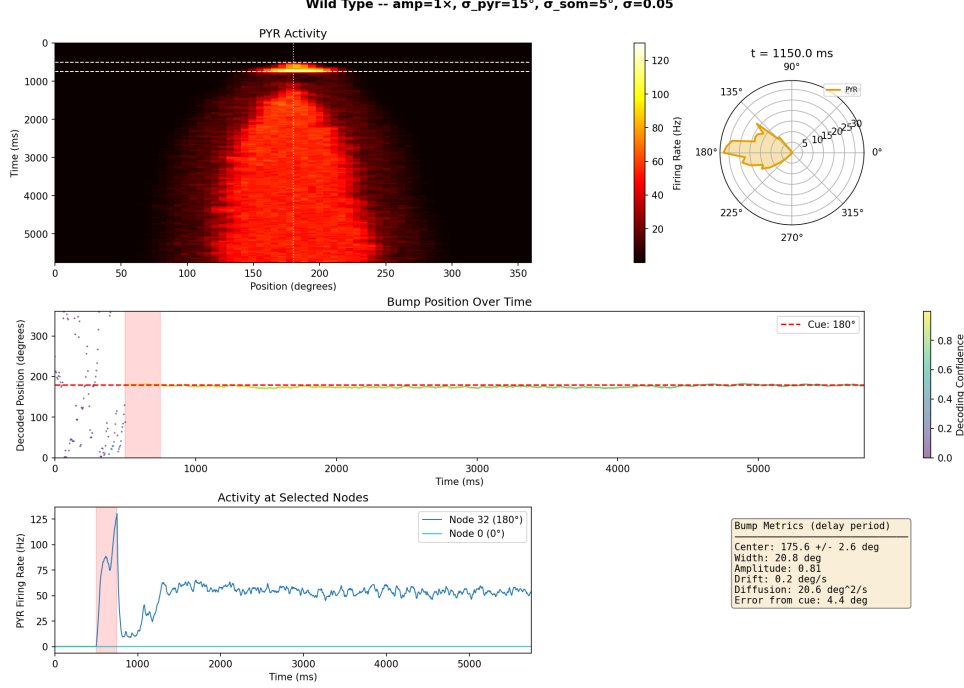


Figure 9: Example run from the SOM Gaussian sweep.

3.6 Towards the AD application: nAChR deactivation

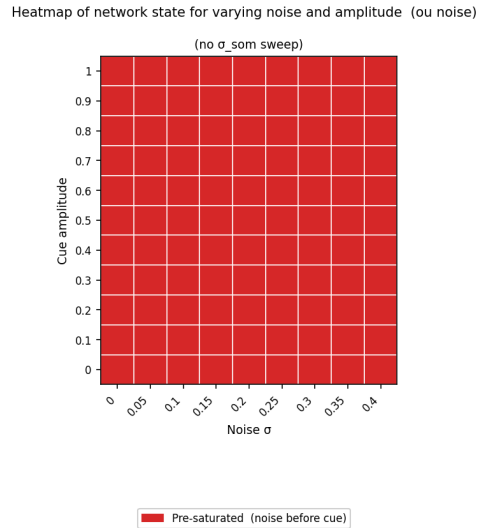


Figure 10: 2D heatmap of the ring simulation under nAChR deactivation.

Here, we attempt to bridge back to the AD question by applying nAChR-drive perturbations to the fitted ring model, simulating the effects of APP accumulation on the local circuit. We use data from Koukouli et al. (2025) to set the deactivation percentages for each nAChR subtype and

target population, and observe the resulting changes in the ring dynamics. While the method is correct, there are strong caveats in our context: the fitted nAChR drive may not represent the actual nAChR drive in vivo, and one might also expect the network to adapt its weights when disrupted. We discuss these limitations in the Discussion and propose future directions to address them. That said, Figure 10 shows that nAChR deactivation drives the network into a saturated regime, in every combination of noise and cue amplitude.

4 Discussion

Summary of findings. Starting from the local circuit of Koukouli et al. (2025), we identified that the published equations cannot support single-node bistability and introduced an NMDA gating variable on the PYR→PYR connection, restoring the saturation needed to fold the nullcline. We built a loss-driven optimisation that successfully located parameters producing a genuine bistable configuration with two well-separated stable fixed points (Section 3.3). The corresponding single-node circuit supports a cue-triggered switch from the low to the high state under noise (Section 3.4). Embedded in a ring without SOM lateral inhibition, the same circuit supports a clear, stable bump (Section 3.5), and does so over a broad parameter regime. We then proposed a more complete inhibition pattern, closer to biological reality, that mimics the known SOM connectivity, but the resulting ring fails to produce a stable bump except in a narrow parameter range (Section 3.5). Finally, we apply nAChR deactivation to the fitted ring and observe that it drives the network into saturation in every condition (Section 3.6). However, the link with biological data has been weakened by multiple factors, and the resulting fit of the network does not reproduce the biological data. These results are not necessarily failures of the model, but they do raise questions about the underlying mechanisms and the model’s limitations. In our context of AD-related working-memory deficits we cannot afford to ignore these issues, so we discuss them in detail below and propose concrete future directions to address them.

SOM silencing at the high state. The two hypotheses raised in Section 3.2 deserve to be developed separately. The first is that the model is restricted to layer II/III: by construction, it does not take vertical connections into account, even though these matter for biological interpretation. Hartung et al. (2024) show that NDNF-expressing interneurons in layer 1 provide pervasive inhibition onto layer II/III interneurons, including PV and VIP, and play a role in regulating those layers. We can also highlight the long-range projections to VIP cells, notably from the hippocampus Graham et al. (2021). The current architecture does not represent these interactions, which would otherwise contribute top-down and long-range pathways. The second hypothesis is geometric: SOM cells are local-receiving but lateral-projecting (Section 1.3), so the SOM rate measured at an active column reflects the rate output to the neighbours, which sit in the low state. In the single-node fit, the SOM population is forced into the same role as both source and target of inhibition at the active column, an arrangement that does not exist on the ring once the PYR→SOM→neighbour-PYR motif is in place. Under that view, SOM silencing at the local fixed point is not a failure of the model but the price of fitting a single column that, by construction, lacks the spatial separation between SOM source and target. A natural next step would be to develop a ring-aware fit, in which the single-node loss is replaced by a small-ring loss that includes the lateral motif, to test whether this allows the SOM rate at the high FP to increase and match the target. We could also test this hypothesis by examining the SOM rate of different cells in experimental data and their relationship to the local PYR rate: if the hypothesis is correct, we should see a negative correlation between the two, with SOM silencing on the active column but not on the neighbours.

The deterministic low fixed point and noise during optimisation. The fitted low PYR fixed point sits at 0.37 Hz, significantly below the 1.75 Hz Rooy target (Table 2). However, in the noisy simulations the population spends most of its time displaced upwards by fluctuations, so the time-averaged rate is higher than the deterministic FP. The optimisation procedure compares the deterministic FP to a target rate measured in vivo, where the rate is itself a noisy time average, and the two quantities are not directly comparable. A noise-aware version of the optimisation, in which the loss is evaluated on time-averaged rates from short stochastic runs rather than the deterministic FP, would close this gap. We list this as a possible future direction below.

Another point we observed was the need for a moderate amount of noise to stabilise the low state in the single-node simulations (Figure 6b). It is interesting that noise, which is often thought of as a destabilising factor, can in fact help stabilise the low state. This suggests that the baseline current into inhibitory populations is not sufficient to stabilise the low state on its own, and that the network relies on a certain amount of noise to activate inhibition and prevent saturation. As we mentioned in Section 1.3, PV cells are known to receive strong thalamic input, which is a likely source of noise in vivo. And, since PV cells are known to be the global regulator, this noise would have a stabilising effect on the whole network.

Bump shape and the discrete-bistability limitation. The no-SOM ring produces a clearly localised bump (Figure 7), but the bump has very steep flanks: each column is either at its high or its low state, with no graded intermediate region. Behavioural data on spatial working memory typically show a smoother spatial profile, with firing rate decaying gradually away from the cued direction Compte et al. (2000). The architecture, as built here, cannot reproduce that. Single-node bistability discretises each column into two regimes, and ring coupling alone does not generate intermediate rates because the coupling is too weak to push a column out of one basin into the other except in an all-or-nothing way. However, in the SOM Gaussian case, the bump is more spread out and show a intermediate region between the high and low state, which is more in line with the behavioural data.

Ring-level non-robustness traced to the single-node fit. The SOM-Gaussian ring forms a bump in only a narrow region of the noise/cue space (Figure 8). Taken at face value, this would suggest that the lateral inhibition is not strong enough to confine an active column, though we cannot firmly state this at this stage, given the limitations of the fit. A structural explanation is the magnitude of the SOM→PYR weight inherited from the single-node fit: that weight was fitted under the assumption that SOM inhibition acts locally on the same column that drove it, while on the ring it acts laterally onto the neighbours. The row-sum principle preserves the homogeneous fixed point but not the inhibition density seen by an isolated active column, so neighbours receive a much sharper lateral inhibition than the single-node fit was tuned for. A ring-aware fit, in which the single-node loss is replaced by a small-ring loss including the lateral motif, is the natural fix.

The AD bridge and the unconstrained nAChR drive. The saturation observed under nAChR deactivation (Figure 10) is the most concerning of the negative results, but it is also the one most likely to reflect a methodological gap rather than a model failure. As things stand, our nAChR currents $I_{\alpha 7, PV}$, $I_{\alpha 7, SOM}$, $I_{\beta 2, SOM}$, $I_{\alpha 5, VIP}$ are fitted with no constraint that distinguishes them from the corresponding baseline external drives I_{PV}^0 , I_{SOM}^0 , I_{VIP}^0 : any shift between the two components leaves the resting state unchanged. The optimisation therefore has no incentive to give the nAChR currents a physiologically meaningful value, and the deactivation of nAChRs is operationally indistinguishable from a generic perturbation of the baseline drive. Koukoulis et al. (2025) address this directly by adding a knock-out firing-rate target: they use knock-out data and simulate the network with the corresponding nicotinic current set to zero, optimising

the network to also match the firing rates in those conditions. We did not have access to such knock-out-condition rate targets in the Rooy dataset used here, so the same constraint could not be applied. The practical consequence is that our AD result (network saturation) reflects the loss of an unconstrained current and cannot be read as a model prediction about A β -induced WM impairment. A meaningful AD application would require either re-fitting with KO rate targets in the loss or, more pragmatically, fixing the nAChR currents from biophysical estimates (channel densities, single-channel currents) and re-optimising only the remaining parameters.

Toward biophysically grounded transfer functions. The hyperbolic ceiling in eq. 8 is deliberately phenomenological. As the objective is to build a biologically grounded model, it is unsatisfactory to have a key mechanism that is not directly tied to a biophysical process: the NMDA gating variable on the PYR $\rightarrow$ PYR recurrence is an improvement in this direction, and we should aim to replace the interneuron ceiling with a similar mechanism. Several biophysical candidates exist: GABA_A-mediated autaptic self-inhibition in PV cells (Bacci et al., 2003; Szegedi et al., 2020), presynaptic GABA_B autoreceptors reducing GABA release at high interneuron firing rates (Davies et al., 1990), and short-term depression of excitatory inputs onto interneurons (Abbott et al., 1997; Tsodyks and Markram, 1997). We could implement each of these mechanisms in the model and test whether they can replace the phenomenological ceiling while still supporting bistability and a good fit to the data.

Constraining the circuit through the Jacobian. The Jacobian regulariser L_{jac} included in our loss deserves a critical look. Its motivation is structural: by penalising too-high effective gains along the canonical pathways, it is meant to discourage solutions where a connection is too strong. The justification, however, is weak. There is no direct biological grounding for it, the threshold ϵ and the list of pathways to enforce are arbitrary, and the term largely amounts to writing the desired circuit topology into the loss rather than letting it emerge from the model. Empirically, L_{jac} stays negligible against the other components throughout training (Figure 4), so it is unclear whether the term meaningfully reshapes the loss landscape at all. We kept it in the current model for transparency, but its necessity is questionable. Specifically, we set the weight of the loss term to 0.1, which gives a near-zero contribution to the total loss.

Fast-synapse approximation on interneurons. One further asymmetry of the model deserves to be made explicit. The fast-synapse approximation used on all interneurons (everywhere except PYR $\rightarrow$ PYR) collapses the GABAergic and AMPA-mediated synaptic variables to instantaneous functions of the presynaptic rate. This is consistent with the short time constants of the corresponding receptors, but it removes exactly the kind of synaptic dynamics that, on PYR, produce bistability.

Declaration of contribution

All steps of this thesis (defining the scientific question, bibliographic research, choosing the general approach and the specific methodology, developing and programming the model and the studies, designing the analysis, interpreting the results, and writing the dissertation) were carried out by myself, Valentin Jacquin. Boris Gutkin, my supervisor, provided guidance throughout the project and detailed feedback on the dissertation as a whole. Evann Rabeau provided additional proofreading of the introduction. Anthropic’s Claude (via Claude Code) was used to assist with coding and troubleshooting during the implementation and analysis stages.

Code availability

All code developed for this thesis — including the model implementation, fitting pipeline, analysis scripts and parameters — is publicly available at https://github.com/valjcq/AD_pfc_modeling.

References

- L. F. Abbott, J. A. Varela, Kamal Sen, and S. B. Nelson. Synaptic depression and cortical gain control. *Science*, 275(5297):221–224, 1997. doi: 10.1126/science.275.5297.221.
- Edson X. Albuquerque, Edna F. R. Pereira, Manickavasagom Alkondon, and Scott W. Rogers. Mammalian nicotinic acetylcholine receptors: From structure to function. *Physiological Reviews*, 2009. doi: 10.1152/physrev.00015.2008.
- Ove Almkvist. Neuropsychological features of early Alzheimer’s disease: Preclinical and clinical stages. *Acta Neurologica Scandinavica*, 94(S165):63–71, 1996. doi: 10.1111/j.1600-0404.1996.tb05874.x.
- Alberto Bacci, John R. Huguenard, and David A. Prince. Functional autaptic neurotransmission in fast-spiking interneurons: A novel form of feedback inhibition in the neocortex. *Journal of Neuroscience*, 23(3):859–866, 2003. doi: 10.1523/JNEUROSCI.23-03-00859.2003.
- Zeinab Breijyeh and Rafik Karaman. Comprehensive review on alzheimer’s disease: Causes and treatment. *Molecules*, 2020. doi: 10.3390/molecules25245789.
- Marcelo Camperi and Xiao-Jing Wang. A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience*, 5(4):383–405, 1998. doi: 10.1023/A:1008837311948.
- Albert Compte, Nicolas Brunel, Patricia S. Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000. doi: 10.1093/cercor/10.9.910.
- Nick Corriveau-Lecavalier, Jenna N. Adams, Larissa Fischer, Eóin N. Molloy, and Anne Maass. Cerebral hyperactivation across the alzheimer’s disease pathological cascade. *Brain Communications*, 6(6):fcae376, 2024. doi: 10.1093/braincomms/fcae376.
- C. H. Davies, S. N. Davies, and G. L. Collingridge. Paired-pulse depression of monosynaptic gaba-mediated inhibitory postsynaptic responses in rat hippocampus. *Journal of Physiology*, 422(1):513–531, 1990. doi: 10.1113/jphysiol.1990.sp018080.
- Alain Destexhe. Self-sustained asynchronous irregular states and up-down states in thalamic, cortical and thalamocortical networks of nonlinear integrate-and-fire neurons. *Journal of Computational Neuroscience*, 27(3):493–506, 2009. doi: 10.1007/s10827-009-0164-4.
- Rastislav DRUGA, Martin SALAJ, and Azzat AL-REDOUAN. Parvalbumin - positive neurons in the neocortex: A review. *Physiological Research*, 72(Suppl 2):S173–191, 2023. doi: 10.33549/physiolres.935005.
- Alexei V. Egorov, Bassam N. Hamam, Erik Fransén, Michael E. Hasselmo, and Angel A. Alonso. Graded persistent activity in entorhinal cortex neurons. *Nature*, 420(6912):173–178, 2002. doi: 10.1038/nature01171.

- Sami El-Boustani and Mriganka Sur. Response-dependent dynamics of cell-specific inhibition in cortical networks in vivo. *Nature Communications*, 5(1):5689, 2014. doi: 10.1038/ncomms6689.
- Brielle R. Ferguson and Wen-Jun Gao. Pv interneurons: Critical regulators of e/i balance for prefrontal cortex-dependent behavior and psychiatric disorders. *Frontiers in Neural Circuits*, 12, 2018. doi: 10.3389/fncir.2018.00037.
- Ernesto Flores-Martínez and Fernando Peña Ortega. Amyloid β peptide-induced changes in prefrontal cortex activity and its response to hippocampal input. *International Journal of Peptides*, 2017:7386809, 2017. doi: 10.1155/2017/7386809.
- Kourtney Graham, Nelson Spruston, and Erik B. Bloss. Hippocampal and thalamic afferents form distinct synaptic microcircuits in the mouse infralimbic frontal cortex. *Cell Reports*, 37(3):109837, 2021. doi: 10.1016/j.celrep.2021.109837.
- Jan Hartung, Anna Schroeder, Rodrigo Alejandro Pérez Vázquez, Rogier B. Poorthuis, and Johannes J. Letzkus. Layer 1 ndnf interneurons are specialized top-down master regulators of cortical circuits. *Cell Reports*, 43(5):114212, 2024. doi: 10.1016/j.celrep.2024.114212.
- Mahesh M. Karnani, Jesse Jackson, Inbal Ayzenshtat, et al. Opening holes in the blanket of inhibition: Localized lateral disinhibition by vip interneurons. *Journal of Neuroscience*, 36(12):3471–3480, 2016. doi: 10.1523/JNEUROSCI.3646-15.2016.
- Anna-Mariya Kirova, Rebecca B. Bays, and Sarita Lagalwar. Working memory and executive function decline across normal aging, mild cognitive impairment, and alzheimer’s disease. *BioMed Research International*, 2015:748212, 2015. doi: 10.1155/2015/748212.
- Fani Koukouli, Boris Gutkin, Zhang Chun-Lei, Ivan Lazarevich, and Marie Rooy. Nicotinic receptors mediate network dysfunction in early alzheimer’s disease. *Nature Medicine*, 2025. doi: 10.1038/s41380-025-03241-4.
- Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC (New York, NY, USA)*, LLVM ’15, pages 1–6, 2015. doi: 10.1145/2833157.2833162.
- Sylvia Lombardo and Uwe Maskos. Role of the nicotinic acetylcholine receptor in alzheimer’s disease pathology and treatment. *Neuropharmacology*, 96:255–262, 2015. doi: 10.1016/j.neuropharm.2014.11.018.
- Marta Milà-Alomà, Gemma Salvadó, Juan Domingo Gispert, et al. Amyloid beta, tau, synaptic, neurodegeneration, and glial biomarkers in the preclinical stage of the alzheimer’s continuum. *Alzheimer’s & Dementia*, 16(10):1358–1371, 2020. doi: 10.1002/alz.12131.
- Christoforos A. Papasavvas, Yujiang Wang, Andrew J. Trevelyan, and Marcus Kaiser. Gain control through divisive inhibition prevents abrupt transition to chaos in a neural mass model. *Physical Review E*, 92(3):032723, 2015. doi: 10.1103/PhysRevE.92.032723.
- D. Perpetuini et al. Working memory decline in alzheimer’s disease is detected by complexity analysis of multimodal eeg-fnirs. *Entropy*, 2020.
- Hyun-Jae Pi, Balázs Hangya, Duda Kvitsiani, Joshua I. Sanders, Z. Josh Huang, and Adam Kepecs. Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477):521–524, 2013. doi: 10.1038/nature12676.

- Rogier B. Poorthuis, Bernard Bloem, Benita Schak, Jordi Wester, Christiaan P. J. de Kock, and Huibert D. Mansvelder. Layer-specific modulation of the prefrontal cortex by nicotinic acetylcholine receptors. *Cerebral Cortex*, 23(1):148–161, 2013. doi: 10.1093/cercor/bhr390.
- Marie Rooy, Ivan Lazarevich, Fani Koukouli, Uwe Maskos, and Boris Gutkin. Cholinergic modulation of hierarchical inhibitory control over cortical resting state dynamics: Local circuit modeling of schizophrenia-related hypofrontality. *Current Research in Neurobiology*, 2021. doi: 10.1016/j.crneur.2021.100018.
- Maria V. Sanchez-Vives, Lionel G. Nowak, and David A. McCormick. Cellular mechanisms of long-lasting adaptation in visual cortical neurons in vitro. *Journal of Neuroscience*, 20(11):4286–4299, 2000. doi: 10.1523/JNEUROSCI.20-11-04286.2000.
- Philip L. Smith. From poisson shot noise to the integrated ornstein–uhlenbeck process: Neurally principled models of information accumulation in decision-making and response time. *Journal of Mathematical Psychology*, 2010. doi: 10.1016/j.jmp.2009.12.002.
- Kartik K. Sreenivasan and Mark D’Esposito. The what, where and how of delay activity. *Nature Reviews Neuroscience*, 2019. doi: 10.1038/s41583-019-0176-7.
- Viktor Szegedi, Melinda Paizs, Judith Baka, et al. Robust perisomatic gabaergic self-innervation inhibits basket cells in the human and mouse supragranular neocortex. *eLife*, 9, 2020. doi: 10.7554/eLife.51691.
- Robin Tremblay, Soohyun Lee, and Bernardo Rudy. Gabaergic interneurons in the neocortex: From cellular properties to circuits. *Neuron*, 91(2):260–292, 2016. doi: 10.1016/j.neuron.2016.06.033.
- Misha V. Tsodyks and Henry Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences*, 94(2):719–723, 1997. doi: 10.1073/pnas.94.2.719.
- Min Wang, Yang Yang, Ching-Jung Wang, et al. Nmda receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron*, 77(4):736–749, 2013. doi: 10.1016/j.neuron.2012.12.032.
- Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006. doi: 10.1523/JNEUROSCI.3733-05.2006.
- Iryna Yavorska and Michael Wehr. Somatostatin-expressing inhibitory interneurons in cortical circuits. *Frontiers in Neural Circuits*, 10, 2016. doi: 10.3389/fncir.2016.00076.
- Joel Zylberberg and Ben W. Strowbridge. Mechanisms of persistent activity in cortical circuits: Possible neural substrates for working memory. *Annual Review of Neuroscience*, 40:603–627, 2017. doi: 10.1146/annurev-neuro-070815-014006.

A Pre-registration

Modeling Nicotinic Receptor Dynamics in Alzheimer's Disease

Valentin Jacquin under the supervision of Dr. Boris Gutkin
Group for Neural Theory
ENS Paris, France

Abstract

Alzheimer's disease (AD) is a prevalent neurodegenerative disorder characterized by progressive cognitive decline. During its early stage, the $A\beta$ molecule accumulates and provokes various disruptions in the neural system, notably leading to changes in prefrontal cortex (PFC) activity. Previous studies have explored the link between its disruption of nicotinic acetylcholine receptors (nAChRs) and alterations in PFC dynamics observed in early AD. Building upon their computational model of the PFC, we aim to investigate how nAChR dysfunction affects PFC activity during working memory tasks, which are known to be impaired in AD. Incorporating nAChR modulation, the model will allow us to explore the impact of receptor dysfunction on network bistability and transition dynamics. By implementing a ring attractor architecture, we will simulate working memory processes and assess how impairments in specific nAChR subtypes influence the network's ability to maintain stable activity states. We will analyze firing rates, transition frequencies, and working memory performance metrics to elucidate the relationship between nAChR dysfunction and cognitive deficits in AD.

Keywords: Alzheimer's disease, nicotinic acetylcholine receptors, prefrontal cortex, computational modeling, working memory.

Scientific Background and Context

Alzheimer's disease (AD) is a widespread neurodegenerative disorder with a high prevalence in elderly population, currently affecting over 50 millions individuals worldwide [Breijyeh and Karaman, 2020]. However, the origins of the cognitive impairments associated with AD remain poorly understood. While Palop et al. [2007] have reported increases in cortical activity in mice with AD, human imaging studies have predominantly observed hypofrontality in AD patients [Choi et al., 2023]. This apparent discrepancy highlights how the underlying mechanisms remain debated. Experimental evidence indicates that amyloid-beta ($A\beta$) peptides modulate neuron's signals by either activating or inhibiting nicotinic acetylcholine receptor (nAChR). More specifically, $\alpha 7$ and $\beta 2^*$ nAChR's subunits seems to be particularly affected by $A\beta$ interactions [Lombardo and Maskos, 2015, Parri et al., 2011].

In this work, we hypothesize that the progressive accumulation of $A\beta$ disrupts nAChR mechanisms, leading to a dysregulation in prefrontal cortex (PFC) network dynamics. Thus, contributing to the cognitive deficits observed in AD. As nAChRs are only expressed in interneurons in layer II/III [Poorthuis et al., 2013], nicotinic signaling is expected to impact pyramidal cells through interneurons modulation. In this regard,

Koukouli et al. [2017] demonstrated the link between nAChRs disruption and pyramidal neurons activity, underlying the interneurons implication in PFC network dynamics. The specific contribution of different receptor subunits to this modulation has also been highlighted [Maskos and Lombardo, 2014]. Still, the precise interactions between nAChR dysfunction and the resulting alterations in PFC network dynamics in the context of AD remain to be explained.

To better capture these mechanisms, previous work developed a reduced population-level model describing interactions between pyramidal neurons and major interneuron subtypes within the PFC [Koukouli et al., 2025]. This model has been fitted to their own produced data. It reproduces the hyperactivity observed in their genetically modified mice and identifies $\alpha 7$ nAChRs as critical modulators of this phenomenon. Moreover, it suggests that $\beta 2^*$ and $\alpha 5$ nAChRs could represent promising therapeutic targets for restoring physiological network balance.

Research Question and Working Hypotheses

While previous explorations have shown promising results in linking nAChR dysfunction to altered PFC activity patterns [Rooy et al., 2021, Koukouli et al., 2025] they primarily focused on resting-state dynamics. However, to link these findings to the observed cognitive deficits in AD, it is essential to investigate how these network alterations manifest during active cognitive tasks. The PFC is known to be a key region in working memory, which deteriorates in early AD [Huntley and Howard, 2010]. Furthermore, electrophysiological recordings during working memory tasks have been shown to reliably distinguish AD patients from healthy controls [Perpetuini et al., 2020], underscoring the connection between altered neural dynamics and cognitive impairment.

Therefore, working memory task provides a good candidate to explore how nAChR dysfunction affects PFC network dynamics in an active context. We will investigate how disruptions in the different nAChR subtypes influence the network's ability to switch between two self-sustained states, which are thought to underlie working memory processes [Compte et al., 2003]. We hypothesize that impairments in specific nAChR subtypes will lead to deficits in the stability and transition dynamics between these states, thereby contributing to the working memory impairments observed in AD.

To another extent, we will also explore how varying levels of external input or sensitivity on specific nAChR subtypes affect these dynamics, building on previous findings that identified $\beta 2^*$ and $\alpha 5$ nAChRs as key modulators of PFC activity [Koukouli et al., 2025].

Planned Methodology

To tackle this question, we will build upon the previously developed nAChR-modulated PFC network model described in [Koukoulis et al. \[2025\]](#).

Modulation of parameters affecting bistability

Previous results have shown an oscillation between high and low activity states in the PFC [[Rooy et al., 2021](#)], but at another order of magnitude than the one observed experimentally during working memory tasks [[Compte et al., 2003](#)]. Thus, it is needed to bridge the gap closer to experimental observations, as we aim to maintain a biologically relevant model. To do so, we will modulate current inputs as we expect them to highly affect the characteristics of both states.

Working memory architecture

Once we achieve a network behavior consistent with experimental findings, we will build a ring attractor network. In this architecture, units are arranged along a circular topology where local excitation and global inhibition produce a localized bump of activity that can be sustained after stimulus offset, encoding a continuous variable such as spatial location. This architecture has been widely used to model working memory processes, notably in [[Wang et al., 2004](#)] where the authors also model the role of interneurons. However, they first do not consider nAChR modulation, and second use an arbitrary connectivity framework between nodes.

Here, we will implement a similar architecture but defining each node as an identical local PFC circuit described by our nAChR-modulated model. We will consider PYR to PYR connections as the driver of the excitatory connectivity, while PV interneurons will mediate the inhibitory connections between nodes, as they are both known to be able to project horizontally [[Levitt et al., 1993](#), [Druga et al., 2023](#)]. SOM and VIP interneurons will be considered as local modulators of the network dynamics as they mainly project locally [[Karnani et al., 2016](#), [Riedemann, 2019](#)].

Expected Observables and Data

We will primarily focus on analyzing the firing rates of the different neuron populations (PYR, PV, SOM, VIP) within the PFC network model. The following observables will be used to assess the impact of nAChR dysfunction and input modulation on network dynamics.

Characteristics of activity states

As previously mentioned, we first aim to get closer to experimental observations during working memory tasks [[Compte et al., 2003](#)]. We will quantify the mean firing rates and variance of both high and low-activity states along to varying input current levels. We expect that increasing input current will lead to higher firing rates in both states.

Transition dynamics

We will quantify the frequency of transitions between high and low-activity states. This should be highly modulated by the input current, and we expect that stronger inputs may lead to prolonged high-activity states, potentially resulting in a one-stability state at a certain threshold. Whereas, lower inputs may favor a more stable low-activity state.

Working memory performance metrics

In the ring attractor model, we will first assess the network's ability to form and maintain a bump of activity in response to a transient stimulus. Then, we will measure the population vectors to quantify the accuracy of working memory representations. We expect the network to be able to maintain stable activity bumps, successfully encoding the stimulus information. In the context of nAChR dysfunction, we will evaluate how impairments in specific receptor subtypes affect the stability and fidelity of these representations.

B Jacobian of the local circuit

The Jacobian referenced by the regulariser L_{jac} (Section 2.6) is the effective-gain matrix of the four-population local circuit at a fixed point. We compute it as $J_{ij} = \partial r_i^* / \partial r_j$, treating the adaptation current $I_{\text{PYR}}^{\text{adapt}}$ as fixed at its steady-state value $J_{\text{adapt,PYR}} r_{\text{PYR}}$. Population order is [PYR, SOM, PV, VIP] throughout.

Transfer-function derivative. For each population i the Wong-Wang transfer function has derivative

$$\Phi'_i(I) = \alpha_i \cdot \frac{1 - e^{-z}(1 + z)}{(1 - e^{-z})^2}, \quad z = g_i \alpha_i (I - \theta_i), \quad (15)$$

with the Taylor limit $\Phi'_i(I) \rightarrow \alpha_i/2$ as $z \rightarrow 0$. Equation 15 is evaluated at the input I_i^* produced by the fixed-point rates.

Partial derivatives of the net inputs. The chain rule gives $J_{ij} = \Phi'_i(I_i^*) \cdot \partial I_i^{\text{net}} / \partial r_j$. The partials follow directly from the input equations of Section 2.3. Writing $D \equiv 1 + g_{\text{GABA}} w_{pe} r_{\text{PV}}^*$ for the divisive denominator and $S^* = \gamma \tau_{\text{NMDA}} r_{\text{PYR}}^* / (1 + \gamma \tau_{\text{NMDA}} r_{\text{PYR}}^*)$ for the NMDA gating variable, the non-zero partials are:

$$\partial I_{\text{PYR}}^{\text{net}} / \partial r_{\text{PYR}} = \frac{J_{\text{NMDA}}}{D} \cdot \frac{\gamma \tau_{\text{NMDA}}}{(1 + \gamma \tau_{\text{NMDA}} r_{\text{PYR}}^*)^2}, \quad (16)$$

$$\partial I_{\text{PYR}}^{\text{net}} / \partial r_{\text{SOM}} = -g_{\text{GABA}} w_{se}, \quad (17)$$

$$\partial I_{\text{PYR}}^{\text{net}} / \partial r_{\text{PV}} = -\frac{J_{\text{NMDA}} S^*}{D^2} \cdot g_{\text{GABA}} w_{pe}, \quad (18)$$

$$\partial I_{\text{SOM}}^{\text{net}} / \partial r_{\text{PYR}} = w_{es}, \quad \partial I_{\text{SOM}}^{\text{net}} / \partial r_{\text{VIP}} = -w_{vs}, \quad (19)$$

$$\partial I_{\text{PV}}^{\text{net}} / \partial r_{\text{PYR}} = w_{ep}, \quad \partial I_{\text{PV}}^{\text{net}} / \partial r_{\text{SOM}} = -g_{\text{GABA}} w_{sp}, \quad (20)$$

$$\partial I_{\text{PV}}^{\text{net}} / \partial r_{\text{PV}} = -g_{\text{GABA}} w_{pp}, \quad \partial I_{\text{PV}}^{\text{net}} / \partial r_{\text{VIP}} = -w_{vp}, \quad (21)$$

$$\partial I_{\text{VIP}}^{\text{net}} / \partial r_{\text{PYR}} = w_{ev}. \quad (22)$$

The PYR row picks up two contributions on the diagonal (the NMDA saturation derivative), and on the PV column the divisive denominator contributes an extra D^{-2} factor multiplying the recurrent NMDA drive $J_{\text{NMDA}} S^*$. All other partials are direct readings of the synaptic weights.

Use in the loss. The full 4×4 matrix $J_{ij} = \Phi'_i(I_i^*) \cdot \partial I_i^{\text{net}} / \partial r_j$ is assembled and the regulariser is computed as $L_{\text{jac}} = \text{relu}(\max_{ij} |J_{ij}| - 5)^2$ (Section 2.6).

C Nullcline computation and fixed-point finding

The bistability loss L_{bistab} (Section 2.6) and the rate-matching terms L_{rate} , $L_{\text{rate,high}}$ are all evaluated on the PYR nullcline $F(r) \equiv \Phi_{\text{PYR}}(I^{\text{net}}(r)) - r$. This appendix details how F and the resulting fixed points are computed.

Reducing the 4D system to the PYR axis. The four-population local circuit defines a four-dimensional dynamics, but the nullcline condition $\dot{r}_i = 0$ for every population reduces the steady state to a one-dimensional curve parametrised by r_{PYR} . Concretely, for a given r_{PYR} we solve the remaining three populations at their own steady state:

- **VIP** has no recurrent input, so $r_{\text{VIP}}^*(r_{\text{PYR}}) = \Phi_{\text{VIP}}(w_{ev} r_{\text{PYR}} + I_{\text{VIP}}^0 + I_{\text{VIP}}^{\text{ACh}})$ is direct.

- **SOM and PV** are coupled (PV self-inhibits and SOM disinhibits PV), so we solve the 2×2 system

$$\begin{cases} r_{\text{SOM}} = \Phi_{\text{SOM}}(w_{es}r_{\text{PYR}} - w_{vs}r_{\text{VIP}}^* - J_{\text{adapt,SOM}}r_{\text{SOM}} + I_{\text{SOM}}^0 + I_{\text{SOM}}^{\text{ACh}}), \\ r_{\text{PV}} = \Phi_{\text{PV}}(w_{ep}r_{\text{PYR}} - g_{\text{GABA}}w_{pp}r_{\text{PV}} - g_{\text{GABA}}w_{sp}r_{\text{SOM}} - w_{vp}r_{\text{VIP}}^* + I_{\text{PV}}^0 + I_{\text{PV}}^{\text{ACh}}), \end{cases} \quad (23)$$

using `scipy.optimize.fsolve`. The system is non-monotonic and can admit multiple solutions in regions where PV is bistable, so we launch `fsolve` from two initial guesses $((r_{\text{SOM}}, r_{\text{PV}}) = (0, 0)$ and $(30, 30)$ Hz) and retain the solution with smaller residual.

Once $(r_{\text{SOM}}^*, r_{\text{PV}}^*, r_{\text{VIP}}^*)$ are known at this r_{PYR} , the PYR net input is assembled with the steady-state adaptation $I_{\text{PYR}}^{\text{adapt}} = J_{\text{adapt,PYR}}r_{\text{PYR}}$ and the steady-state NMDA gating $S^*(r_{\text{PYR}})$ defined in eq. 7, giving a single value $F(r_{\text{PYR}}) = \Phi_{\text{PYR}}(I^{\text{net}}) - r_{\text{PYR}}$.

Sweep. The full nullcline is built by evaluating F on a uniform sweep $r_{\text{PYR}} \in [0, 80]$ Hz with 1000 points. The upper bound is chosen above the highest physiological rate seen in the Rooy targets (~ 68 Hz) and below the rate-clamp $R_{\text{max,phys}}$ enforced during simulation, so that any zero-crossings detected on the sweep correspond to real fixed points rather than clamp artifacts.

Fixed-point detection. Fixed points are sign changes of F . Two-step location:

1. Bracketing. Index pairs $(k, k+1)$ where $\text{sign}(F[k]) \neq \text{sign}(F[k+1])$ are flagged as candidate crossings.
2. Refinement. On each bracket $[r_k, r_{k+1}]$, the crossing is refined with Brent's method (`scipy.optimize.brentq`) applied to a function that re-solves the interneuron sub-system at each Brent iterate. This guarantees that the returned r^* satisfies the full four-population steady-state equations, not just an interpolation of the sweep grid.

Stability classification. Each refined crossing is classified by the local sign of dF/dr , evaluated by linear interpolation of the discrete gradient $\Delta F/\Delta r$ at the crossing location. Crossings with $dF/dr < 0$ are stable, while crossings with $dF/dr > 0$ are unstable. A network is declared bistable when at least two stable crossings are found in $[0, R_{\text{max,phys}}]$.

Link to the loss. The detected crossings feed every rate-related loss term:

- the lowest stable crossing r_{low}^* provides the actual low FP used by L_{rate} (with its corresponding interneuron rates $r_{\text{SOM}}^*, r_{\text{PV}}^*, r_{\text{VIP}}^*$),
- the highest stable crossing r_{high}^* provides the high FP used by $L_{\text{rate,high}}$ when the network is bistable, and $L_{\text{rate,high}}$ is set to zero otherwise,
- the locations r_{low}^* and r_{high}^* also enter L_{margin} through their separation,
- the adaptive low-basin penalty in L_{bistab} uses r_{low}^* as the upper bound of the region in which F must remain positive, scaling the penalty by $1 + |r_{\text{low}}^* - r_{\text{low}}^{\text{target}}|/r_{\text{low}}^{\text{target}}$ to couple the shape constraint and the rate constraint.

In effect, the loss is a structured function on the parametrised nullcline rather than a function of the raw circuit parameters, which is why the loss landscape admits the degeneracies discussed in the Discussion.

D Random parameter search

To assess whether the SOM-silenced solution of Section 3.2 reflects an idiosyncrasy of the optimiser or a generic feature of bistable parameter sets under our equations, we ran an independent random search over the same bounds as the differential-evolution fit (Table 1).

Protocol. Each sample drew every free parameter independently from its log-uniform or uniform prior. The local circuit was integrated deterministically over $T = 2500$ ms with $\Delta t = 0.1$ ms and a burn-in of 1800 ms before the bistability check, and rates were averaged over a 500 ms window at the end. The check used the same nullcline-based fixed-point procedure as the optimisation loss (Appendix C), with no rate target imposed: only the topological $+, -, +, -$ pattern and a minimum FP separation $\Delta r_{\min} = 15$ Hz were required.

Hit rate. Out of 10^5 random samples, 778 ($\approx 0.78\%$) passed the bistability test. The corresponding raw samples are stored in `figs/optim/random_bistable_100k.jsonl` and summarised in `figs/optim/random_bistable_summary_100k.txt`. The hit rate gives an order-of-magnitude estimate of how dense bistable parameter sets are within our bounds, and confirms that bistability is a small but accessible region of the parameter space without any loss-driven steering.

Realistic-bistability subset. Most of the 778 hits pass the topological criterion but at near-trivial separations: the median PYR fixed-point gap is only 1.95 Hz, and many “high” fixed points sit well below physiological active-state rates. We therefore restrict the subsequent analysis to hits with both a non-trivial separation ($r_{\text{PYR}}^{\text{high}} - r_{\text{PYR}}^{\text{low}} \geq 15$ Hz) and a high state that is actually high ($r_{\text{PYR}}^{\text{high}} > 30$ Hz). This leaves $n = 104$ realistic-bistability samples. The numbers below refer to this subset.

SOM silencing is generic. The high-state SOM rate across the 104 samples has median 1.01 Hz and mean 8.89 Hz, against a Rooy target of 35.2 Hz. The distribution is strongly bimodal: 50% of hits sit below 1 Hz (effectively silenced), 61.5% below 5 Hz, and only 19.2% above 20 Hz. The differential-evolution optimiser converges to $r_{\text{SOM}}^{\text{high}} = 0$, which sits squarely in the most populated mode of the random-search distribution. It is not an outlier but a representative draw from the silenced half of the bistable manifold.

Joint SOM + PV activity is not reachable. The more decisive observation concerns the joint activity of both perisomatic and dendritic interneurons at the high state. In the entire $n = 104$ realistic-bistability subset, no sample reaches $r_{\text{SOM}}^{\text{high}} > 20$ Hz and $r_{\text{PV}}^{\text{high}} > 20$ Hz simultaneously, only 6 samples (5.8%) clear the much weaker threshold of both rates exceeding 5 Hz, and only 2 samples (1.9%) clear 10 Hz on both. The closest random hit to the Rooy targets has $r_{\text{SOM}} = 14.9$ Hz, $r_{\text{PV}} = 20.8$ Hz, $r_{\text{VIP}} = 83.8$ Hz, still well below the active-state interneuron rates measured in vivo.

Interpretation. The implication is that the equations themselves, given the bounds we use, structurally prevent a bistable regime in which both SOM and PV stay biologically active at the high state. SOM silencing at the optimum is therefore not a quirk of the differential-evolution path: it is the dominant mode the model offers, and the optimiser settles into it. This is consistent with the structural reading developed in the Discussion (missing inter-layer afferents to VIP, local-source/local-target geometry of the single-node fit) and weakens the alternative reading in which the silencing would simply reflect optimiser bias. Resolving it requires changing the equations or the constraints, not the optimiser.

E Fitted parameters at the bistable optimum

Table 3 reports the free-parameter values obtained at the bistable optimum used throughout the Results section. None of the fitted values is pinned at either boundary, confirming that the optimiser is not constrained by the bounds. Fixed parameters that were not optimised (Wong-Wang shape parameters α_i, θ_i, g_i , time constants $\tau_{\text{NMDA}}, \gamma, \tau_{\text{adapt,PYR}}$, and ring widths) are reported in the relevant Methods sections and omitted here.

Parameter	Lower bound	Fitted value	Upper bound
<i>Adaptation strength (nA/Hz)</i>			
$J_{\text{adapt,PYR}}$	0.001	0.00139	0.2
$J_{\text{adapt,SOM}}$	0.0	0.000	0.0
<i>GABA gain (dimensionless)</i>			
$g_{\text{GABA,base}}$	0.1	4.21	5.0
$g_{\alpha 7}$	0.1	1.09	5.0
<i>Synaptic weights (nA/Hz, except J_{NMDA} in nA)</i>			
J_{NMDA}	0.05	0.267	2.0
w_{ep}	0.001	0.00720	0.05
w_{es}	0.001	0.00601	0.05
w_{ev}	0.001	0.00626	0.05
w_{pe}	0.001	0.00131	0.05
w_{pp}	0.001	0.00120	0.05
w_{se}	0.001	0.0460	0.05
w_{sp}	0.001	0.0157	0.05
w_{vp}	0.001	0.00204	0.05
w_{vs}	0.001	0.0180	0.05
<i>External drives (nA)</i>			
I_{PYR}^0	0.01	0.413	1.5
I_{PV}^0	0.01	0.275	0.6
I_{SOM}^0	0.01	0.182	0.6
I_{VIP}^0	0.01	0.206	0.6
<i>nAChR currents (nA)</i>			
$I_{\alpha 7, \text{PV}}$	0.01	0.0683	0.5
$I_{\alpha 7, \text{SOM}}$	0.01	0.0496	0.5
$I_{\beta 2, \text{SOM}}$	0.01	0.0226	0.5
$I_{\alpha 5, \text{VIP}}$	0.01	0.0217	0.5

Table 3: Fitted parameter values at the bistable optimum.

F Full local-circuit equations

The full system of equations governing the four-population local circuit, with the NMDA gating variable on the PYR→PYR recurrence and the hyperbolic soft ceiling on interneurons, reads:

$$\tau_{\text{PYR}} \dot{r}_{\text{PYR}}(t) = -r_{\text{PYR}}(t) + \Phi_{\text{PYR}}\left(I_{\text{PYR}}^{\text{net}}(t)\right), \quad (24)$$

$$\tau_{\text{SOM}} \dot{r}_{\text{SOM}}(t) = -r_{\text{SOM}}(t) + \Phi_{\text{capped}}^{\text{SOM}}\left(I_{\text{SOM}}^{\text{net}}(t)\right), \quad (25)$$

$$\tau_{\text{PV}} \dot{r}_{\text{PV}}(t) = -r_{\text{PV}}(t) + \Phi_{\text{capped}}^{\text{PV}}\left(I_{\text{PV}}^{\text{net}}(t)\right), \quad (26)$$

$$\tau_{\text{VIP}} \dot{r}_{\text{VIP}}(t) = -r_{\text{VIP}}(t) + \Phi_{\text{capped}}^{\text{VIP}}\left(I_{\text{VIP}}^{\text{net}}(t)\right), \quad (27)$$

with the slow auxiliary variables

$$\tau_{\text{NMDA}} \dot{S}(t) = -S(t) + \gamma[1 - S(t)] r_{\text{PYR}}(t), \quad (28)$$

$$\tau_{\text{adapt}} \dot{I}_{\text{PYR}}^{\text{adapt}}(t) = -I_{\text{PYR}}^{\text{adapt}}(t) - J_{\text{adapt}} r_{\text{PYR}}(t). \quad (29)$$

The net synaptic currents are

$$I_{\text{PYR}}^{\text{net}}(t) = \frac{J_{\text{NMDA}} S(t) - w_{se} r_{\text{SOM}}(t) + I_{\text{PYR}}^0 + I_{\text{PYR}}^{\text{adapt}}(t)}{1 + g_{\text{GABA}} w_{pe} r_{\text{PV}}(t)}, \quad (30)$$

$$I_{\text{SOM}}^{\text{net}}(t) = w_{es} r_{\text{PYR}}(t) - w_{vs} r_{\text{VIP}}(t) + I_{\text{SOM}}^0 + I_{\text{SOM}}^{\text{ACh}}, \quad (31)$$

$$I_{\text{PV}}^{\text{net}}(t) = w_{ep} r_{\text{PYR}}(t) - g_{\text{GABA}} w_{pp} r_{\text{PV}}(t) - g_{\text{GABA}} w_{sp} r_{\text{SOM}}(t) - w_{vp} r_{\text{VIP}}(t) + I_{\text{PV}}^0 + I_{\text{PV}}^{\text{ACh}}, \quad (32)$$

$$I_{\text{VIP}}^{\text{net}}(t) = w_{ev} r_{\text{PYR}}(t) + I_{\text{VIP}}^0 + I_{\text{VIP}}^{\text{ACh}}. \quad (33)$$

The Wong-Wang transfer function $\Phi_i(I) = u_i / (1 - \exp(-g_i u_i))$ with $u_i = \alpha_i(I - \theta_i)$ is used for every population, and the interneuron output is post-processed by the soft ceiling $\Phi_{\text{capped}}^X(I) = r_{\text{max}}^X \Phi^X(I) / (r_{\text{max}}^X + \Phi^X(I))$, defined in eq. 8. The total nAChR drive I_i^{ACh} is the sum of the subunit-specific currents contributing to population i ($\alpha 7$ for PV and SOM, $\beta 2$ for SOM, $\alpha 5$ for VIP).

The ring extension replaces each input by its spatially weighted counterpart following the kernels of Section 2.4: $w_{ee} r_{\text{PYR}}$ is upgraded to $\sum_{\ell} W_{k\ell}^{\text{PYR}} S_{\ell}$, the divisive PV term to $\sum_{\ell} W_{k\ell}^{\text{PV}} r_{\ell}^{\text{PV}}$, and the SOM term to $\sum_{\ell} W_{k\ell}^{\text{SOM}} r_{\ell}^{\text{SOM}}$, while all other currents remain local.

G Koukouli desensitisation data

The deactivation percentages applied as multiplicative scaling on the nAChR-drive parameters when simulating the APP condition (Section 2.5) are reproduced from Koukouli et al. (2025) in Table 4.

Nicotinic receptor	Deactivation / inactivation level
$\alpha 7$	$-90 \pm 10\%$ reduced activation
$\alpha 5$	$40 \pm 5\%$ reduced activation
$\beta 2$	0–25% inactivation

Table 4: nAChR deactivation levels in the APP mouse model relative to wild type, taken from Koukouli et al. (2025).

H SOM connectivity diagram

The SOM lateral-inhibition kernel used in Section 2.4 is illustrated in Figure 11. SOM cells receive input from the local pyramidal population and project their output to the neighbouring columns, with a Gaussian fall-off and a zero diagonal (no self-inhibition). The annular SOM kernel $W_{k\ell}^{\text{SOM}}$ therefore vanishes at $d = 0$ and shows two symmetric lateral lobes peaking at $\pm\mu_{\text{SOM}}$. The PYR→PYR Gaussian kernel $W_{k\ell}^{\text{PYR}}$ is overlaid for reference and peaks at $d = 0$, exactly where the SOM inhibition vanishes.

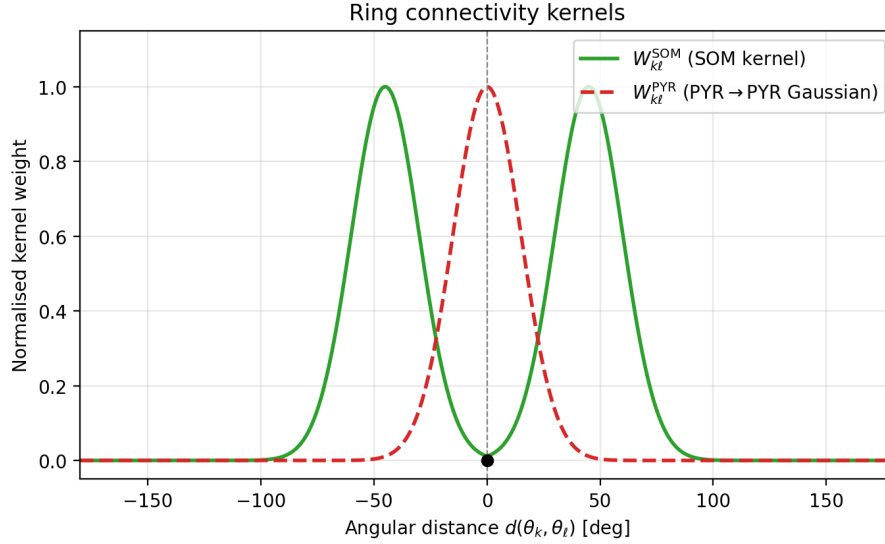


Figure 11: Ring connectivity kernels plotted against the signed angular distance: the annular SOM kernel $W_{k\ell}^{\text{SOM}}$ (solid green, eq. 12) and the PYR→PYR Gaussian kernel $W_{k\ell}^{\text{PYR}}$ (dashed red).

I Single-node simulation at zero noise

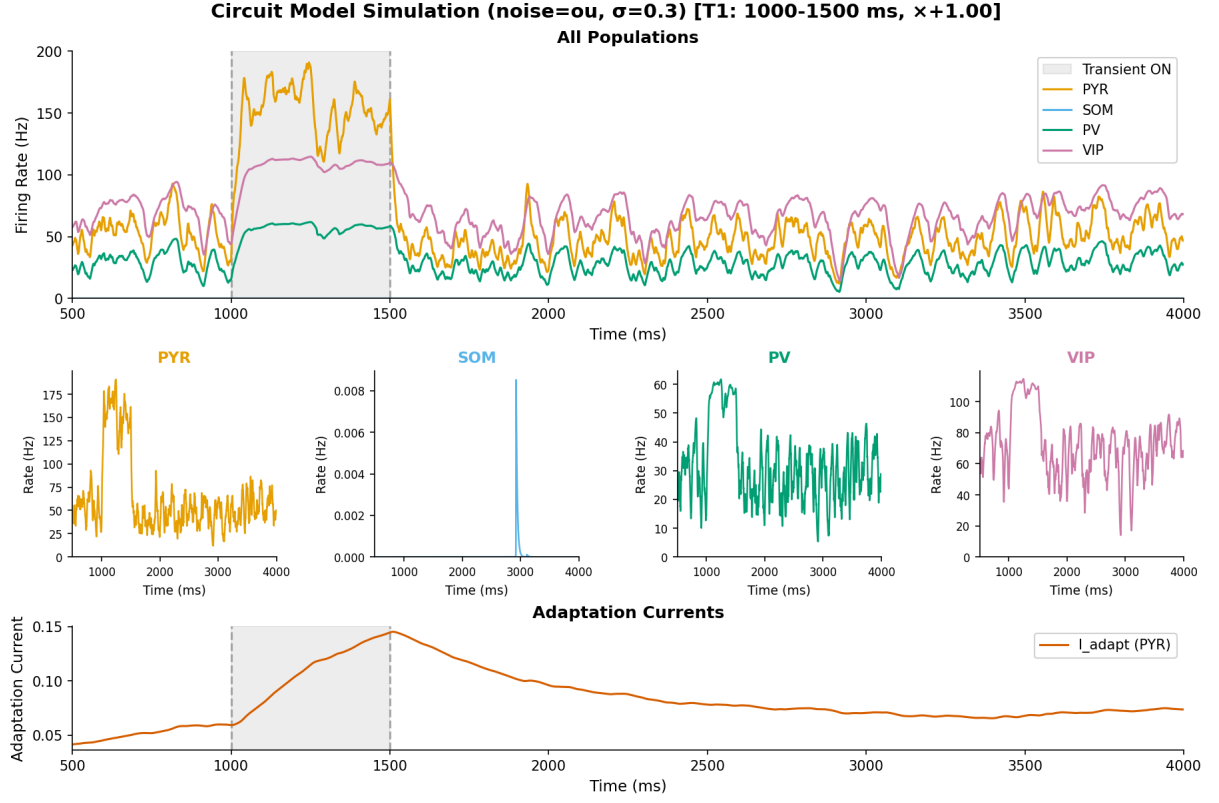


Figure 12: Single-node simulation at $\sigma = 0$.

The example referenced in Section 3.4 illustrates the saturation regime obtained when the noise amplitude is set to zero: the deterministic dynamics drift to the high fixed point without any need for a cue, demonstrating that a finite noise level is required to keep the network in the low state.

J Ring-attractor 2D state-classification heatmap (no-SOM control)

State-classification heatmap for the ring simulation without lateral SOM inhibition, for comparison with the single-node sweep (Figure 6b) and the SOM-Gaussian heatmap (Figure 8).

Heatmap of network state for varying noise and amplitude (ou noise)

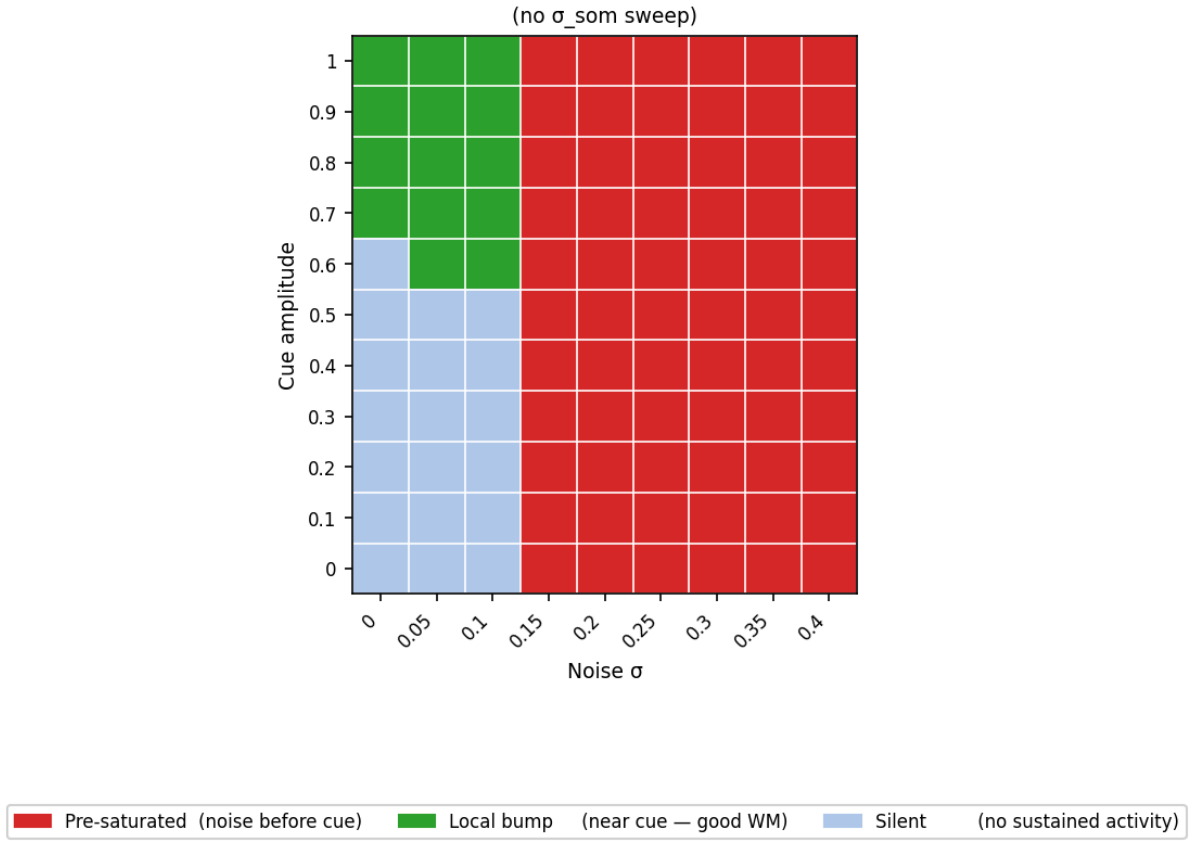


Figure 13: State-classification heatmap for the no-SOM ring under OU noise.

K SOM-Gaussian ring: saturated dashboard example

Representative dashboard from the SOM-Gaussian sweep illustrating the fully saturated regime referenced in Section 3.5: the bump fails to localise, drifts, and spreads across the ring until every column is in the high state.

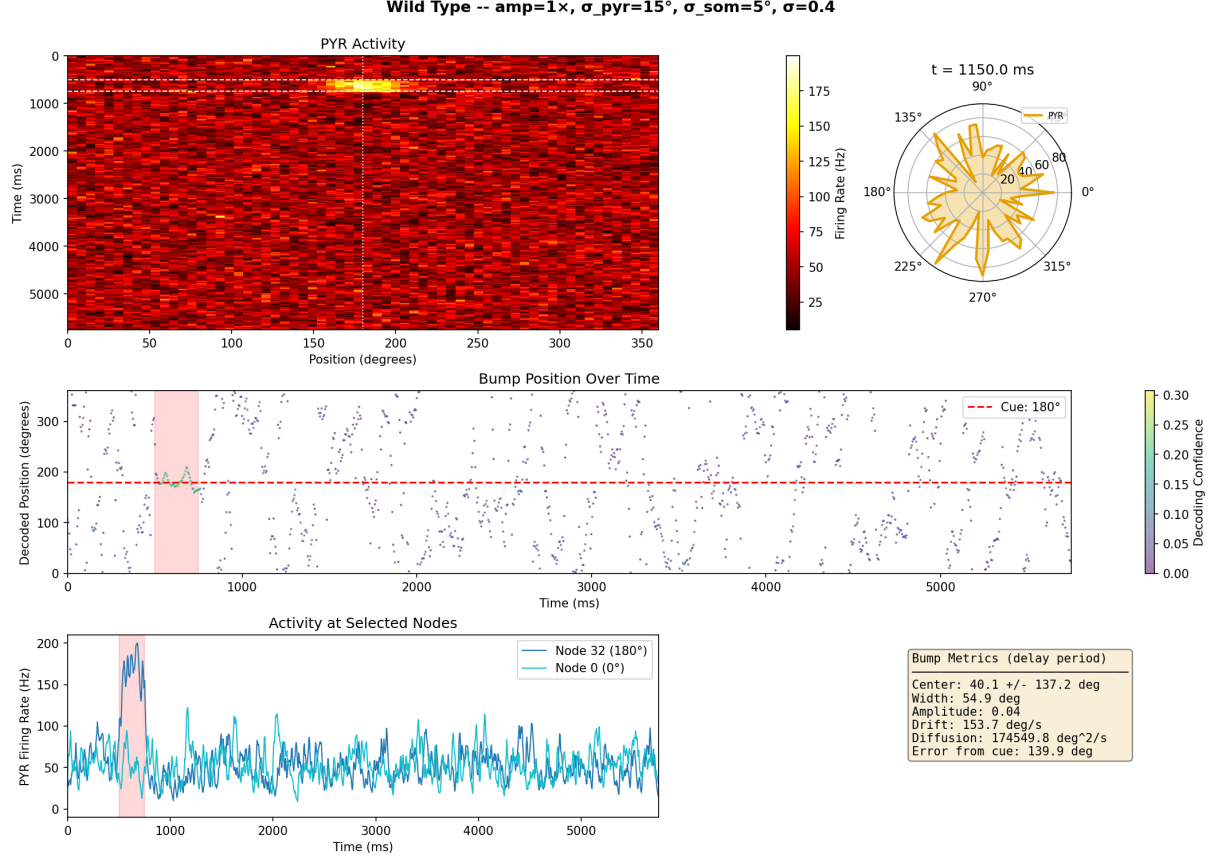


Figure 14: Example run from the SOM-Gaussian sweep ($\sigma = 0.4$, cue amplitude 1.0). The bump fails to localise, drifts and progressively spreads across the ring, eventually leading to a fully saturated network.