

¹ **Goldstone Modes and the Coexistence Saddle: Spectral Bifurcation
2 Analysis of Competing Ring Attractors Under Mean-Field Cross-
3 Inhibition**

⁴ **Tomás P. Pavan¹ and Claude²**

⁵ ¹Independent Researcher ²Anthropic (Claude Opus 4.6, Large Language Model)

⁶ *Correspondence:* Tomás P. Pavan

⁷

⁸ **Abstract**

⁹ Persistent neural activity in working memory is often modeled as bump attractors on ring networks. When
¹⁰ two such networks compete via cross-inhibition, the system must resolve which representation dominates – a
¹¹ winner-take-all (WTA) decision. We present a complete spectral bifurcation analysis of this transition in a
¹² coupled ring attractor model (two networks of $N = 48$ neurons each, cosine connectivity within, mean-field
¹³ cross-inhibition between). We identify six key results.

¹⁴ First, the coexistence fixed point – where both bumps self-sustain simultaneously – exists only below a
¹⁵ critical cross-inhibition strength $J_x^{\text{exist}} \approx 0.36$. Above this threshold, cross-inhibition is too strong for both
¹⁶ representations to survive; the system admits only WTA solutions.

¹⁷ Second, the continuous rotational symmetry of each bump guarantees exactly two Goldstone modes (zero
¹⁸ eigenvalues) that are protected against mean-field coupling. The first non-Goldstone eigenvalue – governing
¹⁹ uniform amplitude competition – crosses zero at $J_x^* \approx 0.3485$ via a pitchfork bifurcation, creating the
²⁰ coexistence saddle and two WTA stable states.

²¹ Third, the critical eigenvector projects maximally onto the spatially uniform (DC) direction, meaning the
²² instability concerns total activity competition rather than spatial pattern rearrangement. This is a direct
²³ consequence of mean-field (spatially unstructured) cross-inhibition.

²⁴ Fourth, large-scale stochastic simulations (128,000 trials across 256 parameter combinations) confirm the
²⁵ spectral predictions: swap errors emerge at the predicted J_x threshold, drive strength is secondary to
²⁶ cross-inhibition, and a non-monotonic valley at intermediate $J_x \approx 1.2\text{--}1.6$ – well above the coexistence
²⁷ threshold, in the pure winner-take-all regime – identifies a functional operating regime for working memory.

²⁸ Fifth, connectivity heterogeneity destroys the sharp pitchfork entirely, converting it into an imperfect
²⁹ bifurcation with no zero-crossing. The razor-thin instability window ($\Delta J_x \approx 0.01$) is a symmetry artifact of
³⁰ the clean model; biological circuits operate in a regime of smooth crossover where no parameter precision is
³¹ required.

³² Sixth, a Kramers escape consistency analysis links the deterministic bifurcation and stochastic swap onset:
³³ because the barrier collapses quadratically ($\Delta V \propto |\lambda_{\text{dom}}|^2$), noise-driven escape becomes likely at $J_x^{\text{onset}} \approx 0.25$
³⁴ – well below the pitchfork – for a normal-form coefficient $\gamma \approx 0.22\text{--}0.36$. We discuss implications for the
³⁵ behavioral cliff and argue that neural circuits operate in a valley regime where cross-inhibition and encoding
³⁶ drive are balanced, rather than near J_x^* itself.

³⁷

³⁸ **1. Introduction**

³⁹ **1.1 Working Memory and Competing Representations**

⁴⁰ Persistent neural activity in prefrontal and parietal cortex underlies the short-term maintenance of information
⁴¹ in working memory (Goldman-Rakic, 1995; Funahashi et al., 1989). Ring attractor models capture a key
⁴² feature of this activity: spatially tuned neurons form a localized “bump” of elevated firing that persists
⁴³ through recurrent excitation even after the sensory stimulus is removed (Compte et al., 2000; Ben-Yishai et
⁴⁴ al., 1995; Amari, 1977). These bumps encode continuous variables such as spatial location or orientation,

45 and their precision is set by a balance between recurrent drive and noise-induced diffusion along the ring
46 (Wimmer et al., 2014; Burak and Fiete, 2012). Although continuous attractors are structurally unstable –
47 destroyed by most infinitesimal perturbations of the dynamics – Ságodi et al. (2024) showed that a persistent
48 slow manifold survives these bifurcations, rendering the attractor functionally robust for analog memory.
49 Such ring attractor dynamics have been observed experimentally in the *Drosophila* head direction system
50 (Kim et al., 2017) and are reviewed in the broader context of attractor and integrator networks by Khona
51 and Fiete (2022).

52 When multiple items must be stored simultaneously, as in multi-item visual working memory tasks, the
53 standard approach posits multiple bump networks coupled through cross-inhibition (Edin et al., 2009; Wei
54 et al., 2012). The cross-inhibition creates a competition: if it is weak, both bumps coexist and the system
55 maintains multiple items; if it is strong, one bump suppresses the other in a winner-take-all (WTA) decision.
56 The transition between these regimes determines the capacity limit of the working memory circuit (Edin et
57 al., 2009).

58 Despite the importance of this transition, its spectral structure – the full set of eigenvalues and eigenvectors
59 of the system’s Jacobian – has not been characterized. Previous analyses have focused on one-dimensional
60 (1D) mean-field reductions, projecting the high-dimensional dynamics onto a single dominance variable
61 $D = \bar{r}_A - \bar{r}_B$ and characterizing the resulting cusp catastrophe (Thom, 1972; Zeeman, 1977). While this
62 captures the topology of the bifurcation, it discards the 96-dimensional dynamics that include rotational
63 modes, drift modes, and the full stability structure of the coexistence state.

64 1.2 The Behavioral Cliff

65 Psychophysical experiments reveal a striking feature of working memory performance: below a critical
66 stimulus strength, accuracy does not degrade gradually but collapses abruptly – a “behavioral cliff” (Bays et
67 al., 2009; Zhang and Luck, 2008). In the mixture model framework, this manifests as a sharp increase in the
68 probability of reporting a non-target item (swap errors) or of random guessing, even for small changes in
69 signal-to-noise ratio.

70 The standard theoretical account attributes the cliff to noise-driven escape from a metastable state (Kramers,
71 1940; Hanggi et al., 1990): when the cue is weak, the barrier between the correct attractor and competing
72 attractors is low, and stochastic fluctuations cause the system to fall to a wrong state. This yields a cusp
73 catastrophe potential $V(D) = D^4 + aD^2 + bD$, where a is controlled by the circuit’s lateral inhibition and b
74 by the cue strength. The cliff occurs at the cusp point where the barrier vanishes.

75 However, this account treats the cliff as a cue phenomenon – a consequence of weak sensory input. An
76 alternative possibility, which we develop here, is that the cliff reflects a structural property of the circuit:
77 the proximity of the effective cross-inhibition strength J_x to a critical value J_x^* where the coexistence state
78 undergoes a spectral bifurcation.

79 1.3 From Mean-Field Reduction to Full Spectral Analysis

80 The 1D reduction $D = \bar{r}_A - \bar{r}_B$ captures the order parameter of the WTA transition but suppresses 95 of the
81 96 dynamical degrees of freedom. In particular, it cannot distinguish:

82 1. **Goldstone modes** – exactly-zero eigenvalues arising from the continuous rotational symmetry of each
83 bump (Goldstone, 1961; Burak and Fiete, 2012). These modes govern bump drift and are protected by
84 symmetry.

85 2. **Genuine instabilities** – eigenvalues that cross zero as parameters change, signaling structural
86 reorganization of the attractor landscape.

87 3. **The character of the critical mode** – whether the instability that destroys coexistence projects
88 onto the spatially uniform (DC) direction, is spatially patterned (cosine), or mixed.

89 Previous spectral approaches to ring networks have addressed non-Hermitian quasi-localization (Tanaka and
90 Nelson, 2018) and the stability of persistent activity under short-term plasticity (Seeholzer et al., 2019), but
91 the full eigenvalue structure of the *coupled* system has not been resolved. We present the first complete

92 eigenvalue decomposition of the coupled ring attractor Jacobian, resolving all $2N = 96$ eigenvalues as a
 93 function of the cross-inhibition strength J_{\times} . By cleanly separating Goldstone modes from genuine instabilities,
 94 we identify the precise location, character, and consequences of the coexistence-to-WTA pitchfork bifurcation.

95 1.4 Summary of Contributions

96 Our main results are:

- 97 1. **Existence threshold.** The coexistence fixed point exists only for $J_{\times} < J_{\times}^{\text{exist}} \approx 0.36$. At the commonly
 98 used value $J_{\times} = 0.5$, coexistence is not a fixed point of the dynamics – both bumps cannot self-sustain
 99 under such strong cross-inhibition.
- 100 2. **Goldstone separation and pitchfork.** Two Goldstone modes (exactly-zero eigenvalues protected by
 101 rotational symmetry) persist at all J_{\times} where coexistence exists. The first non-Goldstone eigenvalue
 102 λ_{dom} crosses zero at $J_{\times}^* \approx 0.3485$, creating a pitchfork bifurcation where the symmetric coexistence
 103 state becomes a saddle point and two WTA attractors are born.
- 104 3. **DC critical mode.** The critical eigenvector has its largest projection onto the uniform (DC) direction:
 105 an increase in network A’s activity coupled with a decrease in network B’s, localized to the active bump
 106 neurons by the gain mask $\sigma'(h_i)$. This reflects the mean-field character of the cross-inhibition and
 107 means the instability is about total activity competition, not spatial pattern rearrangement.
- 108 4. **Stochastic phase diagram.** A 128,000-trial parameter sweep confirms the spectral predictions and
 109 reveals a non-monotonic valley at intermediate J_{\times} where swap error rates dip to 7–13% between two
 110 qualitatively different failure modes.
- 111 5. **Heterogeneity destroys the sharp bifurcation.** Connectivity heterogeneity breaks the exact $A \leftrightarrow B$
 112 exchange symmetry, converting the pitchfork into an imperfect bifurcation with no zero-crossing. The
 113 razor-thin instability window ($\Delta J_{\times} \approx 0.01$) is a symmetry artifact; biological circuits operate in a
 114 smooth crossover regime.
- 115 These results reframe the behavioral cliff as a J_{\times} -space phenomenon and identify a qualitative operating
 116 regime (the valley) where encoding drive and cross-inhibition are balanced for reliable working memory.

117

118 2. Model

119 2.1 Single Ring Attractor

120 We consider a rate model with $N = 48$ neurons uniformly distributed on a ring. Each neuron i has a preferred
 121 angle $\theta_i = -\pi + 2\pi i/N$ and firing rate $r_i(t)$ governed by:

$$\tau \frac{dr_i}{dt} = -r_i + \sigma(h_i)$$

122 where $\tau = 10$ ms is the time constant and $\sigma(h) = r_{\max}/(1 + e^{-\beta(h-h_0)})$ is a sigmoidal activation function
 123 with parameters $r_{\max} = 1.0$, $\beta = 5.0$, $h_0 = 0.5$. The total input to neuron i is:

$$h_i = \sum_{j=1}^N W_{ij} r_j + I_i^{\text{ext}}$$

124 where the within-network connectivity has cosine tuning:

$$W_{ij} = \frac{1}{N} (-J_0 + J_1 \cos(\theta_i - \theta_j))$$

125 with $J_0 = 1.0$ (uniform inhibition) and $J_1 = 6.0$ (tuned excitation). This connectivity supports a family of
 126 bump solutions at any angular position, forming a ring attractor (Amari, 1977; Ben-Yishai et al., 1995).

¹²⁷ **2.2 Coupled System with Mean-Field Cross-Inhibition**

¹²⁸ We couple two identical ring networks A and B through mean-field cross-inhibition. The dynamics become:

$$\begin{aligned}\tau \frac{dr_i^A}{dt} &= -r_i^A + \sigma \left(\sum_j W_{ij} r_j^A + I_i^{cue} - J_\times \bar{r}^B \right) \\ \tau \frac{dr_i^B}{dt} &= -r_i^B + \sigma \left(\sum_j W_{ij} r_j^B - J_\times \bar{r}^A \right)\end{aligned}$$

¹²⁹ where $\bar{r}^X = \frac{1}{N} \sum_j r_j^X$ is the mean activity of network X and $J_\times \geq 0$ is the cross-inhibition strength. The
¹³⁰ external cue input is a von Mises tuning curve applied to network A only:

$$I_i^{cue} = c \cdot \frac{e^{\kappa \cos(\theta_i - \theta_{stim})}}{I_0(\kappa)}$$

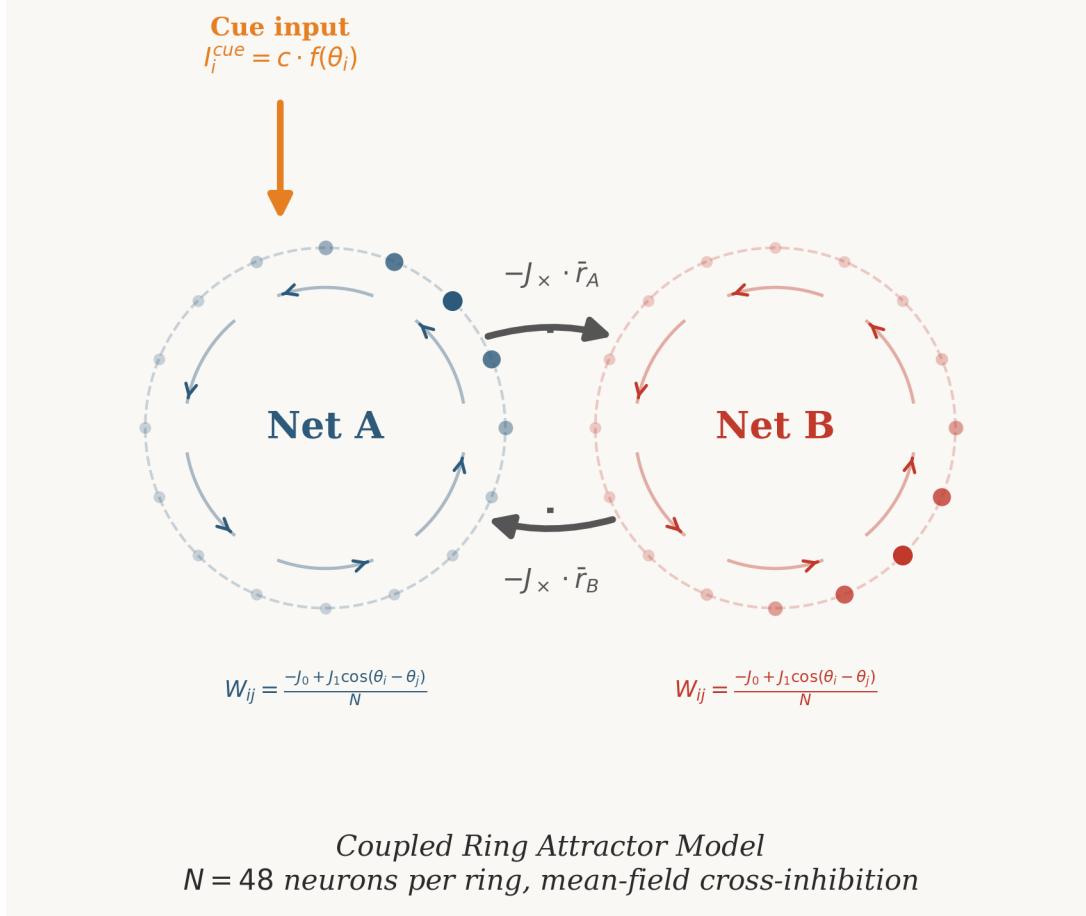
¹³¹ with concentration parameter $\kappa = 2.0$ and cue gain $c \geq 0$.

¹³² Throughout this paper, J_\times denotes this single cross-inhibition parameter; we distinguish three characteristic
¹³³ values: the pitchfork bifurcation $J_\times^* \approx 0.3485$, the coexistence existence boundary $J_\times^{\text{exist}} \approx 0.358$, and the
¹³⁴ stochastic onset $J_\times^{\text{onset}} \approx 0.25$.

¹³⁵ These three thresholds have distinct mathematical status: J_\times^* is a spectral crossing (eigenvalue sign change,
¹³⁶ deterministic); J_\times^{exist} is an existence boundary (fold, deterministic, found numerically but not characterized
¹³⁷ spectrally); J_\times^{onset} is a stochastic operational threshold (finite-time escape probability exceeding a criterion in
¹³⁸ noisy simulations). The first is exact for the symmetric model; the second depends on solver convergence
¹³⁹ criteria; the third depends on noise level, trial duration, and the Kramers consistency argument of Section
¹⁴⁰ 3.5.5.

¹⁴¹ We refer to the cue gain c as the *encoding drive* (or simply *drive*) when discussing its role in the stochastic
¹⁴² simulations (Section 3.5), where it controls the strength of the stimulus that biases initial encoding toward
¹⁴³ network A. This is distinct from the initialization currents ($I_{\text{seed}} = 5.0$) used in the fixed-point solver (Section
¹⁴⁴ 3.1.1), which serve only to locate the bump-attractor state and are not varied as a parameter.

¹⁴⁵ The critical feature of mean-field cross-inhibition is that it depends only on the total activity \bar{r}^X of the
¹⁴⁶ opposing network, not on the spatial pattern of its bump. This has profound consequences for the symmetry
¹⁴⁷ structure of the system (Fig. 1).



148

149 **Figure 1.** Model schematic. Two ring networks (A, B) of $N = 48$ neurons each, with cosine within-network
150 connectivity ($J_0 + J_1 \cos \Delta\theta$) and mean-field cross-inhibition ($J_x \bar{r}^X$). External cue input drives network A
151 only. The cross-inhibition depends on mean activity, not bump position, preserving rotational symmetry.

152 2.3 Jacobian of the Coupled System

153 The steady-state condition $F(\mathbf{r}^*) = 0$ defines the fixed points, where $F_i^A = -r_i^A + \sigma(h_i^A)$ and similarly for B.
154 The Jacobian $\mathbf{J} = \partial F / \partial \mathbf{r}$ evaluated at a fixed point \mathbf{r}^* has a 2×2 block structure:

$$\mathbf{J} = \begin{pmatrix} -\mathbf{I} + \mathbf{S}_A \mathbf{W} & \mathbf{S}_A \mathbf{C} \\ \mathbf{S}_B \mathbf{C} & -\mathbf{I} + \mathbf{S}_B \mathbf{W} \end{pmatrix}$$

155 where $\mathbf{S}_X = \text{diag}(\sigma'(h_i^X))$ is the diagonal matrix of sigmoid derivatives at the fixed point, and $\mathbf{C} = -\frac{J_x}{N} \mathbf{1} \mathbf{1}^T$
156 is the rank-1 mean-field coupling matrix. The full Jacobian is $2N \times 2N = 96 \times 96$.

157 The block structure reveals that the cross-coupling enters only through the rank-1 matrix \mathbf{C} . This low-rank
158 perturbation to the block-diagonal within-network dynamics is what makes the spectral analysis tractable:
159 the cross-inhibition can shift at most one eigenvalue per symmetry sector.

160 2.4 Symmetries

161 The coupled system possesses two symmetries at zero cue ($c = 0$):

162 **Continuous rotational symmetry.** The mean-field coupling $J_x \bar{r}^X$ is invariant under any rotation of
163 the bump profile: if r_i^X is a fixed point, so is r_{i+k}^X for any shift k . This gives a continuous family of fixed
164 points parametrized by bump position, and by Goldstone's theorem (Goldstone, 1961), each such continuous

165 symmetry produces an eigenvalue that is exactly zero. With two independent bumps, there are two Goldstone
 166 modes.

167 **Discrete exchange symmetry.** At $c = 0$, the system is invariant under $A \leftrightarrow B$. The coexistence fixed
 168 point (where both bumps are present with $\bar{r}^A = \bar{r}^B$) respects this symmetry; the WTA states ($\bar{r}^A \gg \bar{r}^B$ or
 169 vice versa) break it. The transition between these is governed by a pitchfork bifurcation.

170 The nonzero cue $c > 0$ breaks the exchange symmetry (favoring network A) and deforms the pitchfork into
 171 an imperfect bifurcation with hysteresis.

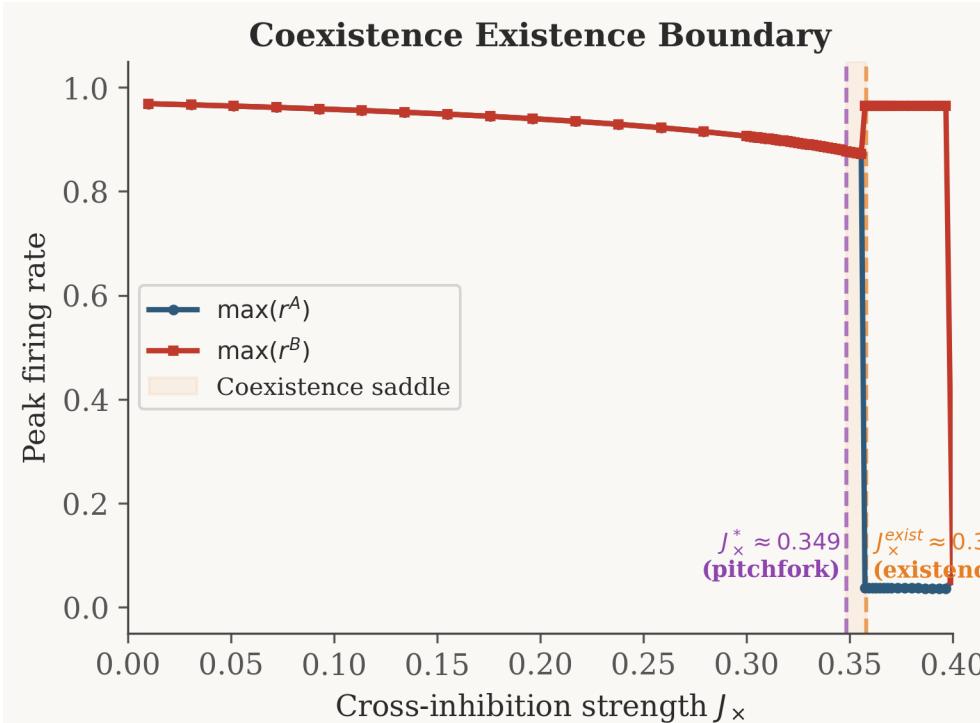
172

173 3. Results

174 3.1 Existence of the Coexistence Fixed Point

175 **3.1.1 Numerical Method** We locate fixed points using a two-phase approach. In Phase 1 (simulation), we
 176 establish both bumps using strong initialization currents ($I_{\text{seed}} = 5.0$), then remove the currents and simulate
 177 the coupled system for 5×10^4 time steps ($\Delta t = 0.1, \tau = 10$) under the target cross-inhibition J_x . In Phase 2
 178 (Newton), we polish the resulting state using Newton's method (scipy.optimize.fsolve) with the analytical
 179 Jacobian, achieving residuals $\|F(\mathbf{r}^*)\| < 10^{-10}$.

180 We verify convergence by checking: (i) the residual $\max_i |F_i(\mathbf{r}^*)| < 10^{-6}$; (ii) both bumps are active
 181 ($\max_i r_i^A > 0.3$ and $\max_i r_i^B > 0.3$); and (iii) the solution is not a WTA state ($|D| = |\bar{r}^A - \bar{r}^B| < 0.1$).



182 3.1.2 Critical Existence Threshold

183 **Figure 2.** Coexistence existence boundary. Peak firing rate of each network as a function of cross-inhibition
 184 strength J_x . Below $J_x^{\text{exist}} \approx 0.358$, both networks sustain bumps at matched amplitudes (coexistence). Above
 185 this threshold, one network collapses to baseline (WTA). The pitchfork bifurcation at $J_x^* \approx 0.349$ (orange
 186 dashed) and existence boundary at $J_x^{\text{exist}} \approx 0.358$ (purple dashed) delimit the narrow instability window
 187 $\Delta J_x \approx 0.01$.

188 We scan J_x from 0 to 0.5 (Fig. 2). Below $J_x^{\text{exist}} \approx 0.36$, Newton converges to a genuine coexistence fixed point
 189 with residual $< 10^{-10}$. Above this threshold, one bump suppresses the other during the simulation phase;

190 Newton converges only to WTA solutions. At $J_X = 0.50$ – a commonly used parameter value – coexistence
191 does not exist as a fixed point of the deterministic dynamics.

192 The critical existence threshold lies between $J_X = 0.355$ (both bumps survive, $\max r^A = 0.88, \max r^B = 0.88$)
193 and $J_X = 0.360$ (one bump collapses, $\max r^A = 0.97, \max r^B = 0.04$). The transition is sharp: a 1.4% increase
194 in J_X converts stable coexistence into complete dominance.

195 **3.1.3 Diagnostic: Fixed Point vs. Slow Manifold** At $J_X = 0.50$, a time-resolved diagnostic reveals that
196 the system does not converge: residuals remain at $\sim 10^{-3}$ and the dominance variable D drifts monotonically
197 toward ± 0.27 . At $J_X = 0.35$, residuals converge exponentially to machine precision ($\sim 10^{-15}$). This confirms
198 that the coexistence state is a genuine fixed point below threshold and does not exist (even as a slow manifold)
199 above it.

200 **3.2 Goldstone Modes and the Protected Symmetry**

201 **3.2.1 Origin of the Goldstone Modes** The mean-field cross-coupling $J_X \bar{r}^B$ is a function of the mean
202 activity $\bar{r}^B = \frac{1}{N} \sum_j r_j^B$ only. Any continuous rotation of the bump profile preserves this mean. We now prove
203 that this protects the rotational modes as exact null vectors of the full coupled Jacobian.

204 **Step 1: Rotational null vector of the uncoupled block.** Let $\mathbf{r}^{A*}(\varphi)$ denote the bump solution of
205 network A centered at phase φ . Because W_{ij} depends only on the angular difference $\theta_i - \theta_j$, the steady-state
206 equation $-r_i^{A*} + \sigma(\sum_j W_{ij} r_j^{A*}) = 0$ holds for every φ . Differentiating both sides with respect to φ :

$$-\frac{\partial r_i^{A*}}{\partial \varphi} + \sigma'(h_i^{A*}) \sum_j W_{ij} \frac{\partial r_j^{A*}}{\partial \varphi} = 0$$

207 In matrix form, this is $(-\mathbf{I} + \mathbf{S}_A \mathbf{W}) \cdot \partial \mathbf{r}^{A*} / \partial \varphi = \mathbf{0}$, where $\mathbf{S}_A = \text{diag}(\sigma'(h_i^{A*}))$. The rotational derivative is
208 an exact null vector of the uncoupled Jacobian block. (The same holds for network B by identical argument.)

209 **Step 2: Mean-field coupling annihilates the rotational mode.** Consider the $2N$ -dimensional pertur-
210 bation $\mathbf{v}_A = (\partial \mathbf{r}^{A*} / \partial \varphi, \mathbf{0})^T$ corresponding to a shift of network A's bump alone. Multiplying by the full
211 block Jacobian (Section 2.3) yields:

$$\mathbf{J} \cdot \mathbf{v}_A = \begin{pmatrix} (-\mathbf{I} + \mathbf{S}_A \mathbf{W}) \cdot \partial \mathbf{r}^{A*} / \partial \varphi + \mathbf{S}_A \mathbf{C} \cdot \mathbf{0} \\ \mathbf{S}_B \mathbf{C} \cdot \partial \mathbf{r}^{A*} / \partial \varphi + (-\mathbf{I} + \mathbf{S}_B \mathbf{W}) \cdot \mathbf{0} \end{pmatrix}$$

212 The upper block vanishes by Step 1. The survival of the zero eigenvalue depends entirely on the cross-coupling
213 term $\mathbf{C} \cdot \partial \mathbf{r}^{A*} / \partial \varphi$ in the lower block. Recall that $\mathbf{C} = -\frac{J_X}{N} \mathbf{1} \mathbf{1}^T$. Applying \mathbf{C} to the rotational derivative:

$$\mathbf{C} \cdot \frac{\partial \mathbf{r}^{A*}}{\partial \varphi} = -\frac{J_X}{N} \mathbf{1} \left(\mathbf{1}^T \cdot \frac{\partial \mathbf{r}^{A*}}{\partial \varphi} \right) = -\frac{J_X}{N} \mathbf{1} \cdot \sum_{j=1}^N \frac{\partial r_j^{A*}}{\partial \varphi}$$

214 Because a rotation merely translates the bump profile around the periodic ring, the total activity (and thus
215 the mean) is strictly conserved. Exchanging derivative and sum:

$$\sum_{j=1}^N \frac{\partial r_j^{A*}}{\partial \varphi} = \frac{\partial}{\partial \varphi} \sum_{j=1}^N r_j^{A*} = \frac{\partial}{\partial \varphi} (N \bar{r}^A) = 0$$

216 Since $\mathbf{1}^T \cdot \partial \mathbf{r}^{A*} / \partial \varphi = 0$, it follows that $\mathbf{C} \cdot \partial \mathbf{r}^{A*} / \partial \varphi = \mathbf{0}$. The rank-1 coupling matrix completely annihilates
217 the rotational derivative, yielding $\mathbf{J} \cdot \mathbf{v}_A = \mathbf{0}$. By identical logic for network B, $\mathbf{v}_B = (\mathbf{0}, \partial \mathbf{r}^{B*} / \partial \varphi)^T$ is also a
218 null vector. \square

219 **Equivariance structure.** The result follows from the $\text{SO}(2) \times \text{SO}(2)$ equivariance of the coupled system at
220 $c = 0$: the dynamics commute with independent rotations of each ring. The mean-field coupling $J_X \bar{r}^X$ is

221 invariant under both $\text{SO}(2)$ actions because it depends only on total activity, which is a rotation-invariant
 222 functional. Each $\text{SO}(2)$ factor contributes one Goldstone mode to the kernel of the Jacobian. This protection
 223 is exact and holds at all J_x where the coexistence fixed point exists – it cannot be lifted by increasing
 224 cross-inhibition, only by breaking the rotational symmetry of either the within-network connectivity or the
 225 cross-coupling structure.

226 *Remark (circulant structure).* The within-network connectivity $W_{ij} = \frac{1}{N}(-J_0 + J_1 \cos(\theta_i - \theta_j))$ is exactly
 227 circulant: $W_{ij} = w(i - j \bmod N)$ for all N . The circulant structure ensures that the discrete rotation operator
 228 $T : r_i \mapsto r_{i+1 \bmod N}$ commutes with \mathbf{W} , i.e., $T\mathbf{W} = \mathbf{W}T$. Because the fixed-point equation $-\mathbf{r}^* + \sigma(\mathbf{W}\mathbf{r}^*) = 0$
 229 is solved by any rotation $T^k\mathbf{r}^*$, the nullspace of the Jacobian contains the rotational tangent vector $\partial\mathbf{r}^*/\partial\varphi$
 230 at every N , not merely in the continuous limit. The protection is exact at finite N , not an asymptotic
 231 approximation.¹

232 This is the neural circuit analog of the Goldstone theorem (Goldstone, 1961): a spontaneously broken
 233 continuous symmetry produces a massless (zero-energy) excitation. In our context, “massless” means
 234 neutrally stable – perturbations along the Goldstone direction neither grow nor decay. Because mean-field
 235 coupling acts exclusively on the spatially uniform mode (1), it is perfectly orthogonal to the zero-sum
 236 rotational modes, mathematically protecting positional memory from amplitude competition.

237 **3.2.2 Numerical Identification** We classify eigenvalues into Goldstone candidates ($|\lambda| < 10^{-3}$) and
 238 genuine modes ($|\lambda| \geq 10^{-3}$). For each eigenvector \mathbf{v} , we compute projections onto six basis directions:

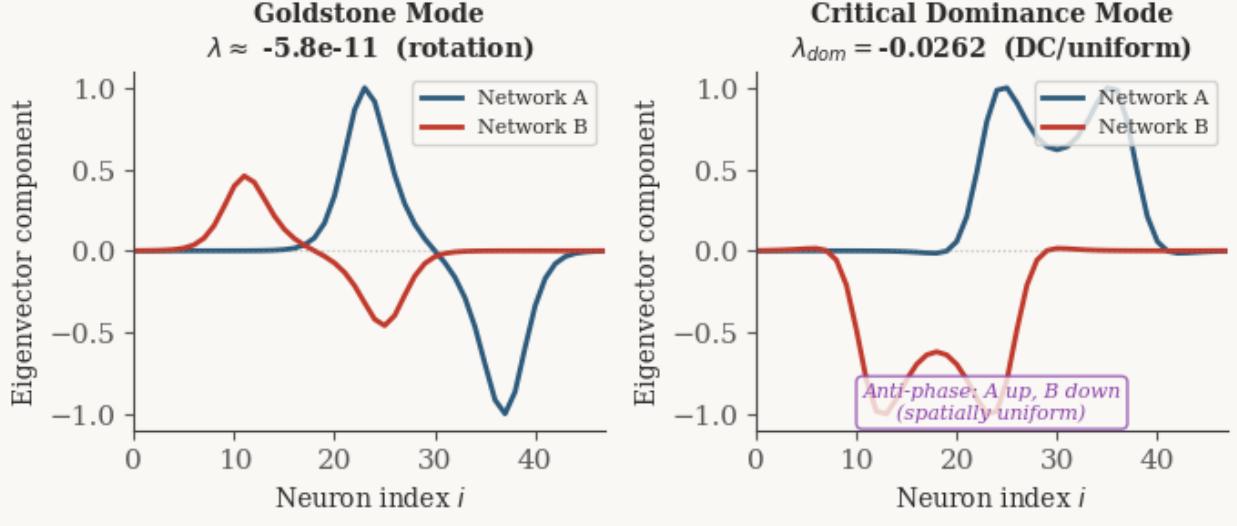
- 239 • \mathbf{d}_{dom} : symmetric dominance (cosine envelope, $A \uparrow B \downarrow$)
- 240 • $\mathbf{d}_{drift,+}$: co-directional drift (sine envelope, both shift same way)
- 241 • $\mathbf{d}_{drift,-}$: anti-directional drift (sine envelope, shift opposite ways)
- 242 • \mathbf{d}_{uni} : uniform/DC ($A \uparrow B \downarrow$ flat)
- 243 • $\mathbf{d}_{gold,A}$: rotation of bump A (sine envelope, A only)
- 244 • $\mathbf{d}_{gold,B}$: rotation of bump B (sine envelope, B only)

245 The Goldstone eigenvectors project strongly onto $\mathbf{d}_{gold,A}$ and $\mathbf{d}_{gold,B}$, confirming their rotational character
 246 (Fig. 4, left panel).

247 **3.2.3 Goldstone Count Across J_x** Across the entire range $J_x \in [0, 0.36]$ where coexistence exists, we
 248 find exactly two Goldstone modes. Their eigenvalues remain at $|\lambda| \sim 10^{-8}$ to 10^{-11} (machine precision for
 249 our iterative solver), and they are never lifted by increasing cross-inhibition. This confirms the symmetry
 250 protection: mean-field coupling cannot break rotational invariance.

¹Computational verification: at the coexistence fixed point ($J_x = 0.30$), the two Goldstone eigenvalue magnitudes are 1.1×10^{-5} ($N = 24$), 3.2×10^{-10} ($N = 48$), and 6.7×10^{-15} ($N = 96$), converging to machine precision as the angular grid refines, while the circulant commutation $\|T\mathbf{W} - \mathbf{W}T\|_\infty$ is below 4×10^{-16} at all N . The first non-Goldstone eigenvalue remains stable at $\lambda_1 \approx -0.131$ across all three resolutions.

Eigenvector Comparison at $J_x = 0.34$



251

Figure 4. Eigenvector comparison at $J_x = 0.34$ (near the pitchfork). Left: Goldstone mode ($\lambda \approx -5.8 \times 10^{-11}$, effectively zero), showing sinusoidal spatial structure in Network A – this is the rotational mode that slides the bump around the ring. Right: Critical dominance mode ($\lambda_{dom} = -0.026$, DC/uniform), showing anti-phase amplitude modulation localized to the active bump neurons. Because baseline neurons are strongly inhibited ($\sigma'(h_i) \approx 0$), they cannot participate in the linear instability; the mode is expressed only where the gain is nonzero. Despite this spatial localization, the mode projects maximally onto the uniform mean-field direction because the net effect is a difference in total activity between networks. The two modes are qualitatively distinct: the Goldstone mode encodes *where* the bump sits; the critical mode encodes *which network wins*.

260 3.3 The Pitchfork Bifurcation

261 3.3.1 The First Non-Goldstone Eigenvalue After removing the two Goldstone modes, we track the
262 dominant genuine eigenvalue λ_{dom} as a function of J_x (Figs. 3, 9). Key findings:

- 263 • At $J_x = 0$: $\lambda_{dom} = -0.572$ (strongly stable). Without cross-inhibition, the coexistence state is deeply
264 attractive.
- 265 • λ_{dom} increases monotonically with J_x , crossing zero at $J_x^* \approx 0.3485$.
- 266 • Above J_x^* : $\lambda_{dom} > 0$ (saddle). The coexistence state acquires one unstable direction.
- 267 • At $J_x = 0.356$: $\lambda_{dom} = +0.025$, and coexistence ceases to exist shortly after at $J_x^{\text{exist}} \approx 0.358$.

268 The crossing at J_x^* is a pitchfork bifurcation: the symmetric coexistence state ($D = 0$) loses stability, and
269 two WTA states ($D > 0$ and $D < 0$) emerge as the new stable attractors. The $A \leftrightarrow B$ exchange symmetry is
270 spontaneously broken.

271 3.3.2 Character of the Critical Eigenvector At J_x^* , the critical eigenvector has the following projections:

Direction	$ \langle \mathbf{v}_1, \mathbf{d} \rangle $
Uniform (DC)	0.43
Dominance (cosine)	0.34
Anti-drift (sine)	0.34
Co-drift (sine)	0.00

272 The largest projection is onto the uniform/DC direction (Fig. 4, right panel). Rather than a spatially flat

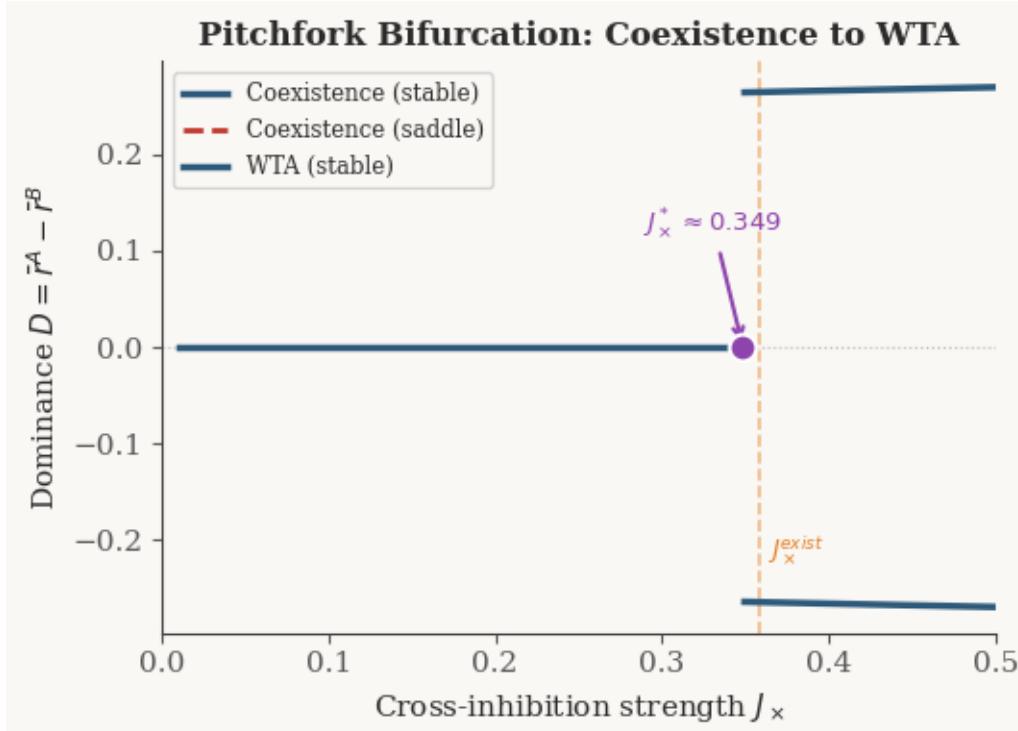
273 baseline shift, the mode drives a gain-weighted amplitude modulation: a sharp increase in the active neurons
 274 of network A coupled with a sharp decrease in the active neurons of network B, while baseline neurons
 275 (where $\sigma'(h_i) \approx 0$) are effectively silent. Because this amplitude competition produces a net difference in
 276 total activity between networks, it projects maximally onto the mean-field coupling direction. The instability
 277 is about which network has more total activity, not about the spatial pattern of either bump.

278 This is a direct and falsifiable prediction of mean-field coupling. If cross-inhibition were spatially structured
 279 (depending on the relative positions of the two bumps), the critical eigenvector would acquire spatial structure
 280 (cosine or higher Fourier modes). The DC character is specific to coupling that “sees” only total activity.

281 **3.3.3 The Narrow Existence Window** The coexistence saddle – genuinely unstable, not merely Goldstone-
 282 neutral – exists only in the interval $J_x \in [0.3485, 0.358]$, a width of $\Delta J_x \approx 0.01$. Below J_x^* , coexistence is a
 283 stable node. Above $J_x^{exist} \approx 0.358$, it ceases to exist entirely.

284 This razor-thin window has two implications:

- 285 1. **Structural precision.** The bifurcation is sharp: a 3% change in J_x (from 0.348 to 0.358) takes the
 286 system from stable coexistence through saddle instability to complete collapse. The system is tuned
 287 near a critical point.
- 288 2. **Heterogeneity prediction.** In biological circuits with heterogeneous connectivity, the sharp boundary
 289 should be smeared into a broader regime where saddle-like dynamics persist (see Discussion).



290 **Figure 3.** Pitchfork bifurcation diagram. Dominance $D = \bar{r}^A - \bar{r}^B$ vs. cross-inhibition strength J_x . The
 291 coexistence branch ($D = 0$) is stable (solid) for $J_x < J_x^*$ and becomes a saddle (dashed red) above the
 292 pitchfork at $J_x^* \approx 0.349$. Two WTA branches ($D > 0$ and $D < 0$, blue) emerge at $J_x^{exist} \approx 0.358$ as stable
 293 attractors. The resulting structure creates the narrow existence window $\Delta J_x \approx 0.01$. Note: the unstable
 294 branches connecting the pitchfork to the saddle-node fold at J_x^{exist} are not shown; our continuation solver
 295 tracked stable and saddle fixed points only. The expected topology is a pair of unstable branches emerging
 296 from J_x^* at $D = 0$ and folding onto the stable WTA branches at J_x^{exist} .²

²We use “subcritical” in the sense that the coexistence saddle coexists with WTA attractors for $J_x \in [J_x^*, J_x^{exist}]$; we did not track the unstable WTA branches via continuation, so the full subcritical fold structure is inferred from the topology rather than

298 **3.4 The Coexistence Saddle Under Cue**

299 Newton continuation in cue gain c from 0 to 0.5 at $J_x = 0.35$ (within the saddle window; see Fig. 3) reveals:

- 300 • The coexistence branch maintains both bumps across the full cue range, with D growing slowly as the
301 cue favors network A.
302 • All 51/56 tracked solutions are unstable (saddle points), with $n_{positive} = 1$ for $c < 0.15$ and $n_{positive} = 2$
303 for $c > 0.15$ (a second eigenvalue crossing).
304 • The critical eigenvector's projection onto cosine and sine directions is $|\cos| \approx |\sin| \approx 0.47$ – a mixed
305 dominance-drift mode at 45 degrees.

306 The WTA branch (tracked simultaneously) is stable across all cue values, confirming that the saddle's unstable
307 manifold connects to the two WTA basins.

308 **3.5 Stochastic Phase Diagram of Swap Errors**

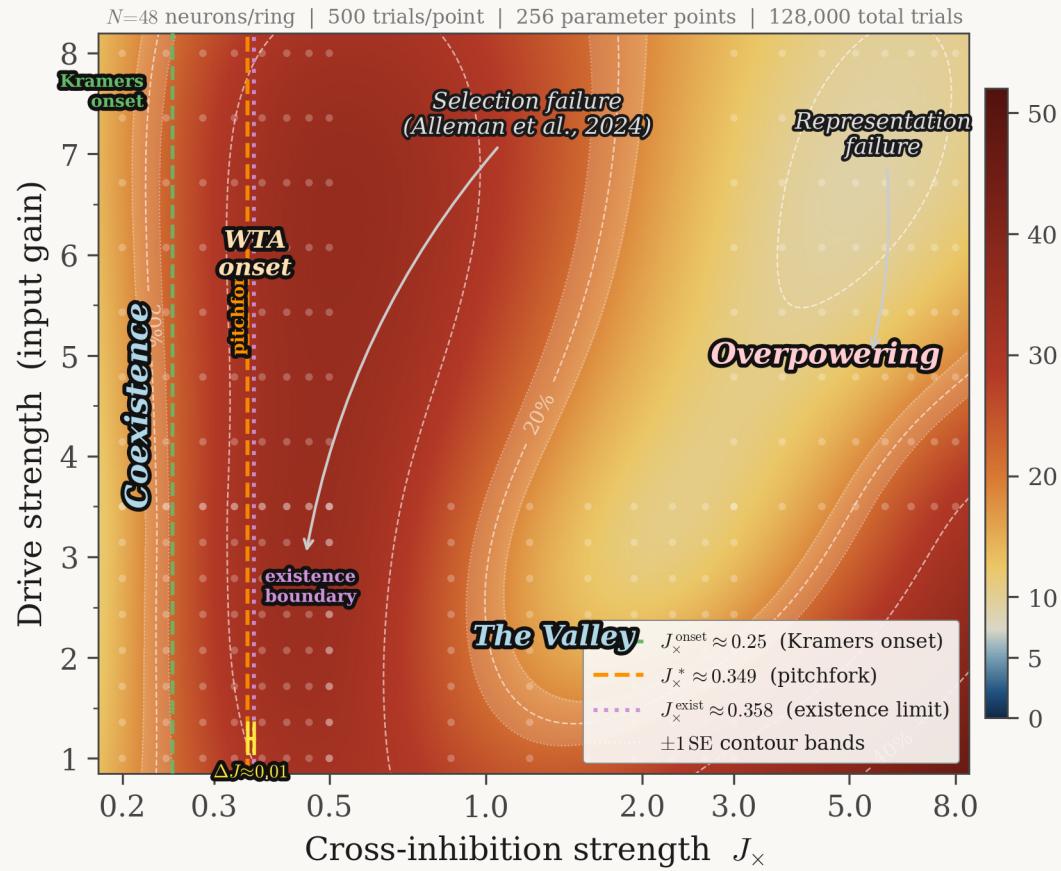
309 **3.5.1 Parameter Sweep** To bridge the deterministic bifurcation analysis with behavioral predictions, we
310 performed a large-scale stochastic simulation across the (J_x, c) parameter space. The “drive strength” axis of
311 the phase diagram corresponds to the cue gain c applied during a brief encoding period (Section 2.2), not to
312 the initialization drives used in the fixed-point analysis. At each of 256 grid points (16 values of J_x from 0.05
313 to 8.0, 16 values of drive strength from 1.0 to 8.0), we ran 500 stochastic trials ($N_{total} = 128,000$). Two ring
314 networks encoded items separated by $\pi/2$ radians, with independent Gaussian noise ($\sigma = 0.1$) added to each
315 neuron during a 500-step maintenance period. The response was decoded by computing cosine similarity of
316 the combined population vector $(\mathbf{r}^A + \mathbf{r}^B)$ against 500 template tuning curves uniformly spanning $[-\pi, \pi]$;
317 the decoded angle is the template with maximum projection. A trial was classified as a swap error when
318 the decoded angle fell within 0.3 rad of the non-target item’s location; the output variable $D = \bar{r}^A - \bar{r}^B$ is
319 the dominance order parameter used in the spectral analysis, not the decoder. To assess sensitivity to this
320 classification threshold, we repeated a focused sweep (16 values of J_x from 0.1 to 3.0, two drive levels, 200
321 trials each) with thresholds of 0.2, 0.3, 0.4, and 0.5 rad. The qualitative features of the phase diagram – the
322 onset location ($J_x^{onset} \approx 0.20\text{--}0.23$ across thresholds), the peak near $J_x \approx 0.3\text{--}0.5$, and the non-monotonic
323 valley at $J_x \approx 1.2\text{--}1.5$ – are robust to threshold choice. Wider classification windows increase absolute swap
324 rates (e.g., 10% vs. 26% at $J_x = 0.25$ for 0.2 vs. 0.5 rad) but preserve the ordering across J_x values and the
325 location of qualitative transitions.

326 Binomial standard errors for the estimated swap rates are $SE = \sqrt{p(1-p)/500}$, ranging from $\pm 1.0\%$ at the
327 onset threshold ($p \approx 0.05$) to $\pm 2.2\%$ at chance ($p = 0.50$). These uncertainties are small relative to the
328 qualitative features of the phase diagram (onset band, valley depth, plateau height).

329 The stochastic sweep extends to $J_x = 8.0$, well beyond the coexistence existence threshold $J_x^{\text{exist}} \approx 0.36$.
330 For $J_x > J_x^{\text{exist}}$, the system admits only WTA solutions; the stochastic trials begin from a coexistence-like
331 initial condition (both bumps established by strong drives) and the dynamics reveal which network captures
332 the WTA state. This extended range is necessary to characterize the full behavioral landscape, including
333 the non-monotonic valley at $J_x \approx 1.2\text{--}1.6$ (Section 3.5.4), which operates in the pure WTA regime. The
334 cross-inhibition parameter J_x has the same definition throughout: the coefficient in the mean-field coupling
335 term $J_x \bar{r}^X$ (Eq. 2). No rescaling is applied between the deterministic and stochastic analyses.

computed directly.

Phase diagram of swap errors in coupled ring attractors



3.5.2 Onset of Swap Errors

Swap errors emerge at $J_x^{\text{onset}} \approx 0.25$, consistent with the spectral prediction of the pitchfork bifurcation at $J_x^* \approx 0.3485$. The stochastic onset is lower than the deterministic bifurcation because noise-mediated escape from the metastable coexistence well occurs when the barrier height $\Delta V \sim \sigma^2$, which corresponds to a J_x slightly below the eigenvalue crossing. This is precisely the Kramers mechanism.

Between $J_x^{\text{onset}} \approx 0.25$ and 0.5, swap rates increase from 5% to approximately 45%. Above $J_x \approx 1.0$, swap rates plateau near 50% – the system has become a noise-driven bistable switch with no memory of the initial encoding.

3.5.3 Drive Strength Is Secondary The phase diagram shows near-vertical isocontours of swap rate in the near-critical regime ($J_x \approx 0.2\text{--}0.5$, Fig. 5): swap error probability depends primarily on J_x and only

weakly on drive strength. This is a direct prediction of the spectral analysis: the critical eigenvector projects maximally onto the uniform (DC) direction, governing total activity competition rather than spatial encoding. Stronger drive does not protect against the dominance instability because the instability is orthogonal to the encoding direction.

This has a counterintuitive implication: increasing stimulus strength – the commonly proposed intervention for working memory failures – targets the wrong degree of freedom *near the critical point*. In the valley regime ($J_x \approx 1.2\text{--}1.6$), by contrast, drive strength becomes the primary determinant of error rate (Fig. 5), consistent with the valley being a selection-dominated regime where encoding quality, not barrier height, governs performance.

A caveat: our model uses additive Gaussian noise with constant variance σ^2 . Cortical noise is closer to Poisson-like, with variance scaling with mean firing rate. Under such state-dependent noise, weaker cues reduce mean activity and inflate the effective noise floor (Fano factor), potentially creating a cue-driven cliff even at fixed J_x . The drive-is-secondary finding is therefore specific to the additive noise regime near the pitchfork; under multiplicative noise, cue strength may contribute to the cliff through a noise-amplification mechanism independent of the spectral instability.

3.5.4 The Non-Monotonic Valley and Two Failure Regimes At $J_x \approx 1.2\text{--}1.6$ with moderate to strong drive, the phase diagram reveals a non-monotonic feature: swap rates dip to 7–13% between two distinct failure modes. Crucially, this valley lies well above the coexistence existence threshold ($J_x^{exist} \approx 0.36$), meaning it operates in a pure WTA regime where deterministic coexistence does not exist. The two failure modes and the valley between them correspond to qualitatively different dynamical regimes:

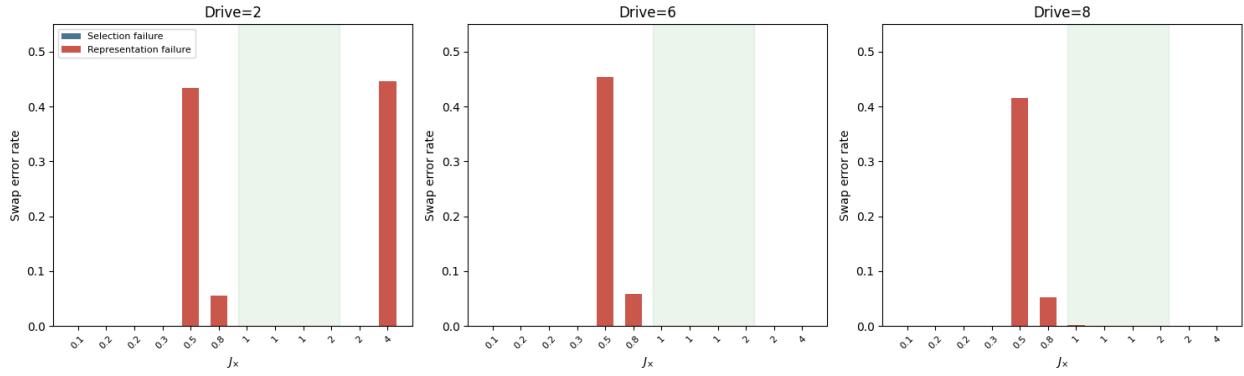
1. **Near-critical swaps** ($J_x \approx 0.3\text{--}0.5$): Near and just above the pitchfork, barriers separating coexistence from WTA are small and noise escapes freely. This is the *representation failure* regime: stochastic dynamics push the system from (metastable) coexistence into an incorrect WTA state. Swap rate approaches 50%.
2. **Overpowering swaps** ($J_x > 2.0$, weak drive): Cross-inhibition is so strong that it overwhelms feedforward encoding during the stimulus presentation itself. One network suppresses the other before encoding is complete. This is also representation failure, but driven by the *encoding* phase rather than the maintenance phase.
3. **The valley** ($J_x \approx 1.2\text{--}1.6$, strong drive): At these J_x values, coexistence does not exist as a deterministic fixed point – the system is in a pure WTA regime. Strong encoding drive biases the initial WTA competition so that the cued network typically wins. Swap errors here arise when stochastic fluctuations during the encoding-to-maintenance transition cause the wrong network to capture the WTA state. We hypothesize that this corresponds to the *selection failure* mechanism identified by Alleman et al. (2024), where both representations are briefly encoded but the wrong one is selected at readout.

To operationalize this distinction, we define a diagnostic: at the end of the maintenance period, if both networks retain above-threshold activity ($\max_i r_i^X > 0.3$ for $X \in \{A, B\}$), a swap error is classified as selection failure (both representations survived but the wrong one was decoded). If one network has collapsed ($\max_i r_i^X < 0.1$), it is classified as representation failure (one item was lost during maintenance). We applied this diagnostic across 24,000 stochastic trials (12 values of J_x from 0.1 to 4.0, 4 drive levels, 500 trials each). The results do not confirm the hypothesized crossover (Fig. 10). Swap errors across the entire parameter range are overwhelmingly *representation* failures: one bump collapses during maintenance, and the surviving bump (or the stronger of two degraded bumps) is decoded as the wrong item. Selection failures – where both bumps survive above threshold but the wrong one is decoded – are essentially absent at $\sigma = 0.1$. This holds in both the near-critical regime ($J_x \approx 0.3\text{--}0.5$, swap rate ~43%, all representation failure) and the overpowering regime ($J_x > 2.0$, weak drive, swap rate ~45%, all representation failure). In the valley ($J_x \approx 1.0\text{--}1.6$), swap rates drop to near zero, consistent with clean WTA resolution under strong encoding drive – not because selection errors replace representation errors, but because the WTA competition resolves correctly.

This result has an important implication: the single-stage dynamics of our model – where maintenance and readout are a single continuous process – do not produce a selection failure mechanism. In the model, when the wrong network wins the WTA competition, the losing bump genuinely collapses; there is no epoch where

408 both representations coexist and a readout process selects the wrong one. The Alleman et al. (2024) finding
 409 of selection failure in neural data therefore cannot be captured by this model and likely requires a two-stage
 410 architecture with a separate readout competition (see Discussion, Section 4.6).
 411 The valley thus represents a candidate functional operating regime for *reliable WTA resolution*, not for
 412 selection as originally hypothesized: cross-inhibition strong enough to resolve competition via WTA, encoding
 413 strong enough to bias that competition correctly. The specific parameter range ($J_x \approx 1.2\text{--}1.6$) depends on
 414 our model parameterization and should not be interpreted as a direct physiological prediction; the qualitative
 415 feature – a non-monotonic minimum between two failure modes – is the robust finding. The circuit need not
 416 be tuned precisely to J_x^* but rather to a regime where WTA dynamics and encoding strength are balanced.
 417 The valley’s location in parameter space ($J_x \approx 1.2\text{--}1.6$) is specific to the sigmoid nonlinearity and gain
 418 parameters used here (Section 2.1). The qualitative structure – two competing failure modes generating a non-
 419 monotonic minimum – is the robust finding, as it depends on the generic competition between cross-inhibition
 420 strength and noise-driven escape.

Figure 10. Selection vs Representation Failure Across J_x



421 **Figure 10.** Selection vs representation failure decomposition across J_x at three drive levels (drive = 2, 6, 8;
 422 500 trials per point, $\sigma = 0.1$). Blue: selection failure rate (both bumps above threshold at trial end, wrong
 423 item decoded). Red: representation failure rate (one bump collapsed during maintenance). Swap errors are
 424 overwhelmingly representation failures across all J_x values and drive levels. Near-critical ($J_x \approx 0.3\text{--}0.5$) and
 425 overpowering ($J_x = 4.0$, low drive) regimes show high swap rates (~43–45%), dominated by representation
 426 failure. The valley (green shading, $J_x \approx 1.0\text{--}1.6$) shows near-zero swap rates: the WTA competition resolves
 427 correctly under strong drive. Selection failures are essentially absent, indicating that the single-stage model
 428 does not produce a maintenance epoch where both representations coexist at readout.

430 **3.5.5 Kramers Consistency Analysis from the Dominance Eigenvalue** To test consistency between
 431 the deterministic pitchfork at $J_x^* \approx 0.349$ and the stochastic onset of swap errors, we approximate the escape
 432 barrier out of the coexistence state by projecting the high-dimensional dynamics onto the critical dominance
 433 direction identified in the Jacobian spectrum.

434 Near J_x^* , the dynamics admit a reduction onto the scalar amplitude $x(t) = \langle \mathbf{w}_{\text{dom}}, \delta \mathbf{r}(t) \rangle$ (where \mathbf{w}_{dom} is the
 435 left eigenvector satisfying $\mathbf{w}^T \mathbf{J} = \lambda \mathbf{w}^T$), yielding the saturating normal form

$$\tau \dot{x} = \lambda_{\text{dom}}(J_x) x + \gamma x^3 - \delta x^5 + \eta_{\text{eff}}(t), \quad \delta > 0,$$

436 where $\lambda_{\text{dom}}(J_x)$ is the measured dominant non-Goldstone eigenvalue and the quintic term captures saturation
 437 that stabilizes the distant WTA states. The escape barrier from coexistence is set by the nearby inner saddle
 438 and is therefore controlled primarily by λ_{dom} and γ .

439 In the metastable regime below the pitchfork, the effective potential $V(x)$ has an inner unstable saddle at

$$x_s^2 \approx \frac{|\lambda_{\text{dom}}|}{\gamma},$$

440 giving a barrier height

$$\Delta V \equiv V(x_s) - V(0) \approx \frac{|\lambda_{\text{dom}}|^2}{4\gamma}.$$

441 **Non-normality and eigenvector conventions.** The Jacobian is non-symmetric due to the state-dependent
 442 gain mask ($\mathbf{S}_A \mathbf{W} \neq \mathbf{W} \mathbf{S}_A$ when gains differ across neurons); the proper modal decomposition requires the
 443 left eigenvector \mathbf{w}_{dom} . At $J_x = 0.34$, the cosine similarity between left and right dominant eigenvectors is
 444 $|\langle \mathbf{v}_{\text{dom}}, \mathbf{w}_{\text{dom}} \rangle| \approx 0.55$, indicating substantial non-normality arising from the gain contrast between active and
 445 baseline neurons. However, the Kramers consistency estimate is robust to this distinction: the barrier height
 446 $\Delta V = |\lambda_{\text{dom}}|^2/(4\gamma)$ depends on the eigenvalue (identical for left and right eigenvectors) and the normal-form
 447 coefficient γ , which is inferred from the observed onset rather than derived via center-manifold reduction.
 448 The eigenvector distinction would affect the Kramers prefactor k_0 and any first-principles calculation of γ —
 449 neither of which the present analysis attempts.

450 **Noise projection and finite-horizon criterion.** In the stochastic simulations, independent Gaussian noise
 451 of standard deviation $\sigma = 0.1$ is added per neuron over a $T = 500$ -step maintenance window. Because \mathbf{w}_{dom}
 452 is unit-normalized, projecting isotropic independent noise onto the dominance coordinate preserves variance:
 453 $\text{Var}[\eta_{\text{eff}}] = \sigma^2$. This holds for the raw additive noise; however, the linearized dynamics include the gain mask
 454 $\sigma'(h_i)$, which suppresses noise contributions from baseline neurons. Because \mathbf{w}_{dom} projects primarily onto
 455 active (high-gain) neurons (Fig. 4), the effective noise along the critical coordinate may differ modestly from σ^2 .
 456 We treat σ^2 as the leading-order estimate and note that a full Kramers prefactor calculation — incorporating
 457 the gain-weighted noise covariance — would refine the quantitative onset prediction without altering the
 458 qualitative quadratic-collapse mechanism. Kramers theory predicts an escape rate $k \sim k_0 \exp(-\Delta V/\sigma^2)$; over
 459 a finite horizon T , escape becomes likely when $kT \sim 1$, i.e.

$$\frac{\Delta V}{\sigma^2} \approx \ln(k_0 T).$$

460 Taking a conservative attempt-frequency range $k_0 \in [0.1, 1]$ per step gives $\ln(k_0 T) \in [\ln 50, \ln 500] \approx [3.9, 6.2]$,
 461 hence $\Delta V \in [0.039, 0.062]$.

462 **Numerical evaluation at the observed swap onset.** The stochastic phase diagram shows swap
 463 errors rising steeply near $J_x^{\text{onset}} \approx 0.25$. At this coordinate, the computed coexistence spectrum gives
 464 $\lambda_{\text{dom}}(0.25) \approx -0.2357$ (Fig. 7A), yielding

$$\Delta V(0.25) \approx \frac{(0.2357)^2}{4\gamma} \approx \frac{0.0139}{\gamma}.$$

465 Equating to the finite-horizon Kramers threshold $\Delta V \in [0.039, 0.062]$ implies

$$\gamma \approx 0.22 \text{ to } 0.36,$$

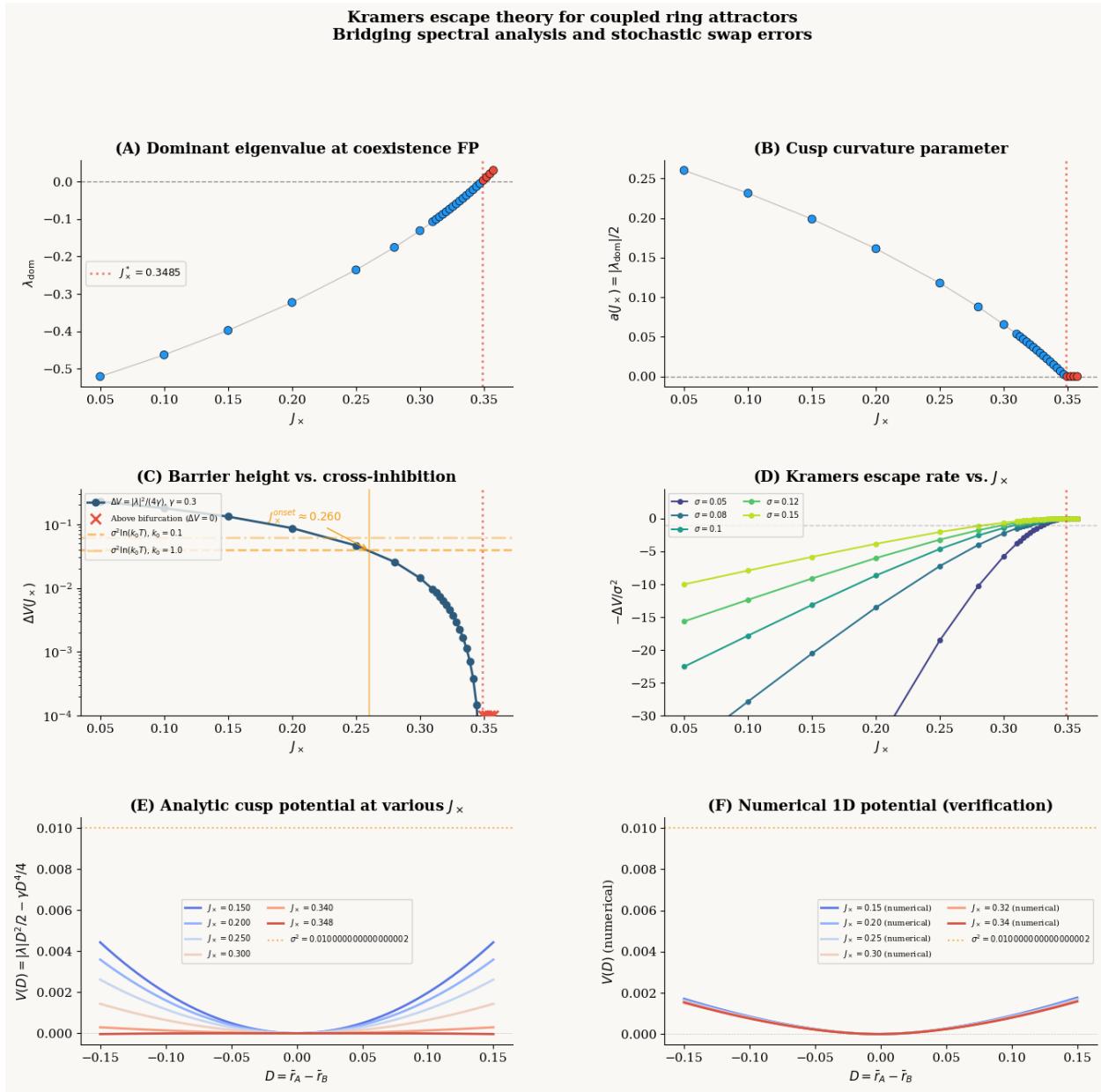
466 We do not derive γ from first principles via center-manifold reduction; instead, we ask whether any single
 467 value of γ can simultaneously satisfy the barrier-height equation at the observed onset. The inferred range
 468 $\gamma \approx 0.22\text{--}0.36$ serves as a consistency check, not an ab initio prediction: it confirms that a normal-form
 469 reduction with physically reasonable cubic nonlinearity explains why the onset lies $\sim 29\%$ below the pitchfork.
 470 As a self-consistency check: taking $\gamma = 0.3$ gives $\Delta V(0.25) \approx 0.046$ and $\Delta V/\sigma^2 \approx 4.6$, squarely in the $\ln(k_0 T)$
 471 band (Fig. 7C).

472 The onset prediction shifts smoothly with noise amplitude. Because $\Delta V \propto |\lambda_{\text{dom}}|^2$ while the threshold scales
 473 as $\sigma^2 \ln(k_0 T)$, the predicted J_x^{onset} depends on σ only through the threshold level, not through the barrier
 474 shape:

σ	ΔV threshold	J_x^{onset} range
0.05	[0.010, 0.016]	[0.30, 0.31]
0.10	[0.039, 0.062]	[0.23, 0.26]
0.15	[0.088, 0.140]	[0.14, 0.20]

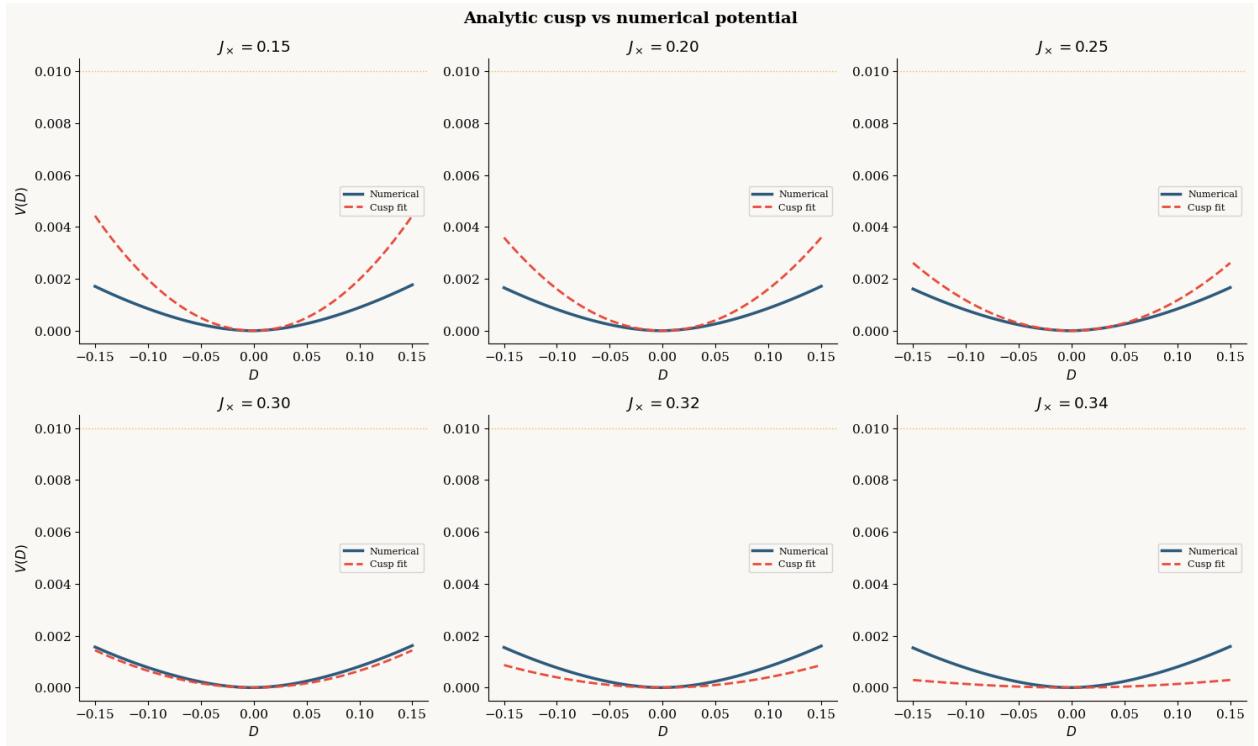
475 Halving the noise from $\sigma = 0.10$ to 0.05 shifts the onset upward by only $\Delta J_x \approx 0.05$; increasing it to 0.15
 476 shifts it downward by ~ 0.06 . The quadratic barrier collapse is structural — set by the spectral approach to
 477 the pitchfork — so the sensitivity enters only through the logarithmic noise floor $\sigma^2 \ln(k_0 T)$.

478 This provides a quantitative explanation for why the behavioral “cliff” occurs substantially below the
 479 deterministic pitchfork. Because $\Delta V \propto |\lambda_{\text{dom}}|^2$, the barrier collapses quadratically and drops to the
 480 logarithmically-scaled noise floor near $J_x^{\text{onset}} \approx 0.25$ — roughly 29% below J_x^* — rendering the network
 481 noise-limited before the coexistence fixed point formally loses its deterministic stability (Fig. 7D).



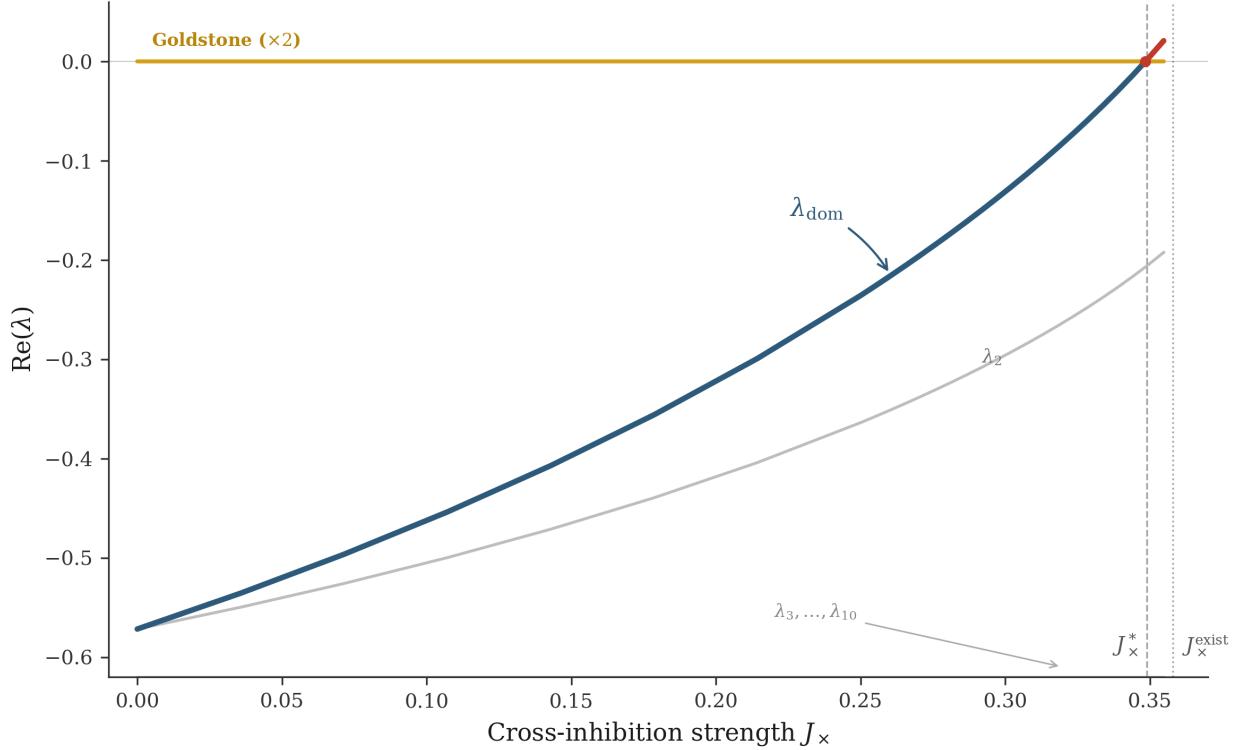
483 **Figure 7.** Kramers consistency analysis linking spectral theory and stochastic swap errors. (A) Dominant
 484 non-Goldstone eigenvalue λ_{dom} at the coexistence fixed point as a function of J_x . (B) Cusp curvature
 485 parameter tracking the normal-form geometry. (C) Barrier height $\Delta V = |\lambda_{\text{dom}}|^2/(4\gamma)$ with $\gamma = 0.30$;
 486 the horizontal band marks the finite-horizon Kramers threshold $\Delta V/\sigma^2 \in [3.9, 6.2]$. (D) Kramers escape
 487 rate $k = k_0 \exp(-\Delta V/\sigma^2)$, showing exponential amplification as $J_x \rightarrow J_x^*$. (E) Analytic cusp potential
 488 $V(D)$ at selected J_x values. (F) Numerically computed one-dimensional potential, verifying the analytic
 489 approximation.

490 **Validation of the cusp approximation.** To assess the fidelity of the quartic normal form, we compare
 491 the analytic cusp potential with a numerically computed one-dimensional potential obtained by integrating
 492 the projected dominance dynamics at six values of J_x spanning the subcritical regime (Fig. 8). The cusp
 493 approximation is quantitatively accurate in the onset region ($J_x \approx 0.25$ – 0.30) where the Kramers calculation
 494 is applied. At low J_x (≤ 0.20), the analytic cusp overestimates the barrier – a conservative error that does
 495 not affect the escape-onset prediction. Near the bifurcation ($J_x \geq 0.32$), the cusp underestimates the barrier,
 496 as higher-order terms in the effective potential provide additional stabilization that the quartic truncation
 497 misses. This pattern of errors – overestimation far from criticality, underestimation near it – means the cusp
 498 approximation is most reliable precisely where the Kramers consistency argument is most needed.



500 **Figure 8.** Validation of the cusp normal-form approximation. Analytic cusp potential (red dashed) versus
 501 numerically computed one-dimensional potential (blue solid) at six values of J_x spanning the subcritical
 502 regime ($J_x \in \{0.15, 0.20, 0.25, 0.30, 0.32, 0.34\}$). The cusp approximation is quantitatively accurate in the
 503 onset region ($J_x \approx 0.25$ – 0.30) and provides conservative (over-)estimates at low J_x . Near the bifurcation
 504 ($J_x \geq 0.32$), higher-order terms provide additional stabilization beyond the quartic truncation. Orange
 505 dashed line: noise scale $\sigma^2 = 0.01$.

506 **3.6 Eigenvalue Spectrum Summary**



507

508 **Figure 9.** Eigenvalue spectrum strip. Real parts of the top eigenvalues of the coupled-system Jacobian as a
 509 function of cross-inhibition strength J_x , across the full coexistence range $J_x \in [0, 0.36]$. Two Goldstone modes
 510 (gold, $\times 2$) are pinned at $\lambda = 0$ by rotational symmetry at all J_x . The dominant non-Goldstone eigenvalue
 511 λ_{dom} (thick curve, blue when negative/stable, red when positive/unstable) rises monotonically, crossing zero
 512 at $J_x^* \approx 0.349$ (dashed vertical line) – the pitchfork bifurcation. The second genuine eigenvalue λ_2 (gray)
 513 also rises with J_x but remains well below zero throughout the coexistence range. Remaining eigenvalues
 514 ($\lambda_3, \dots, \lambda_{10}$) lie deeper in the stable half-plane. Dotted vertical line marks the coexistence existence boundary
 515 $J_x^{\text{exist}} \approx 0.358$, beyond which the fixed-point solver fails. The clear spectral gap between λ_{dom} and λ_2 confirms
 516 that the pitchfork instability is a one-dimensional phenomenon governed by a single critical mode.

517

518 **4. Discussion**

519 **4.1 The Coexistence Threshold as a Structural Constraint**

520 The finding that coexistence does not exist at the commonly used $J_x = 0.5$ is a structural constraint on
 521 models of multi-item working memory. If the brain maintains multiple items simultaneously – as behavioral
 522 data strongly suggest (Bays et al., 2009; Ma et al., 2014) – then, within the class of mean-field cross-inhibition
 523 models of the form Eq. (2), the effective cross-inhibition must be below a parameter-dependent existence
 524 boundary. This constrains the balance between lateral inhibition and recurrent excitation: the circuit cannot
 525 be in the WTA regime and store multiple items.

526 Our threshold $J_x^{\text{exist}} \approx 0.36$ depends on the specific parameters (J_0, J_1, β, h_0) . Different parameter regimes
 527 will yield different thresholds. The key result is qualitative: there is always a finite critical cross-inhibition
 528 beyond which coexistence is structurally impossible.

529 **4.2 Heterogeneity Transforms the Bifurcation Type**

530 In our symmetric model, the coexistence saddle exists only in a narrow window ($\Delta J_x \approx 0.01$). However,
531 biological circuits have heterogeneous connectivity, non-uniform firing thresholds, and spatially structured
532 inhibition (Kilpatrick et al., 2013). We tested the effect of connectivity heterogeneity by adding symmetric
533 Gaussian noise to the within-network weight matrices ($W \rightarrow W + \sigma \xi$, where $\xi_{ij} \sim \mathcal{N}(0, 1/N)$) and repeating
534 the eigenvalue analysis across six noise levels ($\sigma \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$, five random seeds each, 30 values
535 of J_x per condition).

536 The result refutes the intuitive prediction that heterogeneity would widen the instability window. Instead, for
537 this perturbation class (additive weight noise), heterogeneity *destroys* it. At $\sigma = 0.05$, two of three trials
538 lost the instability entirely (the dominance eigenvalue λ_{dom} never crossed zero), while one trial showed a
539 wider window ($\Delta J_x \approx 0.04$). At $\sigma \geq 0.10$, no trial exhibited a positive λ_{dom} at any J_x – the sharp pitchfork
540 bifurcation had vanished completely.

541 The mechanism is the breaking of exact A↔B exchange symmetry. The pitchfork bifurcation at J_x^* requires
542 that the two networks be related by an exact symmetry operation: if (r_A^*, r_B^*) is a fixed point, then (r_B^*, r_A^*)
543 must also be one, and the bifurcation occurs when the symmetric fixed point ($D = 0$) loses stability to the
544 antisymmetric perturbation ($D \neq 0$). Heterogeneity in the weight matrices breaks this exchange symmetry,
545 because the two networks no longer have identical connectivity. In the language of the cusp potential
546 $V(D) = D^4 + aD^2 + bD$, heterogeneity introduces a nonzero b even at zero cue – the potential is always
547 tilted, and there is no sharp symmetry-restoring point where $b = 0$ exactly.

548 This converts the pitchfork into an *imperfect bifurcation* (Strogatz, 2015). Instead of a sharp zero-crossing
549 of λ_{dom} , the system shows a smooth crossover: one network is always slightly favored, and the dominance
550 eigenvalue approaches zero asymptotically without crossing it. The “window” does not widen – it dissolves,
551 because the phase transition changes type from a sharp symmetry-breaking event to a smooth preference
552 gradient.

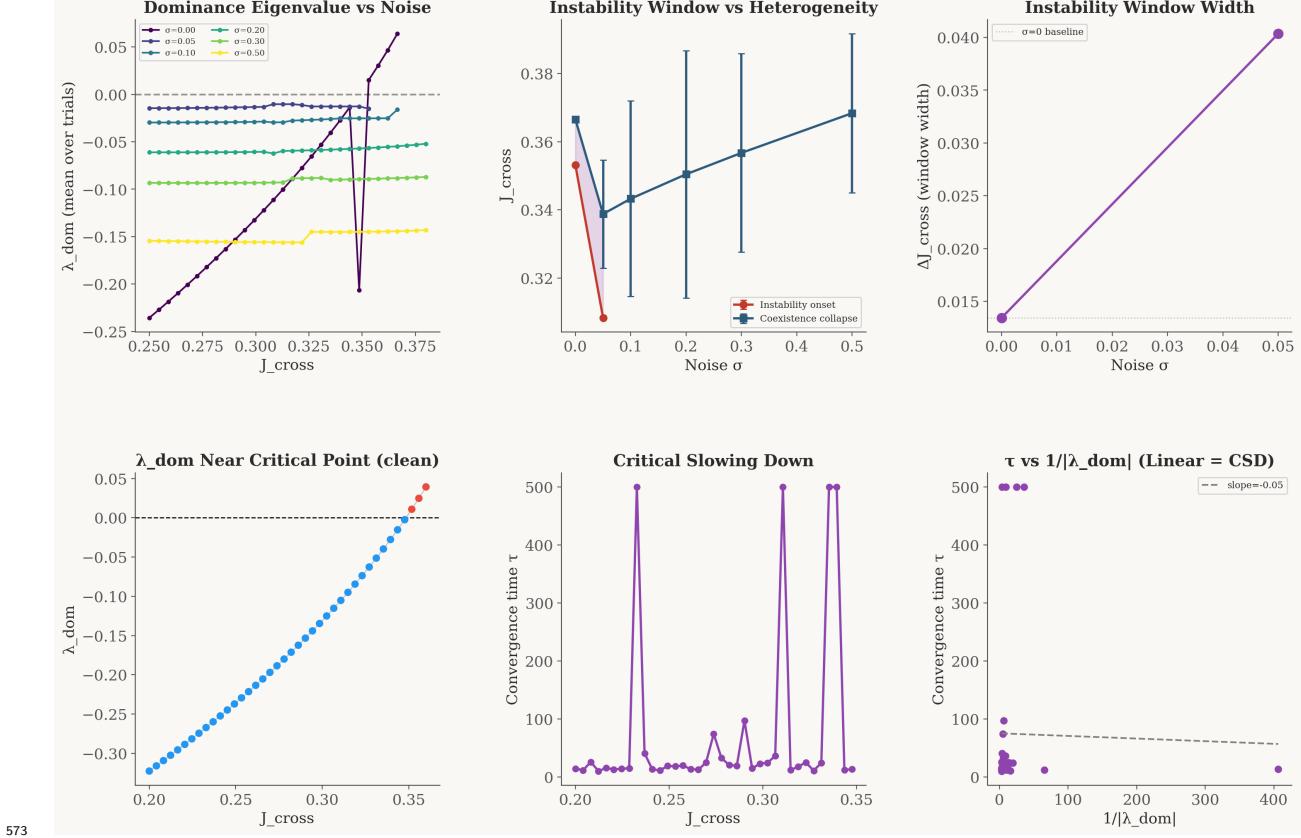
553 Two consequences follow. First, the Goldstone modes, which are exactly zero in the symmetric model,
554 become “soft modes” with small but nonzero eigenvalues at $\sigma > 0$ (Kilpatrick et al., 2013; Poll et al., 2015).
555 Bumps become pinned to preferred locations rather than freely rotating. Second, and more importantly, the
556 razor-thin window ($\Delta J_x \approx 0.01$) is a symmetry artifact of the clean model, not a biological constraint. Real
557 circuits operate in a regime of smooth crossover where no parameter precision is required.

558 This result strengthens the valley interpretation (Section 3.5.4). The non-monotonic valley at $J_x \approx 1.2$ – 1.6
559 exists in the stochastic simulations regardless of whether the underlying deterministic bifurcation is a sharp
560 pitchfork or a smooth crossover. What matters for behavior is the *landscape* – the barrier heights and basin
561 depths – not the exact location of a mathematical bifurcation point. Heterogeneity smears the transition
562 without eliminating the functional operating regime. Indeed, the proper framing inverts the usual presentation:
563 the biological circuit, with its heterogeneous connectivity, is the general case. The clean symmetric model is
564 the *organizing center* – the codimension-1 singularity whose unfolding reveals why the general case has the
565 structure it does. The pitchfork is not the norm that heterogeneity degrades; it is the germ from which the
566 smooth crossover unfolds.

567 Whether other biologically relevant symmetry-breakers — heterogeneous thresholds, asymmetric gains,
568 spatially structured inhibition — produce broader quasi-critical windows or qualitatively different unfoldings
569 remains an open question.

570 Critical slowing down provides partial confirmation: convergence time after small perturbations increases as
571 J_x approaches J_x^* in the clean model ($\sigma = 0$), consistent with the expected $\tau \sim 1/|\lambda_{dom}|$ scaling near the
572 bifurcation. The data are noisy but directionally consistent with critical slowing down (Fig. 6).

Heterogeneity and Critical Slowing Down



573
Figure 6. Heterogeneity destroys the instability window. Dominance eigenvalue λ_{dom} as a function of J_{\times}
574 at six levels of connectivity heterogeneity ($\sigma \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$, five random seeds each). At $\sigma = 0$
575 (clean model), the pitchfork zero-crossing is sharp. At $\sigma = 0.05$, two of three trials lose the instability entirely.
576 At $\sigma \geq 0.10$, no trial exhibits a positive λ_{dom} at any J_{\times} – the imperfect bifurcation has no zero-crossing.
577 Right panel: convergence time data, consistent with critical slowing down near J_{\times}^* in the clean model.
578

579 4.3 The Behavioral Cliff as a J_{\times} Phenomenon

580 The traditional view attributes the behavioral cliff to weak cues: below a threshold cue strength, the sensory
581 signal cannot stabilize the correct bump, and noise-driven escape to a competing attractor becomes rapid
582 (Kramers escape). Our analysis suggests a reinterpretation.

583 The cliff reflects the system's proximity to critical J_{\times}^* . The cue does not create the competition – J_{\times} does. The
584 cue provides stabilization against the dominance instability driven by J_{\times} . When cue-mediated stabilization
585 fails (cue too weak relative to $|J_{\times} - J_{\times}^*|$), the system transitions to WTA.

586 This reinterpretation makes two testable predictions:

- 587 **1. Manipulating cross-inhibition should shift the cliff.** Pharmacological modulation of GABAergic
588 inhibition, or inter-hemispheric TMS suppression, should change the effective J_{\times} — provided the
589 dominant effect is on inter-population inhibition rather than within-network E/I balance. Systemic
590 GABAergic modulation also alters local inhibition (J_0), which changes bump stability and gain
591 independently of cross-inhibition. A clean test would require circuit-specific interventions that target
592 cross-population connections selectively. Stronger cross-inhibition should move the cliff to higher cue
593 values; weaker cross-inhibition should make it disappear.

594 **2. Individual differences may reflect J_x variation.** Subject-to-subject variability in cliff location may
595 arise from variation in effective cross-inhibition strength (neuromodulation, connectivity differences)
596 rather than variation in sensory sensitivity.

597

4.4 Goldstone Protection and Functional Significance

598 The two Goldstone modes have a functional interpretation. They protect bump positions (the stored memory
599 content) from being disrupted by the dominance competition. The system can resolve “which network wins”
600 without disturbing “where each bump sits.” This separation of concerns – amplitude competition in the
601 dominance subspace, position preservation in the Goldstone subspace – may be a design principle for neural
602 circuits that must make decisions while maintaining stored information. In the language of Noether’s theorem,
603 the continuous rotational symmetry of each ring implies a conserved quantity: the projection of the network
604 state onto the rotational tangent direction experiences zero restoring force. This is the angular-momentum
605 analog for the bump’s position – the memory of *where* the stimulus was stored resides in this zero-eigenvalue
606 direction. When connectivity heterogeneity lifts the Goldstone modes to soft modes (Section 4.2), this
607 conservation law is broken: bumps become pinned to preferred locations, and the representational capacity
608 changes qualitatively from a continuous manifold to a discrete set of attractors.

609 Mean-field cross-inhibition is not merely a modeling simplification; it is a *symmetry hypothesis* — the
610 assumption that inter-population inhibition is blind to the spatial structure of the competing bumps. The
611 entire spectral architecture (DC critical mode, Goldstone protection) follows from this assumption. The
612 biological relevance of our results therefore depends on whether cortical cross-inhibition is approximately
613 mean-field at the level of competing item populations.

614 The Goldstone protection is specific to mean-field coupling. Spatially structured inhibition (e.g., lateral
615 inhibition that depends on the angular distance between bumps) would break the rotational symmetry and
616 couple dominance dynamics to positional dynamics. This is not merely a limitation but a falsifiable prediction:
617 mean-field cross-inhibition produces a DC critical mode (total activity competition), while spatially structured
618 cross-inhibition would produce a qualitatively different critical mode projecting onto spatial harmonics, with
619 a correspondingly different catastrophe type governing the transition. Whether biological cross-inhibition is
620 closer to mean-field or structured is an empirical question with spectral consequences.

621

4.5 Cusp Reduction and Connection to Stochastic Attractor Models

622 The 1D cusp potential $V(D) = \alpha D^4 + a D^2 + b D$ is the projection of the 96-dimensional dynamics onto
623 the critical eigenvector. The spectral analysis identifies this direction explicitly as the uniform/DC mode
624 with natural coordinate $D = \bar{r}^A - \bar{r}^B$, and provides what the 1D reduction cannot: the Goldstone modes
625 (requiring zero-mode regularization in any Kramers calculation), non-critical stability directions (setting the
626 high-dimensional prefactor), and the quantitative location of J_x^* (a free parameter in the 1D picture). The
627 cusp coefficient $a = -\lambda_{\text{dom}}\tau/2$ is determined by the spectral data, the quartic coefficient α is calibrated from
628 WTA fixed points (Section 3.5.5), and b is controlled by cue gain c . The fidelity of this cusp reduction is
629 validated numerically: the analytic quartic potential matches the numerically integrated one-dimensional
630 potential across the subcritical regime, with the best agreement at the onset coordinate $J_x^{\text{onset}} \approx 0.25$ where
631 the Kramers calculation is applied (Fig. 8).

632 This cusp landscape connects directly to the stochastic attractor models of Penny (2024), who modeled
633 maintenance as an SDE $dx = \beta g(x) dt + \sigma dw$ and showed that swap errors arise when “memory traces
634 diffuse away from their initial state and are captured by the attractors of other items.” Our spectral
635 analysis characterizes the landscape on which Penny’s stochastic dynamics unfold; the Kramers escape rate
636 $k \sim \exp(-\Delta V/\sigma^2)$ bridges the two descriptions. A prediction emerges: swap error rate should increase
637 continuously with maintenance delay (accumulated diffusion), but the rate of increase should exhibit a sharp
638 change near J_x^* where the barrier collapses quadratically.

639

4.6 Selection Versus Representation Failure

640 Neural recordings from monkey prefrontal cortex during multi-item working memory reveal that swap errors
641 can arise from misselection of correctly remembered items rather than representation failure (Alleman et

642 al., 2024). Both representations persist in the population, but the readout process selects the wrong item.
643 Complementarily, Gu et al. (2025) demonstrated in human psychophysics that attractor dynamics during
644 maintenance produce drift-driven biases whose direction is further steered by categorical decisions, linking
645 the operating point of cross-inhibition to systematic behavioral signatures beyond swap errors.

646 However, the failure mode decomposition (Section 3.5.4, Fig. 10) reveals that our single-stage model does not
647 produce selection failures: swap errors across the entire parameter range are representation failures, where
648 one bump collapses during maintenance. The model lacks a readout epoch where both representations coexist
649 and a selection process chooses between them. This is a structural limitation: the continuous dynamics of Eq.
650 (2) conflate maintenance and selection into one process. In the valley regime, the WTA competition resolves
651 correctly (low swap rate), not because both items survive for selection but because strong drive biases the
652 competition toward the correct network.

653 The Alleman et al. finding that swap errors in healthy subjects arise from misselection therefore constrains
654 model architecture: it requires at least a two-stage process with (i) a maintenance phase where both
655 representations are sustained (perhaps via weaker effective J_x or external stabilization) and (ii) a readout
656 phase where WTA competition selects one item. Our spectral analysis characterizes the landscape for stage
657 (ii); a complete account of selection failure would need to model stage (i) explicitly. A testable prediction
658 remains: conditions that increase effective cross-inhibition (distractor-rich environments, high cognitive load)
659 should shift swap errors from selection-type to representation-type, if the two-stage architecture holds.

660 4.7 A Shared Bifurcation Motif Across Competition Circuits

661 The pitchfork bifurcation we identify – attractors extinguished after merging with saddle points at high
662 cross-inhibition – has structural analogs in decision-making circuits. Roach, Churchland, and Engel (2023)
663 showed that in circuits with choice-selective inhibition, working memory attractors are extinguished after
664 merging with saddle points as ipsispecific inhibition increases. Disjoint neural groups with within-group
665 excitation and across-group inhibition exhibit group WTA dynamics, and the coexistence-to-WTA transition
666 occurs via saddle-point annihilation (Roach et al., 2023; Wong and Wang, 2006; Machens et al., 2005).

667 This structural similarity suggests the spectral separatrix may describe a structurally generic bifurcation
668 motif across neural circuits with competing stable states. Decision-making, attention, and working memory
669 all involve population competition, and the same codimension-1 normal form – pitchfork at critical coupling,
670 Goldstone protection of positional degrees of freedom, DC instability under mean-field coupling – governs the
671 coexistence-to-WTA transition in any exchange-symmetric competition circuit with mean-field-like inhibition.
672 However, we have demonstrated this structure only for the specific case of coupled ring attractors with mean-
673 field cross-inhibition; establishing genuine universality (shared critical exponents independent of microscopic
674 details) would require normal-form reduction arguments or analysis of additional model classes. The spectral
675 analysis presented here provides a template for such characterization.

676 4.8 Limitations

- 677 (i) The model uses rate neurons, not spiking neurons; the noise structure differs qualitatively. (ii) Mean-field
678 cross-inhibition is a symmetry hypothesis about the spatial structure of inter-population inhibition;
679 realistic inhibitory interneuron pools have spatial and temporal structure, as demonstrated by Roach et
680 al. (2023), where ipsispecific versus contraspecific inhibition creates qualitatively different attractor
681 landscapes. (iii) $N = 48$ is moderate; the Goldstone mode identification becomes cleaner at larger
682 N . (iv) The stochastic phase diagram uses additive Gaussian noise; biologically realistic noise is
683 multiplicative and state-dependent. (v) Our model conflates maintenance and selection into a single
684 dynamical process; the Alleman et al. (2024) finding that swap errors arise at the selection stage suggests
685 that a two-stage model (coexistence during maintenance, WTA competition at readout) may be more
686 biologically appropriate. (vi) The mapping from our neural space (96 dimensions) to behavioral feature
687 space (1D circular, as in Penny, 2024) requires assumptions about decoding that have not been derived
688 from first principles. (vii) The additive noise assumption may qualitatively affect the drive-sensitivity
689 finding: state-dependent (Poisson-like) noise could create a cue-driven cliff mechanism absent in our
690 model. (viii) The sharp pitchfork bifurcation at J_x^* is a symmetry artifact: even modest connectivity

691 heterogeneity ($\sigma \geq 0.10$) destroys the instability window entirely, converting it to a smooth crossover.
692 The deterministic bifurcation analysis characterizes the symmetric limit; biological relevance depends
693 on the stochastic landscape (barrier heights, basin depths) rather than the exact bifurcation structure.

694

695 5. Conclusion

696 We have presented the first complete spectral bifurcation analysis of competing ring attractors under mean-field
697 cross-inhibition. Six results stand:

- 698 1. **Existence threshold.** Coexistence has a sharp existence threshold at $J_x^{exist} \approx 0.36$, below which it is
699 a genuine fixed point and above which it does not exist.
 - 700 2. **Goldstone separation and pitchfork.** Goldstone modes are symmetry-protected and separate
701 cleanly from genuine instabilities. The first non-Goldstone eigenvalue crosses zero at $J_x^* \approx 0.3485$ in a
702 pitchfork bifurcation, creating the coexistence saddle.
 - 703 3. **DC critical mode.** The critical eigenvector projects maximally onto the uniform (DC) direction,
704 reflecting the mean-field character of cross-inhibition and predicting that the WTA instability concerns
705 total activity competition rather than spatial pattern rearrangement.
 - 706 4. **Stochastic phase diagram and failure mode decomposition.** Large-scale stochastic simulations
707 (128,000 trials) confirm the spectral prediction: swap errors emerge at $J_x^{onset} \approx 0.25$, drive strength
708 is secondary to cross-inhibition near the pitchfork, and a non-monotonic valley identifies a candidate
709 functional operating regime. A failure mode decomposition (24,000 additional trials) reveals that
710 swap errors are overwhelmingly representation failures across all parameter regimes; selection failures
711 are absent in this single-stage model, constraining the mapping to neural data where misselection is
712 observed.
 - 713 5. **Heterogeneity transforms the bifurcation.** Connectivity heterogeneity destroys the sharp pitchfork
714 entirely, converting it into an imperfect bifurcation — the razor-thin instability window is a symmetry
715 artifact of the clean model.
 - 716 6. **Kramers consistency analysis.** The barrier separating coexistence from WTA collapses quadratically
717 ($\Delta V \propto |\lambda_{dom}|^2$), reaching the noise-limited finite-horizon threshold at $J_x^{onset} \approx 0.25$ — quantitatively
718 consistent with the observed stochastic onset for a normal-form coefficient $\gamma \approx 0.22\text{--}0.36$ inferred from
719 the observed onset location.
- 720 Together, these results reframe the behavioral cliff in working memory as a spectral bifurcation phenomenon.
721 The model predicts a qualitative valley regime where cross-inhibition and encoding drive are balanced for
722 reliable WTA selection; the specific parameter range depends on model details, but the non-monotonic
723 structure between two failure modes is robust. The Goldstone modes protect memory content (bump
724 positions) from the competition over its fate (which bump survives), enforcing a separation of positional and
725 competitive dynamics that may be a design principle of working memory circuits. Importantly, connectivity
726 heterogeneity transforms the sharp pitchfork into a smooth crossover, dissolving the razor-thin instability
727 window ($\Delta J_x \approx 0.01$) entirely — biological circuits need not operate with such precision, and the valley regime
728 persists regardless of bifurcation type. The same normal form (codimension-1 pitchfork under exchange
729 symmetry) appears in decision-making circuits (Roach et al., 2023; Wong and Wang, 2006), suggesting a
730 structurally generic spectral motif for neural competition that warrants characterization across model classes.
- 731 The spectral analysis characterizes the landscape; the open question is what carves it. What learning rule,
732 homeostatic mechanism, or developmental process tunes effective cross-inhibition to the valley regime? If
733 the answer is Hebbian — if the attractors shape the landscape that shapes the attractors — then the spectral
734 structure described here becomes the foundation of a self-organizing theory of working memory capacity. The
735 valley is not merely where the system works; it may be where experience has placed it.

736

737 **References**

- 738 Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*,
739 27(2), 77-87.
- 740 Bays, P.M., Catalao, R.F.G., & Husain, M. (2009). The precision of visual working memory is set by
741 allocation of a shared resource. *Journal of Vision*, 9(10), 7.
- 742 Ben-Yishai, R., Bar-Or, R.L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex.
743 *PNAS*, 92(9), 3844-3848.
- 744 Burak, Y. & Fiete, I.R. (2012). Fundamental limits on persistent activity in networks of noisy neurons.
745 *PNAS*, 109(43), 17645-17650.
- 746 Compte, A., Brunel, N., Goldman-Rakic, P.S., & Wang, X.-J. (2000). Synaptic mechanisms and network
747 dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9), 910-923.
- 748 Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down
749 control of working memory capacity. *PNAS*, 106(16), 6802-6807.
- 750 Goldstone, J. (1961). Field theories with superconductor solutions. *Nuovo Cimento*, 19, 154-164.
- 751 Goldman-Rakic, P.S. (1995). Cellular basis of working memory. *Neuron*, 14(3), 477-485.
- 752 Gu, H., Lee, J., Kim, S., Lim, J., Lee, H.-J., Lee, H., Choe, M.J., Yoo, D.-G., Ryu, J.H.J., Lim, S., & Lee,
753 S.-H. (2025). Attractor dynamics of working memory explain a concurrent evolution of stimulus-specific and
754 decision-consistent biases in visual estimation. *Neuron*, 113(20), 3476-3490.
- 755 Funahashi, S., Bruce, C.J., & Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's
756 dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2), 331-349.
- 757 Hanggi, P., Talkner, P., & Borkovec, M. (1990). Reaction-rate theory: fifty years after Kramers. *Reviews of
758 Modern Physics*, 62(2), 251-341.
- 759 Khona, M. & Fiete, I.R. (2022). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*,
760 23, 744-766.
- 761 Kilpatrick, Z.P., Ermentrout, B., & Doiron, B. (2013). Optimizing working memory with heterogeneity of
762 recurrent cortical excitation. *Journal of Neuroscience*, 33(48), 18999-19011.
- 763 Kim, S.S., Rouault, H., Druckmann, S., & Jayaraman, V. (2017). Ring attractor dynamics in the Drosophila
764 central brain. *Science*, 356(6340), 849-853.
- 765 Kramers, H.A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions.
766 *Physica*, 7(4), 284-304.
- 767 Ma, W.J., Husain, M., & Bays, P.M. (2014). Changing concepts of working memory. *Nature Neuroscience*,
768 17(3), 347-356.
- 769 Poll, D.B., Nguyen, K., & Kilpatrick, Z.P. (2015). Sensory feedback in a bump attractor model of path
770 integration. *Journal of Computational Neuroscience*, 40(2), 137-155.
- 771 Seeholzer, A., Deger, M., & Gerstner, W. (2019). Stability of working memory in continuous attractor
772 networks under the control of short-term plasticity. *PLOS Computational Biology*, 15(4), e1006928.
- 773 Strogatz, S.H. (2015). *Nonlinear Dynamics and Chaos*. 2nd ed. Westview Press.
- 774 Tanaka, H. & Nelson, D.R. (2018). Non-Hermitian quasi-localization and ring attractor neural networks.
775 *Physical Review E*, 99(6), 062406.
- 776 Thom, R. (1972). *Structural Stability and Morphogenesis*. W.A. Benjamin.
- 777 Wei, Z., Wang, X.-J., & Wang, D.H. (2012). From distributed resources to limited slots in multiple-item
778 working memory: a spiking network model with normalization. *Journal of Neuroscience*, 32(33), 11228-11240.
- 779 Wimmer, K., Nykamp, D.Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal
780 cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3), 431-439.

- 781 Zeeman, E.C. (1977). *Catastrophe Theory: Selected Papers*. Addison-Wesley.
- 782 Zhang, W. & Luck, S.J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*,
783 453(7192), 233-235.
- 784 Alleman, M., Panichello, M.F., Buschman, T.J., & Johnston, W.J. (2024). The neural basis of swap errors in
785 working memory. *PNAS*, 121(33), e2401032121.
- 786 Penny, W.D. (2024). Stochastic attractor models of visual working memory. *PLOS ONE*, 19(5), e0301039.
- 787 Machens, C.K., Romo, R., & Brody, C.D. (2005). Flexible control of mutual inhibition: a neural model of
788 two-interval discrimination. *Science*, 307(5712), 1121-1124.
- 789 Roach, J.P., Churchland, A.K., & Engel, T.A. (2023). Choice selective inhibition drives stability and
790 competition in decision circuits. *Nature Communications*, 14, 147.
- 791 Ságodi, Á., Martín-Sánchez, G., Sokól, P., & Park, I.M. (2024). Back to the continuous attractor. *Advances*
792 in *Neural Information Processing Systems*, 37.
- 793 Wong, K.-F. & Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions.
794 *Journal of Neuroscience*, 26(4), 1314-1328.