Supplementary Material

A Developmental Stability Theory of Predictive Hierarchies: Why Some Perceptual Architectures Remain Stable Under Perturbation

Supplementary file 1: Formal Derivations, Robustness Analyses, and Plain-Language Companion

---

Contents

S1 From unstable directions to false percepts (an illustration, not a cortical model)

Instability is not yet hallucination; it is the capacity for it. Near an unstable equilibrium with n_u unstable directions, the local flow can be routed into as many as 2^{n_u} basins — distinct states the system can fall into — and since n_u grows with D (Figure S1A), the number of alternative, potentially false, percepts grows explosively with dimensionality. We illustrate this with a Hopfield network (Figure S1B): with two stored features it has exactly two attractors and no false ones; as features accumulate, spurious attractors proliferate and dominate.

We stress what this does and does not show. Hopfield networks are an idealisation with symmetric weights and binary units; they are not a cortical model. We use the simulation only to demonstrate the qualitative consequence of raising dimensionality — that false attractors multiply super-linearly — not to claim that cortex stores memories this way or computes these fixed points. The cortical claim is the weaker, robust one: whatever the microcircuit, adding independent represented features adds ways for a destabilised hierarchy to settle on something that was never there.

Figure S1: Higher dimensionality yields combinatorially many unstable and false states (qualitative illustration). (A) Number of self-amplifying directions ($\mathrm{Re}\,\lambda > 0$) grows roughly

linearly with D past threshold; higher gain steepens it. (B) Hopfield attractor count ($N = 80$, Monte-Carlo over random starts). At $p = 2$ features: 2 attractors, 0 spurious. As features accumulate, false attractors proliferate and outnumber the true stored ones. Illustrative of the consequence of dimensionality, not a model of cortex. *Alt text: Two plots showing unstable directions and spurious attractors increasing with dimensionality.*

S2 Circular inference: simulations of belief-space divergence

Main-text Eqs. (10)–(11) identify the circular-inference loop gain λ with the stability quantity $g\sigma\sqrt{D}$. Figure S2 shows the consequence by direct simulation: the same fixed quantum of sensory evidence, passed repeatedly around a recurrent loop, settles proportionately at low loop gain but climbs far past what the evidence warrants as λ approaches 1. Panel B plots the equilibrium certainty against λ and confirms that the divergence point coincides exactly with the dynamical stability boundary of main-text Figure 1 — the two are one event described in two coordinate systems, not two mechanisms.

Figure S2: Circular inference converts weak evidence into strong belief. (A) The same evidence, echoed through feedback: low loop gain settles proportionately; higher gain climbs far past what the evidence warrants. (B) Equilibrium certainty $L^\star = L_s/(1-\lambda)$ diverges as $\lambda \to 1$. Because λ rises with D and g, high-dimensional hierarchies reach fixation at lower dysregulation. *Alt text: Two plots showing belief certainty rising and diverging with loop gain.*

S3 Formalising what apical integration is

The key structural claim of apical integration theory is not that feedback exists — every predictive model has feedback — but that feedback acts in a particular *algebraic mode*. In a healthy two-point neuron, apical (context) input does not by itself drive output; it modulates the *gain* applied to basal (feedforward) drive, and this conditionality is what BAC firing implements. Writing b for basal drive, a for apical drive, and $\varphi(x) = \lfloor x \rfloor_+^n$ for the power-law transfer function:

$$r_{\text{intact}} = \varphi(g\,[b(1+\kappa a)])\,, \qquad \text{context is multiplicative.} \tag{S1}$$

Apical decoupling is then not simply "less feedback." It is a change of mode: as the apical compartment loses its conditional coupling ($\kappa \downarrow$), disinhibition and loss of compartmental isolation allow apical input to reach the soma directly, adding an ηa term that drives output on its own:

$$r_{\text{decoupled}} = \varphi\big(g\,[b(1+\kappa' a)] + \eta a\big)\,, \qquad \kappa' < \kappa,\ \eta > 0. \tag{S2}$$

S4 The biophysical bridge: from NMDAR hypofunction to the algebra

Section S3 treats "apical decoupling" as a primitive. That is the weak point, and it is worth repairing directly, because the objection is not that the algebra is wrong but that the algebra is *unmoored*: nothing yet connects it to cellular biology. Schizophrenia has substantial evidence for several classes of abnormality — dopaminergic and serotonergic alterations, enlarged ventricles, reduced prefrontal activation, altered gamma synchrony — none of which is present in all patients. We build the bridge

to NMDA receptor hypofunction not because it is the only well-supported finding, but because it is the one that acts at the cellular locus this model is about (the apical compartment and the interneurons gating it), and because it is the one from which the coefficients κ and η can be *derived* rather than assumed. Other established abnormalities enter the framework elsewhere: dopaminergic alterations map onto the gain g, and structural or connectivity findings onto D and σ_{struct}.

Take the canonical reduced two-point pyramidal neuron (Larkum, 2013; Phillips & Silverstein, 2003). The apical tuft supports a regenerative NMDA/Ca^{2+} dendritic spike; because the NMDA conductance is Mg^{2+}-blocked at rest, the spike ignites only when the tuft is sufficiently depolarised, and back-propagating somatic action potentials supply that depolarisation. Writing the somatic drive as

$$D(b, a) = b + \eta_0 a + A\,\varsigma\left(\frac{a - \theta + \beta b}{s}\right), \qquad \varsigma(z) = \frac{1}{1 + e^{-z}} \tag{S3}$$

the biophysical meaning of each term is fixed and not free: A is the dendritic-spike amplitude, set by apical NMDA conductance; β is how strongly somatic drive primes ignition, i.e. the efficacy of the back-propagating action potential, also NMDA-dependent; θ is the ignition threshold, set by tonic dendritic inhibition from SST interneurons targeting layer 1; and η_0 is a small passive apical leak.

We model the established cellular effects of NMDA receptor hypofunction of severity $1 - f$ at two sites — pyramidal dendrites and the inhibitory interneurons that control them, the latter being the more sensitive population (Gonzalez-Burgos & Lewis, 2012; Lisman et al., 2008) — as follows:

$$\underbrace{A \to fA, \quad \beta \to f\beta}_{\text{(i) pyramidal apical NMDA}}, \qquad \underbrace{\theta \to \theta - \Delta\theta(1 - f)}_{\text{(ii) SST hypofunction}\Rightarrow\text{dendritic disinhibition}} \tag{S4}$$

The first two substitutions are direct consequences of reducing an NMDA-dependent conductance. The third is not: it encodes the further hypothesis that reduced NMDA drive to dendrite-targeting interneurons translates into a lowered effective apical-spike threshold. That is physiologically plausible and is the standard reading of interneuronal NMDAR hypofunction, but it is not mathematically entailed by hypofunction alone. The resulting relationship is therefore conditional on these biophysical assumptions, and we treat $\Delta\theta$ as a parameter whose magnitude is itself an empirical question.

We then *measure* κ and η from Eq. (S3) — they are never assumed — using the 2×2 factorial decomposition that defines them:

$$c_1 = \frac{D(b^*, 0) - D(0, 0)}{b^*}, \quad \eta = \frac{1}{c_1}\frac{D(0, a^*) - D(0, 0)}{a^*}, \quad \kappa = \frac{1}{c_1}\frac{[D(b^*, a^*) - D(b^*, 0)] - [D(0, a^*) - D(0, 0)]}{a^* b^*} \tag{S5}$$

The result (Figure S3) is the required bridge. As f falls, κ collapses monotonically (-103% at $f = 0.25$): losing apical NMDA removes both the spike amplitude and the bAP priming, so somatic drive no longer recruits apical amplification and context ceases to multiply. Simultaneously η rises ($+198\%$ at $f = 0.25$): dendritic disinhibition lowers the ignition threshold, so apical input alone — with no feedforward evidence at all — now crosses it. One lesion, two opposite coefficient changes, exactly the multiplicative-to-additive shift Eqs. (S1)–(S2) posit.

Three consequences deserve emphasis. First, the inference is now a valid deduction from an evidenced premise: *if* NMDAR hypofunction is present, *then* the apical signature follows — so the hypothesis inherits the existing evidence base for NMDA receptor hypofunction in schizophrenia rather than requiring an independent one of its own. Second, the bridge is itself falsifiable in a new

way: it predicts a *non-monotonic* dose–response for the leak, maximal at moderate NMDA antagonism, which is directly testable and which no purely descriptive account predicts. Third, it predicts a *double dissociation* — pyramidal-selective NMDA loss should reduce contextual modulation *without* producing leak, so observing reduced CMI with no leak would indicate pyramidal-restricted pathology rather than refuting the framework.

Figure S3: The biophysical bridge (model-conditional). Within this reduced model, $g_{\mathrm{NMDA}} \downarrow$ yields $\kappa \downarrow$, $\eta \uparrow$. (A) In the intact model, apical drive alone barely depolarises the soma until very large; the same drive accompanied by basal input ignites the dendritic spike early — context is conditional. Under NMDA hypofunction apical drive alone now drives the soma. (B) Coefficients measured from Eq. (S5) as hypofunction increases: κ falls monotonically to zero while η rises severalfold. Note η is non-monotonic, peaking at moderate hypofunction. (C) Dissociation of the two sites: pyramidal NMDA loss alone reduces η. The additive leak requires interneuron-mediated dendritic disinhibition. *Alt text: Three panels showing dendritic coefficients changing with NMDA hypofunction.*

S5 Why the leak must be non-monotonic, and how far the result generalises

An analytic optimum. The non-monotonicity is not a numerical accident; it follows from a competition between the two sites. Lowering f has opposing effects on the leak: disinhibition lowers the threshold θ, letting apical input cross it (pushing η up), while loss of pyramidal NMDA shrinks the spike amplitude A (pulling η down). Writing the numerator of η from Eq. (S3) at $b = 0$ as $\eta_0 a^* + f A_0 \varsigma(u(f))$ with $u(f) = (a^* - \theta_0 + \Delta\theta(1-f))/s$, and using $\varsigma' = \varsigma(1-\varsigma)$,

$$\frac{\partial \eta}{\partial f} \propto \varsigma(u) - f\varsigma'(u)\frac{\Delta\theta}{s} = 0 \iff \varsigma(u^*) = 1 - \frac{s}{f\Delta\theta} \tag{S6}$$

so the leak is maximal when the apical gate has opened to the fraction $1 - s/(f\Delta\theta)$ of its range, and an interior optimum exists at all if and only if

$$f\Delta\theta > s \tag{S7}$$

i.e. if and only if disinhibition per unit hypofunction exceeds the sharpness of the gate. Equation (S6) tracks the numerically located leak optimum across randomly drawn parameter sets with $r = 0.86$ (Figure S4A). This is a genuinely new prediction: there is an optimal degree of NMDAR hypofunction for maximal ascending leak, and Eq. (S7) says when it should be observable at all.

Robustness. To test whether the qualitative result depends on our choice of parameters, we drew 10,000 parameter sets from physiologically plausible ranges (spike amplitude, bAP priming, threshold, gate sharpness, passive leak, disinhibition magnitude, and both probe levels) and recomputed the coefficients over the whole hypofunction range. The qualitative pattern is highly robust (Figure S4B): κ falls with hypofunction in 98.4% of sets, η rises above its intact value in 100%, and η shows an interior (non-monotonic) peak in 100%. The predicted optimum sits at moderate hypofunction (median $f = 0.57$; Figure S4C).

We can therefore say something stronger than before, and still only what is warranted: the *direction* of both coefficient changes, and the *existence* of an interior leak optimum, are structural properties of this model class rather than artefacts of a tuned parameter set. This establishes the proposed relationship *within the model*; its generality across alternative biophysical implementations

remains an empirical and model-comparison question. A multi-compartment conductance-based model with explicit Mg^{2+} kinetics, or one in which interneuronal hypofunction acts primarily on gain rather than threshold, could in principle behave differently.

Figure S4: The coefficient changes are structural within this model class. (A) The analytic optimum condition $\varsigma(u^*) = 1 - s/(f\Delta\theta)$ predicts the numerically located leak peak across randomly drawn parameter sets ($r = 0.86$). (B) Across 10,000 physiologically plausible parameter sets: κ falls with NMDAR hypofunction in 98.4%, η rises in 100%, and η is non-monotonic in 100%. (C) Distribution of the NMDAR function f at which leak is maximal: the optimum lies at moderate hypofunction (median $f = 0.57$). *Alt text: Scatter plot, bar chart and histogram showing robustness across parameter sets.*

S6 Heterogeneity as a measurable mixture: stratification rather than a single mechanism

If schizophrenia is a heterogeneous syndrome, the useful question is not "is the apical mechanism *the* cause?" but "in what fraction of patients is it operative, and can that fraction be measured?" The identifiability analysis of §S8 supplies the instrument: four candidate mechanisms — intact-like, apical decoupling, gain increase, and E/I noise — occupy separable regions of the three-measure space, so a mixed cohort is a mixture over those regions and its composition can be estimated.

We tested this directly. Simulating a heterogeneous cohort with true proportions (15%, 40%, 25%, 20%) across the four mechanisms, measuring the three observables per participant with realistic trial noise, and classifying against reference centroids, the recovered proportions were (18%, 41%, 25%, 16%) — the apical fraction, which is the clinically decisive one, was recovered closely (0.41 vs. 0.40), as was the gain fraction (0.248 vs. 0.250) (Figure S5B). The residual error is confined to the intact-versus-noise boundary, which these three measures separate weakly — an honest limitation that additional measures (e.g. trial-to-trial response variability) would be required to resolve. A power analysis indicates that a cohort of $N \approx 100$ suffices to estimate the apical-subtype fraction to within $\pm 10\%$ (Figure S5C).

Two consequences follow. First, the framework's predictions become *subtype-specific* rather than syndrome-wide: only the apical subgroup should show the joint CMI/leak-slope signature, only that subgroup should show the non-monotonic dose–response of §S5, and only that subgroup should respond preferentially to interventions targeting apical/NMDA discipline rather than gain. Second, the same instrument provides a principled answer to why previous studies of contextual modulation in schizophrenia have produced mixed results: if the apical subtype is 40% of an unselected cohort, a study measuring CMI alone in an unstratified sample would find a diluted, inconsistent effect — which is close to what the literature reports. The framework thus explains an existing pattern of inconsistent findings as a predictable consequence of unstratified sampling, and specifies the design that would resolve it.

Figure S5: Two refinements: the leak is conditional, and heterogeneity is measurable. (A) The conjunction condition: an ascending leak requires both decoupling ($\eta > 0$) and apical drive sufficient to cross the somatic threshold ($a > \theta_s/\eta$). Modest decoupling under ordinary expectation produces loss of contextual amplification (disorganisation) without false percepts. (B) Recovery of mechanism proportions in a simulated heterogeneous cohort ($N = 200$, 60 replications). The apical and gain fractions are recovered closely. (C) Power analysis: $N \approx 100$ suffices to estimate the apical-subtype fraction to $\pm 10\%$. *Alt text: Three panels showing leak conditions, mechanism*

recovery and required sample size.

S7 Two theorems that make the hypothesis discriminating

Formalisation earns its keep only if it forbids something. Two elementary results do the work; both are proved directly from Eqs. (S1)–(S2). Define the contextual modulation index and the ascending leak:

$$\mathrm{CMI}(b,a) = \frac{r(b,a) - r(b,0)}{r(b,0)}, \qquad L(a) = \frac{r(0,a)}{r(b_0,0)} \tag{S8}$$

Result 1 (gain-invariance). *CMI is invariant to the global gain g.* Because φ is a power law, $r \mapsto g^n r$ under a gain change, and the factor cancels in the ratio (S8). Hence no change in PV-mediated gain — the mechanism most often proposed as the alternative — can produce a change in CMI. Any observed reduction in CMI must come from the coupling, not the gain.

Result 2 (leak requires additivity). *Under purely multiplicative coupling the ascending leak is identically zero:* setting $b = 0$ in Eq. (S1) gives $r = \varphi(0) = 0$ for every a, so $L(a) \equiv 0$. A non-zero leak that grows with context strength is therefore impossible without an additive apical term $\eta > 0$. Random E/I noise also produces spurious responses, but noise is by construction independent of a, so it yields a leak that is non-zero but flat in a: $dL/da = 0$. Hence the slope dL/da, not the leak itself, is the diagnostic quantity.

Together these give a signature that belongs to apical decoupling alone (Figure S6): **reduced CMI and positive leak-slope simultaneously**. A gain lesion cannot lower CMI (Result 1); a noise lesion cannot produce a context-dependent leak (Result 2); only the multiplicative-to-additive shift produces both. Because CMI and L are both ratios and hence gain-blind, a third measurement — absolute responsivity $r(b_0, 0)$ — is required to separate an intact circuit from a pure gain increase.

Figure S6: Apical integration formalised, and its unique signature. (A) Input–output curves. Intact: context multiplies drive, and with no feedforward drive there is no output. Decoupled: the additive term produces output with *no evidence present.* (B) CMI is mathematically invariant to global gain — the PV-gain curve lies exactly on the intact curve — but falls under apical decoupling. (C) Ascending leak is identically zero under multiplicative coupling, rises with context under apical decoupling, and is flat in a under E/I noise. *Alt text: Three panels contrasting intact and decoupled neuron input-output behaviour.*

S8 Identifiability: the experiment that would settle it

A unique signature is only useful if it survives measurement noise. We therefore simulated cohorts of virtual participants under four mechanisms — intact, apical decoupling, PV-gain increase, and E/I noise increase — with realistic trial-to-trial variability and heterogeneous parameters, measured the three observables, and attempted to *recover* the generating mechanism (leave-one-out nearest-centroid classification). Recovery accuracy was 90%, with apical decoupling recovered at 100% and no confusion whatever between apical decoupling and either rival. The residual error is entirely the intact-versus-noise distinction, which these three measures separate only weakly — an honest limitation of the design rather than of the theory.

Crucially, all three observables map onto paradigms that already exist in human work. CMI corresponds to contextual modulation of perception — contour integration, surround suppression, perceptual closure — which is measurable psychophysically and is already known to be impaired

in schizophrenia. The leak slope corresponds to the rate of percepts reported when *no* stimulus is present as a function of learned cue strength, which is exactly what conditioned-hallucination paradigms measure (Powers, Mathys, & Corlett, 2017). Responsivity corresponds to evoked response amplitude to a driven stimulus. Nothing here requires a new technology; it requires running the three measures *in the same participants* and reading the joint pattern.

The logical architecture. Each link has a different status:

1. **What is known.** Schizophrenia → NMDAR hypofunction. Supported, but indirectly.
2. **What the model derives.** NMDAR hypofunction → $\kappa \downarrow$, $\eta \uparrow$. Conditional on the biophysical assumptions of §S4; robust across 10,000 parameter sets within this model class (§S5).
3. **What follows mathematically.** $\kappa \downarrow$, $\eta \uparrow \rightarrow$ CMI $\downarrow$ and $dL/da > 0$. Proved (Results 1–2); no additional assumptions.
4. **What remains unknown.** Whether patients actually exhibit CMI $\downarrow$ together with $dL/da > 0$. Not established by any existing study.
5. **How to find out.** Measure the three observables in the same participants, under a dose manipulation.

Two further consequences follow. First, DST's own parameters inherit the same identifiability: σ is estimated by CMI, the leak slope estimates the additive component that produces $\lambda > 0$, and responsivity estimates g — so the framework's terms are not free parameters but measurable quantities. Second, the design doubles as a treatment-stratification tool: patients whose deficit loads on the leak slope should respond to agents restoring apical/NMDA discipline, whereas those loading on responsivity should respond to gain reduction (D2 blockade), a prediction that is directly testable in a trial.

S9 Multisensory reliability weighting: predicted volatility

The stiffness assay described in main-text §3.9 predicts a specific ordering not in the *mean* weight assigned to each modality — which optimal cue combination fixes for everyone — but in its *trial-to-trial variance*. A hierarchy operating with a large stability margin updates its weights sluggishly and consistently; one operating near $S = 0$ updates them erratically, because small precision fluctuations produce large excursions when the leading eigenvalue sits close to zero. Figure S7 simulates this under a precision-weighted update rule for the three predicted groups. Volatility separates them by an order of magnitude while the mean weight does not separate them at all, which is why the variance rather than the mean is the diagnostic quantity.

Figure S7: Predicted volatility of sensory reliability weighting. Simulated trial-by-trial weight estimates under a precision-weighted update in an audiovisual spatial-localisation task. A stiff prior (predicted for the congenitally blind hierarchy) yields low volatility (Var = 0.001); a moderately flexible one (healthy sighted) yields Var = 0.002; a loosely controlled one (predicted for high-risk and late-blind individuals) fluctuates strongly (Var = 0.010). Volatility, not mean, is diagnostic. *Alt text: Time series of estimated sensory weights showing three volatility levels.*

S10 Mixed parameter regimes

The stability margin is a *scalar* computed from a *vector* of couplings, so distinct parameter profiles can yield the same margin while producing quite different phenomenology. Four regimes follow directly.

Table S1: Mixed parameter regimes — how one scalar margin accommodates both phenotypes.

Regime	g	σ_S, σ_C	σ_T, σ_Σ	D	Phenotype
1. Autism only	high	low	normal/low	any	rigid, illusion-resistant; never crosses
2. Autism → psychosis	rises late	low	high	high	autistic first, psychotic later
3. Early-onset SZ with autistic features	high early	low	high	high	thin margin; childhood crossing
4. Typical schizophrenia	rises in adolescence	high	high	high	adolescent crossing

Regime 1 — autism without psychosis. High g (inflexible high error precision) with uniformly low coupling: σ_S, σ_C low and σ_T, σ_Σ normal or low. The product $g\sigma\sqrt{D}$ stays below 1 for any physiological gain, so the hierarchy is rigid, detail-focused and illusion-resistant, and never crosses.
Regime 2 — autistic development with later psychotic symptoms. The clinically important case. Here the σ components move in opposite directions: the integrative components (σ_S, σ_C) remain low — preserving the autistic phenotype of social-inference difficulty, weak central coherence and illusion resistance — while σ_T and/or σ_Σ are elevated, and D is high. Such a person sits inside the stable region during childhood but carries a structural term that is large in the temporal and spatial channels. The normal adolescent rise in g then multiplies against that term, and the trajectory crosses. DST thus predicts a specific developmental signature: autistic features first, psychotic features later, with illusion-resistance preserved into the psychotic phase — a combination puzzling for a diametric model but the expected consequence of a split σ-vector meeting a later gain rise.
Regime 3 — early-onset schizophrenia with autistic features. An early, broad neurodevelopmental perturbation raises D and several σ components at once while also impairing the integrative ones. This produces a thin margin (hence childhood or early-adolescent onset) alongside autistic-like social and integrative deficits — the profile most often seen in very-early-onset cases with premorbid neurodevelopmental abnormality.

The general point is that DST is not a one-dimensional continuum with autism at one end and schizophrenia at the other. It is a product of a gain and a coupling *vector*, and co-occurrence is what happens when the components of that vector are pushed in opposite directions — which is precisely why the two conditions can share a genetic and developmental substrate while presenting as opposites on illusion tasks.

S11 Plain-language companion to every equation

This section reproduces, in full and in order, the plain-language commentary that accompanies each equation of the main text. Every equation in the theory can be read without any mathematics by following this companion alongside the main article. Cross-references are to main-text equation numbers unless prefixed "S".

On Equations (1)–(2): the hierarchy as a stack of committees

Think of the brain as a stack of committees. Each sends its best guess downward and receives complaints (error signals) from below. Equation (1) is a single "unhappiness score" for the stack: large when complaints are loud and the committee is confident about them. Equation (2) says every committee keeps adjusting its guess to lower the unhappiness. The crucial knob is Π, the confidence on each complaint: turn it up and a complaint reshapes beliefs violently; turn it down and the same complaint is ignored. Everything below is about what happens when that knob is mis-set in a hierarchy built in a particular shape.

On Equation (4): mesh versus highways

Equation (4) splits the brain's wiring into two kinds. Most of it is a dense mesh of many weak connections (the "bulk"); a little of it is a handful of very strong highways — hub neurons and dominant feedback loops. Our stability rule $g\sigma\sqrt{D} < 1$ is a statement about the mesh: instability builds up from many small loops, so it grows with how many features D there are. But if a single strong highway is what tips the brain over, then the number of features stops mattering and only that highway's strength does. We are betting on the mesh — and we are telling you exactly which experiment (does instability track feature-count, or track one dominant pathway?) would prove us wrong. Saying where a theory breaks is how you know it is a theory and not a story.

On Equations (5)–(6): the room that screams

Imagine a room of microphones and speakers. One pair is fine. Add more and more, all faintly wired together, and eventually the room erupts into a feedback shriek — not because any one microphone got louder, but because there are now so many loops that some combination is guaranteed to run away. Equation (5) is exactly this: D is the number of pairs (features), g is the master volume (gain), and σ is how tangled the wiring is. The room screams the instant $g\sigma\sqrt{D}$ crosses 1. A low-dimensional channel like touch is a small room with few loops; a high-dimensional one like vision is a stadium of loops, perpetually one notch from disaster.

On Equation (7): measuring dimensionality

"Dimensionality" sounds vague until you measure it the way physicists measure the effective size of anything wobbling: count the directions it actually moves in, weighting big movements more than tiny ones. Equation (7) does this for brain activity. If a region's activity is really just one thing going up and down, $D_{\text{eff}} \approx 1$; if it independently varies in fifty ways, $D_{\text{eff}} \approx 50$. So "how many dimensions does vision have?" has an answer — you record the cortex and compute the participation ratio D_{eff} of the population covariance — and the theory's claim becomes an experiment rather than an assertion.

On Equations (8)–(9): the tug-of-war

Whether a prediction error reshapes your beliefs depends on a tug-of-war: how confident you are in the incoming evidence (Π_e) versus how confident you are in what you already believe (Π_p). Equation (8) is just that tug-of-war written as a fraction. You can make evidence win in two opposite ways: shout the evidence louder (raise Π_e — what dopamine does) or stop trusting your existing beliefs (lower Π_p — what psychedelics do). Both hand more control to the bottom-up stream, which is why both can destabilise a richly wired hierarchy — but because one adds confidence and the other removes it, they should feel different and behave differently, and the theory will hold them to that.

On Equations (10)–(11): the echo chamber

Equation (10) is an echo chamber. You make an assertion; the room echoes back a fraction λ of it; the brain counts that echo as a second, independent voice agreeing, and so raises its confidence in the original assertion — which is then re-emitted more strongly on the next cycle. The louder re-emission is not a new utterance from outside; it is the same belief, now held with greater certainty because its own echo was mistaken for corroboration. Equation (11) is the important part: the reflectiveness of the walls is not a knob we get to set — it *is* the very quantity $g\sigma\sqrt{D}$. So the point at which a whisper of evidence builds into unshakeable conviction ($\lambda = 1$) is the exact same point at which the perception itself becomes unstable. The runaway belief and the runaway dynamics are one event, described twice.

On Equation (12): naming the cables

Saying "the wiring is tangled" (σ) is only useful if you can say why. Equation (12) unpacks the tangle into separate cables you can inspect one at a time: events bound too loosely in time (σ_T), guesses about other people's hidden intentions (σ_S), objects grouped wrongly in space (σ_Σ — the perceptual disorganisation clinicians recognise), words and ideas linked by loosened meaning (σ_M — the substrate of thought disorder), and the brain's built-in cross-talk (σ_C). The list is deliberately open: add a cable if you find one. Because each is separately measurable, σ cannot be dismissed as a fudge factor, and the theory can be tested several independent ways.

On Equation (13): the safety margin

Equation (13) is the whole of the general theory on one line. S is a safety margin. Two things eat into it: how elaborately wired the perceptual world is ($\sigma\sqrt{D}$, largely fixed in childhood) and how loud error-confidence is turned up (g, set by state and chemistry). Fragility appears only when the product exceeds one. Nothing here mentions any diagnosis — yet dropping real conditions onto this single map explains a surprising amount.

On Equation (14): what blindness does and does not protect against

Congenital cortical blindness is hypothesised to be protective against schizophrenia, and existing evidence supports this hypothesis. Blindness does not improve mental health in general, and the specificity here needs care. In congenital or early blindness, rates of depression and anxiety are not elevated — some studies of blind children and adolescents report comparable or even lower rates

than sighted peers. In acquired, later-life vision loss they clearly are elevated, but that elevation is understood as a reaction to profound life change — relearning daily tasks, altered work, relationships, and independence — rather than a direct perceptual consequence of blindness, and it is not the cross-modal hierarchical instability of schizophrenia. So DST claims something narrow: congenital cortical blindness selectively spares the $g\sigma\sqrt{D}$ instability that produces schizophrenia, while mood and anxiety — which do not depend on that product, and which rise after late vision loss for reactive reasons — are left untouched by the protective mechanism.

On Equation (15): the camera shutter

Does vision cause a wide binding window, or does schizophrenia? Both — but at different times and for different reasons. Think of the window as a camera's shutter speed. A camera built for a dim, slow scene (vision) ships with a slow shutter — good engineering, set at the factory (development). Separately, a fault in the electronics (dopamine) can smear the shutter at runtime (illness). The factory setting decides how much damage the runtime fault does: a camera shipped with a fast shutter (the blind hierarchy) barely smears even when faulty. This predicts a clean dissociation — congenitally blind people should have a narrow baseline TBW that resists widening, while sighted high-risk individuals have a wide baseline that widens further.

On Equation (16): why a voice seems to come from outside

Ask why a voice seems to come from outside. The brain never receives signals stamped "world" or "me" — it has to work it out, and it uses three clues: is the signal unusually sharp at the sensory level (the world is sharp, imagination is vague)? does it have a location in space? and did I predict it, as I predict my own speech and movements? Equation (16) is just those three clues added up. In psychosis a false state arises low in the sensory system with the volume turned up, which mimics clue one; it gets pinned to a place, which supplies clue two; and it was never predicted by any plan of mine, which removes clue three. All three point the same way, so the brain concludes — reasonably, on its own evidence — that something out there is speaking. Move the same false state one storey up and it has no sharpness and no location, so it stays inside; but if the "did I do that?" check still fails, it becomes a thought that feels inserted. Disorganisation is the separate case where many states wobble at once and nothing binds cleanly.

On Equations (S1)–(S2): the volume knob and the second microphone

Equations (S1)–(S2) say something sharp in one symbol. In a healthy neuron, context is a volume knob: it can amplify what the senses deliver, but turn the knob with no sensory input and nothing comes out — $b = 0$ gives $r = 0$ no matter how strong the expectation. That single algebraic fact is what stops expectations from being mistaken for evidence. Apical decoupling changes the knob into a second microphone: context now adds directly, so a strong enough expectation produces output with no sensory input at all. This is why the same lesion causes two symptoms that look unrelated. Losing the volume knob means real objects are no longer bound and amplified by their context — that is disorganisation. Gaining the second microphone means expectation alone can generate a percept — that is hallucination. One change of arithmetic, two clinical faces.

On Equations (S3)–(S5): what the bridge buys

This is the step that was missing, and it is worth saying plainly what it buys. Before, we asserted that the apical mechanism breaks in schizophrenia — an assumption with no direct evidence. Now we start from something the field already has substantial evidence for (NMDA receptors underperform in schizophrenia) and show that the breakage *follows*. The gate that decides whether context can act alone is *made of* NMDA receptors, so weakening them necessarily changes the arithmetic. And the two effects pull in opposite directions for a clean reason: the receptors on the pyramidal cell are what let context *amplify* real input, so losing them weakens amplification; the receptors on the inhibitory neurons are what *hold the gate shut*, so losing them lets context through unaccompanied. Turn down one dial, and context simultaneously gets worse at its proper job and better at the job it should never do.

On Equations (S6)–(S7): why worse is not always worse

Why should a worse lesion ever produce fewer false percepts? Because two things happen at once as NMDA receptors fail. Removing inhibition unlocks the gate, so expectation gets through more easily — that pushes hallucination-like leak up. But the leak is itself carried by an NMDA-dependent dendritic event, so past a point there is not enough current left to generate it — that pulls the leak back down. The leak therefore peaks somewhere in the middle, and Eq. (S6) says exactly where. This is why the framework predicts that a moderate NMDA blockade should be more psychotogenic than a very heavy one, which is a testable dose–response claim rather than a story.

On §S6: measuring the mixture

This reframes what the theory is for. Psychiatry has spent decades asking which single abnormality "explains schizophrenia," and has repeatedly found that every candidate is present in some patients and absent in others. Rather than adding another candidate, this framework offers a way to *measure the mixture*: run three behavioural tests in a hundred people and estimate what fraction of them have a circuit in each of four mechanistically distinct states. That is a different kind of claim — not "this is the cause of schizophrenia," but "here is an instrument that says how much of each mechanism a given cohort, or a given patient, actually has." For a syndrome now understood as heterogeneous, that is the more useful object, and it is directly actionable: the subtype assignment predicts which treatment axis (g versus σ) should work for that individual.

On §S7–S8: what formalisation can and cannot do

This is the part we want to be scrupulous about. We have not shown that apical integration is broken in schizophrenia — we cannot, from a desk. What we have shown is that the idea is not vague: it makes an arithmetic commitment, and that commitment forbids specific combinations of results. If someone measures contextual modulation, conditioned hallucinations, and evoked responsivity in the same patients and finds reduced context sensitivity with a *flat* leak-slope, apical decoupling is wrong and a noise or gain account is right. If they find reduced context sensitivity *together with* a leak that grows as expectation grows, then no gain or noise mechanism can account for it, and apical decoupling is the only one of the four candidates left standing. That is what it means to make a theory decidable, and it is the strongest honest claim available: not proof, but a decisive experiment that anyone can now run.

If you remember one thing: the mathematics here is not being used to prove that something is true about the brain. It is being used to make an idea *sharp enough to be tested* — to say

precisely what should be seen if the idea is right, precisely what should be seen if it is wrong, and exactly which measurements would tell the difference. A theory that explains everything and forbids nothing is not much use. The work in this section is about making sure this one forbids things.

On Equations (18)–(20): the join that makes one argument

This is the join that makes the whole paper one argument rather than four. Equation (18) says a neuron's influence on its neighbours has two parts: one that only operates when real sensory evidence is present, and one that operates regardless. Ask what is left when you remove all sensory input — what the network can whisper to itself in an empty room — and only the second part survives. That leftover is exactly the "tangled wiring" σ that the stability criterion has been about from the start. So the dendritic leak and the coupling term are not two ideas; they are the same quantity measured at two levels of description.

On Equation (21): the weir and the riverbed

Picture the margin S as a water level above a weir. Puberty lowers the level (more dopamine, a fully built visual hierarchy); the age at which it first dips below the weir is the age of onset — early if the person began with little headroom (childhood schizophrenia), late if they began with a lot. What happens next depends on whether going under the weir erodes the riverbed. If not, the level rises again and the person recovers. If each flood cuts the bed deeper (maladaptive plasticity raising $\sigma\sqrt{D}$), the baseline sinks, floods come more easily, and function declines. Treating early — keeping the person above the weir — is how you stop the riverbed from eroding.

Supplementary References

All references cited in this Supplementary Material appear in the main article's reference list.