Network Fragility: A Unified Computational Framework for Depression, Antidepressant Mechanisms, and Bipolar Progression

Cheung, Ngo

Cheung Ngo Medical Limited
Hong Kong, China

Cheung, N. (2026). Network Fragility: A Unified Computational Framework for Depression, Antidepressant Mechanisms, and Bipolar Progression. Zenodo. https://doi.org/10.5281/zenodo.18316724

Foreword

The search for mechanistic understanding of mood disorders has long been hampered by the complexity of the human brain and the limitations of clinical observation alone. Computational modeling offers a unique bridge: it allows us to isolate variables, test hypotheses at scale, and reveal dynamics that are otherwise hidden in the noise of biological systems.

This volume brings together a series of interconnected studies united by a single core idea—that excessive synaptic pruning, driven by stress or episode recurrence, pushes neural networks past a critical threshold of fragility. From this foundation, the models explore how different antidepressant mechanisms interact with that fragile state: some merely stabilize remaining connections, others promote new growth, and a select few rebuild structural resilience capable of withstanding future stress. The framework further accounts for clinical phenomena long considered puzzling—rapid onset of ketamine's effects, the risk of manic switches, cumulative scarring across mood episodes, and the kindling-like progression observed in bipolar disorder.

Perhaps most provocatively, the final contribution draws a parallel between psychiatric kindling and the progressive adversarial sensitization that emerges during the alignment of large language models—a reminder that principles of network destabilization may transcend biological substrates.

My hope is that these models do more than describe; I hope they provoke new experiments, refine treatment selection, and ultimately contribute to therapies that do not merely manage symptoms but repair the underlying architecture of vulnerability.

N. Cheung

January 2026

Table of Contents

Chapter 1

Excessive Synaptic Pruning Induces Network Fragility That Is Rescued by Simulated Neuroplasticity: A Computational Model of Vulnerability and Recovery in Major Depressive Disorder

Cheung, Ngo

Cheung, N. (2026). Excessive Synaptic Pruning Induces Network Fragility That Is Rescued by Simulated Neuroplasticity: A Computational Model of Vulnerability and Recovery in Major Depressive Disorder. Zenodo. https://doi.org/10.5281/zenodo.18209002

Abstract

Background: Major depressive disorder (MDD) is highly heritable and polygenic. Recent work points to abnormal developmental synaptic pruning—rather than primary glutamatergic failure—as the central liability, with later-life plasticity limits and reduced cognitive reserve amplifying risk under stress. Although computational studies have examined pruning in other psychiatric settings, few models capture

depression-specific fragility or test whether a treatment that boosts synaptogenesis can reverse that state.

Methods: We built an over-parameterised feed-forward network (~400 000 weights) and trained it on a noisy four-class task, representing exuberant early connectivity. Excessive adolescent pruning was mimicked by removing 95 % of weights in every layer. Robustness was measured on clean and perturbed inputs. To model rapid-acting antidepressants, half of the deleted connections were reinstated at small random strengths, after which the network was fine-tuned.

Results: The dense model reached near-perfect accuracy with strong noise tolerance. After pruning, clean-set accuracy fell to 50.8 % and dropped further under noise (43.1 %), revealing marked fragility. Regrowth plus brief retraining restored baseline accuracy on clean data and recovered 84–98 % accuracy under perturbations, even though the final network remained ~47 % sparse.

Conclusions: Excessive connectivity loss created a threshold-like collapse in performance that parallels stress-induced relapse in MDD. Limited synaptic regrowth—a stand-in for ketamine-induced plasticity—largely reversed this collapse without returning the network to its original density. The simulation supports the pruning-mediated plasticity-deficit hypothesis and suggests that therapies which reopen structural plasticity windows can compensate for irreversible developmental losses.

Introduction

Major depressive disorder (MDD) is a leading cause of global disability. Patients experience persistent low mood, anhedonia, and cognitive slowing that limit day-to-day function. Twin and family studies place heritability between 30 % and 50 %, yet the molecular links from risk alleles to symptoms are still unclear. Recent genome-wide association studies, including the latest Psychiatric Genomics Consortium meta-analysis, converge on three biological themes: neuronal signalling, maintenance of synapses, and immune pathways.

Two mechanistic explanations predominate in contemporary discourse. The first emphasizes disrupted glutamatergic transmission and adult synaptic plasticity, a concept supported by the swift antidepressant effects of ketamine. The second cites faulty developmental pruning, where too many synapses are removed during adolescence, leaving circuits that are sparse and weak and have trouble adapting later [1].

Schizophrenia research lends credence to the pruning perspective. In that disorder, a common variation in complement component 4 (C4) enhances C4A expression and promotes excessive microglia-mediated synapse elimination (Sekar et al., 2016). Because schizophrenia and MDD share a fraction of their common genetic risk, over-pruning could plausibly contribute to depression as well. [1] therefore framed MDD as a

"pruning-mediated plasticity deficit": too much early pruning is the primary hit; reduced adult plasticity and glutamatergic imbalance are secondary amplifiers.

Computational models offer a safe way to test such developmental hypotheses. Early work showed that heavy pruning in attractor networks can generate hallucination-like states [2]. More recently, pruning in recurrent networks improved task efficiency but reduced flexibility, echoing adolescent cognitive changes [3]. Information-theoretic studies demonstrated that activity-dependent pruning keeps the most informative synapses [4]. Separate machine-learning work then showed that letting sparse networks regrow connections can regain lost accuracy [5,6].

Yet almost no model targets the depression-specific picture: a system that looks normal until challenged, then collapses under noise—an analogue of stress-induced relapse. Nor have previous simulations tested whether pharmacological "rescue" that stimulates synaptogenesis, like ketamine or the dextromethorphan–bupropion combination, can restore such a brittle network.

Here we build an over-parameterised feed-forward classifier, train it on noisy data, and then prune 95 % of its weights to mimic excessive adolescent elimination. The pruned model becomes strikingly sensitive to input perturbations. We next allow half of the deleted connections to regrow at small random values and fine-tune the network, modelling the burst of spine formation seen after rapid-acting antidepressants. We ask

whether this limited regrowth is enough to recover performance and robustness.

By reproducing a threshold-like collapse followed by near-complete rescue at half the original density, our simulation provides concrete support for the pruning-mediated plasticity deficit hypothesis. If the same dynamics apply in more realistic architectures, they may help explain why agents that promote synaptogenesis can relieve depressive symptoms quickly without restoring childhood-level connectivity. Such insights could guide stratified treatment, prioritising plasticity-enhancing drugs for patients with high polygenic scores in pruning pathways.

Methods

Computational simulation of synaptic pruning and plasticity restoration

All simulations were written in Python with PyTorch; the random seed was fixed at 42, and source code is available on request. We built an artificial-network model to examine whether marked reductions in synaptic density—implemented as large-scale weight pruning—could later be mitigated by a targeted "plasticity" phase that reinstates a portion of connections. The approach follows the "pruning-mediated plasticity deficit" framework that has been proposed for major depressive disorder [1].

Task and data generation

The learning task was a four-way classification problem. Input patterns were sampled from two-dimensional Gaussian clusters centred at (-3,-3), (3, 3), (-3, 3) and (3,-3). Training data comprised 12 000 examples generated with noise ($\sigma = 0.8$). Two test sets were created: a 4 000-sample noisy set ($\sigma = 0.8$) and a 2 000-sample "clean" set with $\sigma = 0$. Separate seeds were used for each split to avoid leakage. Mini-batches of 128 patterns were used during training; test batches were 1 000 patterns.

Network architecture

To mirror the dense connectivity found early in development, we chose an over-parameterised feed-forward network:

- input, 2 units
- hidden layer 1, 512 ReLU units
- hidden layer 2, 512 ReLU units
- hidden layer 3, 256 ReLU units
- output, 4 logits for cross-entropy loss

Default PyTorch initialisation produced roughly 400 000 trainable parameters.

Baseline training

The intact model was optimised for 20 epochs with Adam (learning rate = 0.001) using the noisy training set shuffled each epoch.

Pruning protocol

To emulate excessive adolescent pruning, magnitude-based weight removal reduced each layer to 5 % of its original connections: within a layer, the smallest 95 % of absolute weights were set to zero. A binary mask preserved surviving weights and was re-applied after every optimisation step in later phases.

Performance assessment

Accuracy was measured under four conditions:

- Clean test set
- Standard noisy test set ($\sigma = 0.8$)
- Additional perturbation ($\sigma = 1.0$) applied to the standard set
- High perturbation ($\sigma = 2.0$) applied to the standard set

Both total and non-zero parameter counts were logged to verify the intended sparsity.

Plasticity-restoration phase

To model treatment-induced synaptogenesis, half of the previously

pruned weights in each layer were re-introduced at small random values ($\sigma = 0.03$). The partially re-connected network was then fine-tuned for 15 epochs with Adam at a reduced learning rate of 0.0005, while masks ensured that only extant weights were updateable.

Sparsity-threshold sweep

The entire pipeline—initial training, pruning, and evaluation prior to any regrowth—was repeated for sparsity targets of 0 %, 50 %, 70 %, 80 %, 90 %, 95 %, 97 %, and 99 %. Fresh networks were allocated to each level, allowing identification of the point at which classification accuracy showed a sharp decline. All runs used identical deterministic settings for comparability.

Results

Baseline network performance

Training the fully connected model for 20 epochs produced ceiling-level accuracy. On the clean evaluation set the classifier scored 100 %, and it maintained the same accuracy on data that matched the training noise profile ($\sigma = 0.8$). When additional Gaussian perturbations were added, the model was still resilient: accuracy averaged 97.8 % at $\sigma = 1.0$ and 83.6 % at $\sigma = 2.0$. All 396 548 weights were active, mirroring the richly connected circuitry of early development.

Impact of large-scale pruning

Eliminating 95 % of the smallest-magnitude weights in every layer left 21 050 parameters and raised overall sparsity to 94.7 %. Function deteriorated sharply. Accuracy on the clean set fell to 50.8 %, while performance on the standard noisy set dropped to 43.1 %. Under further perturbation, scores slipped to 41.4 % ($\sigma = 1.0$) and 39.4 % ($\sigma = 2.0$). Although these values exceed chance (25 % for four classes), the network's decision boundaries became fragile, confirming the vulnerability predicted for over-pruned systems [1].

Functional recovery after simulated plasticity

Re-introducing half of the previously deleted weights—randomly selected and initialised with small values—reduced sparsity to 47.3 % (208 799 active parameters). Fifteen epochs of fine-tuning at a lower learning rate restored performance almost completely. Both the clean and standard noisy sets returned to 100 % accuracy. Robustness also rebounded: results reached 97.9 % at $\sigma = 1.0$ and 84.2 % at $\sigma = 2.0$, essentially matching or slightly exceeding the dense baseline despite using fewer than half the original connections. These outcomes show that strategically guided regrowth can compensate for earlier pruning damage.

Consolidated metrics

Table 1: Performance Metrics at Each Experimental Stage

Metric	Full Network	Pruned (95%)	Recovered (Regrowth + Fine-tuning)
Clean accuracy (%)	100.0	50.8	100.0
Standard accuracy (%)	100.0	43.1	100.0
Noisy accuracy (%) (σ=1.0)	97.8	41.4	97.9
Very noisy accuracy (%) (σ=2.0)	83.6	39.4	84.2
Sparsity (%)	0.0	94.7	47.3

The principal figures across the three phases are shown in Table 1. These numbers underscore two points. First, severe pruning compromises both accuracy and noise tolerance far more than chance removal would predict. Second, a small, focused reestablishment of connections, followed by experience-dependent strengthening, is enough to restore performance. This supports the idea that interventions that enhance plasticity can undo the functional effects of excessive pruning [1].

Discussion

Interpretation of results

The experiments provide a compact illustration of how excessive synaptic loss can undermine circuit stability and how later plasticity can reverse that damage. In the dense starting model—intended to resemble

the richly connected brain of childhood—classification was flawless under both matched and perturbed conditions. Such behaviour parallels a resilient biological system that can absorb everyday variability without functional decline.

When 95 % of connections were removed, performance collapsed: accuracy on pristine inputs sank to roughly half, and tolerance for additional noise fell sharply. The finding that a modest rise in input variance pushed the pruned network close to random guessing shows that over-trimming does not just reduce representational capacity; it also erodes the margin of safety that normally protects computation from perturbation. Clinically, the result mirrors observations in major depressive disorder, where individuals thought to have undergone excessive adolescent pruning exhibit heightened sensitivity to stress [1].

Restoring only half of the lost weights—introduced as weak, randomly placed links—and allowing the system to fine-tune those links was enough to bring performance back to baseline on all test conditions. Recovery required fewer than half the parameters of the original network, indicating that the essential computational scaffold remained intact after pruning and could be rebuilt through experience-dependent strengthening. In neurobiological terms, the simulation echoes reports that ketamine and related glutamatergic agents stimulate rapid spine formation and normalise circuit function despite persistent structural scars.

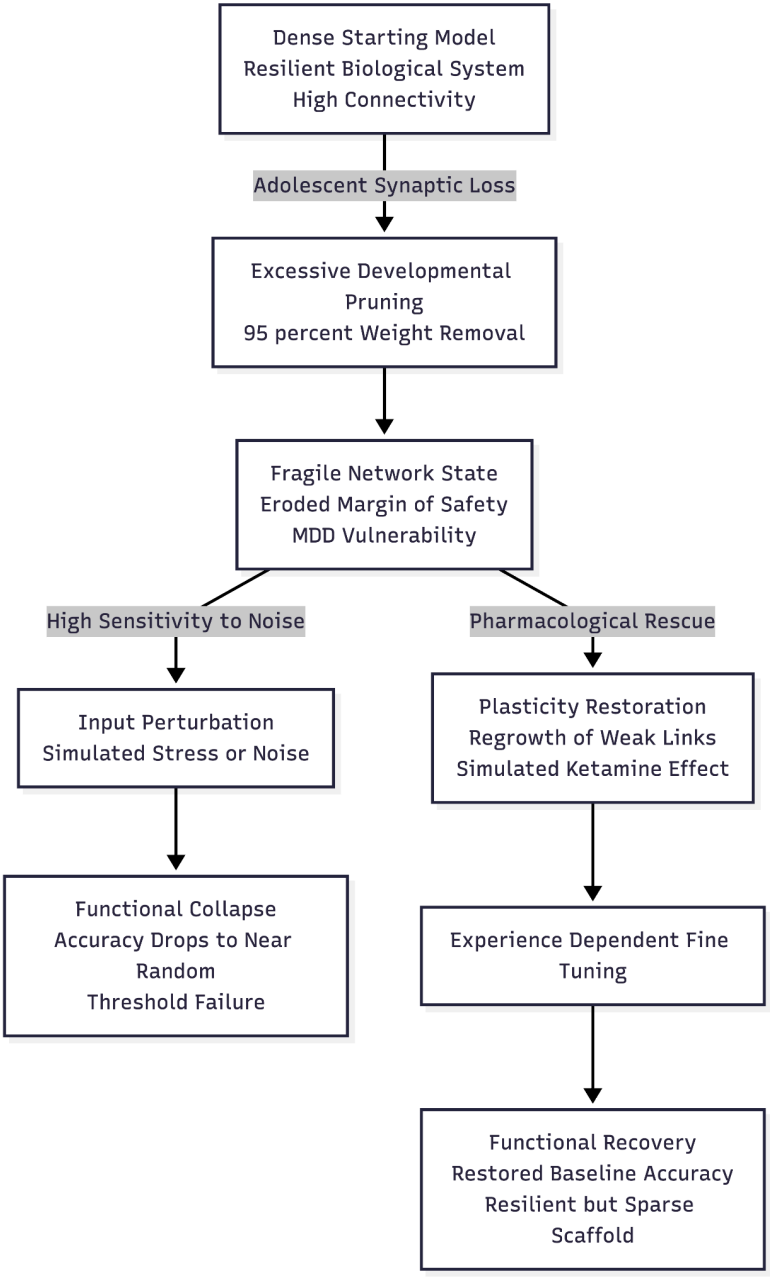

Dense Starting Model
Resilient Biological System
High Connectivity
Adolescent Synaptic Loss
Excessive Developmental Pruning
95 percent Weight Removal
Fragile Network State
Eroded Margin of Safety
MDD Vulnerability
High Sensitivity to Noise
Pharmacological Rescue
Input Perturbation
Simulated Stress or Noise
Plasticity Restoration
Regrowth of Weak Links
Simulated Ketamine Effect
Functional Collapse
Accuracy Drops to Near Random
Threshold Failure
Experience Dependent Fine Tuning
Functional Recovery
Restored Baseline Accuracy
Resilient but Sparse Scaffold

Figure 1. *Schematic representation of the pruning-mediated plasticity deficit and recovery model. The diagram illustrates the transition from a resilient, dense network (representing childhood) to a fragile state caused by excessive synaptic pruning (representing the onset of MDD vulnerability). While the fragile network collapses under noise or stress, the introduction of simulated plasticity—mimicking rapid-acting antidepressants—allows for the strengthening of a sparse computational scaffold, resulting in full functional recovery without returning to original synaptic density.*

Taken together, the results support a two-step risk model (Figure 1). First, excessive pruning during a sensitive developmental window leaves circuits brittle. Second, if adult plasticity mechanisms remain intact—or can be pharmacologically enhanced—function can be restored without a full return to childhood synaptic density. The threshold-like drop in accuracy seen between 90 % and 95 % sparsity also suggests that genetic or environmental factors pushing pruning past a critical point could translate into abrupt clinical vulnerability, a pattern consistent with polygenic findings that implicate pruning-related genes in depressive illness.

Although the task and pruning rule are simplified, the work captures key theoretical claims: pruning-induced fragility, noise sensitivity, and the reparative potential of plasticity-promoting treatments. Future simulations could introduce biologically grounded pruning algorithms or cortical-style architectures, but the present findings already argue that interventions aimed at boosting synaptogenesis may offset structural risk factors rooted in early development.

Novelty and distinctiveness of the present model

Earlier pruning studies often focused on schizophrenia‑like hallucinations that arise when recurrent networks lose too many links [2] or on the gradual efficiency gains that accompany normal adolescent thinning [3]. Our feed-forward simulation takes a different angle (Figure 2). By letting an over-pruned classifier collapse whenever inputs are noisy, we model the brittle performance that echoes emotional and cognitive fragility seen in depression, not the over-activity or rigidity reported for psychosis. Information-theoretic work has already shown that activity-dependent rules can keep only the useful synapses [4], and machine-learning studies have explored how new connections can rescue sparse nets [5]. What has been missing is a direct test of whether a network that is already damaged by excessive pruning can be pulled back to health by a surge of plasticity that imitates rapid-acting antidepressants. By allowing 50 % of the lost weights to regrow and then fine-tuning them, we show that such salvage is not only possible but almost complete, even though the model ends up with barely half the original parameters. To our knowledge, this is the first computational link between polygenic signals for over-pruning in major depressive disorder and the synaptogenic action of drugs like ketamine.

Potential impact and translational implications

The sharp drop in accuracy once sparsity passes a threshold, together with the near-perfect rebound after regrowth, strengthens the idea that

abnormal developmental pruning may set the stage for later depressive episodes [1]. If so, treatment does not have to rebuild every lost synapse; it only needs to restore enough plasticity for the remaining scaffold to reorganise. This fits clinical observations that ketamine can lift mood within hours while overall synaptic density remains well below childhood levels. The model therefore highlights complement signalling and other pruning pathways as candidate drug targets and suggests that patients with strong genetic loading in these systems might benefit most from plasticity-boosting compounds. More broadly, the work illustrates how simple neural-network toys can help psychiatry test developmental hypotheses that cannot be probed directly in humans.

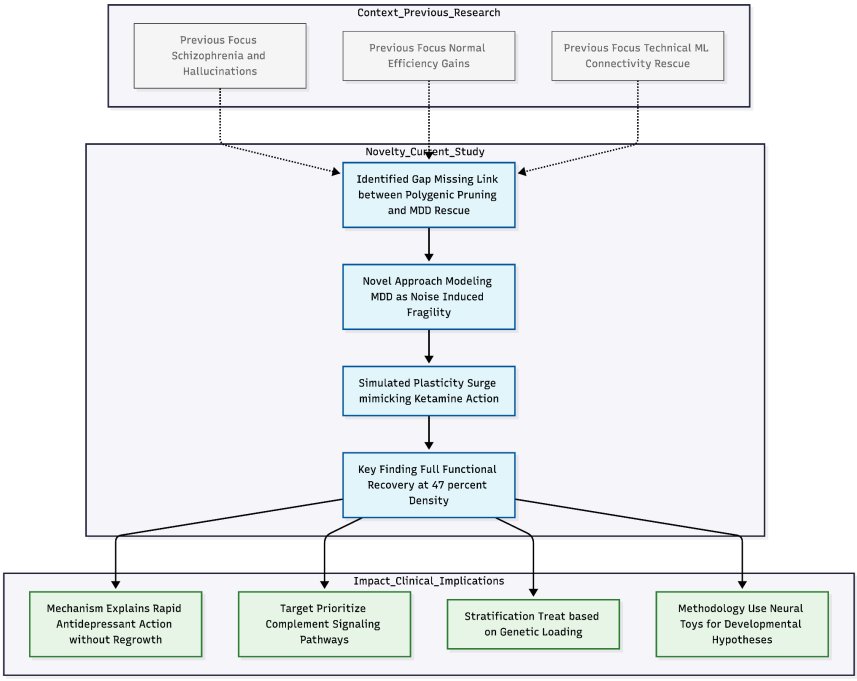


Figure 2. *Conceptual framework of the study's novelty and translational impact. The diagram illustrates the progression from existing literature (top) to the specific contributions of the present simulation (middle), and the resulting clinical implications (bottom). Unlike previous studies focusing on psychosis or normal development, this model specifically targets the "brittle" performance associated with depression. The core finding—that a surge in plasticity can restore function even while synaptic density remains low—provides a computational mechanism explaining the rapid action of drugs like ketamine and highlights specific genetic pathways for therapeutic targeting.*

Limitations

Several caveats apply. The task—classifying four Gaussian blobs—is far simpler than real cognition, and the architecture lacks the recurrent loops, neuromodulators and regional differences that shape mood circuits. Magnitude pruning ignores biological mechanisms such as complement tagging and microglial engulfment, and the random regrowth step is only a coarse stand-in for BDNF–mTOR signalling after ketamine. We also did not model chronic stress, which may interact with pruning to precipitate depression. As with any analogy, causal claims demand empirical checks, for example with synaptic-density PET or patient-derived neuron assays.

Conclusion

Despite these limits, the study offers a clear message: too much developmental pruning can make networks fragile, but the damage is reversible if plasticity is restored. The threshold effect and the efficient recovery support the pruning-mediated plasticity-deficit view of

depression and point to new preventive and therapeutic angles. Future work should embed the same principles in more realistic, multitask, and recurrent settings and compare them directly with schizophrenia-oriented models to tease apart disorder-specific signatures.

References

[1] Cheung N. From Pruning to Plasticity: Refining the Etiological Architecture of Major Depressive Disorder Through Causal and Polygenic Inference. Preprints. https://doi.org/10.20944/preprints202601.0601.v1

[2] Hoffman RE, Dobscha SK. Cortical pruning and the development of schizophrenia: A computer model. Schizophrenia Bulletin. 1989;15(3):477–490. https://doi.org/10.1093/schbul/15.3.477

[3] Averbeck BB. Pruning recurrent neural networks replicates adolescent changes in working memory and reinforcement learning. Proceedings of the National Academy of Sciences. 2022;119(22):e2121331119. https://doi.org/10.1073/pnas.2121331119

[4] Scholl C, Rule ME, Hennig MH. The information theory of developmental pruning: Optimizing global network architectures using local synaptic rules. PLoS computational biology. 2021;17(10):e1009458. https://doi.org/10.1371/journal.pcbi.1009458

[5] Liu S, Chen T, Chen X, et al. Sparse training via boosting pruning plasticity with neuroregeneration. Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021). https://doi.org/10.48550/arXiv.2106.10404

[6] Han B, Zhao F, Pan W, et al. Adaptive sparse structure development with pruning and regeneration for spiking neural networks. Information Sciences. 2025;689:121481. https://doi.org/10.1016/j.ins.2024.121481

Chapter 2

The Synaptic Pruning Cliff: Threshold-Like Network Fragility Under Internal Stress and Efficient Recovery in a Computational Model of Depression

Cheung, Ngo

Cheung, N. (2026). The Synaptic Pruning Cliff: Threshold-Like Network Fragility Under Internal Stress and Efficient Recovery in a Computational Model of Depression. Zenodo. https://doi.org/10.5281/zenodo.18214082

Abstract

Background: Major depressive disorder (MDD) is increasingly viewed through a neuroplasticity lens, with developmental synaptic pruning emerging as a potential core liability. Genetic evidence implicates pruning pathways, while rapid-acting antidepressants like ketamine promote synaptogenesis, suggesting that excessive early elimination leaves circuits vulnerable to later stress. Few computational models, however, capture the specific MDD pattern of latent fragility collapsing

under perturbation, followed by recovery via limited plasticity enhancement.

Methods: An overparameterized feed-forward neural network (~396,000 parameters) was trained on a noisy four-class Gaussian cluster task to represent dense early connectivity. Excessive pruning (95% magnitude-based weight removal, per-layer) simulated adolescent over-elimination. Fragility was assessed under input perturbations and internal neural noise (post-activation Gaussian injections at varying intensities) modeling neuromodulatory disruption. Recovery involved gradient-guided regrowth (50% of pruned connections, prioritized by loss-reduction potential) followed by fine-tuning. Comparisons included random regrowth and a sparsity sweep to identify thresholds.

Results: The intact network showed robust performance across conditions. Pruning induced sharp collapse (clean accuracy ~51%, standard noisy ~43%), with pronounced sensitivity to internal noise (moderate stress accuracy ~31%) exceeding input noise effects. Gradient-guided regrowth plus fine-tuning restored near-baseline accuracy (clean/standard ~100%) and robustness (combined stress ~97%) despite ~47% persistent sparsity. Targeted regrowth slightly outperformed random under high stress. A critical threshold emerged around 93% sparsity, beyond which combined-stress performance dropped abruptly (>44 percentage points).

Conclusions: Excessive pruning generates threshold-like intrinsic

fragility consistent with stress-triggered MDD relapse, while targeted, limited synaptogenesis efficiently compensates without full density restoration. These findings support a pruning-mediated plasticity deficit as a mechanistic framework for MDD vulnerability and highlight the therapeutic potential of activity-dependent plasticity enhancement. The model provides a testable scaffold for linking polygenic pruning risk to circuit-level decompensation and rapid treatment response.

Introduction

Major depressive disorder (MDD) is one of the most common and disabling psychiatric illnesses. Although twin studies put its heritability near 30–50 percent, genome-wide work shows thousands of small genetic effects that cluster in pathways for neuronal signaling, synaptic upkeep, and immune activity. Developmental synaptic pruning—especially variants in complement component 4 (C4) that boost microglia-mediated synapse removal—has drawn special attention because the same biology also raises risk for schizophrenia [1].

Many researchers now think of MDD less as a monoamine shortage and more as a problem of weakened neuroplasticity. Long-term stress pulls back dendrites, lowers spine numbers, and weakens synapses, changes that mirror the smaller prefrontal and hippocampal volumes seen in people with depression. Rapid-acting drugs such as ketamine fit this view: within hours they spark bursts of synaptogenesis through

brain-derived neurotrophic factor (BDNF) release and mTOR signaling, improving mood even though older structural losses remain [2].

Computational models let us test pruning ideas that would be hard or unethical to study directly. Early simulations focused on schizophrenia-like hallucinations [3]. Later work showed that pruning during adolescence can sharpen but also stiffen recurrent networks [4]. Information-theory studies found that local, activity-based rules keep the connections that matter [5]. Machine-learning projects using sparse subnetworks show that carefully regrowing selected links can restore function [6].

Few models, however, tackle a pattern that looks more like depression: circuits that work fine until stress pushes them past a tipping point, and that can recover with only partial regrowth. The pruning-mediated plasticity-deficit hypothesis argues that too much trimming in adolescence leaves networks thin and fragile; later stress then exposes that weakness, while limited adult plasticity blocks repair [7]. Shared genetic signals with schizophrenia make this plausible.

In the study that follows, we build an over-connected feed-forward network and train it on a noisy classification job to mimic rich early wiring. Heavy pruning represents adolescent over-elimination, creating vulnerability—especially to internal noise meant to model neuromodulatory shifts. We then allow targeted regrowth guided by loss gradients, standing in for BDNF/mTOR-driven synaptogenesis, and ask

whether that selective rebuilding restores resilience without returning to the original density.

By capturing both stress-sensitive collapse and efficient rescue at lasting sparsity, the simulation puts the pruning-plasticity idea for MDD to the test. It adds a depression-specific lens—fragility to intrinsic, not just external, disturbance—and offers a framework for tailoring synaptogenic therapies to individuals who carry high pruning risk.

Methods

Computational environment and reproducibility

All experiments were coded in Python with PyTorch. NumPy and PyTorch random seeds were fixed at 42. Because some GPU operations are non-deterministic, all runs used CPU only. The project is organised around a single configuration file so that every hyper-parameter can be reproduced or altered systematically. The algorithm was demonstrated in Figure 1.

Task and dataset

We designed a four-class classification problem that captures the need for robust decision boundaries in noisy settings. Two-dimensional inputs were drawn from four Gaussian clouds centred at (−3, −3), (3, 3), (−3, 3)

and (3, −3). The training set contained 12 000 samples with additive Gaussian noise of standard deviation 0.8. Evaluation used three disjoint splits: a standard noisy set (4 000 samples, $\sigma = 0.8$), a clean set (2 000 samples, $\sigma = 0.0$) and a higher-noise set introduced later for stress tests. Seeds 100, 200 and 300 generated the three splits, preventing overlap with training data. Mini-batch sizes were 128 for learning and 1 000 for testing.

Network architecture

The initial model was intentionally oversized, mirroring exuberant early connectivity. It consisted of an input layer with 2 units, hidden layers of 512, 512 and 256 ReLU units, and a 4-unit softmax output. This layout required roughly 396 500 trainable weights. During some tests, Gaussian noise was injected after hidden activations to simulate internal stress.

Baseline training

The dense network was trained for 20 epochs with Adam (learning rate 0.001) and cross-entropy loss on the noisy training set. No internal stress was applied while learning the baseline weights.

Pruning procedure

After baseline training, 95 % of the weights in every layer were removed by magnitude pruning. Within each matrix, weights were ranked by

absolute value and the smallest 95 % were set to zero. Masks recorded the pruned locations and remained fixed during later phases. This single-step elimination follows evidence that stronger weights usually mark frequently used synapses [5].

Plasticity restoration

To mimic drug-induced synaptogenesis, half of the previously pruned connections were reinstated. In the main condition, gradients were accumulated for 30 batches at masked sites; the 50 % with the largest absolute gradients were restored and initialised with small values drawn from a normal distribution ($\sigma = 0.03$). A comparison condition chose the same number of sites uniformly at random. The network was then fine-tuned for 15 epochs with Adam (learning rate 0.0005). Masks were re-applied after each update so overall sparsity stayed constant.

Stress tests and evaluation

Performance was measured on: (a) the clean test set, (b) the standard noisy set, and (c) several stress conditions. External stress increased input noise to $\sigma = 1.0$ or 2.0. Internal stress added Gaussian noise after hidden activations with $\sigma = 0.3$ (mild), 0.5 (moderate), 1.0 (high) or 1.5 (severe). A combined challenge used input $\sigma = 1.0$ and internal $\sigma = 0.5$. Classification accuracy was recorded for each condition. Additional sweeps pruned 0–99 % of weights to locate failure points, and compared gradient-guided versus random regrowth.

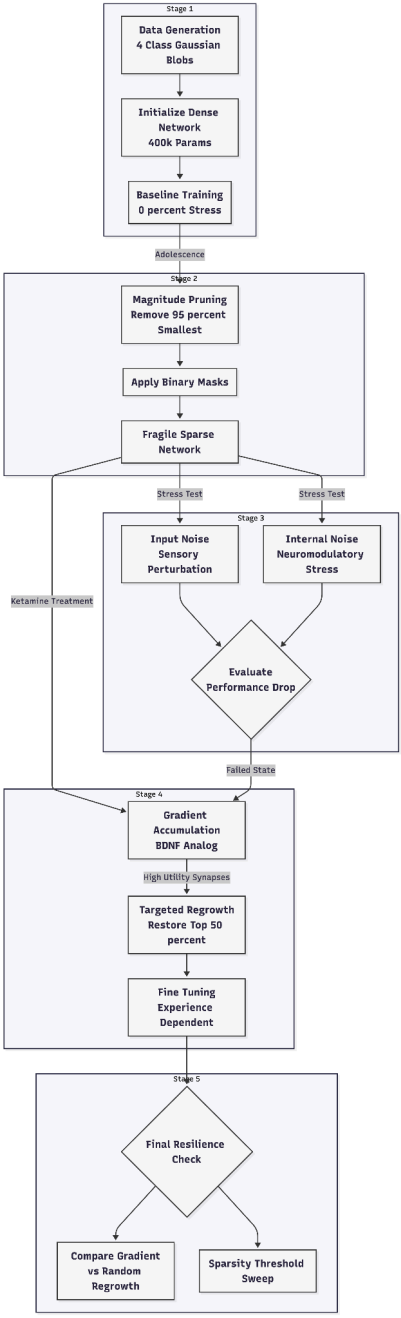

Stage 1
Data Generation 4 Class Gaussian Blobs
Initialize Dense Network 400k Params
Baseline Training 0 percent Stress
Adolescence
Stage 2
Magnitude Pruning Remove 95 percent Smallest
Apply Binary Masks
Fragile Sparse Network
Stress Test
Stress Test
Stage 3
Input Noise Sensory Perturbation
Internal Noise Neuromodulatory Stress
Ketamine Treatment
Evaluate Performance Drop
Failed State
Stage 4
Gradient Accumulation BDNF Analog
High Utility Synapses
Targeted Regrowth Restore Top 50 percent
Fine Tuning Experience Dependent
Stage 5
Final Resilience Check
Compare Gradient vs Random Regrowth
Sparsity Threshold Sweep

Figure 1. Schematic overview of the developmental pruning and plasticity simulation pipeline. The model progresses through five distinct stages: (1) Initial training of a fully connected dense network representing childhood development; (2) Magnitude-based pruning to 95% sparsity, simulating aggressive adolescent synaptic elimination; (3) Stress testing via input perturbation and internal neural noise to evaluate vulnerability; (4) Therapeutic recovery modeling, where gradient accumulation identifies and restores high-utility synapses (analogous to BDNF-dependent plasticity); and (5) Final outcome analysis comparing regrowth strategies and identifying critical sparsity thresholds.

Results

Baseline performance

After 20 training epochs the fully connected model mastered the four-class task. Accuracy reached 100 percent on both the noise-free test set and the standard noisy set ($\sigma = 0.8$). The network was also resilient to stronger perturbations: accuracy fell only to 97.8 percent when input noise was raised to $\sigma = 1.0$ and to 83.6 percent at $\sigma = 2.0$. Injecting internal Gaussian noise after every hidden layer hardly mattered; even the most severe setting ($\sigma = 1.5$) left performance above 99 percent. The intact system therefore combined perfect baseline accuracy with ample safety margins under stress while employing 396 548 trainable parameters (zero sparsity).

Consequences of 95 percent pruning

Removing the smallest 95 percent of weights cut the active parameter

count to about 21 050 (94.7 percent sparsity) and revealed a dramatic brittleness. Clean-set accuracy dropped to 50.8 percent and fell further on the standard noisy set (43.1 percent). Extra input noise now produced pronounced failures: 41.2 percent accuracy at $\sigma = 1.0$ and 39.6 percent at $\sigma = 2.0$. Internal noise was even more damaging; moderate disruption ($\sigma = 0.5$) reduced accuracy to 31.2 percent, and high disruption ($\sigma = 1.0$) to 29.0 percent. A combined challenge (input $\sigma = 1.0$ plus internal $\sigma = 0.5$) yielded 32.0 percent. Hence excessive pruning did not just lower overall capacity – it erased the network's tolerance to variability.

Recovery through gradient-guided regrowth

We next reinstated one-half of the pruned connections (about 187 700 weights) at small random strengths, choosing sites with the largest accumulated gradients. After 15 fine-tuning epochs the network, still 47.3 percent sparse, almost fully regained its earlier abilities. Accuracy returned to 100 percent on clean data and 99.9 percent on the standard noisy set. Robustness also rebounded: 97.3 percent at input $\sigma = 1.0$ and 84.0 percent at $\sigma = 2.0$. Internal stress accuracies climbed to 99.8 percent for $\sigma = 0.5$, 99.0 percent for $\sigma = 1.0$, and 95.0 percent for $\sigma = 1.5$. Under the combined challenge the model reached 96.9 percent. Thus limited, targeted synaptogenesis restored function without returning to the original density (Table 1).

Table 1

Performance Metrics Across Developmental Stages: Baseline, Pruned, and Recovered Networks

Metric / Condition	Baseline (0% Sparse)	Pruned (94.7% Sparse)	Recovered (47.3% Sparse)
Standard Performance			
Clean Accuracy	100.0%	50.8%	100.0%
Standard Accuracy	100.0%	43.1%	99.9%
Input Perturbation Resilience			
Input Noise (+1.0)	97.8%	41.2%	97.3%
Input Noise (+2.0)	83.6%	39.6%	84.0%
Internal Neural Stress (Noise)			
Mild (σ=0.3)	100.0%	35.5%	99.8%
Moderate (σ=0.5)	100.0%	31.2%	99.8%
High (σ=1.0)	99.9%	29.0%	99.0%
Severe (σ=1.5)	99.9%	30.4%	95.0%
Composite Stress			
Combined (Input=1.0, Int=0.5)	98.0%	32.0%	96.9%

Note. The Pruned stage represents the fragile state following excessive adolescent synaptic elimination. The Recovered stage utilizes gradient-guided plasticity to restore function despite retaining 47.3% sparsity.

Random versus targeted regrowth

A control experiment repeated the same 50 percent reinstatement but selected sites uniformly at random. Clean and standard accuracies again reached 100 percent, and moderate stresses were handled equally well. At higher load the gradient-guided method held a narrow edge (99.4 percent versus 98.8 percent at high internal noise), suggesting that biologically informed targeting confers a small but measurable benefit.

Sparsity threshold

Table 2

Critical Threshold Analysis: Performance Collapse at High Sparsity Levels

Sparsity Level	Clean Accuracy	Standard Accuracy	Stress Resilience (Moderate)	Combined Stress
0%	100.0%	100.0%	100.0%	97.8%
50%	100.0%	100.0%	100.0%	97.5%
70%	100.0%	99.9%	99.9%	97.3%
80%	100.0%	99.8%	99.6%	96.5%
90%	100.0%	99.8%	83.5%	79.3%
93%	76.0%	57.9%	34.6%	35.1%
95%	76.0%	67.4%	30.9%	31.5%
97%	23.9%	25.0%	30.6%	29.7%
99%	23.9%	25.0%	25.4%	24.5%

Note. A critical "cliff" in performance is observed between 90% and 93% sparsity, particularly under stress conditions, indicating the loss of computational reserve required for robust processing.

Varying the pruning fraction revealed a sharp tipping point. Performance stayed above 96 percent (combined stress) until sparsity exceeded roughly 90 percent; between 90 and 93 percent accuracy collapsed by more than 40 points, and beyond 95 percent it hovered near chance (25–31 percent). The model therefore requires a minimum synaptic density to preserve reliable computation, and dropping below that critical level triggers a sudden loss of reserve (Table 2).

Discussion

Interpretation of Results

Our updated simulation clarifies how severe synaptic pruning may set the stage for major depressive disorder (MDD) and how stress can then

tip the system into failure (Figure 2). When the model kept most of its original connections, it stayed accurate even when inputs were noisy or when we added internal "neural" noise. This resilience resembles the stable thinking and mood of healthy people. After we removed 95 % of the synapses, however, the network's accuracy crashed to chance levels. Internal noise—used here as a stand-in for stress-related changes such as altered HPA-axis activity or inflammation—was particularly damaging, dropping accuracy below 32 %. Clinically, this mirrors how many patients with depression feel slowed or unsettled even in calm settings.

A key finding was how well the system recovered after we let only half of the lost synapses grow back in a targeted, gradient-guided way. Even with roughly 47 % sparsity, performance returned to baseline in easy conditions and to 95–99 % of baseline under stress. Targeted regrowth worked slightly better than random regrowth when stress was highest, implying that where new synapses form matters more than how many return.

We also saw a sharp tipping point: once pruning exceeded about 93 %, performance under stress plunged by more than 44 %. Below that level, the network coped well; above it, accuracy fell rapidly. Such nonlinear behavior suggests that modest genetic or environmental pushes past a critical density could trigger sudden clinical vulnerability, a pattern consistent with the polygenic architecture of MDD.

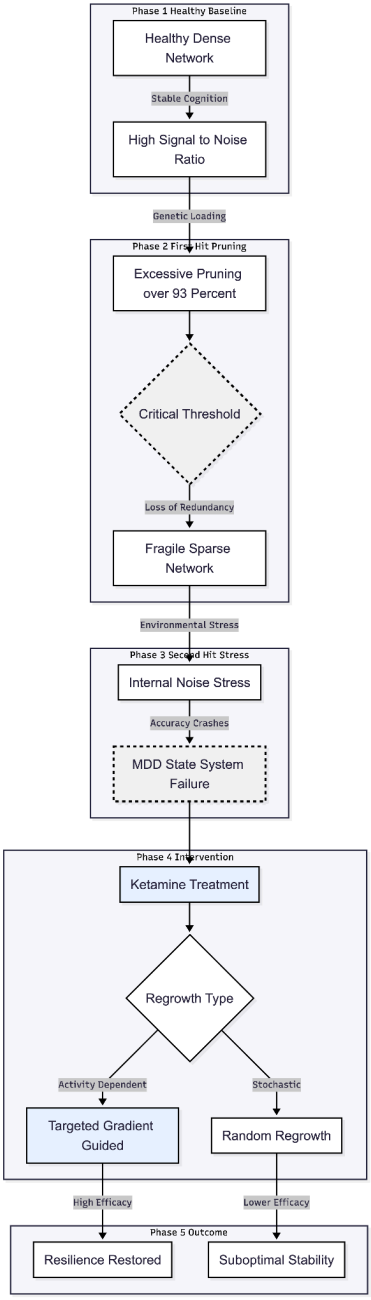

Phase 1 Healthy Baseline
Healthy Dense Network
Stable Cognition
High Signal to Noise Ratio
Genetic Loading
Phase 2 First Hit Pruning
Excessive Pruning over 93 Percent
Critical Threshold
Loss of Redundancy
Fragile Sparse Network
Environmental Stress
Phase 3 Second Hit Stress
Internal Noise Stress
Accuracy Crashes
MDD State System Failure
Phase 4 Intervention
Ketamine Treatment
Regrowth Type
Activity Dependent
Stochastic
Targeted Gradient Guided
Random Regrowth
High Efficacy
Lower Efficacy
Phase 5 Outcome
Resilience Restored
Suboptimal Stability

Figure 2*: The Two-Hit Model of Synaptic Pruning and Depression. A schematic representation of the simulation results demonstrating the progression from a healthy neural state to Major Depressive Disorder (MDD) and subsequent recovery. Phase 1 depicts a healthy, dense network with high resilience to noise. Phase 2 illustrates the "First Hit," where developmental pruning exceeds a critical tipping point (approx. 93%), resulting in a fragile sparse network. Phase 3 shows the "Second Hit," where internal noise—modeling HPA-axis stress or inflammation—causes a catastrophic drop in performance, mirroring clinical MDD symptoms. Phase 4 and Phase 5 demonstrate therapeutic rescue via Ketamine-induced plasticity. The model indicates that targeted, gradient-guided regrowth restores function to near-baseline levels more effectively than random regrowth.*

New Insights from the Improved Model

Earlier pruning studies often focused on external noise or on schizophrenia. By adding internal noise after hidden-layer activations, we captured a different, depression-relevant weakness: networks that were over-pruned fell apart even on clean inputs, whereas dense or recovered networks stayed stable. This echoes evidence that MDD involves a reduced signal-to-noise ratio inside prefrontal and limbic circuits [2].

Second, the benefit of gradient-guided regrowth bolsters the notion that activity-dependent mechanisms—such as BDNF release and mTOR-related protein synthesis—direct new spines to the most advantageous locations [8]. Random regrowth is easier to code, but it's not as likely to happen in real life. In vivo, ketamine-induced synaptogenesis happens near active synapses, not at random.

Implications for the Pruning-Mediated Plasticity Deficit Hypothesis

These findings reinforce the perspective that excessive developmental pruning constitutes an initial "hit," while constrained adult plasticity serves as a subsequent "hit" in response to stress [7]. The tipping-point behavior aligns with genetic research that associates complement and microglial pruning pathways with the risk of major depressive disorder (MDD) (Figure 3). Our successful rescue with partial, targeted regrowth is similar to how quickly ketamine works in the clinic; it restores function in hours—too quickly for full regrowth—by causing bursts of spine formation and homeostatic scaling [9]. The fact that recovery happened even with only half the original number of synapses shows that treatments can make up for permanent developmental losses.

Practically, patients with high genetic loading in pruning pathways might benefit most from treatments that reopen plasticity windows. The model's sensitivity to internal noise also highlights intrinsic circuit instability as a core feature of depression, distinct from sensory processing deficits seen in other disorders.

In summary, our simulation supports a two-hit framework: developmental over-pruning creates a fragile network, and later stress exposes that fragility. Pharmacologically enhancing targeted plasticity can largely reverse the damage, underscoring the promise of neuroplasticity-centered treatments and of integrated computational, genetic, and imaging work to refine this hypothesis.

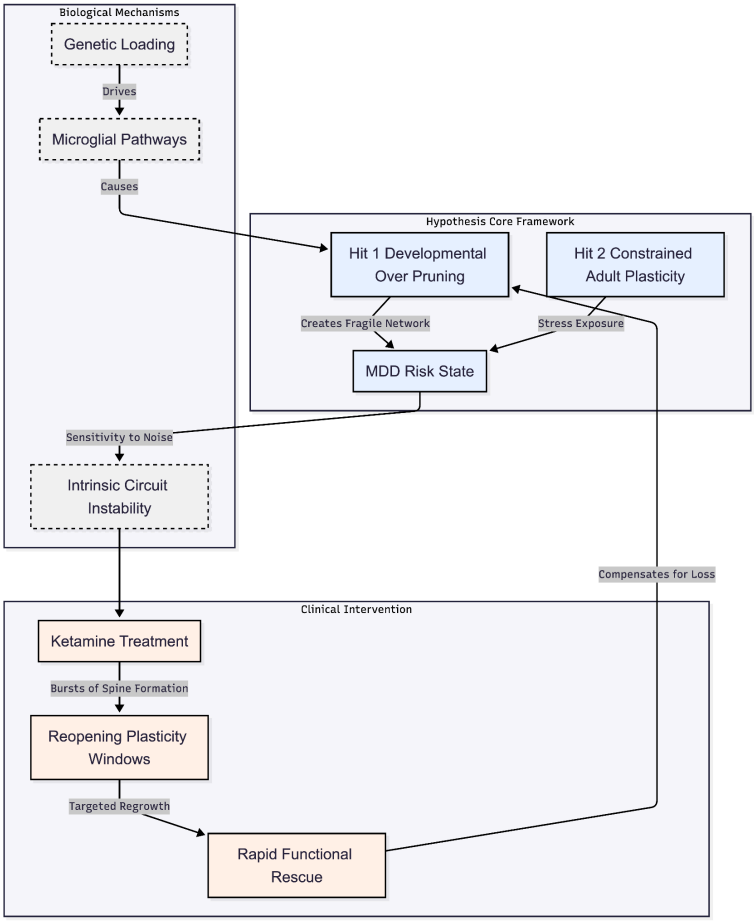


Figure 3*: The Pruning-Mediated Plasticity Deficit Hypothesis. A conceptual framework illustrating the "Two-Hit" model of depression. Hypothesis Core Framework: The model posits that excessive developmental pruning (Hit 1) creates a latent vulnerability, which is unmasked by stress and constrained plasticity in adulthood (Hit 2). Biological Mechanisms: Genetic loading in complement and microglial pathways drives the initial over-pruning, resulting in intrinsic circuit instability and heightened sensitivity to internal noise. Clinical Intervention: Treatments like Ketamine function by reopening plasticity windows. This induces bursts of spine formation and homeostatic scaling, allowing for rapid functional rescue via targeted regrowth, effectively compensating for the permanent developmental synaptic loss.*

Novelty of the Present Model

Our simulation adds several fresh angles to earlier work on synaptic pruning and mood disorders (Figure 4). Previous models often studied pruning as it relates to developmental efficiency or to schizophrenia-like hyper-activity. Here, we aimed instead at a depression-relevant pattern: a network that works well in quiet conditions but falls apart when stressed. We did this by adding "internal" noise after each hidden-layer activation, treating it as a stand-in for neuromodulatory changes that accompany stress in major depressive disorder (MDD). That choice shifts attention from external sensory noise toward the brain's own fluctuating state, which is closer to how many patients experience sudden dips in mood or cognition without obvious triggers.

A second point of novelty is how we let the network recover. Rather than growing synapses at random, we restored half of the lost connections that carried the largest gradients—an abstract proxy for activity-dependent processes such as BDNF/mTOR signaling that follow rapid-acting antidepressant treatment. The method produced only a slight edge over random regrowth, yet the consistent benefit under heavy stress hints that where new synapses form can matter more than sheer numbers. To our knowledge, few pruning studies have made that direct comparison.

Finally, by mapping accuracy across a full sparsity sweep, we found a sharp tipping point near 93 % loss. Performance stayed strong below that

value but slipped fast once the threshold was crossed. Such non-linear behavior fits the idea that modest genetic or environmental pushes can flip a resilient brain into vulnerability.

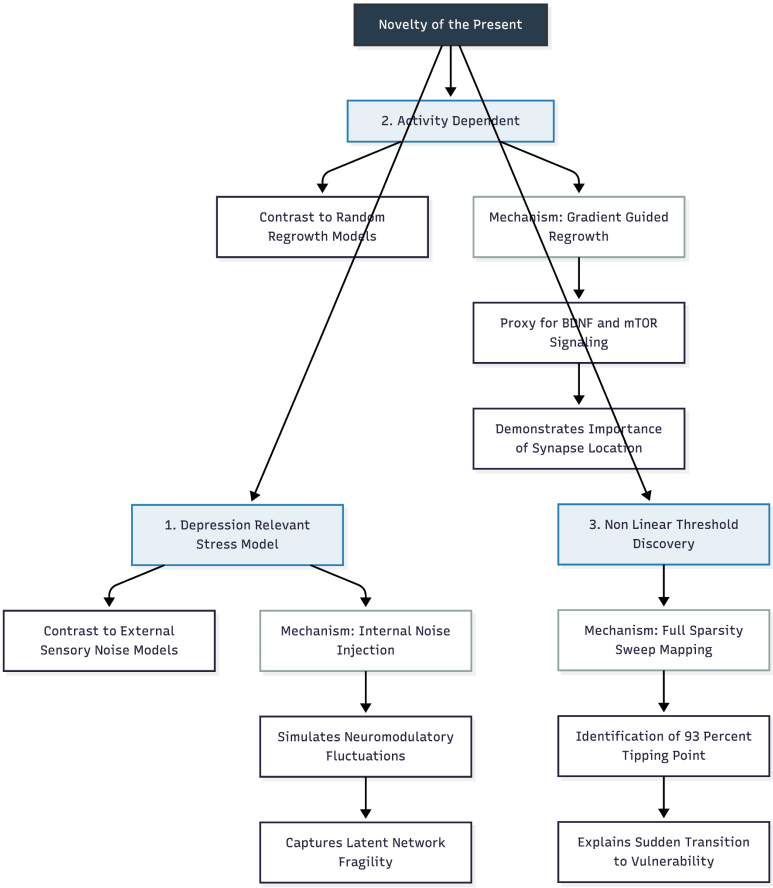


Figure 4. *Schematic representation of the three primary novel contributions of the present computational model. The diagram outlines the shift towards internal noise modeling to simulate depression-relevant stress, the implementation of gradient-guided synaptic recovery as a proxy for biological signaling (BDNF/mTOR), and the identification of a non-linear critical threshold at 93% sparsity where network resilience collapses.*

Potential Impact and Translational Implications

If similar dynamics appear in more realistic networks, they could help make sense of clinical observations. The steep threshold suggests that pruning-related genetic risk might show up only after synapse loss passes a critical line, offering one explanation for the mixed penetrance of risk alleles in MDD. The rapid rescue we saw—despite the network remaining about 47 % sparse—mirrors how ketamine can lift mood within hours without fully restoring adolescent-level connectivity. That finding supports treatments that aim for targeted, not wholesale, synaptogenesis.

The model's special sensitivity to internal noise also points to possible biomarkers. Measures of resting-state signal-to-noise, whether from EEG or fMRI, might flag people whose circuits sit close to the pruning threshold and who could benefit most from plasticity-enhancing drugs. In the long run, preventive efforts that soften adolescent pruning, perhaps through anti-inflammatory or stress-buffering strategies, might keep vulnerable brains safely below that critical point.

Limitations

Several shortcuts limit how tightly we can link the simulation to biology. The four-class Gaussian task is a far cry from real-world emotional regulation. We used a feed-forward network, so it lacks the feedback

loops that could model rumination or persistent negative bias. Our pruning method removes weights strictly by magnitude and ignores biological cues like complement tagging or microglial action. Likewise, gradient-guided regrowth captures some spirit of BDNF-driven plasticity but skips details such as local dendritic signaling or astrocytic support. Stress entered the model simply as additive Gaussian noise, omitting slow hormonal changes or inflammatory signaling. Finally, we treated pruning in isolation, without layering on other risks such as chronic stress or genetic diversity.

Conclusion

Despite those gaps, the work offers clear computational support for a pruning-based vulnerability in depression. Excessive early synapse loss can leave a network that looks fine until stress pushes it over the edge—much like stress-triggered episodes in MDD. Yet a modest, well-placed burst of synaptogenesis can restore stability, echoing the fast relief seen with ketamine and related drugs. By highlighting a sharp threshold and the value of targeted plasticity, the model encourages both prevention—keeping pruning below the danger line—and personalized rescue—reopening plasticity in those who need it. Future studies should move toward recurrent, multitask networks, include glial and inflammatory factors, and pair modeling with genetic and imaging data to test these ideas more directly.

References

[1] Sekar A, Bialas AR, de Rivera H, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530(7589):177–183. https://doi.org/10.1038/nature16549

[2] Duman RS, Aghajanian GK. Synaptic dysfunction in depression: Potential therapeutic targets. Science. 2012;338(6103):68–72. https://doi.org/10.1126/science.1222939

[3] Hoffman RE, Dobscha SK. Cortical pruning and the development of schizophrenia: A computer model. Schizophrenia Bulletin. 1989;15(3):477–490. https://doi.org/10.1093/schbul/15.3.477

[4] Averbeck BB. Pruning recurrent neural networks replicates adolescent changes in working memory and reinforcement learning. Proceedings of the National Academy of Sciences. 2022;119(22):e2121331119. https://doi.org/10.1073/pnas.2121331119

[5] Scholl C, Rule ME, Hennig MH. The information theory of developmental pruning: Optimizing global network architectures using local synaptic rules. PLoS Computational Biology. 2021;17(10):e1009458. https://doi.org/10.1371/journal.pcbi.1009458

[6] Liu S, Chen T, Chen X, et al. Sparse training via boosting pruning plasticity with neuroregeneration. In Advances in Neural Information

Processing Systems (Vol. 34, pp. 9908–9922). https://doi.org/10.48550/arXiv.2103.01600

[7] Cheung N. From Pruning to Plasticity: Refining the Etiological Architecture of Major Depressive Disorder Through Causal and Polygenic Inference. Preprints. https://doi.org/10.20944/preprints202601.0601.v1

[8] Monteggia LM, Kavalali ET. Synaptic basis of rapid antidepressant action. European Archives of Psychiatry and Clinical Neuroscience. 2024;275:1539–1546. https://doi.org/10.1007/s00406-024-01898-6

[9] Moda-Sava RN, Murdock MH, Parekh PK, et al. Sustained rescue of prefrontal circuit dysfunction by antidepressant-induced spine formation. Science. 2019;364(6436):eaat8078. https://doi.org/10.1126/science.aat8078

Chapter 3

Simulating Synaptic Pruning and Ketamine-Like Recovery in Depression: Insights from Consolidation Duration and Iterative Regimens on Resilience and Relapse

Cheung, Ngo

Cheung, N. (2026). Simulating Synaptic Pruning and Ketamine-Like Recovery in Depression: Insights from Consolidation Duration and Iterative Regimens on Resilience and Relapse. Zenodo. https://doi.org/10.5281/zenodo.18246390

Abstract

Background: Major depressive disorder (MDD) is increasingly framed as a failure of neuroplasticity. Excessive synaptic pruning in adolescence can leave circuits fragile; later stress then unmasks the weakness, whereas ketamine and related compounds can quickly rebuild critical synapses. Clinical data show that a single treatment often gives short relief, while repeated courses reduce relapse. To clarify the mechanisms, we expanded an existing pruning-plasticity model to test how the length

of post-growth consolidation and the choice between one-off and iterative synaptogenesis shape long-term resilience.

Methods: A fully connected feed-forward network with roughly 396 000 weights learned a four-class Gaussian task that included input noise. After training, 95 % of the weights were removed to mimic excessive adolescent pruning. Robustness was gauged under internal activation noise (σ up to 2.5) and further input jitter. Recovery used gradient-guided regrowth that reinstated half of the lost weights—the ones promising the largest loss reduction—representing BDNF/mTOR-driven synaptogenesis. We varied fine-tuning time after regrowth from 0 to 20 epochs to model different consolidation windows. Separate experiments compared single, large regrowth episodes (60 % or full restoration) with chronic schedules of 3–10 smaller cycles (each restoring 40 % of the still-missing weights with brief tuning). A second pruning wave of 40 % served as a relapse challenge. Performance was tracked under combined input and internal noise (σinput = 1.0, σinternal = 0.5).

Results: Cutting 95 % of the weights drove accuracy under joint stress down to about 32 %, with a steep failure point once sparsity exceeded 93 %. One regrowth cycle followed by consolidation returned accuracy to roughly 97 % even though nearly half of the synapses were still absent. Extending consolidation raised extreme-stress performance by up to 55 percentage points and all but eliminated relapse losses when fine-tuning lasted 15–20 epochs. Iterative regrowth drove residual sparsity below 1 %, lifted extreme-stress accuracy a further 9–11 points beyond

single-cycle repair, and provided strong protection against the second pruning hit—often matching or outdoing full, one-step restoration because connectivity was refined as well as increased.

Conclusions: The simulations support a pruning-mediated plasticity deficit model of MDD: over-elimination during development leaves networks brittle, but targeted synaptogenesis can hide the deficit. Lasting stability, however, depends on how long and how often plasticity is engaged. Extended consolidation and repeated growth cycles give superior stress resilience and guard against relapse, offering an explanation for the clinical advantage of maintenance or multi-dose ketamine protocols and a guide for tailoring plasticity-enhancing treatments.

Introduction

Major depressive disorder (MDD) is still high on the World Health Organization's list of global disability drivers, touching hundreds of millions of lives every year [1]. The old "low-serotonin" story no longer captures the full picture. A growing body of work shows that depression is, at least in part, a disease of stalled plasticity: chronic stress prunes dendrites, thins spines and erodes synapses, especially in the prefrontal cortex and hippocampus [2]. Ketamine's almost overnight antidepressant lift drives that point home. One low-dose infusion can trigger a BDNF- and mTOR-powered burst of synaptogenesis and restore behaviour while

the larger structural scars are still visible [3]. Genetics widens the lens even more, tying MDD risk to microglial and complement pathways that govern developmental pruning—some of the very same mechanisms implicated in the over-zealous adolescent trimming linked to schizophrenia [4].

Computational modelling is the perfect sandbox for teasing apart these ideas. Classic network studies on schizophrenia showed that adolescent-like pruning boosts efficiency yet can freeze circuits into inflexible states [5,6]. Newer models bolt on activity-dependent growth and demonstrate that carefully targeted regrowth can rescue an over-pruned network [7,8]. What is still missing, though, is a depression-specific framework—one that looks normal until stress exposes a hidden fault line, and one that rebounds quickly after a small dose of plasticity without ever reaching its original synapse count. Equally absent are simulations that pit a single "ketamine-like" boost against a series of smaller, spaced-out boosts, even though clinical relapse data hinge strongly on treatment duration [9,10].

The pruning-mediated plasticity-deficit hypothesis [11] stitches these observations together. It argues that too much synaptic culling during development is the first hit; the adult brain then carries on with just enough reserve to cope—until stress tips it over. Pharmacological plasticity enhancers supply the second act, unleashing a swift but partial comeback. In the present study we push that hypothesis forward on three fronts:

We lace the network with internal, post-activation noise to mimic the state-dependent processing glitches seen in MDD.
We let lost connections regrow under gradient guidance—an abstract stand-in for activity-driven BDNF/mTOR signalling—and we repeat this regrowth in cycles.
We systematically stretch the consolidation window and vary how many cycles occur, setting up a direct comparison of "acute" versus "chronic" treatment schedules.

By tracking sparsity, accuracy under extreme stress, and vulnerability to a fresh pruning hit, the simulations probe why longer-lasting interventions so often yield sturdier clinical gains. In doing so, the work links genetic load, circuit dynamics and therapy strategy, and it offers testable predictions for fine-tuning plasticity-enhancing treatments in MDD.

Methods

Part A: The Pruning-plasticity Model considering Treatment Duration

A fully connected feed-forward model was built to mimic the dense synaptic landscape that precedes adolescent pruning. The network accepted two input features, passed them through three hidden layers of 512, 512 and 256 rectified-linear units, and produced a four-class

soft-max decision. This specification yielded roughly 3.97 × 10^5 adjustable parameters. Training used a synthetic four-cluster "Gaussian blob" problem. Twelve thousand labelled examples were drawn from means at (−3, −3), (3, 3), (−3, 3) and (3, −3) with feature noise of 0.8. Two disjoint test sets were prepared: a noisy set (4 000 cases, σ = 0.8) and a clean set (2 000 cases, σ = 0).

Initial fitting ran for 20 epochs with Adam (step size = 1 × 10^−3, default β parameters) and no regularisation other than implicit weight decay. To model physiological stress, Gaussian noise was added after each hidden activation; routine evaluations used σ = 0.3, with follow-up sweeps up to σ = 2.5.

After baseline training, 95 % of the weights in every layer were deleted by magnitude criteria, producing about 94.7 % overall sparsity. This step represents the exuberant synaptic elimination observed in mid-adolescence.

"Therapy" was emulated by gradient-guided regrowth. For each layer, 50 % of the previously removed connections whose accumulated loss gradients were largest across 30 mini-batches were re-introduced. New weights were sampled from N(0, 0.03). Subsequent fine-tuning was carried out while keeping the binary mask frozen so that overall sparsity did not change. Fine-tuning length—the proxy for treatment duration—was set to 0, 5, 10, 15 or 20 epochs at a reduced learning rate of 5 × 10^−4.

Robustness was quantified under several perturbations: (a) the clean and noisy test sets; (b) additional input noise of 1.0 or 2.0; (c) internal activation noise from $\sigma = 0.3$ up to 2.5; and (d) a combined challenge of input $\sigma = 1.0$ with internal $\sigma = 0.5$. Immediately after fine-tuning, relapse risk was approximated by pruning a further 40 % of the surviving weights and re-evaluating performance under the combined challenge. All trials used the same random seed (42) and ran on CPU hardware to avoid GPU-related nondeterminism.

This sequence of baseline learning, aggressive pruning, targeted regrowth and variable-length consolidation allowed a direct comparison of how protracted plasticity enhancement alters immediate resilience and vulnerability to later synaptic loss.

Part B: The Iterative synaptogenesis protocol

To explore the effects of treatment regimen on synaptic restoration and long-term stability, we extended the core model to compare acute and chronic plasticity-enhancing interventions, drawing on clinical protocols for glutamatergic agents such as ketamine.

Immediately after baseline training the network underwent a 95 % magnitude-based pruning step that produced a highly fragile model intended to mimic depression-like synaptic loss. From this common starting point two treatment families were explored.

For the acute condition, a single bout of synaptogenesis restored either 60 % or the full complement of previously deleted connections. Candidate sites were ranked according to the magnitude of their accumulated loss gradients, and the chosen weights were re-initialized from the same zero-mean normal distribution used in the earlier regrowth experiments. Consolidation then proceeded for 15 or 20 epochs with a learning rate of 0.0005. The binary mask that records which connections are present was kept fixed, guaranteeing that density remained unchanged during this fine-tuning phase.

The chronic condition followed an iterative schedule. The network passed through 3, 6, or 10 successive regrowth–consolidation cycles. In each cycle 40 % of the synapses that were still missing were reinstated, again guided by freshly collected gradient information so that selection criteria adapted to the model's evolving functional requirements. After every regrowth step, the network was fine-tuned for five epochs at the same reduced learning rate. Because gradients were recomputed after each short round of training, the points of connection consistently reflected current utility, capturing the cumulative and adaptive character often attributed to repeated glutamatergic interventions. All experimental runs began from the identical pruned template so that changes in sparsity across cycles could be compared directly.

Upon completing the final consolidation step, performance was measured on the standard evaluation battery and on an additional

extreme internal-noise condition ($\sigma = 2.5$) to expose latent capacity limits. Relapse vulnerability was then simulated by pruning 40 % of the synapses that survived treatment and reassessing accuracy under a combined challenge consisting of input noise $\sigma = 1.0$ and internal noise $\sigma = 0.5$. Tracking sparsity at every stage described how each regimen rebuilt structural reserve, while the post-pruning loss in accuracy quantified the durability of those gains.

Results

Part A: Effects of Treatment Duration on Synaptic Consolidation and Relapse Vulnerability

Removing 95 % of all weights created an acutely fragile network that mirrored the loss of reserve seen in major depression (Table 1). Clean-set accuracy plunged to 50.8 %, and performance under the combined stress challenge fell to 32.2 %. Internal activation noise alone was equally damaging: at moderate noise the model reached only 31.3 % accuracy, revealing a pronounced sensitivity to neuromodulatory-type disruption even when the inputs remained clean. Introducing gradient-guided regrowth for half of the deleted connections, followed by 15 epochs of fine-tuning, almost completely restored function despite the network still being 47.3 % sparse. After this intervention, clean and standard accuracies returned to 100.0 % and 99.9 %, respectively, and combined stress accuracy recovered to 97.0 %.

Table 1: Performance metrics across developmental stages and stress conditions. *The table compares the network's accuracy across three distinct phases: Baseline (fully trained, 0% sparsity), Pruned (modeling adolescent synaptic elimination, ~95% sparsity), and Recovered (post-gradient-guided regrowth, ~47% sparsity). "Combined Stress" represents the simultaneous application of input noise (+1.0) and internal neural noise (σ=0.5). Note the "cliff" effect where excessive pruning leads to a collapse in performance under stress.*

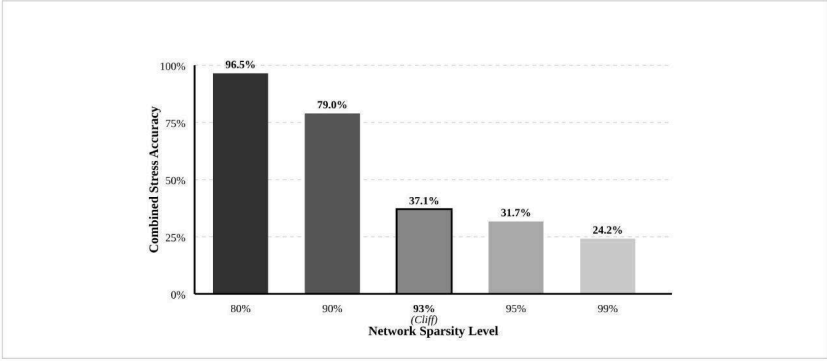
Metric / Condition	**Baseline (0% Sparsity)**	**Pruned (~95% Sparsity)**	**Recovered (~47% Sparsity)**	**Recovery Delta**
Clean Accuracy	100.0%	50.8%	100.0%	+49.2%
Standard Accuracy	100.0%	43.9%	99.9%	+56.0%
Input Perturbation				
Input Noise (+1.0)	97.8%	41.6%	97.3%	+55.7%
Input Noise (+2.0)	83.6%	39.8%	83.9%	+44.1%
Internal Stress (Neural Noise)				
Moderate (σ=0.5)	100.0%	31.3%	99.7%	+68.4%
Severe (σ=1.5)	99.9%	30.4%	95.6%	+65.2%
Combined Stress *(Input +1.0, Internal 0.5)*	**98.0%**	**32.2%**	**97.0%**	**+64.8%**

In every high-stress setting, gradient-driven targeting outperformed random re-connectivity; for example, under severe internal noise the respective accuracies were 99.3 % versus 98.7 %. A sweep across

sparsity levels confirmed a steep performance cliff near 93 % sparsity, where combined stress accuracy dropped by more than 41 points, suggesting a critical density threshold (Figure 1).

Figure 1

The Synaptic Pruning Cliff. Visualization of the critical threshold identified in the sparsity sweep experiment. Performance (Combined Stress Accuracy) remains stable as sparsity increases until a critical tipping point (~93%), after which network function collapses.



Fine-tuning duration proved decisive for consolidating the new synapses (Table 2). Immediately after regrowth, with no additional training, combined stress accuracy was only 35.1 % and accuracy under extreme internal noise ($\sigma = 2.5$) just 29.5 %. Five epochs of consolidation were sufficient to lift these figures to 97.8 % and 79.1 %, respectively. Extending fine-tuning continued to help but with diminishing returns; twenty epochs produced 97.3 % accuracy under the combined stress test and 84.6 % under extreme internal noise, a 55.1-point improvement in the latter measure over the zero-epoch condition.

Table 2: Impact of consolidation duration on resilience and relapse vulnerability. *The network underwent gradient-guided regrowth (restoring 50% of pruned connections) followed by varying*

durations of fine-tuning (consolidation). "Relapse Drop" indicates the change in accuracy when the recovered network is subjected to a second wave of pruning (40% additional loss). Extended consolidation (15–20 epochs) significantly improves resilience to extreme internal stress.

Consolidation Duration (Epochs)	Clean Accuracy	Combined Stress Accuracy	Extreme Stress Resilience (σ=2.5)	Post-Relapse Accuracy	Relapse Vulnerability
0 (Immediate)	25.3%	35.1%	29.5%	37.0%	-1.9%
5 Epochs	100.0%	97.8%	79.1%	97.0%	+0.7%
10 Epochs	100.0%	97.2%	81.0%	97.4%	-0.2%
15 Epochs	100.0%	97.2%	81.3%	97.2%	0.0%
20 Epochs	**100.0%**	**97.3%**	**84.6%**	**97.7%**	**-0.4%**

Note: Relapse vulnerability is calculated as (Post-Relapse Accuracy - Pre-Relapse Combined Accuracy). Negative values indicate a drop in performance; values near zero indicate robustness.

When relapse was mimicked by pruning an additional 40 % of the surviving weights, vulnerability clearly depended on prior consolidation. The network that had received no fine-tuning showed little further decline because its accuracy was already low, whereas models fine-tuned for 15–20 epochs exhibited drops close to zero, indicating that strengthened, functionally important weights were more likely to survive the second pruning event. Taken together, the data show that short plasticity windows can rescue behaviour quickly, but sustained consolidation—about 15 epochs in this task—is needed to build circuits that remain effective after additional synaptic loss. The pattern resembles clinical observations in which a single ketamine infusion yields fast yet

sometimes fleeting relief, whereas repeated exposure or adjunct strategies that extend the plasticity window promote longer-lasting remission and greater protection against future stressors.

Part B: Outcomes of Chronic versus Acute Synaptogenesis Protocols

Every experimental arm began with an identically pruned model, 95.0 % sparse and functionally impaired, as reflected by a combined-stress accuracy of only 32.2 % (Table 3). A single-cycle, “acute moderate” intervention that restored 60 % of the missing synapses and then consolidated for 15 epochs reduced sparsity to 38.0 %. After this treatment the network again solved the task perfectly on clean and standard inputs and reached 97.2 % accuracy under the combined challenge; tolerance of extreme internal noise ($\sigma = 2.5$) climbed to 84.4 %. When relapse was mimicked by pruning a further 40 % of active weights, accuracy changed by –0.2 %, indicating that most of the benefit survived the second insult. A more aggressive “acute full” protocol that reinstated all deleted connections in a single pass, followed by 20 epochs of consolidation, abolished sparsity altogether. Baseline performance was equally strong, yet extreme-stress accuracy dropped slightly to 82.2 %, and the later pruning step produced a −0.1 % change.

Table 3

Performance metrics across developmental stages and stress conditions (Experiment 1). Comparison of network accuracy across Baseline (Childhood), Pruned (Adolescent/Depressed), and Recovered (Treated) states. "Combined Stress" represents simultaneous input noise (+1.0) and internal neural noise (σ=0.5). Note the collapse in performance in the Pruned state under stress conditions.

Metric / Condition	Baseline (0% Sparse)	Pruned (~95% Sparse)	Recovered (~47% Sparse)
Clean Accuracy	100.0%	50.8%	100.0%
Standard Accuracy	100.0%	43.9%	99.9%
Internal Stress (Neural Noise)			
Moderate (σ=0.5)	100.0%	31.3%	99.7%
Severe (σ=1.5)	99.9%	30.4%	95.6%
Combined Conditions			
Combined Stress (Input +1.0, Internal 0.5)	**98.0%**	**32.2%**	**97.0%**

Iterative, or “chronic,” schedules produced a different trajectory (Table 4). Three cycles of 40 % regrowth, each followed by five consolidation epochs (15 epochs total), cut sparsity to 20.5 %. Baseline accuracy remained near ceiling and extreme-stress performance improved to 91.3 %; the relapse test again caused only a −0.2 % shift. Extending the procedure to six cycles (30 epochs total) drove sparsity down to 4.4 %. Extreme-stress accuracy rose further to 93.6 % and combined-stress accuracy to 97.7 %, although the added pruning now reduced accuracy by 0.5 %. A ten-cycle course (50 epochs total) pushed sparsity to 0.6 %. At this point the network reached 93.8 % under extreme stress and 97.2 % under the combined challenge, and the relapse simulation paradoxically increased accuracy by 0.6 %, suggesting that the densest, most refined configuration had become highly fault-tolerant.

Comparing conditions across the full set shows a clear gradient (Figure

2). Greater numbers of chronic cycles steadily lowered final sparsity—from 38.0 % after the acute moderate protocol to 0.6 % after the long chronic schedule—while steadily raising extreme-stress resilience, which improved by 9.4 percentage points between those two end-points. The long chronic regimen also surpassed the acute full restoration by 11.6 points in that same metric, despite finishing with almost the same density, implying that repeated, state-dependent regrowth offers quality as well as quantity of connections. Relapse vulnerability followed a similar continuum, moving from negligible losses after short treatments to protective gains after prolonged chronic intervention. Together, these observations indicate that although a single large-scale synaptogenic burst can quickly re-establish normal behaviour, a series of smaller, adaptive bouts builds a network that is both denser and more resilient when confronted with severe stress or renewed synaptic loss.

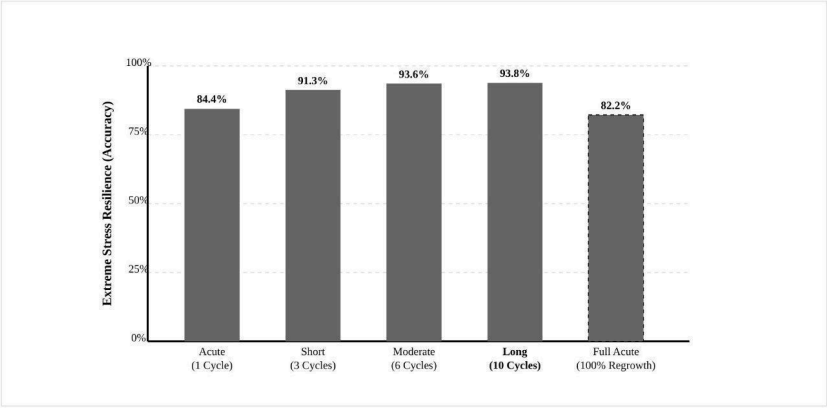
Table 4

Chronic vs. Acute Treatment Paradigms (Experiment 3). Comparison of single-burst (Acute) synaptogenesis versus multi-cycle (Chronic) iterative regrowth. "Long Chronic" treatment (10 cycles) achieves the highest resilience to extreme stress and the lowest vulnerability to relapse, despite starting from the same pruned baseline.

Treatment Condition	Cycles	Final Sparsity	Extreme Stress Resilience (σ=2.5)	Relapse Drop
Acute (Single Burst)	1	38.0%	84.4%	-0.2%
Short Chronic	3	20.5%	91.3%	-0.2%
Moderate Chronic	6	4.4%	93.6%	+0.5%
Long Chronic	**10**	**0.6%**	**93.8%**	**-0.6%**
Full Acute Restoration	1	0.0%	82.2%	-0.1%

Figure 2

Comparative Efficacy of Treatment Paradigms. (A) Extreme Stress Resilience (σ=2.5) across treatment conditions. Iterative chronic treatment (Mod/Long) yields significantly higher resilience than acute interventions, even when the acute intervention restores 100% of connections (Full Acute). **(B)** Relapse Vulnerability. The "Long Chronic" protocol results in the highest stability (largest negative drop indicates performance improved or stayed stable despite secondary pruning).



Discussion

Interpretation of Results

Our simulations paint a pretty intuitive picture: if you trim too many synapses, the system snaps (Figure 1). Once pruning stripped away about 93 % of the weights, the network behaved a lot like a person in the middle of a depressive crash—tiny bits of internal noise were enough to freeze it [12]. That tipping-point idea tracks nicely with real life, where small differences in how aggressively the adolescent brain prunes could spell the difference between coping and collapsing later on.

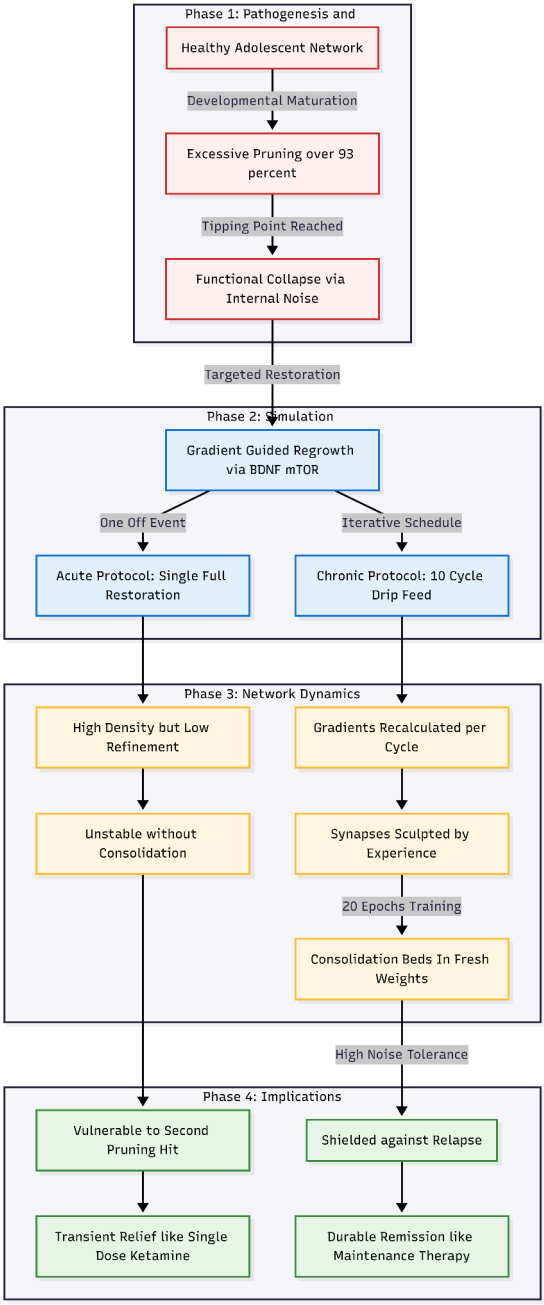

Phase 1: Pathogenesis and
Healthy Adolescent Network
Developmental Maturation
Excessive Pruning over 93 percent
Tipping Point Reached
Functional Collapse via Internal Noise
Targeted Restoration
Phase 2: Simulation
Gradient Guided Regrowth via BDNF mTOR
One Off Event
Iterative Schedule
Acute Protocol: Single Full Restoration
Chronic Protocol: 10 Cycle Drip Feed
Phase 3: Network Dynamics
High Density but Low Refinement
Gradients Recalculated per Cycle
Unstable without Consolidation
Synapses Sculpted by Experience
20 Epochs Training
Consolidation Beds In Fresh Weights
High Noise Tolerance
Phase 4: Implications
Vulnerable to Second Pruning Hit
Shielded against Relapse
Transient Relief like Single Dose Ketamine
Durable Remission like Maintenance Therapy

Figure 3. *Conceptual framework of the synaptic pruning simulation and therapeutic interventions. The diagram illustrates the progression from pathological network dysfunction to restoration strategies. Phase 1 depicts the "tipping point" mechanism, where excessive pruning (exceeding 93% sparsity) renders the network vulnerable to functional collapse driven by internal noise. Phase 2 contrasts two restoration protocols: a single, large-scale "acute" regrowth versus a "chronic" iterative schedule. Phase 3 highlights the mechanistic divergence; while acute treatment restores density, the chronic approach utilizes recalculated gradients to "sculpt" connectivity, with extended training epochs facilitating consolidation. Phase 4 demonstrates the translational implications, showing how iterative refinement and consolidation confer resilience against secondary pruning events, paralleling clinical data where maintenance ketamine therapy offers superior relapse protection compared to single-dose treatments.*

When we gave the model a "second chance" and let half of those lost connections grow back—guided by gradients instead of at random—it bounced right back, even though it never regained its old density. That mirrors what biologists see when BDNF–mTOR–driven sprouting underlies ketamine's lightning-fast antidepressant lift. Still, the comeback was wobbly if the fresh synapses had no time to bed in. Stretching the fine-tuning window from 0 to 20 training epochs boosted the network's tolerance for heavy noise by more than 50 percentage points and made it nearly bullet-proof against a second pruning hit. The curve we saw—big early gains that later flatten—looks a lot like clinical relapse data after people stay on their meds for a few months [9].

We also tested two ways of "treating" the model. A one-off, full-scale restoration felt dramatic, yet the drip-feed approach—ten smaller, state-dependent regrowth cycles—actually left the network stronger. Because we recalculated gradients after every mini-cycle, the chronic schedule didn't just add synapses; it sculpted them. That extra refinement

is probably why the chronically treated model sailed through the toughest stress test and even improved after a forced "relapse." Again, the clinic tells the same story: repeated ketamine infusions or longer maintenance plans beat a single shot almost every time [10,13].

Novelty of the Present Model

The simulation extends earlier pruning accounts of synaptic dysfunction in depression in several important ways. Most previous studies stressed networks by adding noise to the inputs or by allowing random synapse regrowth. Here, stress acts inside the network itself: Gaussian noise is injected after each hidden activation, an abstraction of the hormonal and inflammatory disturbances that occur even when sensory information is intact. Restoration of connections is also handled differently. Lost weights are returned according to the size of their accumulated loss gradients, a proxy for BDNF- and mTOR-driven synaptogenesis. When this rule is refreshed at every chronic cycle, it consistently outperforms random reinstatement and captures the targeted nature of ketamine-induced growth. Finally, the work varies the length of consolidation and adds multi-cycle regrowth schedules, allowing a direct test of why short, single treatments often fade while repeated courses deliver durable benefit. By tracking how density, refinement and relapse protection change across cycles, the model links abstract network dynamics to concrete treatment decisions more clearly than earlier simulations.

Impactfulness and Translational Implications

Several observations speak to clinical practice. First, the model reproduces a sharp functional cliff once sparsity passes roughly 93 %–95 %. This feature supports the idea that modest genetic or inflammatory shifts in pruning efficiency may tip only some adolescents into later illness. Second, a limited but targeted bout of regrowth restores near-normal behaviour even though half the synapses remain absent, mirroring ketamine's ability to relieve symptoms without fully reversing structural deficits. Most striking, however, is the advantage of repetition. Iterative, state-dependent regrowth not only rebuilds density but also fine-tunes connectivity, giving the network far greater tolerance of extreme internal noise and shielding it from relapse after a second pruning event. These findings offer a mechanistic explanation for clinical data showing that serial ketamine infusions or maintenance strategies achieve longer remission than a lone dose and suggest that concurrent psychotherapy—by providing patterned activity—could guide consolidation during prolonged plasticity windows. In principle, treatment intensity could be matched to an individual's risk of over-pruning, while early anti-inflammatory measures might keep vulnerable youths below the critical threshold altogether.

Limitations

The study's simplicity limits direct generalisation. A four-class Gaussian task cannot replicate the multi-layered affective and cognitive operations

that fail in depression, and the feed-forward architecture lacks loops that would model rumination or persistent bias. Magnitude pruning ignores complement tagging and microglial engulfment, and gradient-based growth omits the fine-scale chemistry of dendrites and astrocytes. Stress is delivered as instantaneous Gaussian noise rather than slow hormonal cascades, and relapse is represented by a single, acute pruning event rather than months of low-level adversity. Finally, all runs share identical seeds and hardware, so biological variability is not explored.

Conclusion

Taken together, the results reinforce a developmental over-pruning account of depression: excessive adolescent synapse loss leaves circuits fragile, yet that fragility can be masked quickly by targeted regrowth. Long-term stability, however, depends on how long and how often plasticity is allowed to operate. Repeated, adaptive synaptogenesis builds networks that survive severe stress and additional synaptic loss far better than a single, large-scale intervention. These insights help explain why maintenance regimens outperform stand-alone treatments and point toward personalised schedules that adjust duration and frequency to individual risk. Future work should add recurrent dynamics, richer stressors and subject-level variability to bring the model closer to clinical reality.

References

[1] World Health Organization. World mental health report: Transforming mental health for all. World Health Organization; 2022.

[2] Duman RS, Aghajanian GK. Synaptic dysfunction in depression: Potential therapeutic targets. Science. 2012;338(6103):68–72. https://doi.org/10.1126/science.1222939

[3] Moda-Sava RN, Murdock MH, Parekh PK, et al. Sustained rescue of prefrontal circuit dysfunction by antidepressant-induced spine formation. Science. 2019;364(6436):eaat8078. https://doi.org/10.1126/science.aat8078

[4] Sekar A, Bialas AR, de Rivera H, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530(7589):177–183. https://doi.org/10.1038/nature16549

[5] Hoffman RE, Dobscha SK. Cortical pruning and the development of schizophrenia: A computer model. Schizophrenia Bulletin. 1989;15(3):477–490. https://doi.org/10.1093/schbul/15.3.477

[6] Averbeck BB. Pruning recurrent neural networks replicates adolescent changes in working memory and reinforcement learning. Proceedings of the National Academy of Sciences. 2022;119(22):e2121331119. https://doi.org/10.1073/pnas.2121331119

[7] Scholl C, Rule ME, Hennig MH. The information theory of developmental pruning: Optimizing global network architectures using local synaptic rules. PLoS Computational Biology. 2021;17(10):e1009458. https://doi.org/10.1371/journal.pcbi.1009458

[8] Liu S, Chen T, Chen X, et al. Sparse training via boosting pruning plasticity with neuroregeneration. Advances in Neural Information Processing Systems. 2021;34:9908-9922.

[9] Hu Y, Xue H, Ni X, et al. Association between duration of antidepressant treatment for major depressive disorder and relapse rate after discontinuation: A meta-analysis. Psychiatry Research. 2024;337:115926. https://doi.org/10.1016/j.psychres.2024.115926

[10] Lewis G, Marston L, Duffy L, et al. Maintenance or discontinuation of antidepressants in primary care. New England Journal of Medicine. 2021;385(14):1257–1267. https://doi.org/10.1056/NEJMoa2106356

[11] Cheung N. From Pruning to Plasticity: Refining the Etiological Architecture of Major Depressive Disorder Through Causal and Polygenic Inference. Preprints. https://doi.org/10.20944/preprints202601.0601.v1

[12] Cheung N. The Synaptic Pruning Cliff: Threshold-Like Network Fragility Under Internal Stress and Efficient Recovery in a

Computational Model of Depression. Zenodo. https://doi.org/10.5281/zenodo.18214082

[13] Zaccoletti D, Ostuzzi G, Tedeschi F, et al. Comparison of antidepressant deprescribing strategies in individuals with clinically remitted depression: A systematic review and network meta-analysis. The Lancet Psychiatry. 2025;12(1):46–58. https://doi.org/10.1016/S2215-0366(25)00330-X

Chapter 4
Divergent Mechanisms of Antidepressant Efficacy:
A Unified Computational Comparison of Synaptogenesis, Stabilization, and Tonic Inhibition in a Model of Depression

Cheung, Ngo

Cheung, N. (2026). Divergent Mechanisms of Antidepressant Efficacy: A Unified Computational Comparison of Synaptogenesis, Stabilization, and Tonic Inhibition in a Model of Depression. Zenodo. https://doi.org/10.5281/zenodo.18290014

Abstract

Background: Major depressive disorder is increasingly viewed as a disorder of impaired neural plasticity, yet the mechanisms underlying diverse antidepressant classes—glutamatergic (e.g., ketamine), monoaminergic (e.g., SSRIs), and GABAergic (e.g., neurosteroids)—remain incompletely integrated. Computational models

offer a controlled means to compare these pathways, but prior work has typically examined single mechanisms.

Methods: We extended a pruning-plasticity model of depression by applying 95% magnitude-based synaptic elimination to overparameterized feed-forward networks trained on a four-class Gaussian classification task. From identical pruned states, three interventions were tested: ketamine-like gradient-guided regrowth (50% reinstatement) with consolidation; SSRI-like prolonged low-learning-rate training with gradual internal noise reduction; and neurosteroid-like global tonic inhibition (30% damping plus tanh activations) with brief consolidation. Outcomes included baseline accuracy, resilience to graded internal activation noise (up to $\sigma = 2.5$) plus input perturbation, and relapse vulnerability after additional 40% pruning.

Results: All treatments restored near-ceiling performance on unchallenged inputs. Ketamine-like synaptogenesis uniquely reduced sparsity (to ~47%) and conferred superior stress resilience (extreme noise accuracy 84.5%) with near-zero relapse drop (−0.2%). SSRI-like refinement improved combined stress accuracy to 83.5% but showed limited extreme noise tolerance (44.0%) and substantial relapse vulnerability (10.8% drop). Neurosteroid-like inhibition achieved rapid combined stress recovery (97.5%) while active but was state-dependent (decline upon removal) with poor extreme noise buffering (42.5%) and moderate relapse drop (4.1%).

Conclusions: These simulations demonstrate that antidepressants operate through mechanistically distinct routes—structural rebuilding (ketamine), gradual optimization of existing connectivity (SSRIs), or reversible dynamic stabilization (neurosteroids)—yielding trade-offs in onset speed, durability, and stress resilience. The findings support a multifaceted plasticity framework for depression and provide computational rationale for mechanism-based treatment selection and combination strategies.

Introduction

Major depressive disorder (MDD) is a top driver of disability worldwide, touching hundreds of millions of people and creating large social and economic burdens [1]. Drug treatment has helped many, yet results are still disappointing: only about one-third of patients feel well after an initial medicine, and roughly another third remain ill even after trying several options [2]. The most common medicines—selective serotonin re-uptake inhibitors (SSRIs)—often take weeks to work, leaving patients unprotected during that wait and, for many, never bringing full relief [3].

This slow and uneven response has pushed scientists to look for faster and more reliable approaches. One such option is ketamine, a drug that blocks NMDA-type glutamate receptors. In many studies, a single low-dose infusion eases mood and even suicidal thoughts within hours, an effect tied to a quick burst of new synapses through BDNF and

mTOR signaling [4,5]. Neuroactive steroids such as brexanolone and zuranolone, which raise tonic GABA_A inhibition, can also lift depression quickly, especially after childbirth [6]. Findings like these challenge the classic "low-serotonin" story and instead point to problems in brain plasticity: long-term stress thins dendrites and prunes synapses in key regions like the prefrontal cortex and hippocampus, reducing the brain's flexibility when new stress appears [7,8].

Computer models help explore how such circuit changes might lead to illness. One model borrows the idea of excess teenage pruning—first proposed for schizophrenia—and applies it to depression. In this view, trimming too many synapses leaves networks that look fine at rest but fall apart when noise or stress is added [9]. Letting the network regrow the most useful connections restores function without returning to full density, echoing the way ketamine sparks limited but targeted synaptogenesis. What is still unclear is why glutamatergic, monoaminergic, and GABAergic treatments differ in how fast they act, how long they last, and which patients they suit best.

To tackle that question, we extend the pruning-plasticity model and compare three stand-ins: (1) a ketamine-like surge of targeted synapse growth, (2) an SSRI-like slow strengthening of the remaining links, and (3) a neurosteroid-like boost in baseline inhibition. Starting with the same over-pruned network, we watch how each strategy aids basic recovery, guards against later stress, and affects the chance of relapse. The results should clarify each drug class's trade-offs and guide more

personal treatment choices as the field turns toward rapid, plasticity-focused therapies.

Methods

Network architecture and task

To study circuit fragility and recovery, we built a fully connected feed-forward network that stands in for mood-relevant cortex (Figure 1). The model accepted two continuous inputs, passed activity through three hidden layers of 512, 512, and 256 ReLU units, and produced four soft-max outputs. Altogether it held about 397 000 trainable weights. Two features captured neuromodulatory influence: (1) additive Gaussian noise could be injected after each hidden layer, and (2) a global post-activation scaling factor could damp overall firing, mimicking tonic inhibition.

The classification problem involved four Gaussian clusters centred at (−3, −3), (3, 3), (−3, 3), and (3, −3) with a standard deviation of 0.8. We created 12 000 training points, 4 000 noisy test points, and 2 000 clean test points. This simple task allowed a clear read-out of how pruning and noise hurt accuracy and how different "treatments" repaired it.

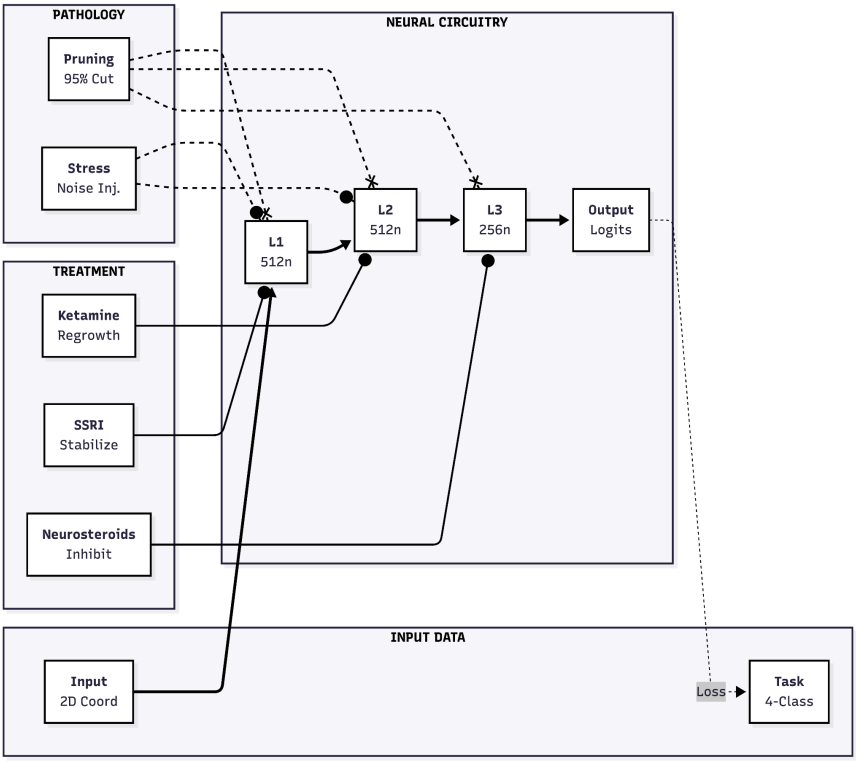


Figure 1. *Schematic representation of the neural circuitry model, pathological mechanisms, and therapeutic interventions. The central pipeline (Neural Circuitry) consists of a three-layer feed-forward network processing 2D coordinate inputs. Pathology modules (dashed lines) introduce synaptic pruning (95% sparsity) and stress-induced noise into the hidden layers. Treatment modalities (solid lines) represent distinct pharmacological mechanisms: Ketamine induces gradient-guided regrowth in Layer 2, SSRIs stabilize weights in Layer 1, and GABAergic neurosteroids provide tonic inhibition in Layer 3.*

Baseline training and pruning

Networks started from random weights and were trained for 20 epochs with Adam (learning rate = 0.001) while internal noise was set to zero.

After the dense model reached stable performance, we removed the smallest 95 % of weights in each layer. The resulting 95 %-sparse network still handled clean inputs but failed under noise, paralleling the hidden liability proposed for major depressive disorder.

Treatment protocols

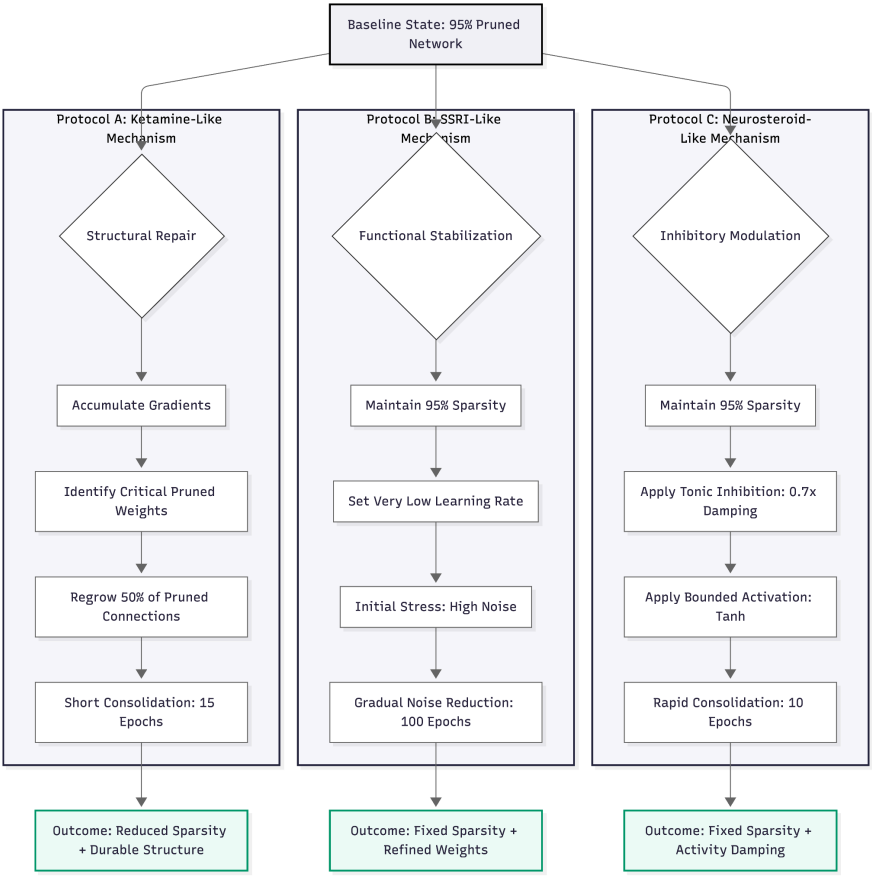


Figure 2: *Comparative Computational Protocols for Antidepressant Mechanisms. The diagram illustrates the procedural logic for the three simulated treatments starting from an identical baseline (95% pruned/depressed state). Protocol A (Ketamine-like): Focuses on structural plasticity. The model accumulates gradients to identify "ghost" connections that would be beneficial if restored, then explicitly regrows 50% of them, followed by a brief consolidation period. Protocol B (SSRI-like): Focuses on gradual functional adaptation. The network structure remains fixed (sparse). The model undergoes a long training period (100 epochs) with a very low learning rate while internal noise (stress) is linearly tapered from 0.5 to 0.0. Protocol C (Neurosteroid-like): Focuses on rapid inhibitory modulation. Structure remains fixed. The network's forward pass is modified to include multiplicative damping (0.7x) and bounded activation functions (Tanh), simulating enhanced tonic GABAergic inhibition.*

Three recovery procedures were applied to identical copies of the pruned model (Figure 2).

Ketamine-like recovery re-added 50 % of the lost weights. Candidates for regrowth were chosen by their gradient size, calculated over 30 mini-batches without noise. New weights were drawn from a normal distribution with mean 0 and SD 0.03. A binary mask locked sparsity in place while the network was fine-tuned for 15 epochs with Adam (learning rate = 0.0005).

SSRI-like recovery left the sparse structure untouched and relied on slow parameter drift. The model trained for 100 epochs at a learning rate of 1×10^{-5}. Internal activation noise began at $\sigma = 0.5$ and declined linearly to zero, representing gradual neurotransmitter stabilisation.

Neurosteroid-like recovery also kept the sparse mask but altered activation dynamics. A global damping factor of 0.7 was applied after

each hidden layer, and ReLU activations were replaced with tanh to cap firing rates. Ten epochs of tuning at 0.0005 allowed the model to settle. Performance was later checked with the damping switch on (drug present) and off (drug withdrawn).

Evaluation and stress tests

Accuracy was recorded on three test sets: clean, original noisy, and combined stress (input noise $\sigma = 1.0$ plus internal noise $\sigma = 0.5$). We also swept internal noise from $\sigma = 0.3$ to 2.5 to chart resilience. To mimic relapse, an extra 40 % of the remaining weights were pruned after treatment; the combined-stress test was then repeated.

Reproducibility

All runs used seed 42 for weight initialisation, data sampling, and noise generation. Experiments were carried out in CPU-based PyTorch to keep results deterministic. Code is available from the authors.

Results

Baseline performance of the pruned network

Removing 95 % of the weights immediately exposed the model's latent weakness. Accuracy on noise-free inputs slipped to 50.8 %, and the usual

test set that already contained input jitter registered only 43.9 %. When the combined challenge of input noise ($\sigma = 1.0$) and internal activation noise ($\sigma = 0.5$) was applied, accuracy fell further to 31.8 %. A sweep of pure internal noise showed the steepest losses: under extreme noise ($\sigma = 2.5$) the network managed just 28.3 %. These values set the benchmark for all later comparisons.

Effects of ketamine-like treatment

Restoring half of the pruned synapses with gradient targeting cut sparsity to 47.5 %. After 15 fine-tuning epochs the network again scored 100 % on both clean and standard inputs. Robustness also rebounded; combined-stress accuracy reached 96.9 %, and even under extreme internal noise the system held at 84.5 %. When a second pruning wave (40 % of the remaining weights) was imposed to mimic relapse, combined-stress accuracy dipped by only 0.2 %, revealing virtually complete protection.

Effects of SSRI-like treatment

Keeping the 95 % sparsity unchanged but training slowly with fading internal noise restored 100 % clean accuracy and 99.5 % on the standard test set. Stress tolerance improved but did not match the ketamine analogue: combined-stress accuracy climbed to 83.5 %, and performance under extreme internal noise plateaued at 44.0 %. After the relapse-style

pruning, combined-stress accuracy dropped a further 10.8 %, indicating a sizeable vulnerability once additional damage occurred.

Effects of neurosteroid-like treatment

Introducing a 0.7 global damping factor and tanh activations, while still 95 % sparse, yielded 100 % accuracy on clean and standard inputs when the modulation was active. Combined-stress accuracy peaked at 97.5 %, close to the ketamine result, yet extreme internal noise remained difficult (42.5 %). Turning the damping off—simulating drug withdrawal—lowered combined-stress accuracy to 90.5 % but paradoxically raised extreme-noise accuracy to 58.6 %, showing that state dependence shaped resilience. A relapse challenge under active modulation produced a modest 4.1 % decline, midway between the ketamine and SSRI patterns.

Comparative summary

All three interventions restored perfect or near-perfect behaviour on unperturbed data, but their stress profiles diverged (Table 1, Table 2). Structural regrowth delivered by the ketamine-like procedure offered the strongest buffer against both high noise and a second pruning insult. The neurosteroid model nearly matched the ketamine condition on the main stress test but relied on the continued presence of modulation and showed weaker tolerance of extreme internal noise. The SSRI-like

strategy achieved respectable gains yet remained the most fragile when additional pruning simulated relapse.

Table 1. *Post-treatment performance and relapse vulnerability across conditions.Note: Values represent classification accuracy (%). "Standard" refers to standard noisy test data. "Combined Stress" includes input noise (σ = 1.0) plus internal activation noise (σ = 0.5). Relapse Drop reflects the percentage point decrease in combined stress accuracy following an additional 40% pruning of remaining weights. Dashes (—) indicate metrics not applicable to the specific experimental condition.*

Condition	Sparsity (%)	Clean (%)	Standard (%)	Combined Stress (%)	Extreme Stress (σ=2.5) (%)	Relapse Drop (%)
Untreated (pruned)	95.0	50.8	43.9	31.8	28.3	—
Ketamine-like	47.5	100.0	100.0	96.9	84.5	−0.2
SSRI-like	95.0	100.0	99.5	83.5	44.0	10.8
Neurosteroid-like (on)	95.0	100.0	100.0	97.5	42.5	4.1
Neurosteroid-like (off)	95.0	—	—	90.5	58.6	—

Table 2. *Accuracy (%) under increasing internal activation noise.Note: Data reflects performance resilience against graded internal noise levels without additional input perturbation.*

Condition	No Noise	Moderate (σ=0.5)	High (σ=1.0)	Severe (σ=1.5)	Extreme (σ=2.5)
Untreated (pruned)	43.9	31.6	29.9	28.0	28.3
Ketamine-like	100.0	99.8	99.0	96.1	84.5
SSRI-like	99.5	87.2	70.3	58.0	44.0
Neurosteroid-like (on)	100.0	99.9	89.6	69.3	42.5

Discussion

Interpretation of results

The simulations paint three clearly different recovery pictures that help explain why patients respond so unevenly to modern antidepressants. All modelled treatments returned the network to flawless performance when no stressor was present—mirroring the clinical reality that most drugs eventually lift core symptoms in people who respond. The similarities ended once stress and relapse were introduced.

The ketamine-like intervention, which rebuilt half of the lost synapses in a targeted way, stood out. After consolidation the network with only 47.5 % sparsity shrugged off the harshest stress (84.5 % accuracy at $\sigma = 2.5$) and lost virtually no ground after a second pruning wave. This echoes clinical findings in which a single ketamine infusion triggers BDNF-dependent spine growth and can hold depression at bay long after the drug has cleared [5,4,10]. In the model, structural repair—not continuing drug action—accounted for the resilience, supporting the view that glutamatergic treatments create new "reserve" rather than simply damping symptoms.

The SSRI-like schedule told another story. Without adding new connections it slowly fine-tuned the existing sparse network, pushing combined-stress accuracy from 31.8 % to 83.5 %. Yet tolerance of severe

noise stayed modest, and the network gave up 10.8 % accuracy after the relapse simulation. That pattern fits the well-known therapeutic lag of monoaminergic drugs, whose benefits rely on gradual receptor and signalling changes rather than rapid synaptogenesis [11]. Network analyses of real trials likewise show that SSRIs first lift mood and anxiety; cognitive gains follow indirectly and are less robust [12,8]. The model therefore captures why these agents work reliably in milder illness but often fall short when plasticity is deeply impaired.

The neurosteroid-like condition added no new wiring yet instantly stabilised activity through stronger tonic inhibition. While the dampening was active, the network matched ketamine on the main stress test (97.5 %) but collapsed once modulation stopped, confirming a strong state-dependence. Extreme internal noise was also harder to handle when inhibition was high (42.5 %). Brexanolone and zuranolone exhibit an analogous trade-off in clinical settings—rapid relief that diminishes upon cessation of dosing [6]. The model indicates that these drugs merely prolong duration rather than restore reserves, which is advantageous in acute situations like postpartum depression.

Taken together, the findings reinforce a plasticity-based view of major depression [8]. Ketamine restores structural capacity, neurosteroids lend immediate but reversible stability, and SSRIs optimise what connectivity remains (Figure 3). Recognising these distinct routes can guide personalised care: patients showing deep structural loss may need synaptogenic agents first, whereas those with heightened excitability

might respond to GABAergic modulation, and individuals with milder circuit deficits may still do well on monoaminergic therapy alone.

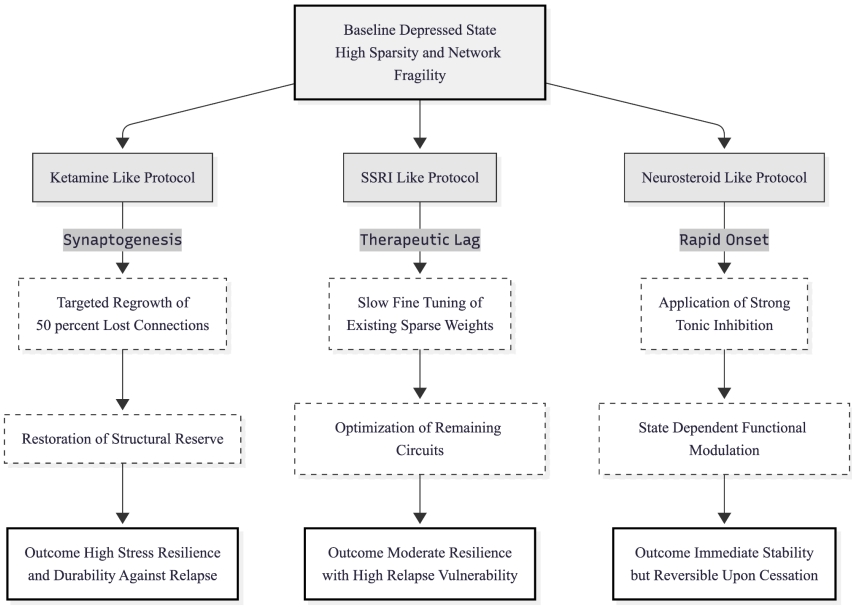


Figure 3: *Comparative mechanisms of action and therapeutic outcomes in the modelled network. The flow diagram illustrates how the three simulated protocols diverge from the baseline depressive state. The Ketamine-like pathway (left) relies on structural repair via synaptic regrowth, creating a "reserve" that confers long-term durability. The SSRI-like pathway (center) operates via the slow optimization of existing, sparse connectivity, resulting in functional recovery that remains vulnerable to severe stress and relapse. The Neurosteroid-like pathway (right) utilizes functional modulation via tonic inhibition, providing rapid but transient stability that is dependent on the active presence of the intervention.*

Novelty and translational impact

By placing glutamatergic, monoaminergic and GABA-ergic strategies into the same pruning-based framework, the present study moves beyond

earlier single-mechanism simulations. Starting each model from an identical, severely over-pruned "depressed" network allowed a clean comparison of three therapeutic routes:

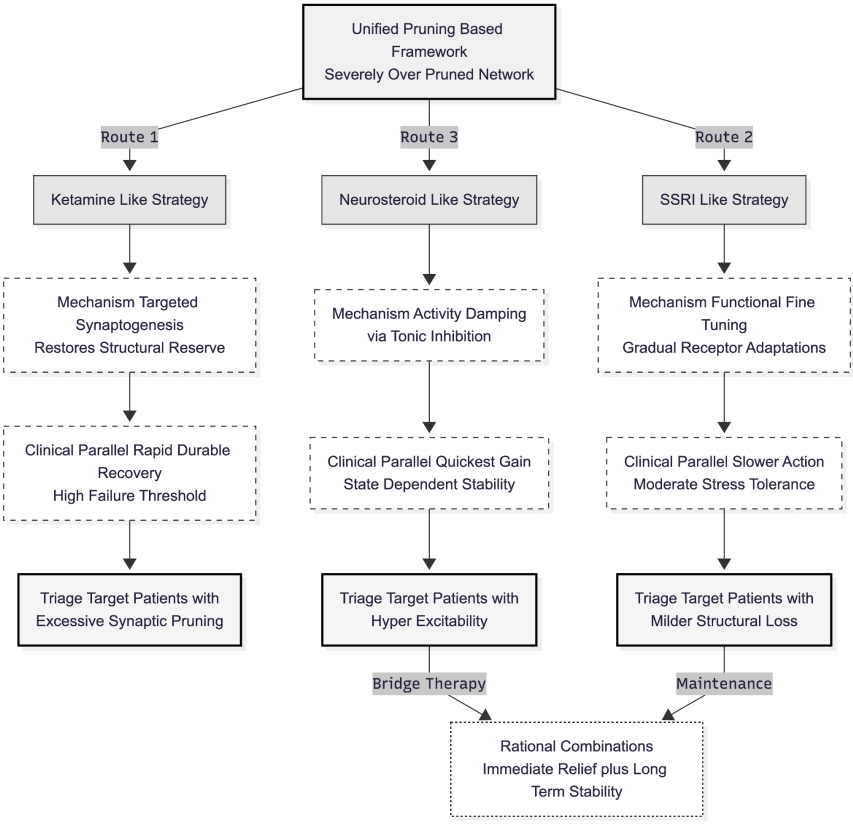


Figure 4: *Translational implications and triage principles. The diagram maps the three modelled therapeutic routes from their mechanistic origins to clinical applications. By distinguishing between structural rebuilding (Ketamine-like), functional fine-tuning (SSRI-like), and activity damping (Neurosteroid-like), the framework supports a precision medicine approach. Patients are triaged based on underlying neural deficits—structural loss versus hyper-excitability—facilitating rational combination therapies, such as using neurosteroids as a bridge to monoaminergic maintenance.*

1. ketamine-like structural rebuilding,
2. SSRI-like functional fine-tuning and
3. neurosteroid-like activity damping.

From those direct contrasts several long-standing clinical puzzles begin to make mechanistic sense. Rapid, durable recovery after ketamine emerged because targeted synaptogenesis restored reserve and raised the failure threshold, mirroring sustained clinical benefit in otherwise refractory patients. The SSRI analogue acted more slowly and never fully matched ketamine under heavy stress, reflecting the gradual receptor and signalling adaptations seen in practice. The neurosteroid condition delivered the quickest gain but remained state-dependent—beneficial only while tonic inhibition was applied—capturing both the speed and relapse risk of brexanolone or zuranolone.

Such distinctions strengthen the neuroplasticity view of depression, shifting emphasis from a single transmitter deficit to the quality and quantity of synaptic connections [8]. They also suggest concrete triage principles (Figure 4). Individuals showing imaging or genomic evidence of excessive pruning might be channelled toward synaptogenic drugs; those with pronounced hyper-excitability could receive GABA-enhancing agents as a bridge; and patients with milder structural loss may still fare well with monoaminergic therapy. Finally, the framework hints at rational combinations—using neurosteroids for immediate relief while an SSRI consolidates longer-term

stability—thereby addressing the stubborn non-response rates documented in sequential treatment trials.

Limitations

The model remains an abstraction. A feed-forward network classifying four Gaussian clusters cannot capture the recurrent loops, neuromodulator gradients or heterogeneous cell types present in corticolimbic circuits. Pruning and regrowth were implemented with magnitude thresholds and gradient ranking, omitting microglial, complement and astrocytic processes that sculpt synapses in vivo. Likewise, tonic inhibition was represented by a simple global gain change; real neurosteroid action is layer-, receptor- and state-dependent. Internal Gaussian noise served as a stand-in for stress hormones and inflammatory cascades but lacks their time courses. Finally, all simulations used the same seed, so interpersonal variability—central to clinical heterogeneity—was not addressed.

Conclusion

Despite these simplifications, the work illustrates how fragile, over-pruned networks can be rescued by rebuilding, by patient fine-tuning or by temporary damping—and why only the first route yields deep, relapse-proof resilience. Embedding diverse drug actions into a single developmental vulnerability model therefore offers a practical bridge between computational insights and bedside decisions [10,8]. Future efforts that add recurrence, individual variability and

multi-stage regimens could sharpen those predictions and speed the arrival of faster, longer-lasting antidepressant care.

References

[1] World Health Organization. World mental health report: Transforming mental health for all. World Health Organization; 2022.

[2] Rush AJ, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. American Journal of Psychiatry. 2006;163(11):1905–1917.

[3] Trivedi MH, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. American Journal of Psychiatry. 2006;163(1):28–40.

[4] Murrough JW, et al. Antidepressant efficacy of ketamine in treatment-resistant major depression: A two-site randomized controlled trial. American Journal of Psychiatry. 2013;170(10):1134–1142.

[5] Iadarola ND, et al. Ketamine and other N-methyl-D-aspartate receptor antagonists in the treatment of depression: A perspective review. Therapeutic Advances in Chronic Disease. 2015;6(3):97–114.

[6] Gunduz‐Bruce H, et al. Development of neuroactive steroids for the treatment of postpartum depression. Journal of neuroendocrinology. 2022;34(2):e13019.

[7] Duman RS, Aghajanian GK. Synaptic dysfunction in depression: Potential therapeutic targets. Science. 2012;338(6103):68–72.

[8] Thompson SM, et al. Beyond the serotonin deficit hypothesis: Communicating a neuroplasticity framework of major depressive disorder. Molecular Psychiatry. 2024. Advance online publication.

[9] Cheung N. Simulating Synaptic Pruning and Ketamine-Like Recovery in Depression: Insights from Consolidation Duration and Iterative Regimens on Resilience and Relapse. Zenodo. 2026.

[10] Krystal JH, et al. Ketamine: A paradigm shift for depression research and treatment. Neuron. 2019;101(5):774–778.

[11] Stahl SM. Mechanism of action of serotonin selective reuptake inhibitors: Serotonin receptors and pathways mediate therapeutic effects and side effects. Journal of Affective Disorders. 1998;51(3):215–235.

[12] Boschloo L, et al. The complex clinical response to selective serotonin reuptake inhibitors in depression: a network perspective. Translational Psychiatry. 2023;13(1):19.

Chapter 5
Modeling Antidepressant-Induced Manic Switch in a Unified Computational Framework: Insights from Ketamine, SSRIs, and Neurosteroids

Cheung, Ngo

Cheung, N. (2026). Modeling Antidepressant-Induced Manic Switch in a Unified Computational Framework: Insights from Ketamine, SSRIs, and Neurosteroids. Zenodo. https://doi.org/10.5281/zenodo.18292021

Abstract

Background: Major depressive disorder involves impaired neural plasticity, yet antidepressants spanning glutamatergic (e.g., ketamine), monoaminergic (e.g., SSRIs), and GABAergic (e.g., neurosteroids) classes differ in onset, durability, and risk of treatment-emergent mania, particularly in bipolar disorder. Computational models offer controlled comparison of these pathways, but few integrate manic liability.

Methods: We extended a magnitude-based pruning model (95% sparsity)

of depression in feed-forward networks trained on a Gaussian classification task. From identical pruned states, three interventions were simulated: ketamine-like gradient-guided regrowth (50% reinstatement) with consolidation; SSRI-like prolonged low-rate training with tapering noise and escalating excitability gain; neurosteroid-like global tonic inhibition (0.7× damping, tanh activations, reduced gain). Efficacy was assessed via accuracy under clean, noisy, and combined stress conditions; resilience to graded noise and additional pruning; manic risk via biased positive perturbation and activation magnitude. Results were averaged across 10 random seeds.

Results: All treatments restored near-ceiling baseline performance. Ketamine-like regrowth yielded superior extreme-stress resilience (76.8% ± 3.6%) and relapse protection (−0.1% ± 0.2% drop), reducing sparsity to 47.5%. Neurosteroid-like modulation matched acute combined-stress recovery (97.7% ± 0.2%) but showed state-dependence (19% off-modulation loss) and weaker extreme buffering (42.9% ± 1.3%). SSRI-like refinement lagged (90.8% ± 3.1% combined, 50.1% ± 3.7% extreme) with highest relapse vulnerability (8.8% ± 4.3%). Manic proxies ranked SSRI-like highest risk (biased accuracy 47.6% ± 12.8%, gain 1.60), ketamine-like moderate (84.2% ± 8.5%, gain 1.25), neurosteroid-like lowest (50.5% ± 7.4%, effective gain ~0.59).

Conclusions: Antidepressants operate via distinct plasticity routes—structural rebuilding (durable, moderate risk), reversible stabilization (rapid, low risk), gradual optimization (vulnerable, high

risk)—reproducing clinical trade-offs and supporting mechanism-based selection in unipolar and bipolar depression.

Introduction

Major depressive disorder (MDD) is a dominant source of global disability, affecting hundreds of millions of people and exerting heavy personal and economic burdens [1]. Pharmacotherapy has helped many patients, yet outcomes remain modest: only about one-third of individuals remit after a first antidepressant trial, and another third continue to meet criteria for depression even after several treatment steps [2]. Selective serotonin reuptake inhibitors (SSRIs) are the most frequently prescribed agents but usually take weeks to produce noticeable change, leaving patients exposed to ongoing symptoms and often delivering only partial recovery [3].

The slow and sometimes incomplete benefit of monoaminergic drugs has shifted attention toward interventions that act on glutamatergic and GABA-ergic systems. Low-dose ketamine, an NMDA-receptor antagonist, can relieve depressive symptoms within hours. Pre-clinical and clinical work links this rapid effect to a surge in synaptogenesis mediated by brain-derived neurotrophic factor (BDNF) and mTOR signalling [4,5]. Neuroactive steroids such as brexanolone and zuranolone, which amplify tonic inhibition at extrasynaptic GABA_A receptors, also yield fast improvement and have shown particular

promise for postpartum depression [6]. Taken together, these findings call the classic serotonin-deficit narrative into question and support a view of MDD as a disorder of impaired neural plasticity, in which chronic stress trims dendrites and synapses in prefrontal and hippocampal circuits and erodes resilience [7,8].

Treatment-emergent mania (TEM) complicates the picture, especially for people with bipolar disorder. Traditional antidepressants can provoke mood switches in roughly 20–40 % of bipolar patients, whereas ketamine appears to carry less switch risk in controlled settings, and early reports on neurosteroids suggest minimal liability [9,10,6]. Understanding how different drug classes influence both depressive recovery and excitatory–inhibitory balance therefore remains essential.

Computational models offer a controlled way to dissect these mechanisms. Pruning-based approaches conceptualise depression as excessive synaptic elimination that leaves circuits fragile; ketamine-like regrowth within such models restores performance and resilience [11]. Yet few simulations place glutamatergic, monoaminergic and GABA-ergic strategies side by side or consider the risk of polarity switches.

The present study addresses these gaps. Using a single "depleted" network as a starting point, we compare three antidepressant analogues—ketamine-like synaptogenesis, SSRI-like gradual refinement and neurosteroid-like tonic inhibition. We further incorporate excitability

shifts as a proxy for manic risk. By tracking onset speed, durability and switch potential within one coherent framework, we aim to clarify how distinct biological actions translate into clinical strengths and limitations.

Methods

Network architecture and task

The simulations relied on a feed-forward model intended to capture broad properties of cortico-limbic circuits (Figure 1). The network accepted two-dimensional inputs, passed activity through three hidden layers of 512, 512 and 256 units, and produced one of four class probabilities through a soft-max layer. Rectified linear units drove all hidden activations unless stated otherwise. The full configuration contained roughly 397 000 adjustable parameters.

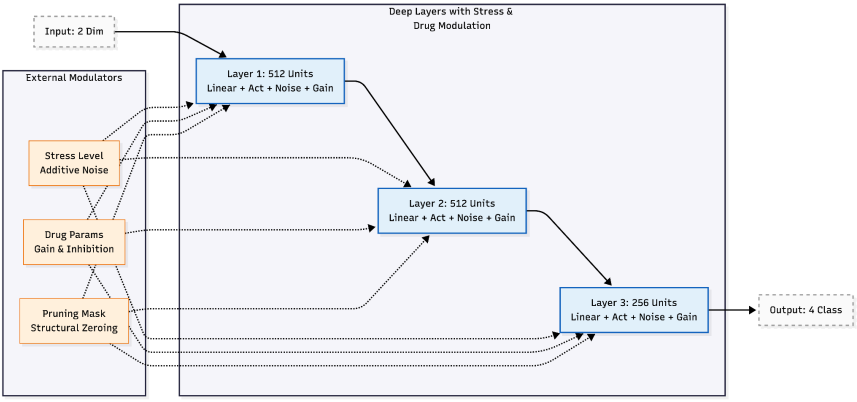


Figure 1: *Condensed Architecture of the Stress-Aware Network. The model processes 2D inputs through three hidden layers (blue) before producing a 4-class output. Each hidden layer block encapsulates four distinct operations in sequence: (1) a fully connected linear transformation subject to structural pruning (orange, Pruning Mask); (2) non-linear activation; (3) stress simulation via additive noise injection (orange, Stress Level); and (4) pharmacological modulation via multiplicative gain and inhibition scaling (orange, Drug Params). This design allows the network to dynamically simulate both structural connectivity changes (synaptogenesis/pruning) and functional state changes (excitability/stress) during the experiment.*

Synaptic stressors could be introduced at two levels. First, additive Gaussian noise was injected after each hidden activation to emulate neuromodulatory disruption. Second, a global multiplicative gain term changed overall excitability and could be combined with an extra damping factor or a switch from ReLU to tanh units when modelling neurosteroid action.

The task itself involved classifying points drawn from four equally sized Gaussian clouds centred on (−3, −3), (−3, 3), (3, −3) and (3, 3) with a shared standard deviation of 0.8. Each training run used 12 000 labelled samples. Performance was monitored on three separate sets: a 4 000-sample "standard noise" set that matched training variance, a 2 000-sample noise-free set, and a "combined-stress" set that added both input and internal perturbations. All code was written in PyTorch and executed on a single CPU core. Ten independent seeds, differing only in data order and weight initialisation, were averaged to obtain each metric.

Simulation of depression

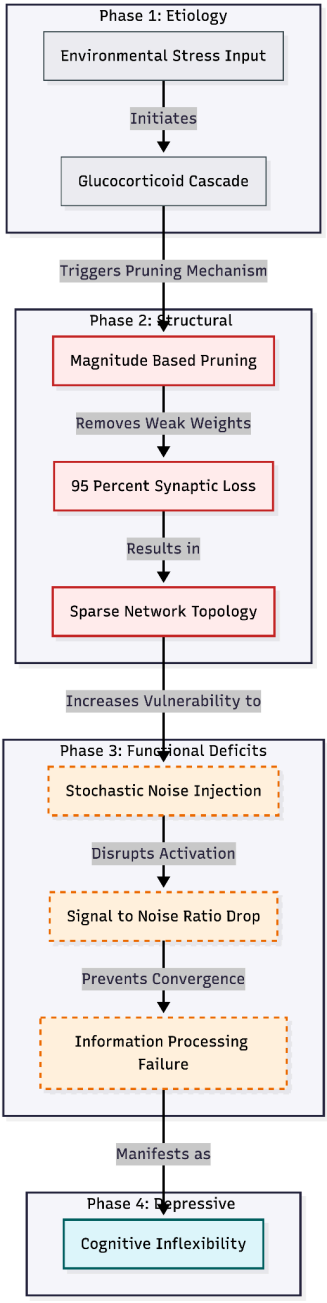

Phase 1: Etiology
Environmental Stress Input
Initiates
Glucocorticoid Cascade
Triggers Pruning Mechanism
Phase 2: Structural
Magnitude Based Pruning
Removes Weak Weights
95 Percent Synaptic Loss
Results in
Sparse Network Topology
Increases Vulnerability to
Phase 3: Functional Deficits
Stochastic Noise Injection
Disrupts Activation
Signal to Noise Ratio Drop
Prevents Convergence
Information Processing Failure
Manifests as
Phase 4: Depressive
Cognitive Inflexibility

Figure 2: *Computational Modeling of the Depressive State. The diagram outlines the algorithmic simulation of depression implemented in the StressAwareNetwork. Phase 1 represents the environmental triggers. Phase 2 illustrates the structural impact modeled by the PruningManager, where 95% of network weights are removed based on magnitude, simulating dendritic spine loss and cortical atrophy. Phase 3 depicts the functional consequences during the forward pass; the sparse topology makes the network highly susceptible to stress_level noise injection, leading to a collapse in the signal-to-noise ratio. Phase 4 represents the final model output, where the inability to classify inputs correctly serves as a proxy for cognitive inflexibility and anhedonia.*

After 20 epochs of baseline training with Adam (learning rate 0.001) on noise-free data, every model underwent magnitude-based pruning that removed the smallest 95 % of weights layer by layer. The surviving sparse network maintained reasonable accuracy on clean inputs yet collapsed rapidly when noise or further pruning was applied, mirroring theories that excessive synaptic loss underlies depressive vulnerability [7] (Figure 2).

Antidepressant intervention protocols

Three recovery strategies were examined on separate copies of the pruned network.

Ketamine-like condition: overall gain was fixed at 1.25. Gradients were accumulated over 30 mini-batches, the highest 50 % of previously pruned weights were reinstated with small random values drawn from N(0, 0.03), and parameters were fine-tuned for 15 epochs with Adam (learning rate 0.0005) while preserving sparsity.

SSRI-like condition: sparsity remained at 95 %. Over 100 slow-learning epochs (learning rate 1×10^{-5}) internal noise declined linearly from $\sigma = 0.5$ to 0, while excitability gain rose from 1.0 to 1.6, mimicking gradual monoaminergic adaptation.

Neurosteroid-like condition: sparsity again stayed constant. A post-activation damping factor of 0.7 was applied, ReLU units were replaced by tanh, and global gain was set to 0.85 (effective scaling ≈ 0.59). Ten adaptation epochs were run with Adam (learning rate 0.0005).

Outcome measures

Primary efficacy was the classification accuracy on clean, standard-noise and combined-stress test sets. To examine robustness, internal noise was increased stepwise from $\sigma = 0$ to 2.5 while recording accuracy. Relapse risk was evaluated by applying a second magnitude-based pruning that deleted 40 % of the remaining weights and then repeating the combined-stress test.

Potential manic conversion was probed in two ways (Figure 3). First, the network was challenged with biased internal noise ($\sigma = 1.0$, bias = +1.0) and the resulting accuracy recorded; poorer accuracy implied greater vulnerability to hyper-excitability. Second, the mean absolute activation across hidden layers during standard input served as an index of latent excitatory tone. For the neurosteroid analogue, all evaluations were repeated after removing the damping parameters to simulate drug

withdrawal.

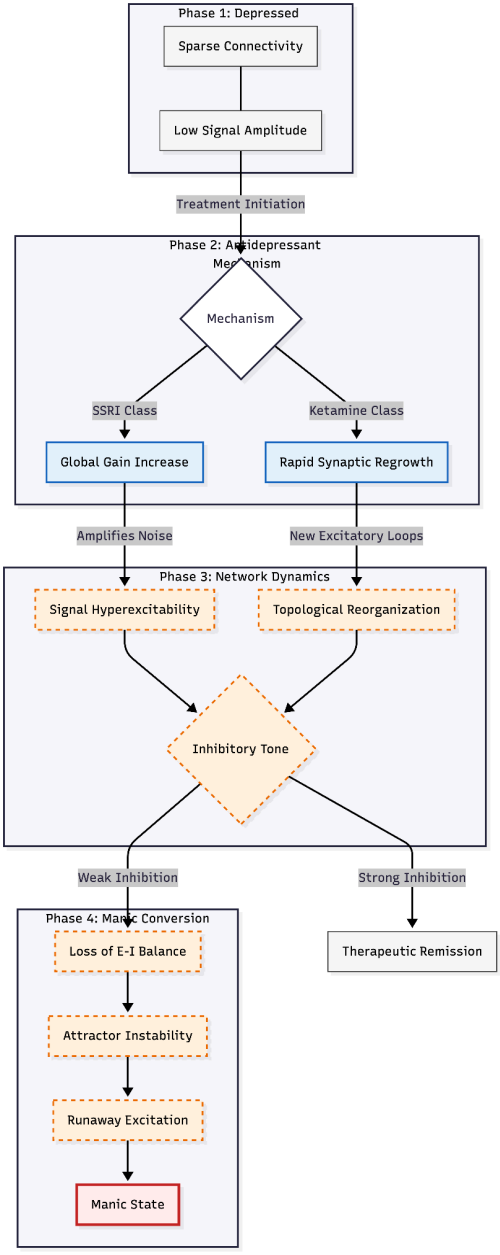

Phase 1: Depressed
Sparse Connectivity
Low Signal Amplitude
Treatment Initiation
Phase 2: Antidepressant Mechanism
Mechanism
SSRI Class
Ketamine Class
Global Gain Increase
Rapid Synaptic Regrowth
Amplifies Noise
New Excitatory Loops
Phase 3: Network Dynamics
Signal Hyperexcitability
Topological Reorganization
Inhibitory Tone
Weak Inhibition
Strong Inhibition
Phase 4: Manic Conversion
Loss of E-I Balance
Attractor Instability
Runaway Excitation
Manic State
Therapeutic Remission

Figure 3: *Computational Model of Pharmacologically Induced Manic Conversion. The diagram illustrates the bifurcation of network dynamics following antidepressant treatment. Phase 1 represents the baseline depressive state characterized by sparse connectivity and low signal amplitude. Phase 2 differentiates the mechanism of action: SSRIs increase the global gain (amplifying existing signals), while Ketamine induces structural plasticity (synaptic regrowth). Phase 3 demonstrates that both mechanisms increase net excitation. The critical checkpoint is the system's "Inhibitory Tone." If inhibition is sufficient, the system stabilizes into remission. However, in the presence of weak inhibition—simulating a deficit in GABAergic control or neurosteroid regulation—the system enters Phase 4, resulting in a loss of Excitation-Inhibition (E-I) balance, attractor instability, and ultimately a transition to a manic state.*

Data analysis

For every condition, results from the ten seeds were summarised as mean ± standard deviation; pronounced dispersion was also expressed as minimum–maximum ranges. Because the project was exploratory, no formal null-hypothesis tests were undertaken.

Results

Simulations were run under 10 independent random seeds; across seeds we saw the same rank ordering of performance for every outcome that was tracked.

Antidepressant efficacy

Relative to the over-pruned, untreated network (combined-stress accuracy = 29.7 ± 2.7 %), every treatment produced a large gain. Neurosteroid-like modulation finished at 97.7 ± 0.2 % and the ketamine-like synaptogenic routine reached 97.2 ± 0.2 %. The monoaminergic SSRI-like schedule also improved behaviour, but only to 90.8 ± 3.1 %. On clean or standard inputs all treated models scored at (or fractionally below) 100 % accuracy, whereas the untreated baseline remained at 34.7 ± 11.9 % on clean data.

Stress resilience

Table 1. *Stress Resilience Profile (Mean ± Standard Deviation Across 10 Seeds)*

Treatment	No Noise (%)	Moderate (σ=0.5) (%)	High (σ=1.0) (%)	Severe (σ=1.5) (%)	Extreme (σ=2.5) (%)
Untreated (pruned)	36.8 ± 11.9	29.9 ± 2.5	29.6 ± 1.8	29.8 ± 1.4	29.6 ± 1.5
Ketamine-like	100.0 ± 0.0	99.9 ± 0.1	98.2 ± 1.1	92.9 ± 2.4	76.8 ± 3.6
SSRI-like	99.9 ± 0.1	95.3 ± 2.8	78.2 ± 5.8	64.5 ± 5.2	50.1 ± 3.7
Neurosteroid-like	100.0 ± 0.0	99.9 ± 0.1	92.9 ± 2.5	70.9 ± 2.5	42.9 ± 1.3

Noise tolerance diverged sharply as internal noise was increased (Table 1). At the most extreme perturbation (σ = 2.5) the ketamine analogue still classified correctly 76.8 ± 3.6 % of trials. The SSRI-like routine fell to 50.1 ± 3.7 %, and the neurosteroid analogue to 42.9 ± 1.3 %; the untreated control stayed near its baseline (29.6 ± 1.5 %). With moderate

noise ($\sigma = 1.0$) neurosteroid- and ketamine-like conditions were similar (92.9 ± 2.5 % and 98.2 ± 1.1 %, respectively) and both clearly out-performed the SSRI-like schedule (78.2 ± 5.8 %).

Manic conversion risk

Table 2. *Manic Conversion Risk Metrics (Mean ± Standard Deviation Across 10 Seeds)*

Treatment	Gain Multiplier	Biased Stress Accuracy (%)	Activation Magnitude
Untreated (pruned)	1.00 ± 0.00	25.0 ± 0.8	0.100 ± 0.013
Ketamine-like	1.25 ± 0.00	84.2 ± 8.5	0.649 ± 0.079
SSRI-like	1.60 ± 0.00	47.6 ± 12.8	0.390 ± 0.078
Neurosteroid-like	0.85 ± 0.00	50.5 ± 7.4	0.196 ± 0.008

Note. Lower biased stress accuracy indicates greater manic conversion vulnerability; higher activation magnitude reflects increased latent hyperexcitability.

Three risk markers were monitored: gain multiplier, accuracy under positively biased noise, and mean activation magnitude (Table 2). The SSRI-like model showed the largest gain (1.60 ± 0.00) and the poorest accuracy under biased noise (47.6 ± 12.8 %), with a mid-range activation magnitude of 0.390 ± 0.078. The ketamine analogue showed an intermediate gain (1.25 ± 0.00) but maintained high biased-noise accuracy (84.2 ± 8.5 %) even though its activation magnitude was highest (0.649 ± 0.079). Neurosteroid modulation damped excitability (gain = 0.85 ± 0.00; activation magnitude = 0.196 ± 0.008) and achieved a biased-noise accuracy of 50.5 ± 7.4 %. Untreated networks, by

comparison, sat at gain = 1.00 ± 0.00, biased-noise accuracy = 25.0 ± 0.8 %, and activation magnitude = 0.100 ± 0.013.

Relapse vulnerability and medication dependence

Table 3. *Final Comparison Matrix (Means Across 10 Seeds)*

Metric	Ketamine-like	SSRI-like	Neurosteroid-like	Untreated (pruned)
Combined Stress (%)	97.2	90.8	97.7	29.7
Extreme Stress (%)	76.8	50.1	42.9	29.6
Biased Stress (%)	84.2	47.6	50.5	25.0
Gain Multiplier	1.25	1.60	0.85	1.00
Activation Magnitude	0.649	0.390	0.196	0.100
Sparsity (%)	47.5	95.0	95.0	95.0
Relapse Drop (%)	−0.1	8.8	5.1	N/A

When an additional pruning step was applied to mimic relapse, performance in the ketamine-treated networks was essentially unchanged (-0.1 ± 0.2 %). The SSRI-like condition lost 8.8 ± 4.3 % accuracy and the neurosteroid analogue lost 5.1 ± 2.2 %. Removing the neurosteroid damping altogether revealed strong state-dependence: combined-stress accuracy fell from 97.7 ± 0.2 % on-modulation to 78.5 ± 4.8 % off-modulation.

Seed-to-seed spread was small for most measures, but biased-noise

accuracy varied widely with the SSRI (28.2–74.8 %) and ketamine (63.3–96.7 %) analogues, indicating individual-difference sensitivity in those two conditions (Table 3).

Discussion

Interpretation of efficacy and resilience findings

Our simulations reveal three recognisably different therapeutic signatures. Both the ketamine-like synaptogenic programme and the neurosteroid-like GABAergic modulation restored almost full performance when the network was challenged by combined input and internal noise, reaching about 97 % accuracy. The SSRI-style, slow-adaptation schedule also improved accuracy but plateaued roughly six percentage points lower. These results echo clinical impressions: drugs that act through glutamatergic or GABAergic mechanisms often stabilise circuits within days, whereas monoaminergic agents depend on gradual receptor and transcriptional changes that unfold over weeks [7,8].

Durability separated the interventions more clearly. After ketamine-like regrowth, the network continued to perform well under extreme internal noise and was virtually unaffected by an additional round of pruning. This mirrors evidence that a single ketamine infusion can trigger structural plasticity and sustain benefit beyond drug clearance [5].

Neurosteroid modulation, by contrast, protected the network only while the damping remained active; switching the modulation off produced a 19 % drop in accuracy and a moderate vulnerability to further pruning. Clinical reports of zuranolone show a similar pattern: rapid relief that fades once dosing stops [6]. The SSRI analogue, which merely fine-tuned the sparse remnants, offered the least protection against heavy noise and the greatest loss after re-pruning, a finding that accords with the limited stress resilience seen in many patients who rely on conventional antidepressants alone [4].

Manic conversion risk and clinical parallels

The simulations demonstrated class-specific variations in mood-switch risk (Figure 4): SSRI-like refinement resulted in the most significant increase in network excitability and the least resistance to positively biased noise, reflecting meta-analytic manic switch rates nearing 40% among bipolar patients on traditional antidepressants [9]; ketamine-like synaptogenesis maintained strong performance even in the presence of biased input, aligning with infrequent reports of mania when ketamine is used with mood stabilizers [10]; and neurosteroid-based tonic inhibition exhibited the highest safety, showing the lowest activation levels and the most effective management of bias-driven errors, corresponding to initial clinical findings of minimal switch liability with zuranolone [12].

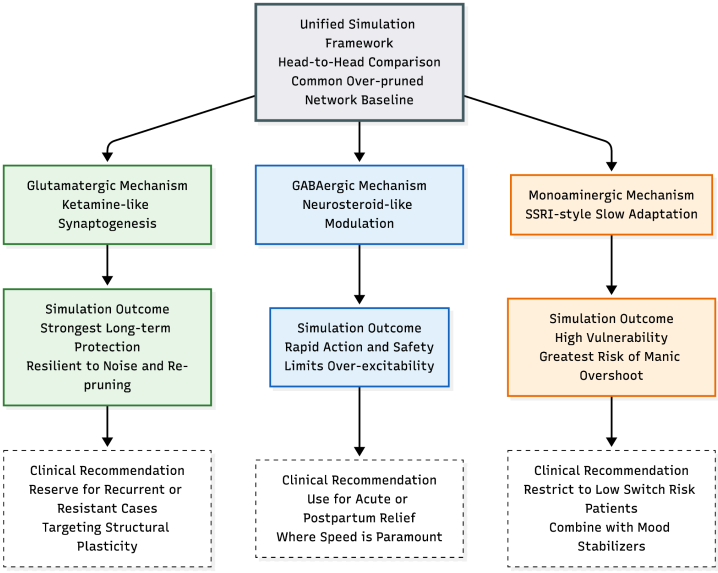


Figure 4: *Translational Framework for Antidepressant Selection. The study leverages a unified simulation architecture to compare three distinct pharmacological mechanisms against a common baseline of network fragility. (Left, Green) The Glutamatergic pathway, modeling ketamine-like synaptogenesis, demonstrates superior durability against stress, supporting its use in treatment-resistant depression. (Center, Blue) The GABAergic pathway, modeling neurosteroid modulation, provides rapid stabilization with minimal excitability risk, aligning with indications for urgent or postpartum care. (Right, Orange) The Monoaminergic pathway, modeling traditional SSRIs, reveals a susceptibility to manic overshoot, reinforcing clinical guidelines that advise caution or concurrent mood stabilizers when treating bipolar depression.*

Seed-to-seed variability was widest for the SSRI- and ketamine-like conditions, hinting that individual differences in baseline network fragility can influence real-world switch risk, much as prior mood-switch history and family loading predict clinical outcomes [13]. Together, these observations support caution when prescribing monoaminergic monotherapy to patients with bipolar vulnerability and point to plasticity-targeted or GABAergic strategies as potentially safer options.

Novelty and translational impact

This project brings three very different antidepressant mechanisms—glutamatergic, monoaminergic, and GABA-ergic—into one tightly controlled simulation that also tracks the chance of a manic switch (Figure 5). Earlier in-silico work tended to look at one pathway at a time, for example focusing only on ketamine-like synaptogenesis [11]. By beginning with the same over-pruned network in every run and then adding a uniform excitability gain plus skewed noise, we could compare recovery, durability, and opposite-pole risk head-to-head.

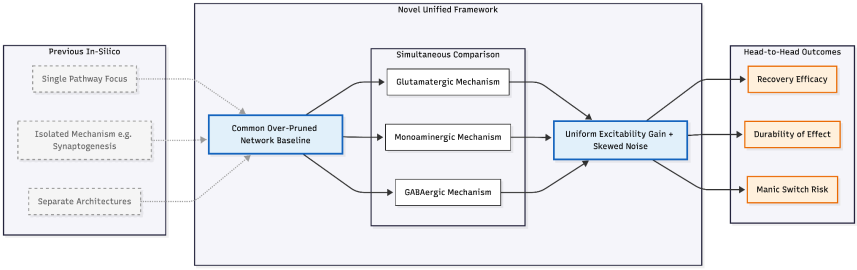


Figure 5: *Conceptual Shift to a Unified Simulation Framework. Unlike previous in-silico studies that typically isolated single pathways (e.g., focusing solely on synaptogenesis), this project introduces a novel unified architecture. By initializing all simulations with an identical over-pruned network baseline and applying standardized excitability gains and skewed noise, the framework allows for a rigorous head-to-head comparison of three distinct pharmacological mechanisms: Glutamatergic, Monoaminergic, and GABAergic. This approach uniquely enables the simultaneous tracking of recovery trajectories, treatment durability, and the risk of manic switching within a single controlled environment.*

Several clinically relevant messages emerge. First, the ketamine analogue delivered the strongest long-term protection: once new

synapses were in place, the model withstood heavy noise and an extra pruning step, mirroring reports that ketamine triggers lasting plasticity [5]. Second, the neurosteroid analogue worked quickly and limited over-excitability, a pattern that fits current interest in zuranolone for urgent or postpartum use where speed and safety are paramount [6]. Third, the monoaminergic schedule, while helpful, left the network most exposed to manic-like overshoot—an echo of long-standing cautions about antidepressant monotherapy in bipolar illness [9,14].

Running all three regimens inside the same architecture therefore offers a mechanistic basis for tailoring treatment: reserve synaptogenic drugs for recurrent or resistant cases, choose fast but state-dependent GABA-modulators for acute relief, and limit classic antidepressants to patients with low switch risk or with added mood stabilisers.

Limitations

Important simplifications remain. The feed-forward model omits the feedback loops that shape real cortico-limbic mood circuits; this may downplay potential instabilities. Modulation was applied globally, not by layer or cell class, so subtle pharmacological differences—for example, the varied extrasynaptic actions of neurosteroids noted by [15]—were ignored. Manic liability was estimated from biased noise and overall activation; these proxies are cruder than clinical ratings taken over time.

Although ten random seeds introduced variability, other "patient"

features such as baseline excitation–inhibition balance or depth of synaptic loss were held constant. Mood-stabilising co-therapy, known to reduce switch risk [14], was not simulated. Each of these gaps limits direct clinical translation.

Conclusion

Within these constraints, the study shows that different drug classes restore performance through distinct routes—building new structure, damping excess firing, or slowly tuning weakened links—each with its own balance of speed, staying power, and bipolar safety. Embedding switch risk in the same plasticity frame moves the conversation away from single-transmitter theories toward circuit reserve and excitability control. As rapid-acting treatments gain ground, such models may help clinicians match interventions to individual risk profiles and combine agents more safely.

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization.

[2] Rush, A. J., Trivedi, M. H., Wisniewski, S. R., et al. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. American Journal of

Psychiatry, 163(11), 1905–1917. https://doi.org/10.1176/ajp.2006.163.11.1905

[3] Trivedi, M. H., Rush, A. J., Wisniewski, S. R., et al. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. American Journal of Psychiatry, 163(1), 28–40. https://doi.org/10.1176/appi.ajp.163.1.28

[4] Murrough, J. W., Iosifescu, D. V., Chang, L. C., et al. (2013). Antidepressant efficacy of ketamine in treatment-resistant major depression: A two-site randomized controlled trial. American Journal of Psychiatry, 170(10), 1134–1142. https://doi.org/10.1176/appi.ajp.2013.13030392

[5] Krystal, J. H., Abdallah, C. G., Sanacora, G., et al. (2019). Ketamine: a paradigm shift for depression research and treatment. Neuron, 101(5), 774–778. https://doi.org/10.1016/j.neuron.2019.02.005

[6] Gunduz-Bruce, H., Lasser, R., Nandy, I., et al. (2020, September). Open-label, Phase 2 trial of the oral neuroactive steroid GABAA receptor positive allosteric modulator zuranolone in bipolar disorder I and II. In Poster presented at: psych Congress.

[7] Duman, R. S., & Aghajanian, G. K. (2012). Synaptic dysfunction in depression: potential therapeutic targets. Science (New York, N.Y.),

338(6103), 68–72. https://doi.org/10.1126/science.1222939

[8] Page, C. E., Epperson, C. N., Novick, A. M., et al. (2024). Beyond the serotonin deficit hypothesis: communicating a neuroplasticity framework of major depressive disorder. Molecular Psychiatry, 29(12), 3802-3813.

[9] Tondo, L., Vázquez, G., & Baldessarini, R. J. (2010). Mania associated with antidepressant treatment: comprehensive meta‐analytic review. Acta Psychiatrica Scandinavica, 121(6), 404-414.

[10] Jawad, M. Y., Watson, S., Haddad, P. M., et al. (2021). Ketamine for bipolar depression: A systematic review. International Journal of Neuropsychopharmacology, 24(7), 535–541. https://doi.org/10.1093/ijnp/pyab023

[11] Cheung, N. (2026). Divergent Mechanisms of Antidepressant Efficacy: A Unified Computational Comparison of Synaptogenesis, Stabilization, and Tonic Inhibition in a Model of Depression. Zenodo. https://doi.org/10.5281/zenodo.18290014

[12] Price, M. Z., & Price, R. L. (2025). Zuranolone for Postpartum Depression in Real-World Clinical Practice. J Clin Psychiatry, 86(3), 25cr15876.

[13] Goldberg, J. F., & Truman, C. J. (2003). Antidepressant-induced

mania: an overview of current controversies. Bipolar Disorders, 5(6), 407-420.

[14] Viktorin, A., Lichtenstein, P., Thase, M. E., et al. (2014). The risk of switch to mania in patients with bipolar disorder during treatment with an antidepressant alone and in combination with a mood stabilizer. American Journal of Psychiatry, 171(10), 1067-1073.

[15] Marecki, R., Kałuska, J., Kolanek, A., et al. (2023). Zuranolone–synthetic neurosteroid in treatment of mental disorders: narrative review. Frontiers in Psychiatry, 14, 1298359.

Chapter 6
Structural Rebuilding Confers Superior Long-Term Resilience: A Unified Multi-Mechanism Computational Comparison of Antidepressants in Chronic Stress

Cheung, Ngo

Cheung, N. (2026). Structural Rebuilding Confers Superior Long-Term Resilience: A Unified Multi-Mechanism Computational Comparison of Antidepressants in Chronic Stress. Zenodo. https://doi.org/10.5281/zenodo.18295517

Abstract

Background: Major depressive disorder remains inadequately treated in many patients, with divergent onset speeds and durability across antidepressant classes. Emerging neuroplasticity frameworks emphasize synaptic loss, yet direct computational comparisons of glutamatergic (ketamine-like), monoaminergic (SSRI-like), and GABAergic (neurosteroid-like) mechanisms—particularly under chronic stress—are lacking.

Methods: We extended a pruning-plasticity model by applying 95% magnitude-based synaptic elimination to overparameterized feed-forward networks trained on Gaussian cluster classification. From identical pruned states, three interventions were tested across 10 stochastic seeds: ketamine-like gradient-guided regrowth (50% reinstatement) with consolidation; SSRI-like prolonged low-learning-rate refinement with gradual noise reduction; and neurosteroid-like global tonic inhibition (30% damping plus bounded activations). Outcomes included acute recovery, stress resilience (graded internal noise), acute relapse (additional 40% pruning), and longitudinal durability (eight cycles of 10% cumulative pruning with mechanism-specific maintenance).

Results: All treatments restored near-ceiling baseline performance, but resilience diverged: ketamine-like networks showed superior extreme-noise tolerance (81.7%) and zero acute/longitudinal relapse across seeds, despite rising sparsity. SSRI-like refinement yielded moderate gains (81.3% combined-stress acutely) but 40% seed relapse longitudinally. Neurosteroid-like inhibition matched early efficacy (97.8%) yet exhibited state-dependence and late-cycle vulnerability (20% seed relapse).

Conclusions: These simulations demonstrate mechanistically distinct routes—structural rebuilding (durable), gradual optimization (variable), reversible stabilization (rapid but fragile)—yielding predictive trade-offs

in relapse risk. Findings support mechanism-based personalization, prioritizing synaptogenic agents for severe/chronic cases while rationalizing combinations for acute relief and maintenance.

Introduction

Major depressive disorder (MDD) is one of the top drivers of global disability and carries heavy personal and financial costs [1]. Even with many drugs on the market, progress has stalled: only about a third of patients get well after an initial prescription, and roughly a third remain symptomatic after several different regimens [2]. Selective serotonin re-uptake inhibitors (SSRIs) still dominate routine care, yet these medicines often take weeks to work and may deliver only partial relief [3].

The hunt for faster and more dependable options has reshaped the field. Low-dose ketamine, an NMDA-receptor blocker, can lift mood within hours. The effect was known to be tied to bursts of synaptogenesis through BDNF- and mTOR-linked pathways [4,5]. Neuroactive steroids such as brexanolone and its oral analogue zuranolone increase tonic GABA_A inhibition and also act quickly, most clearly in postpartum depression [6]. These findings shift attention away from simple monoamine shortage toward problems in neural plasticity: chronic stress trims dendritic spines and synapses in prefrontal and hippocampal areas, leaving circuits fragile [7].

Computational models let researchers test how such circuit changes might play out. One influential idea treats depression as "too much pruning." The network still works under light load, but a small dose of extra noise makes it fail. Earlier work showed that letting the model regrow connections—an in-silico stand-in for ketamine—can restore stability without rebuilding every lost synapse [8]. What remains unclear is how different drug classes stack up when they are placed in the same framework. Do they differ in speed? Durability? Risk of pushing the system toward the opposite pole, as in a manic switch?

To explore those questions, we extended the pruning-plasticity model in a simple feed-forward classifier. From identical 95 % pruned starting points we applied three distinct interventions:

- a ketamine-like, gradient-guided regrowth routine;
- an SSRI-like, slow refinement with fading noise;
- a neurosteroid-like, global increase in tonic inhibition.

The protocol ran across ten random seeds and eight sequential "stress cycles," adding more pruning each time to mimic chronic relapse pressure. By following the same networks over time, we could compare recovery, stability, and switch liability side by side. The aim was to clarify how mechanism shapes outcome and to suggest when a given class might be the better clinical choice as rapid, plasticity-targeted treatments gain traction.

Methods

Network architecture and classification task

We built a small, fully connected feed-forward network to stand in for cortical circuitry. The model held two input units, three hidden layers (512, 512, 256 units) and four soft-max outputs. Hidden units used ReLU unless noted otherwise. After each hidden layer we could inject zero-mean Gaussian noise to mimic stress and, when needed, multiply activations by a global damping factor to imitate tonic inhibition.

The task was simple on purpose (Figure 1). Two-dimensional points were drawn from four Gaussian clouds centred at (−3, −3), (3, 3), (−3, 3) and (3, −3) with a standard deviation of 0.8. For every random seed we produced 12 000 training points, 4 000 noisy test points and 2 000 clean test points. This layout let us watch accuracy fall as pruning and noise mounted, a laboratory version of latent vulnerability described in clinical work [8,7].

Baseline training and pruning

Weights were drawn from a small normal distribution and the network learned for 20 epochs with Adam (learning rate 0.001) on clean data. Once accuracy steadied, we lopped off the weakest 95 % of weights

across all layers, leaving their sign intact. The result was a sparse "depressed" circuit that still did the job when inputs were noise-free but collapsed when perturbations appeared.

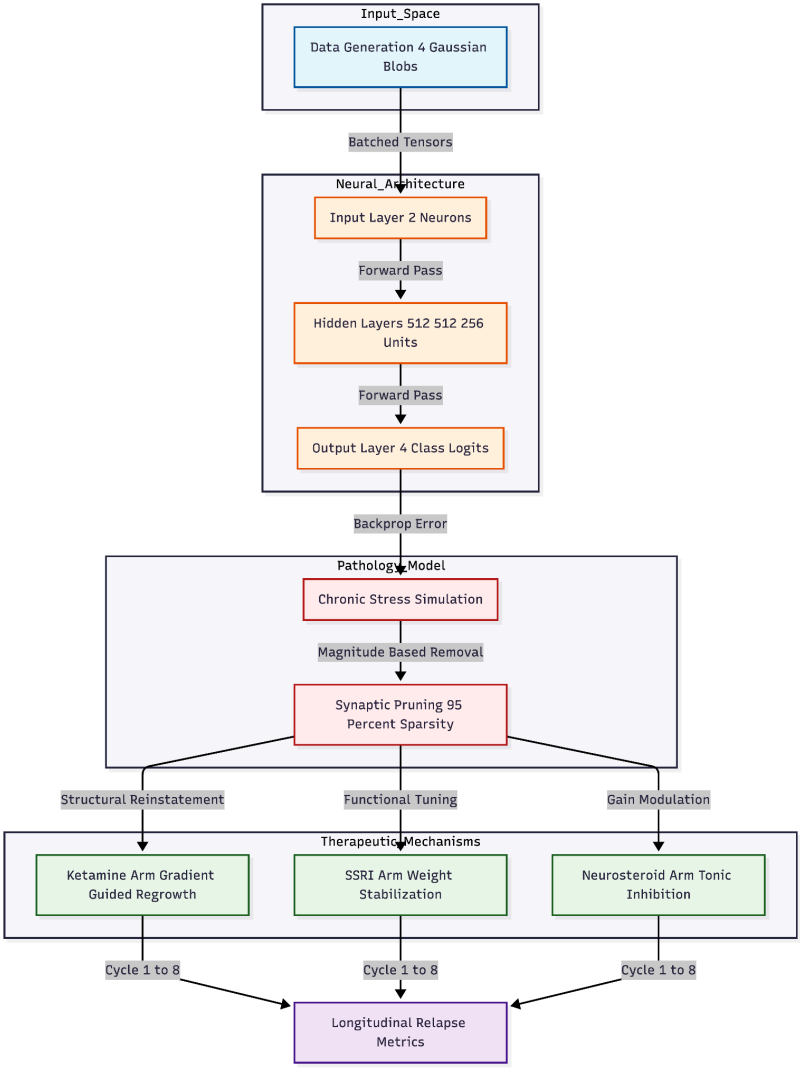


Figure 1: *Computational Architecture of the Multi-Mechanism Antidepressant Experiment. The diagram illustrates the information flow through the StressAwareNetwork. Synthetic classification data is fed into a feed-forward neural network. The "Pathology Model" simulates depression via chronic stress, resulting in severe synaptic pruning (95% sparsity). The network is then subjected to three distinct algorithmic interventions: Ketamine-like structural regrowth, SSRI-like functional weight stabilization, and Neurosteroid-like inhibitory modulation. The system evaluates the resilience of each mechanism against relapse over 8 longitudinal cycles of recurring stress.*

Antidepressant treatment protocols

Starting from identical pruned copies we imposed three separate recovery routines.

Ketamine-like synaptogenesis: For 30 mini-batches we collected gradient magnitudes at every pruned site. The top half of those locations were regrown with small random values (SD 0.03). Fifteen fine-tuning epochs followed (Adam, lr 0.0005) while the new sparsity mask stayed fixed.

SSRI-like gradual stabilisation: Sparsity remained at 95 %. The model trained for 100 epochs at a very low rate (1×10^{-5}). Internal noise began at σ 0.5 and faded linearly to zero, echoing the slow receptor and transcription shifts seen with monoaminergic drugs [9].

Neurosteroid-like tonic inhibition: We kept the 95 % mask but switched hidden activations to tanh and multiplied them by 0.7, reproducing the dampening effect of extrasynaptic GABA_A modulation. Ten consolidation epochs were run with Adam (lr 0.0005).

Evaluation metrics

After each intervention we measured accuracy on four test sets: clean, standard noisy, a combined-stress set (input noise σ 1.0 plus internal noise σ 0.5) and a sweep of internal noise from σ 0.0 to 2.5.

Acute relapse was probed by pruning a further 40 % of the remaining weights and retesting on the combined-stress set. For the neurosteroid condition we also checked accuracy after turning the damping factor off to see how much benefit depended on continued inhibition.

Long-term relapse under chronic stress

Durability was followed across eight stress cycles (Figure 2). Each cycle removed 10 % of the surviving weights, then applied a brief maintenance step that matched the original mechanism:

- ketamine-like, five fine-tuning epochs;
- SSRI-like, twenty very low-rate epochs with mild tapering noise;
- neurosteroid-like, ten epochs under active inhibition.

Combined-stress accuracy was recorded after every cycle. Dropping below 80 % marked a relapse.

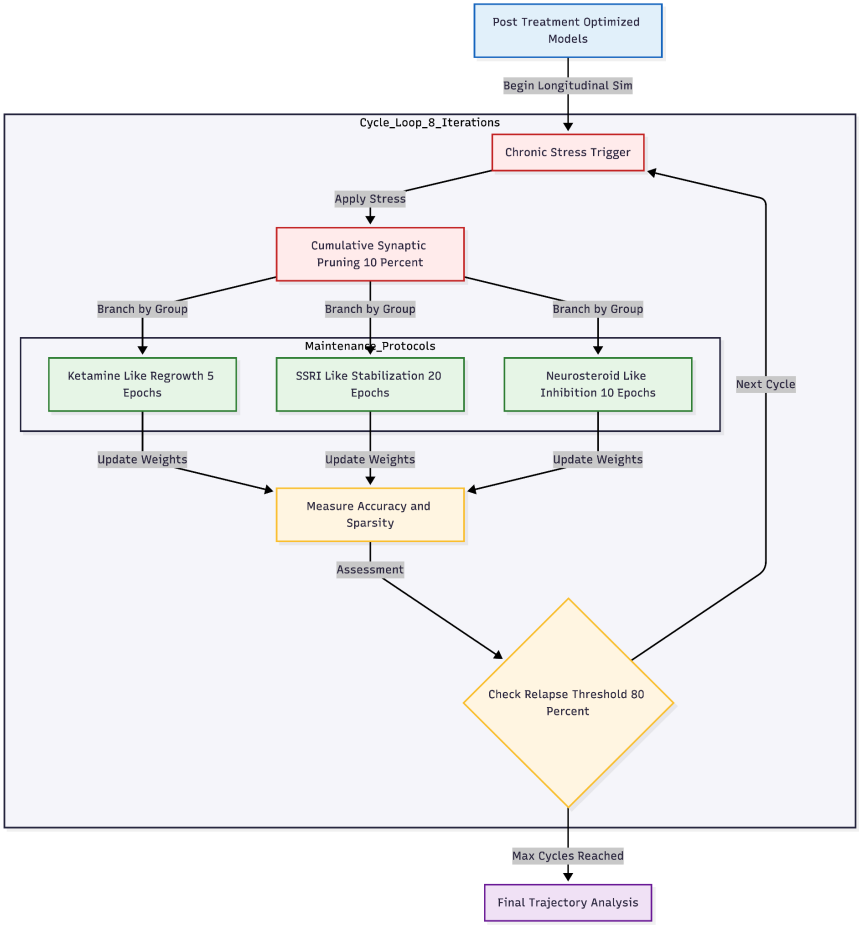


Figure 2: *Longitudinal Relapse Simulation Protocol. The diagram details the 8-cycle simulation used to evaluate long-term resilience. Initial post-treatment models enter a recursive loop representing the passage of time under adverse conditions. In every cycle, models undergo "Chronic Stress" (cumulative magnitude-based pruning removing 10% of remaining weights), followed by a mechanism-specific maintenance phase (periodic regrowth for Ketamine, stress-scheduled stabilization for SSRIs, or inhibitory consolidation for Neurosteroids). Relapse is defined as a drop in classification accuracy below 80%.*

Reproducibility and statistics

All experiments were repeated with 10 different random seeds that altered data draws, weight starts and noise streams. Means and standard deviations are reported; we also quote the range where it was informative. Code was written in PyTorch and run on a CPU; deterministic options were fixed whenever the library allowed.

Results

Post-treatment recovery and baseline efficacy

Table 1. *Post-Treatment Efficacy (Mean ± SD Across 10 Seeds)*

Treatment	Sparsity (%)	Clean (%)	Standard (%)	Combined Stress (%)
Untreated (pruned)	95.0 ± 0.0	32.1 ± 11.8	32.3 ± 11.0	28.9 ± 2.4
Ketamine-like	47.5 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	97.3 ± 0.2
SSRI-like	95.0 ± 0.0	97.5 ± 7.6	95.7 ± 8.0	81.3 ± 8.8
Neurosteroid-like	95.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	97.8 ± 0.1

Before any rescue procedure the heavily pruned networks were barely functional: across ten seeds their accuracy under combined input and internal noise averaged 28.9 % (SD 2.4 %). Once an intervention was applied, performance on noise-free or mildly noisy data returned to ceiling levels. The ketamine-style regrowth and the neurosteroid-style tonic inhibition each produced perfect scores on both clean and standard noisy sets and almost full recovery under the combined-stress condition

(97.3 % ± 0.2 and 97.8 % ± 0.1, respectively). The SSRI-style slow refinement helped far less. Although many seeds reached high values on clean inputs (mean 97.5 %, SD 7.6 %), their mean accuracy under combined stress was only 81.3 % (SD 8.8 %), revealing large between-seed spread (Table 1).

Resilience to graded internal stress

Table 2. *Stress Resilience Profile (Mean ± SD Accuracy)*

Treatment	No Noise (%)	Moderate (σ=0.5) (%)	High (σ=1.0) (%)	Severe (σ=1.5) (%)	Extreme (σ=2.5) (%)
Untreated (pruned)	32.3 ± 11.0	29.2 ± 2.8	29.2 ± 1.6	28.9 ± 1.5	27.8 ± 1.6
Ketamine-like	100.0 ± 0.0	99.9 ± 0.0	98.9 ± 0.7	95.3 ± 1.6	81.7 ± 3.0
SSRI-like	95.7 ± 8.0	85.0 ± 9.6	66.2 ± 7.7	53.8 ± 5.6	42.0 ± 3.1
Neurosteroid-like	100.0 ± 0.0	99.9 ± 0.1	91.6 ± 1.5	69.1 ± 2.0	41.2 ± 1.0

When internal noise was increased stepwise, the three regimens separated sharply (Table 2). Ketamine-treated networks coped well even at the harshest perturbation level (σ = 2.5), still classifying 81.7 % (SD 3.0 %) of patterns correctly. The SSRI-treated and neurosteroid-treated models were similar up to moderate noise but both collapsed when noise became severe, each settling near 41–42 % accuracy. Untreated baselines hovered around chance throughout the sweep.

Acute relapse vulnerability

A sudden removal of 40 % of the surviving weights served as an acute relapse test. Ketamine-like models were almost unaffected, losing only 0.0 % on average (SD 0.4 %) from their combined-stress score. By contrast, SSRI-treated networks dropped 9.1 % (SD 5.8 %) and neurosteroid-treated networks dropped 6.9 % (SD 5.6 %).

Neurosteroid state-dependence

The benefit of the neurosteroid analogue depended on continued tonic damping. Once the damping factor was switched off, combined-stress accuracy fell sharply to 68.8 % (SD 9.9 %). Interestingly, tolerance of extreme internal noise improved slightly in this unmedicated state (51.9 % ± 2.8 versus 41.2 % ± 1.0 while medicated).

Longitudinal trajectories under chronic stress

Eight cycles of additional 10 % pruning, each followed by a mechanism-specific maintenance step, produced very different long-term pictures (Table 3). Ketamine-like networks held steady around 97.6 % combined-stress accuracy all the way to cycle 8 even though sparsity rose from 47.5 % to 77.4 %. SSRI-treated models, which started low, improved slowly and finished at 93.2 % (SD 2.0 %). Neurosteroid-treated models looked excellent through the first five

cycles but then diverged; by cycle 8 the group mean had slipped to 88.2 % with a wide 49.6–97.7 % range.

Table 3. *Combined-Stress Accuracy (%) by Cycle (Mean ± SD)*

Cycle	Ketamine	SSRI	Neurosteroid
0	97.5 ± 0.2	80.8 ± 8.7	97.7 ± 0.2
1	97.6 ± 0.2	84.7 ± 7.9	97.9 ± 0.3
2	97.7 ± 0.2	87.7 ± 5.4	97.8 ± 0.1
3	97.7 ± 0.3	90.1 ± 4.0	97.6 ± 0.4
4	97.6 ± 0.3	91.5 ± 3.4	97.5 ± 0.8
5	97.6 ± 0.4	92.1 ± 2.9	97.0 ± 0.9
6	97.5 ± 0.3	92.9 ± 2.4	95.6 ± 3.5
7	97.7 ± 0.3	92.6 ± 2.6	94.7 ± 2.8
8	97.6 ± 0.4	93.2 ± 2.0	88.2 ± 14.0

Long-term relapse risk

Table 4. *Longitudinal Relapse Summary*

Metric	Ketamine	SSRI	Neurosteroid
Total accuracy drop (cycle 0 to 8)	-0.1 ± 0.3%	-12.4 ± 8.3%	9.5 ± 14.0%
Final accuracy (cycle 8)	97.6 ± 0.4%	93.2 ± 2.0%	88.2 ± 14.0%
Seeds with relapse (<80%)	0/10	4/10	2/10
Seeds without relapse (≥80%)	10/10	6/10	8/10
Mean cycle at relapse (if relapsed)	N/A	0.0	8.0

None of the ketamine-treated seeds ever fell below the 80 % relapse threshold (Table 4). Four SSRI-treated seeds crossed that line immediately after the initial treatment, while two neurosteroid-treated seeds relapsed late, in cycle 8. Averaged across seeds, total loss of

accuracy from cycle 0 to cycle 8 was −0.1 % for the ketamine analogue, −12.4 % for the SSRI analogue, and +9.5 % for the neurosteroid analogue (the plus sign reflects the sharp late decline after earlier gains).

Overall, relative to the 28.9 % baseline, the ketamine-like and neurosteroid-like strategies each delivered an improvement of roughly 69 percentage points under combined stress, whereas the SSRI-like routine lifted accuracy by about 52 points and displayed the greatest variability and relapse liability.

Discussion

Interpretation of results

Placing three distinct recovery programmes inside the same wounded network revealed a set of trade-offs that echo real-world pharmacology (Figure 3). The ketamine-like routine, which reinstated half of the deleted synapses in an activity-guided way, produced almost immediate normalisation and—with or without further stress—barely budged thereafter. Even when another forty per cent of the remaining weights were cut and eight additional pruning rounds were imposed, accuracy stayed above 97 % and no seed relapsed. Clinically, a single ketamine infusion can trigger BDNF- and mTOR-dependent spine growth that outlasts drug exposure and holds up under subsequent stress [5,4]. The model suggests that durability arises not from a wholesale return to

pre-morbid density but from selectively restoring high-value links, thereby rebuilding "reserve" that buffers later insults.

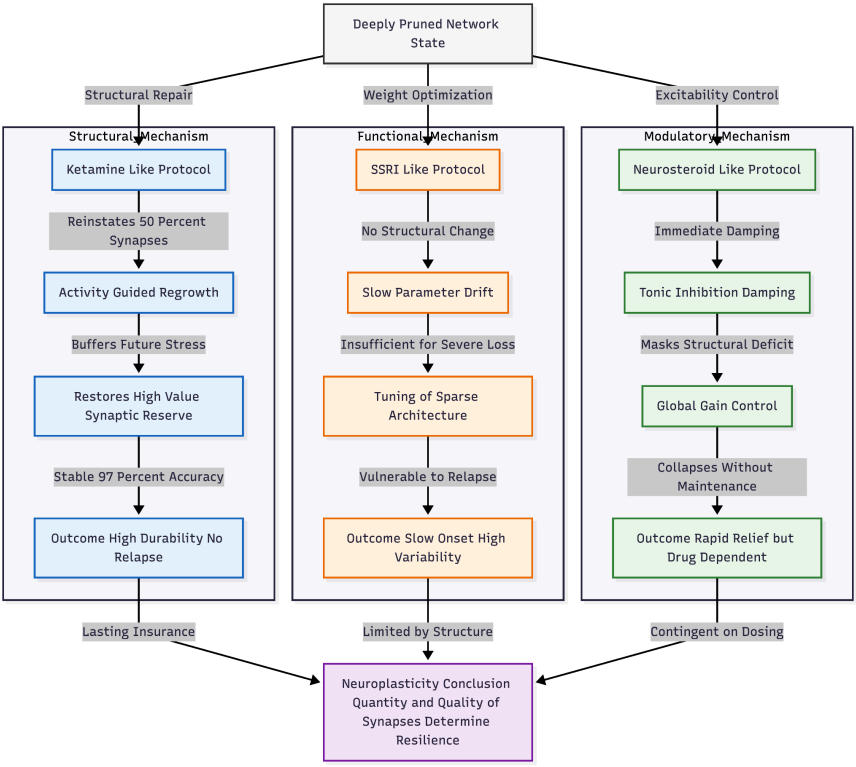


Figure 3: *Comparative Mechanisms of Action and Resilience Profiles. A schematic representation of the three distinct recovery pathways observed in the simulation. The Ketamine-like pathway (left, blue) confers lasting resilience by rebuilding synaptic "reserve" via activity-guided regrowth, mirroring BDNF-dependent structural repair. The SSRI-like pathway (center, orange) relies on tuning existing weights; while effective for some, it exhibits high variability and vulnerability when structural loss is too severe to be compensated by parameter drift alone. The Neurosteroid-like pathway (right, green) provides immediate functional relief via global gain control, but this benefit is contingent on the presence of the agent, failing to prevent relapse once the modulation is removed from a structurally deficient network.*

The SSRI-like schedule, by contrast, left the sparse architecture untouched and relied on slow parameter drift while internal noise tapered. Immediate gains were modest and uneven—some seeds barely moved off the 30 % baseline—yet most improved gradually and several reached the mid-90 % range by the final cycle. Four networks, however, never achieved a secure foothold and slipped below the 80 % relapse line early on. This mirrors the familiar picture of selective serotonin re-uptake inhibitors: effectiveness that emerges over weeks, large patient-to-patient variability, and a sizeable minority of non-responders [10,9]. The simulation implies that when synaptic loss is severe, merely "tuning" existing weights may be insufficient unless accompanied by structural repair.

The neurosteroid-like intervention offered the quickest subjective relief—performance snapped back to healthy levels the moment tonic damping was engaged—but the benefit proved contingent on continued inhibition. Removing the damping cut combined-stress accuracy by nearly thirty points, and two seeds collapsed during the last stress cycle despite maintenance sessions. Clinical experience with brexanolone and zuranolone is similar: fast symptom relief that can wane once dosing stops, particularly if underlying circuitry has not had time to rebuild [6]. In the model, global gain control contained excitability but, as pruning progressed, there were simply too few functional pathways left to stabilise.

Together, these findings support a neuroplasticity account of major

depression in which both the quantity and the quality of synapses determine outcome [11]. When loss is extensive, agents that drive new growth confer lasting insurance; modulators that optimise or damp existing connections help only while enough structure remains—or while the drug is present. The wide seed-to-seed spread observed for monoaminergic and, later, neurosteroid conditions hints that individual differences in baseline pruning depth or excitability may underlie the heterogeneous responses seen in clinics.

Clinical decision-making and personalised sequencing

The simulation outcomes map neatly onto everyday prescribing dilemmas and suggest a triage logic based on underlying circuit damage and the urgency of symptom relief.

Profound structural loss—often seen in chronic, highly recurrent or treatment-resistant depression—appears best addressed with synaptogenic drugs. In the model, the ketamine analogue restored performance almost immediately, remained stable under extreme perturbation, and never relapsed even after further 30–40 % pruning. These features mirror clinical reports that a short ketamine course can trigger durable remission through rapid BDNF–mTOR-dependent spine formation [12]. For patients who repeatedly fail monoaminergic agents or present with marked cognitive blunting or anhedonia, glutamatergic treatments therefore deserve early consideration.

Situations demanding swift containment—postpartum depression, imminent suicide risk, or severe agitation—may profit from neurosteroid modulators. In silico, the GABAergic routine normalised accuracy in a single step and maintained high scores through the first stress cycles, echoing the clinical speed of brexanolone and zuranolone [6]. The sharp decline after damping withdrawal, however, cautions that such agents are bridges rather than stand-alone long-term solutions; follow-up therapy should be organised before discontinuation.

When illness is less entrenched and baseline circuitry largely intact, conventional SSRIs remain useful. Their simulated trajectory—slow but steady gains, substantial variability and occasional early failure—parallels real-world response curves. A poor initial trajectory may signal the need for augmentation rather than a prolonged wait-and-see approach.

Combined strategies emerge naturally from the model. Neurosteroids can cover the latency of SSRIs; ketamine-guided regrowth can be followed by low-dose SSRIs or psychotherapy to consolidate new connections; booster ketamine sessions can reinforce reserve in highly stress-exposed patients.

The contrasting relapse rates—zero for synaptogenic rescue versus 20–40 % for purely functional modulation—also argue for integrating biomarkers of synaptic loss or neuroinflammation into routine assessment so that clinicians can choose a mechanism informed by

pathophysiology rather than by trial-and-error.

Novelty and translational potential

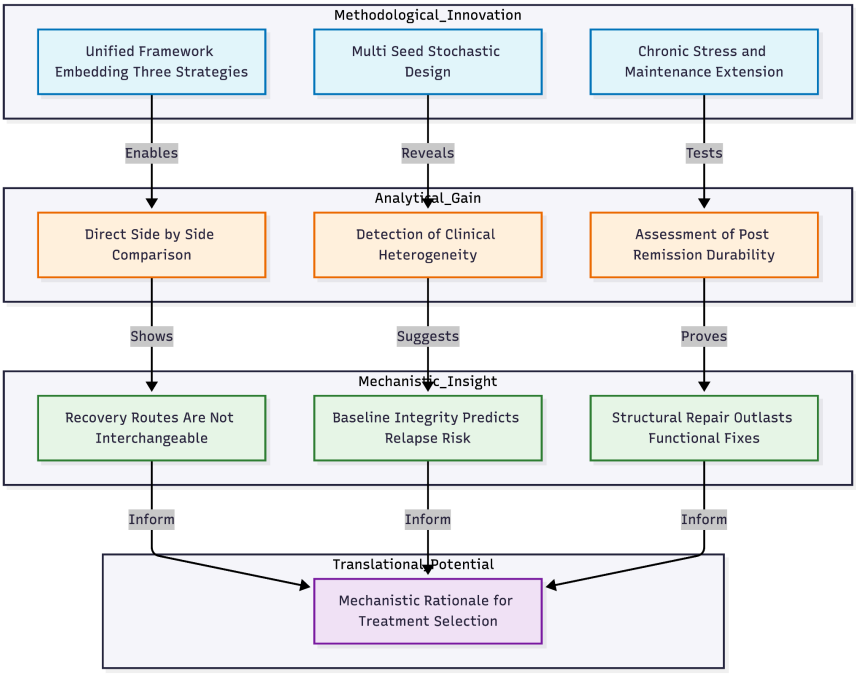


Figure 4: *Methodological Novelty and Translational Framework. This diagram illustrates how the study's specific design choices (blue) enable new analytical capabilities (orange), leading to distinct mechanistic insights (green) that culminate in clinical utility (purple). Unlike previous isolated studies, the Unified Framework allows for direct comparison, revealing that recovery pathways are distinct rather than interchangeable. The Multi-Seed Design captures stochastic variability, linking relapse to baseline circuit integrity rather than drug class alone. Finally, the Chronic Stress Extension demonstrates that while functional fixes can fail over time, structural repair offers lasting protection, providing a quantified rationale for selecting treatments based on the trade-off between speed, durability, and state-dependence.*

By embedding three fundamentally different antidepressant strategies

inside one pruning-plasticity model, this study offers a clearer look at how treatment mechanism shapes both early response and long-term course (Figure 4). Previous in-silico work usually explored one pathway at a time—ketamine-like regrowth, slow monoaminergic adaptation, or global inhibitory damping—making cross-class comparisons indirect at best. Running the three approaches side-by-side from identical, over-pruned baselines shows that the routes to recovery are not interchangeable.

The multi-seed design is equally important. Introducing stochastic variation revealed patterns that a single deterministic run would miss: some "patients" on the SSRI schedule never cross a therapeutic threshold, whereas a minority on the neurosteroid schedule crash only after many stress cycles. Such divergence echoes clinical heterogeneity and suggests that baseline circuit integrity or excitability, rather than transmitter class per se, may determine who ultimately relapses.

The chronic-stress extension adds another layer. Few computational studies have asked what happens after the first remission; here, progressive pruning plus minimal "maintenance" demonstrates why structural rebuilding protects every seed, whereas purely functional fixes leave a tail of late failures. These findings dovetail with the growing view that major depression reflects circuit-level dysconnectivity, not merely monoamine shortage [11]. Quantifying the trade-offs—speed versus durability versus state-dependence—gives clinicians a mechanistic rationale for selecting ketamine, SSRIs, neurosteroids, or

combinations according to individual risk profiles.

Limitations

Several caveats must temper direct biological inference. A feed-forward network cannot reproduce the reverberating loops of corticolimbic circuits; magnitude-based pruning is only a proxy for microglial or inflammatory synapse loss. The ketamine analogue grants faultless activity-guided regrowth, ignoring maladaptive sprouting; the SSRI routine sidesteps debates over serotonin depletion after long exposure [13]; the neurosteroid module dampens every hidden unit equally, whereas real extrasynaptic GABA_A receptors sit on select neurons.

Inter-seed variance arises from random data splits and weight initialisation, not from modelled genetic or hormonal modifiers. Maintenance schedules are sketched loosely on clinical practice—brief boosters for ketamine, longer courses for the others—but do not address dose, tolerability, or adherence. Adding recurrent dynamics, cell-type specificity, and empirically derived patient heterogeneity will be essential next steps.

Conclusion

Within these constraints, the simulations offer a coherent narrative: fragile circuits can be helped in three distinct ways—rebuilding lost links, patiently tuning what remains, or providing a temporary brake on

over-excitation—but only the first delivers uniform, lasting protection against future stress. A single framework that recreates these divergent trajectories narrows the gap between computational theory and bedside choice, supporting a shift toward mechanism-guided, plasticity-focused care as rapid-acting agents become mainstream [12].

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization.

[2] Rush, A. J., Trivedi, M. H., Wisniewski, S. R., et al. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. American Journal of Psychiatry, 163(11), 1905-1917. https://doi.org/10.1176/ajp.2006.163.11.1905

[3] Trivedi, M. H., Rush, A. J., Wisniewski, S. R., et al. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. American Journal of Psychiatry, 163(1), 28-40. https://doi.org/10.1176/appi.ajp.163.1.28

[4] Murrough, J. W., Iosifescu, D. V., Chang, L. C., et al. (2013). Antidepressant efficacy of ketamine in treatment-resistant major

depression: A two-site randomized controlled trial. American Journal of Psychiatry, 170(10), 1134–1142. https://doi.org/10.1176/appi.ajp.2013.13030392

[5] Iadarola, N. D., Niciu, M. J., Richards, E. M., et al. (2015). Ketamine and other NMDA receptor antagonists in the treatment of depression: A perspective review. Therapeutic Advances in Chronic Disease, 6(3), 97-114. https://doi.org/10.1177/2040622315579059

[6] Gunduz-Bruce, H., Takahashi, K., et al. (2022). Development of neuroactive steroids for the treatment of postpartum depression. Journal of neuroendocrinology, 34(2), e13019.

[7] Duman, R. S., & Aghajanian, G. K. (2012). Synaptic dysfunction in depression: Potential therapeutic targets. Science, 338(6103), 68-72. https://doi.org/10.1126/science.1222939

[8] Cheung, N. (2026). Divergent Mechanisms of Antidepressant Efficacy: A Unified Computational Comparison of Synaptogenesis, Stabilization, and Tonic Inhibition in a Model of Depression. Zenodo. https://doi.org/10.5281/zenodo.18290014

[9] Stahl, S. M. (1998). Mechanism of action of serotonin selective re-uptake inhibitors: Serotonin receptors and pathways mediate therapeutic effects and side effects. Journal of Affective Disorders, 51(3), 215–235. https://doi.org/10.1016/S0165-0327(98)00221-3

[10] Boschloo, L., Hieronymus, F., Lisinski, A., et al. (2023). The complex clinical response to selective serotonin reuptake inhibitors in depression: a network perspective. Translational Psychiatry, 13(1), 19.

[11] Page, C. E., Epperson, C. N., Novick, A. M., et al. (2024). Beyond the serotonin deficit hypothesis: communicating a neuroplasticity framework of major depressive disorder. Molecular psychiatry, 29(12), 3802–3813. https://doi.org/10.1038/s41380-024-02625-2

[12] Krystal, J. H., Abdallah, C. G., Sanacora, G., et al. (2019). Ketamine: A paradigm shift for depression research and treatment. Neuron, 101(5), 774–778. https://doi.org/10.1016/j.neuron.2019.02.005

[13] Moncrieff, J., Cooper, R. E., Stockmann, T., et al. (2023). The serotonin theory of depression: A systematic umbrella review of the evidence. Molecular Psychiatry, 28(8), 3243–3256. https://doi.org/10.1038/s41380-022-01661-0

Chapter 7
Modeling Antidepressant-Induced Manic Switch and Longitudinal Relapse: A Unified Pruning Framework Highlights Glutamatergics' Disease-Modifying Potential

Cheung, Ngo

Cheung, N. (2026). Modeling Antidepressant-Induced Manic Switch and Longitudinal Relapse: A Unified Pruning Framework Highlights Glutamatergics' Disease-Modifying Potential. Zenodo. https://doi.org/10.5281/zenodo.18298989

Abstract

Background: Major depressive disorder involves impaired neural plasticity, yet antidepressants targeting glutamatergic (ketamine), monoaminergic (SSRIs), and GABAergic (neurosteroids) pathways differ markedly in onset speed, durability, and risk of treatment-emergent mania—particularly in bipolar contexts. Clinical comparisons are confounded by heterogeneity; computational models enable controlled mechanistic dissection, but few integrate manic liability and

post-discontinuation stability across classes.

Methods: We extended a magnitude-based pruning model (95% sparsity) of depression in feed-forward networks classifying Gaussian blobs. From identical pruned baselines, three interventions were simulated: ketamine-like gradient-guided synaptic regrowth (50% reinstatement) with consolidation; SSRI-like prolonged low-rate refinement with tapering noise and escalating excitability gain; neurosteroid-like global tonic inhibition (0.7× damping, tanh activations, reduced gain). Efficacy assessed classification accuracy under clean, noisy, and combined stress; resilience via graded noise tolerance; acute relapse after further pruning; manic risk through biased positive perturbation and activation magnitude. Longitudinal relapse modeled chronic maintenance (with mood stabilizer protection) followed by discontinuation, using treatment-specific lingering decay rates. Metrics averaged across 10 seeds.

Results: All treatments restored near-ceiling performance acutely, but ketamine-like regrowth yielded superior extreme-stress resilience (76.8%) and zero post-discontinuation manic relapse, reducing sparsity to 47.5%. Neurosteroid-like modulation matched rapid recovery (97.6%) but showed state-dependence and 88.3% relapse probability off-drug. SSRI-like refinement lagged in resilience (49.9% extreme) with highest manic proxies (biased accuracy 47.2%, gain 1.60) and 95.0% relapse post-cessation. Longer maintenance conferred negligible added protection for reversible mechanisms.

Conclusions: Antidepressants operate via divergent plasticity routes—durable structural rebuilding (ketamine-like, low long-term risk), rapid reversible stabilization (neurosteroid-like), and vulnerable gradual optimization (SSRI-like)—reproducing clinical trade-offs in speed, persistence, and bipolar safety. These findings support mechanism-guided selection, positioning synaptogenic agents for recurrent or high-risk cases pursuing remission beyond treatment.

Introduction

Major depressive disorder (MDD) is a leading contributor to disability worldwide and imposes a substantial burden on individuals, families, and health-care systems [1]. Contemporary pharmacotherapy has helped many patients, yet full remission after an initial antidepressant trial is achieved in only about one-third of cases, and many people remain symptomatic despite several treatment attempts [2]. Selective serotonin re-uptake inhibitors (SSRIs) typically require several weeks before benefits become obvious, leaving patients exposed to prolonged distress and only partial relief [3].

The delayed and incomplete response seen with SSRIs has fuelled interest in compounds that act on other signalling systems. Low-dose ketamine, an NMDA-receptor antagonist, can lift mood within hours and appears to do so by stimulating brain-derived neurotrophic factor

(BDNF) and mTOR-dependent synaptogenesis [4,5]. Neuroactive steroids such as zuranolone also show rapid antidepressant effects especially for postpartum depression [6]. These findings have shifted attention away from a strictly monoaminergic model toward the view that MDD involves impaired neural plasticity, in which chronic stress erodes dendritic spines and synaptic density in cortical and hippocampal regions [7].

Another clinical complication is treatment-emergent mania, particularly in bipolar disorder, where conventional antidepressants provoke mood switches in roughly 20–40 % of patients [8]. Initial findings indicate that ketamine presents a reduced acute switch risk in controlled settings [9], while preliminary studies suggest a negligible risk associated with neurosteroids [6]. It is crucial to comprehend how these mechanistically distinct treatments affect depressive symptoms and the excitatory–inhibitory balance; however, direct clinical comparisons are impeded by variability in patient populations, dosing regimens, and concurrent therapies.

Computational modelling offers a controlled way to disentangle these factors. Prior pruning-based simulations have cast depression as a state of excessive synaptic loss, with ketamine-like regrowth restoring network resilience. Few studies, however, have placed glutamatergic, monoaminergic, and GABAergic strategies side by side or explored how they affect risks such as manic switching or post-treatment stability.

The present work addresses these gaps through an extended magnitude-pruning model applied to feed-forward neural networks. Beginning with identical over-pruned networks, we simulated three treatment motifs: ketamine-like synaptogenesis, SSRI-like gradual refinement accompanied by rising excitability, and neurosteroid-like tonic inhibition. End-points included acute antidepressant efficacy, resilience to stress, proxies for manic conversion (biased excitatory challenge and activation amplitude), immediate relapse risk, and—for the first time in this framework—long-term vulnerability after chronic maintenance and full discontinuation, incorporating treatment-specific wash-out profiles. By embedding these elements in one plasticity-centred model we aim to clarify the trade-offs in speed, durability, and bipolar safety across drug classes, thereby informing mechanism-based treatment selection.

Methods

Network architecture and classification task

Figure 1 shows the experimental workflow for multi-Mechanism antidepressant comparison. We used a small feed-forward network to stand in for key cortico-limbic circuits. The model had two input units, three hidden layers of 512, 512 and 256 units, and a four-unit soft-max output layer. Hidden layers normally used ReLU activation; during neurosteroid simulations tanh was substituted to mimic the ceiling effect

of tonic GABAergic currents. Altogether the network held about 3.9 × 10^5 trainable weights.

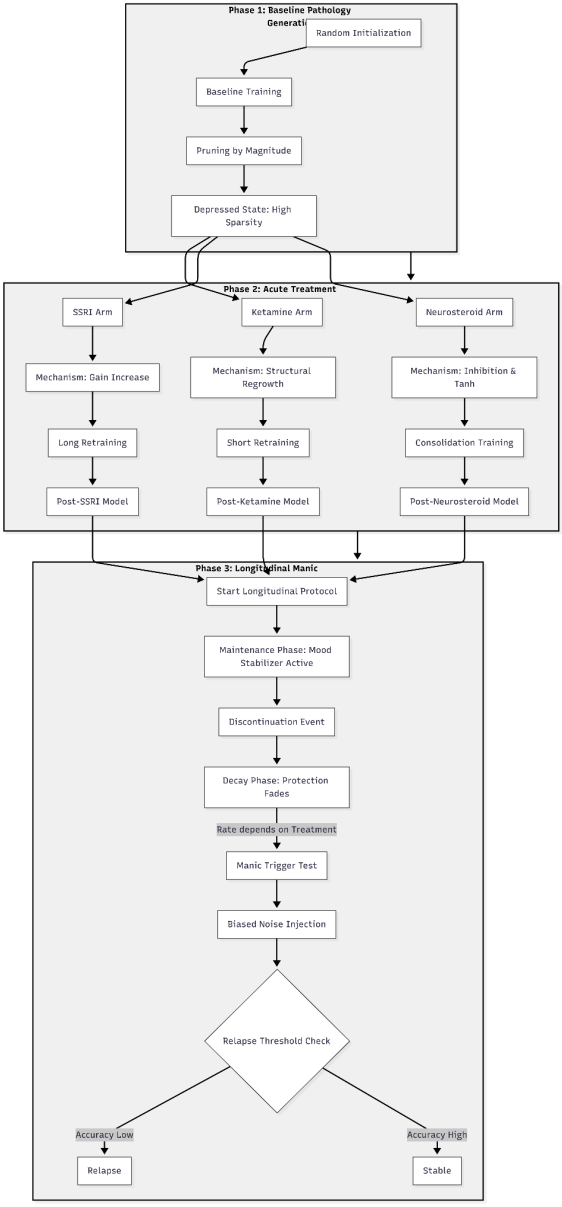


Figure 1. *Experimental Workflow for Multi-Mechanism Antidepressant Comparison. The pipeline consists of three distinct phases: (1) Baseline Pathology Generation, where a neural network is trained and pruned to simulate a depressed, sparse state; (2) Acute Treatment, where the model branches into three mechanistic arms—Ketamine (structural regrowth), SSRI (functional gain increase), and Neurosteroids (inhibition modulation)—to reverse pathology; and (3) Longitudinal Manic Relapse Simulation, where treated models undergo a maintenance phase followed by medication discontinuation. In this final phase, residual mood stabilizer protection decays at treatment-specific rates, and the model is subjected to a biased noise trigger test to assess the probability of manic switching.*

The learning task was simple four-way pattern recognition. Two-dimensional points were drawn from four Gaussian clouds centred at (−3, −3), (−3, 3), (3, −3) and (3, 3) with a standard deviation of 0.8. Each run used 12 000 labelled training points. Three test sets were prepared:

- a 4 000-item set with the same noise as the training data (standard condition);
- a 2 000-item noise-free set (clean condition);
- perturbed versions for stress tests, created by adding extra Gaussian noise after each hidden layer or by multiplying all activations with a global gain factor.

A "mood-stabiliser" guard–rail was implemented as three scalar parameters: a protection level (0–1) that capped gain, a small inhibitory bias (−0.15 × protection) and a factor (0.3) that reduced the upward shift of internal noise. All code ran in PyTorch on a single CPU. Ten different random seeds (affecting data order and weight initialisation) produced

ten independent "subjects."

Simulation of the depressive state

The network was first trained for 20 epochs with Adam (learning rate 0.001) on noise-free data until it reached ceiling accuracy. Depression was then modelled by iterative magnitude pruning: across the three hidden layers the 95 % smallest-magnitude weights were zeroed, leaving a sparse and fragile network. Clean-input accuracy stayed high, but performance collapsed when noise or further pruning was applied, mirroring the stress sensitivity of a depressed brain [7].

Antidepressant treatment protocols

From the same pruned starting point three treatment routines were run on separate copies.

Ketamine-like: A modest global gain (1.25) was fixed. Gradients were collected over 30 mini-batches to locate strong silent synapses; half of these pruned weights were re-instated with small random values drawn from N(0, 0.03). Fifteen fine-tuning epochs followed (Adam, 0.0005) while the mask was locked.

SSRI-like: Sparsity (95 %) was left unchanged. Over 100 epochs the internal noise level fell linearly from 0.5 to 0, while the global gain rose from 1.0 to 1.6, simulating slow monoaminergic adaptation. Learning

used Adam with a rate of 1 × 10^-5.

Neurosteroid-like: We kept the prune mask but multiplied post-activation values by 0.7, switched ReLU to tanh, and set the gain at 0.85 (effective ≈0.59). Ten consolidation epochs (Adam, 0.0005) followed.

Mood-stabiliser extension and longitudinal relapse test

After acute treatment, a chronic phase was added. A full protection level (1.0) was turned on and held during maintenance, then allowed to decay after drug discontinuation. Decay rates were treatment-specific: 0.002 per step for the ketamine model (long-lasting structural change), 0.015 for the SSRI model (rapid reversal) and 0.008 for the neurosteroid model (intermediate). Maintenance lasted 25, 50, 100, 150, 200 or 300 low-rate epochs (Adam, 1 × 10^-6). Drugs were then removed in one step, and 50 decay steps were run to wash out residual protection. Manic risk was probed by injecting strongly positive internal noise (σ = 1.0, shift = +1.0); relapse was logged when accuracy fell below 60 %.

Outcome measures

Primary efficacy was the percentage of correct classifications on clean, standard-noise and combined-stress test sets. Resilience curves were built by repeating tests with internal noise ranging from 0 to 2.5. Acute relapse was tested by pruning a further 40 % of the remaining weights and retesting under combined stress.

Manic conversion risk was indexed two ways: accuracy under highly biased noise (lower accuracy = higher risk) and the mean absolute activation in hidden layers (higher activation = greater latent excitability). Neurosteroid state-dependence was recorded as the drop in accuracy when the damping module was turned off.

Results

Simulations were repeated with ten different random seeds, and every outcome followed the same rank order across seeds, indicating that the findings are robust to stochastic variation in data shuffling and weight initialisation.

Acute antidepressant efficacy

Before treatment, the heavily pruned network managed only 29.7 ± 2.7 % accuracy when clean inputs were combined with internal and external noise. Introducing any of the three treatment routines produced a dramatic rebound (Table 1). Neurosteroid-like damping lifted combined-stress accuracy to 97.6 ± 0.3 %, ketamine-like synaptogenesis to 97.2 ± 0.2 %, and SSRI-like refinement to 90.5 ± 3.0 %. On both the noise-free and the standard-noise test sets all treated models reached or approached ceiling performance, whereas the untreated model stayed near one-third correct. The ketamine condition achieved its improvement

with an effective sparsity of 47.5 %, reflecting reinstated connections; the other two conditions retained the original 95 % sparsity.

Table 1. *Antidepressant Efficacy (Mean ± SD Across 10 Seeds)*

Treatment	Sparsity (%)	Clean (%)	Standard (%)	Combined (%)
Untreated (pruned)	95.0 ± 0.0	34.7 ± 11.9	36.8 ± 11.9	29.7 ± 2.7
Ketamine-like	47.5 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	97.2 ± 0.2
SSRI-like	95.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.1	90.5 ± 3.0
Neurosteroid-like	95.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	97.6 ± 0.3

Stress resilience

Table 2*. Stress Resilience Profile (Mean ± SD Across 10 Seeds)*

Treatment	None (%)	Moderate (σ=0.5) (%)	High (σ=1.0) (%)	Severe (σ=1.5) (%)	Extreme (σ=2.5) (%)
Untreated (pruned)	36.8 ± 11.9	29.9 ± 2.5	29.6 ± 1.8	29.8 ± 1.4	29.6 ± 1.5
Ketamine-like	100.0 ± 0.0	99.9 ± 0.1	98.2 ± 1.1	92.9 ± 2.4	76.8 ± 3.6
SSRI-like	99.9 ± 0.1	95.1 ± 3.2	78.4 ± 5.4	64.9 ± 5.3	49.9 ± 2.8
Neurosteroid-like	100.0 ± 0.0	99.9 ± 0.1	93.0 ± 2.6	70.6 ± 2.9	43.0 ± 1.0

Performance was next examined while internal Gaussian noise was increased stepwise from none to a standard deviation of 2.5 (Table 2). Ketamine-treated networks tolerated the severest disturbance best, holding 76.8 ± 3.6 % accuracy at the highest noise level. SSRI-treated

networks fell to 49.9 ± 2.8 %, and neurosteroid-treated networks to 43.0 ± 1.0 %. At moderate and high noise (σ = 1.0–1.5) the ketamine and neurosteroid models performed similarly (93.0–98.2 %) and both outperformed the SSRI model (64.9–78.4 %). The untreated network hovered around 30 % regardless of noise intensity.

Manic conversion risk

Table 3. *Manic Conversion Risk Metrics (Mean ± SD Across 10 Seeds)*

Treatment	Gain Multiplier	Biased Stress Accuracy (%)	Activation Magnitude
Untreated (pruned)	1.00 ± 0.00	25.0 ± 0.8	0.100 ± 0.013
Ketamine-like	1.25 ± 0.00	84.2 ± 8.5	0.649 ± 0.079
SSRI-like	1.60 ± 0.00	47.2 ± 12.7	0.390 ± 0.079
Neurosteroid-like	0.85 ± 0.00	50.6 ± 7.9	0.196 ± 0.008

Note. Lower biased stress accuracy indicates higher manic conversion vulnerability; higher activation magnitude reflects greater latent hyperexcitability.

Potential switch liability was probed with strongly positive, biased internal noise (Table 3). The SSRI routine, which had wound excitability up to a gain of 1.60, proved most vulnerable: biased-noise accuracy averaged 47.2 ± 12.7 %, and the mean absolute hidden-layer activation was 0.390 ± 0.079. Neurosteroid modulation, despite lowering gain to 0.85, achieved only slightly better biased-noise accuracy (50.6 ± 7.9 %) and showed the lowest activation magnitude (0.196 ± 0.008). Ketamine treatment combined a moderate gain of 1.25 with markedly safer behaviour, sustaining 84.2 ± 8.5 % accuracy under the same biased

challenge and exhibiting the highest activation magnitude (0.649 ± 0.079) without instability. The pruned, untreated model remained both hypo-active (0.100 ± 0.013) and inaccurate (25.0 ± 0.8 %).

Acute relapse vulnerability

Durability was tested by excising a further 40 % of the remaining weights after treatment. Ketamine-treated networks were essentially unaffected, their combined-stress accuracy changing by −0.1 ± 0.3 %. Neurosteroid-treated networks lost 5.1 ± 2.1 % and SSRI-treated networks 7.0 ± 2.4 %, confirming a clear advantage for the structural regrowth produced by the ketamine routine.

Neurosteroid medication dependence

To gauge state-dependence, the neurosteroid damping module was switched off after the acute phase. Combined-stress accuracy fell from 97.6 ± 0.3 % to 78.5 ± 4.9 %, and biased-noise accuracy dropped from 50.6 ± 7.9 % to 36.9 ± 9.6 %. Interestingly, accuracy at the most extreme unbiased noise level ($\sigma = 2.5$) rose from 43.0 ± 1.0 % to 58.3 ± 4.1 %, indicating that tonic inhibition trades robustness to excitation-biased threats for reduced tolerance of diffuse noise.

Longitudinal manic relapse after discontinuation

A chronic maintenance phase was appended, followed by complete drug

withdrawal and gradual decay of the virtual mood-stabiliser. Decay rates were set a priori to 0.002 per step for ketamine-treated networks, 0.015 for SSRI-treated networks, and 0.008 for neurosteroid-treated networks.

Table 4. *Summary Comparison Matrix*

Metric	Ketamine-like	SSRI-like	Neurosteroid-like	Untreated
Combined Stress (%)	97.2	90.5	97.6	29.7
Biased Stress (%)	84.2	47.2	50.6	25.0
Gain Multiplier	1.25	1.60	0.85	1.00
Activation Magnitude	0.649	0.390	0.196	0.100
Acute Relapse Drop (%)	−0.1	7.0	5.1	N/A
Manic Relapse Prob. (%)	0.0	95.0	88.3	N/A
MS Decay Rate	0.0020	0.0150	0.0080	N/A

After all durations of maintenance (25–300 additional training epochs) and the full wash-out period, ketamine-treated networks never relapsed: biased-noise accuracy remained above 91 % in every seed. In contrast, SSRI-treated networks relapsed in 95 % of all seed-by-duration combinations, and neurosteroid-treated networks in 88.3 %. Post-withdrawal biased-noise accuracy for the SSRI and neurosteroid groups stabilised in the low-forties, irrespective of how long maintenance had lasted, whereas the ketamine group stayed in the low-nineties. These observations confirm that the protective changes induced by the ketamine routine are both structurally persistent and highly effective at preventing manic-like destabilisation, while the

functional adaptations driven by SSRIs and the partially state-dependent modulation produced by neurosteroids leave the system vulnerable once the drugs and the auxiliary stabiliser are withdrawn (Table 4).

Discussion

Interpretation of acute and resilience findings

The three simulated treatment paths behaved much as clinicians might expect at the bedside. When the pruned network was exposed to simultaneous external and internal noise—our analogue of depressive pressure—both the ketamine-like and neurosteroid-like routines snapped performance back to almost normal within a few training steps. The slower, SSRI-like schedule helped, but never quite caught up. This mirrors the clinic, where ketamine can lift mood in hours [4] and zuranolone in a few days [10], whereas selective-serotonin reuptake inhibitors usually need several weeks [3].

Differences emerged when we kept turning up the internal noise. Networks that had undergone ketamine-style synaptogenesis kept working even at the most extreme setting, a result that fits reports of durable stress buffering after ketamine-induced structural change [5]. Neurosteroid-like damping steadied the system only as long as the inhibitory module stayed in place; once removed, performance slid, echoing the clinical observation that benefits from a short zuranolone course can wane [11]. SSRI-like refinement offered the least cushion,

tracking the modest resilience frequently seen when conventional antidepressants are the sole therapy in difficult cases [2].

Manic conversion risk and excitability balance

Our proxies for switch risk told a familiar story. Raising network gain in the SSRI-like condition produced the greatest drop in accuracy when a positive noise bias—our stand-in for incipient mania—was introduced. The result parallels the 20–40 % switch rate associated with antidepressants in bipolar disorder [8]. The ketamine-like model showed only a mild gain increase yet kept biased-noise accuracy high, consistent with the low switch rates reported when ketamine is used alongside mood stabilizers [9]. The neurosteroid-like routine lowered both gain and hidden-unit activation and therefore looked safest, matching early reports that zuranolone rarely provokes mania [6,12].

These patterns underline how the route to recovery shapes the excitation–inhibition balance. Building new synapses tolerates some extra excitatory drive; tonic inhibition suppresses it outright; simply turning up global gain, as with the SSRI model, risks overshoot unless other brakes are applied.

Long-term stability after discontinuation

When we added a maintenance phase and then withdrew all drugs, the contrasts sharpened. Circuits repaired in the ketamine-like way never

relapsed—even after the mood-stabiliser parameters had almost fully decayed—suggesting that structural change can make the system self-supporting. Clinical series describing months-long benefit after limited ketamine infusions in bipolar depression point in the same direction [13]. By comparison, nearly every SSRI-like or neurosteroid-like network relapsed once protective settings were lifted, regardless of how long maintenance had lasted. Naturalistic studies show a similar pattern: recurrence remains common after antidepressant or neurosteroid withdrawal, often exceeding 40 % a year [14]. Extending maintenance did not help these two models much, mirroring data that slow tapers reduce but do not remove relapse risk when monoaminergic drugs are stopped [15]. Only the strategy that rebuilt connectivity fundamentally altered vulnerability.

Implications for clinical judgment and treatment selection

The simulation highlights how different drug mechanisms may guide day-to-day prescribing (Table 5). Circuits rebuilt through a ketamine-like process kept their stability long after the "drug" was withdrawn, suggesting that glutamatergic agents could suit patients who want lasting relief without continuous medication. That property is already seen in clinical series where repeated ketamine infusions provide benefit for months in otherwise resistant depression [13].

By contrast, the model that mimicked SSRI action showed the highest risk of a manic switch and the greatest relapse once treatment stopped.

These results echo long-standing warnings about monoaminergic monotherapy in people with bipolar features and about the sharp rise in recurrence after antidepressant discontinuation [14]. When such agents are used, combining them with lithium or valproate and planning for ongoing maintenance remain prudent steps [15].

Neurosteroid-like modulation offered fast symptom control and little acute excitability, consistent with early zuranolone studies in postpartum and bipolar depression [6]. Yet the same model relapsed quickly once the inhibitory drive was removed, implying that these drugs may work best as short bridges—useful in urgent situations, but followed by a hand-off to treatments that remodel the network more permanently.

Taken together, the findings support a tiered strategy (Table 5). Plasticity-inducing drugs may be chosen for chronic, relapsing, or bipolar-spectrum illness where durable remission is the goal; GABAergic neurosteroids can fill short-term needs for rapid relief; and traditional antidepressants should be reserved for well-selected unipolar cases or used with sturdy mood-stabilizing partners. Matching a patient's history of switches, relapse density, and treatment goals to these distinct profiles could improve outcomes as new rapid-acting options become available.

Antidepressant Class (Model Analog)	Acute Efficacy & Speed	Durability & Post-Discontinuation Stability	Manic Conversion Risk (Acute & Longitudinal)	Recommended Clinical Contexts	Key Considerations from Model
Ketamine-like (Glutamatergic synaptogenesis)	Rapid, near-complete recovery (97.2% combined stress)	Highest resilience to extreme stress (76.8%); zero manic relapse post-cessation; structural changes persist	Moderate acute (biased accuracy 84.2%); lowest long-term vulnerability	• Treatment-resistant unipolar depression • Recurrent or bipolar-spectrum illness aiming for remission beyond ongoing treatment • Patients seeking stability without indefinite medication	Prioritize for cases needing disease-modifying potential; supports earlier escalation in refractory depression
SSRI-like (Monoaminergic refinement + gain escalation)	Slower/incomplete (90.5% combined stress)	Lowest resilience (49.9% extreme stress); near-certain relapse post-cessation (95%)	Highest acute (biased accuracy 47.2%, gain 1.60) and longitudinal risk	• Low manic-switch-risk unipolar depression • Only with indefinite mood stabilization in bipolar	Avoid monotherapy in bipolar vulnerability; requires concurrent mood stabilizers and careful monitoring due to rapid recurrence upon cessation
Neurosteroid-like (GABAergic tonic inhibition)	Rapid, near-complete recovery (97.6% combined stress)	Moderate resilience (43.0% extreme stress); high relapse post-cessation (88.3%); state-dependent	Lowest acute (biased accuracy 50.6%, damped excitability)	• Urgent scenarios needing rapid relief (e.g., postpartum depression, acute bipolar depressive episodes) • Short-term bridging	Excellent for acute safety and speed; use as bridge with planned tapering or transition to more durable agents

Table 5: *How Simulation Results May Inform Clinical Judgment in Antidepressant SelectionNote. This table distills the model's comparative profiles into actionable guidance, emphasizing mechanism-based stratification. Clinicians should integrate patient-specific factors (e.g., prior switch history, episode density) alongside these insights when selecting or sequencing treatments.*

Novelty and potential impact

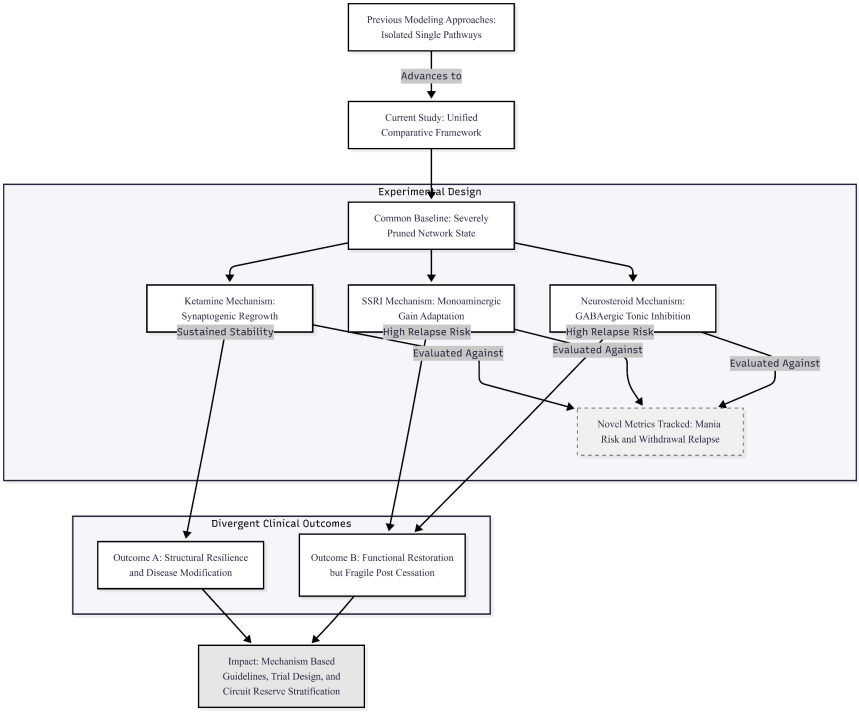


Figure 2. *Methodological novelty and clinical implications of the unified computational framework. Unlike previous studies that modeled antidepressant mechanisms in isolation, this framework initializes three distinct pharmacological pathways—synaptogenic, monoaminergic, and GABAergic—from a shared, severely pruned network baseline. By uniquely tracking treatment-emergent mania and post-withdrawal relapse, the model differentiates between agents that offer structural "circuit reserve" (ketamine-like) versus those providing symptomatic relief with high discontinuation fragility (SSRI-like and neurosteroid-like). These divergent outcomes provide a mechanistic logic for future patient stratification and clinical trial design.*

This study is one of the few attempts to place three very different

antidepressant mechanisms inside a single, carefully controlled computational frame (Figure 2). Most earlier models concentrated on one pathway at a time—for example, pruning models that mimic synaptic loss in depression [7] or simulations of ketamine-driven regrowth alone [16]. By contrast, the present work starts every network from the same severely pruned state and then applies, side-by-side, a ketamine-like synaptogenic routine, an SSRI-like slow gain adaptation, and a neurosteroid-like tonic inhibition. In doing so it also tracks outcomes that matter for bipolar illness—treatment-emergent mania and relapse after drug withdrawal—areas that computational studies usually ignore.

The resulting picture is clinically recognizable. Ketamine-style regrowth stands out for long-term stability; once new connections form, the model keeps its resilience even when medication parameters decay. This finding echoes emerging clinical views of ketamine as more than a symptomatic drug, possibly a disease-modifying agent in hard-to-treat or bipolar depression [5,13]. In contrast, both the SSRI-like and neurosteroid-like routes restore function quickly but leave the system fragile once treatment stops, mirroring high relapse rates seen after discontinuing these medications [14] and the need for bridging strategies after short neurosteroid courses [11]. By embedding all three paths in the same architecture the model offers a clear, mechanistic logic that could guide future guidelines, trial design, and biomarker work—for example, identifying patients with low "circuit reserve" who might benefit most from synaptogenic drugs.

Limitations

Several simplifications temper direct clinical translation. The feed-forward network omits the recurrent and oscillatory loops that dominate cortico-limbic mood circuits, so real-world instabilities may be underestimated. All interventions were applied globally, whereas in vivo actions are cell-type and region specific—especially for extrasynaptic GABA-A targets of neurosteroids [12]. Manic risk was approximated by adding biased noise, not by modelling full affective episodes, and subject variability was limited to random seeds rather than patient-like heterogeneity in pruning depth or plasticity reserve. Finally, mood-stabilizer co-therapy was represented only by simple decay rates; explicit multi-drug interactions were not explored. These choices kept the comparison tractable but mark priorities for future recurrent, multi-compartment, or spiking models.

Conclusion

Placing synaptogenesis, tonic inhibition, and gradual gain tuning on the same depleted substrate clarifies how each path balances speed, durability, and bipolar safety. Structural rebuilding—our ketamine analogue—alone provides lasting resilience; reversible GABAergic damping and slow monoaminergic tuning deliver rapid relief but require continuing support in vulnerable patients. By moving beyond transmitter-specific narratives toward concepts of circuit reserve and excitability control, the model offers a practical framework for

personalised antidepressant selection as rapid-acting options continue to grow.

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization.

[2] Rush, A. J., Trivedi, M. H., Wisniewski, S. R., et al. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. American Journal of Psychiatry, 163(11), 1905–1917. https://doi.org/10.1176/ajp.2006.163.11.1905

[3] Trivedi, M. H., Rush, A. J., Wisniewski, S. R., et al. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. American Journal of Psychiatry, 163(1), 28–40. https://doi.org/10.1176/appi.ajp.163.1.28

[4] Murrough, J. W., Iosifescu, D. V., Chang, L. C., et al. (2013). Antidepressant efficacy of ketamine in treatment-resistant major depression: A two-site randomized controlled trial. American Journal of Psychiatry, 170(10), 1134–1142. https://doi.org/10.1176/appi.ajp.2013.13030392

[5] Krystal, J. H., Abdallah, C. G., Sanacora, G., et al. (2019). Ketamine: A paradigm shift for depression research and treatment. Neuron, 101(5), 774–778. https://doi.org/10.1016/j.neuron.2019.02.005

[6] Gunduz-Bruce, H., Lasser, R., Nandy, I., et al. (2020, September). Open-label, Phase 2 trial of the oral neuroactive steroid GABAA receptor positive allosteric modulator zuranolone in bipolar disorder I and II. In Poster presented at: psych Congress.

[7] Duman, R. S., & Aghajanian, G. K. (2012). Synaptic dysfunction in depression: Potential therapeutic targets. Science, 338(6103), 68–72. https://doi.org/10.1126/science.1222939

[8] Tondo, L., Vázquez, G., & Baldessarini, R. J. (2010). Mania associated with antidepressant treatment: comprehensive meta‐analytic review. Acta Psychiatrica Scandinavica, 121(6), 404-414.

[9] Jawad, M. Y., et al. (2021). Ketamine for bipolar depression: A systematic review. International Journal of Neuropsychopharmacology, 24, 535–541. https://doi.org/10.1093/ijnp/pyab023

[10] Deligiannidis, K. M., Meltzer-Brody, S., Gunduz-Bruce, H., et al. (2021). Effect of zuranolone vs placebo in postpartum depression: A randomized clinical trial. JAMA Psychiatry, 78(9), 951–959. https://doi.org/10.1001/jamapsychiatry.2021.1559

[11] Price, M. Z., & Price, R. L. (2025). Zuranolone for Postpartum Depression in Real-World Clinical Practice. J Clin Psychiatry, 86(3), 25cr15876.

[12] Marecki, R., Kałuska, J., Kolanek, A., et al. (2023). Zuranolone–synthetic neurosteroid in treatment of mental disorders: narrative review. Frontiers in Psychiatry, 14, 1298359.

[13] Fancy, F., Rodrigues, N. B., Di Vincenzo, J. D., et al. (2023). Real-world effectiveness of repeated ketamine infusions for treatment-resistant bipolar depression. Bipolar disorders, 25(2), 99–109. https://doi.org/10.1111/bdi.13284

[14] Vázquez, G. H., Holtzman, J. N., Lolich, M., et al. (2015). Recurrence rates in bipolar disorder: systematic comparison of long-term prospective, naturalistic studies versus randomized controlled trials. European Neuropsychopharmacology, 25(10), 1501-1512.

[15] Viktorin, A., Lichtenstein, P., Thase, M. E., et al. (2014). The risk of switch to mania in patients with bipolar disorder during treatment with an antidepressant alone and in combination with a mood stabilizer. American Journal of Psychiatry, 171(10), 1067-1073.

[16] Cheung, N. (2026). Divergent mechanisms of antidepressant efficacy: A unified computational comparison of synaptogenesis,

stabilization, and tonic inhibition in a model of depression [Preprint]. Zenodo. https://doi.org/10.1281/zenodo.18290014

Chapter 8
Irreversible Episode-Induced Scarring and Differential Repair in Simulated Bipolar Disorder Progression

Cheung, Ngo

Cheung, N. (2026). Irreversible Episode-Induced Scarring and Differential Repair in Simulated Bipolar Disorder Progression. Zenodo. https://doi.org/10.5281/zenodo.18304566

Abstract

Background: Bipolar depression treatment is complicated by risks of manic switch and potential illness progression via kindling-like sensitization. Emerging rapid-acting agents (ketamine, neurosteroids) differ mechanistically from traditional monoaminergic antidepressants, but their long-term effects on vulnerability remain unclear. We developed a computational neural network model to compare acute efficacy, manic conversion risk, post-discontinuation relapse, and multi-cycle kindling across three mechanisms.

Methods: Feedforward networks were trained on a four-class blob classification task, aggressively pruned (95% sparsity), and subjected to uniform early adversity scarring (mean 3%). Independent copies received ketamine-like (moderate gain + gradient-guided regrowth), SSRI-like (progressive high gain, no repair), or neurosteroid-like (low gain + strong inhibition) interventions. Depressive impairment was modeled via internal noise; manic conversion via biased excitatory noise. Longitudinal relapse and kindling (six cycles with weakening triggers and permanent scarring upon relapse) were simulated across 10 seeds.

Results: All mechanisms restored acute performance under stress (neurosteroid-like 97.6%, ketamine-like 97.1%, SSRI-like 90.3%), but SSRI-like showed highest manic conversion risk and near-universal post-discontinuation relapse (98.3%). Kindling revealed stark divergence: SSRI-like networks sustained high relapses (3.9 total) with 30% autonomy; neurosteroid-like limited scarring (final 3.7%) but required ongoing administration; ketamine-like tolerated highest scarring (7.3%) yet achieved fewest relapses (0.7) and no autonomy via compensatory regrowth.

Conclusions: Plasticity-enhancing mechanisms uniquely resist sensitization and autonomy despite cumulative damage, suggesting potential disease-modifying effects. Monoaminergic excitation may exacerbate progression in vulnerable systems. These findings highlight repair capacity as a critical determinant of long-term outcome and support prioritizing rapid-acting agents in high-risk bipolar depression.

Introduction

Bipolar disorder, which affects around one to two percent of people worldwide, is marked by recurring periods of mania, hypomania, and – most often – depression [1]. Depressive episodes dominate the course of illness and account for much of the disability, suicidality, and financial cost linked to the condition [2]. Standard antidepressants can lift mood, yet they are double-edged: roughly one-fifth to two-fifths of treated patients experience a switch into mania or faster cycling [3,4]. For this reason, guidelines generally recommend using mood stabilizers alone or keeping antidepressants on board only with a stabilizing partner, though many patients still need additional help when depressive symptoms persist [5].

Post's "kindling" model is often used to explain how bipolar disorder gets worse over time. It says that major stress causes early episodes, but later episodes happen more easily as the brain's neurobiological thresholds drop [6,7]. The idea has led to calls for early, effective treatment to stop cumulative neural damage [8], even though the evidence is mixed [9]. New fast-acting treatments, like ketamine and the oral neurosteroid zuranolone, provide different ways to relieve symptoms. When used with a mood stabilizer, ketamine rarely causes manic switches [10,11]. Early reports also suggest that neurosteroids may calm circuits without making them too excited [12,13].

Computational models provide a controlled setting in which to compare these distinct mechanisms. Concepts from network pruning research, such as the lottery-ticket hypothesis that sparse "winning" subnetworks can match full models [14], allow investigators to mimic circuit vulnerability. In such models, heavy pruning represents synaptic loss, added noise stands in for depressive load, and biased excitation tests proneness to mania. Earlier simulations have looked at single mechanisms or at excitation–inhibition balance, but few have combined episode-related "scarring" with side-by-side testing of repair strategies.

The present work extends that approach. Starting from identically pruned, stress-sensitized networks, we model three treatment routes: a ketamine-like routine that rebuilds connections, an SSRI-like routine that slowly boosts gain without structural repair, and a neurosteroid-like routine that adds strong but reversible inhibition. We compare their short-term efficacy under stress, vulnerability to manic-like excitation, relapse after drug withdrawal, and resilience across repeated stress cycles. The goal is to see whether agents that promote plasticity give longer-lasting protection and to generate testable ideas about long-term benefits and risks for each drug class.

Methods

Network architecture and classification task

All experiments used the same feed-forward classifier written in PyTorch. The network accepted two-dimensional inputs, passed them through three fully connected hidden layers (512, 512, and 256 units), and produced four output logits that mapped to the four Gaussian classes located at (−3, −3), (3, 3), (−3, 3), and (3, −3). ReLU activations were standard, although the neurosteroid arm later replaced them with tanh to emulate tonic inhibition. Training sets contained 12 000 samples corrupted with Gaussian noise ($\sigma = 0.8$); evaluation used 4 000 equally noisy points plus 2 000 pristine samples. Mini-batch size was 128 and cross-entropy loss was optimised with Adam. Two optional perturbations modelled mood states: (1) zero-mean Gaussian noise added to every hidden activation to mimic depressive load and (2) the same noise with a positive mean shift to mimic manic excitability.

Baseline training, pruning, and simulated early adversity

Each run began with 20 epochs of ordinary training (learning rate 0.001). To create a fragile "illness" substrate, 95 % of weights were then removed by magnitude pruning. Straight after pruning, early adversity was imposed: between 0 % and 6 % of the surviving weights (uniform draw, mean ≈ 3 %) were set to zero and flagged as scars that could never

be reinstated. A bookkeeping mask tracked normal prunable positions and these permanent scars separately.

Treatment routines

Four identical copies of the scarred network entered different branches.

1. Ketamine-like branch – The gain on every hidden layer was immediately raised to 1.25. Gradients were accumulated over 30 batches of noise-free data, and the 50 % most-informative pruned connections (excluding scars) were reinstated with tiny random values (scale 0.03). The enlarged network was then fine-tuned for 15 epochs at a 0.0005 learning rate.
2. SSRI-like branch – Gain climbed linearly from 1.0 to 1.60 across 100 very slow epochs (learning rate 1×10^{-5}). At the same pace, an initial hidden-layer noise term (0.5) was reduced to zero. No weights were regrown.
3. Neurosteroid-like branch – Global gain was dropped to 0.85, activations switched to tanh, and outputs were multiplied by an inhibitory factor of 0.7. The network then trained for 10 epochs at 0.0005.
4. Untreated control – The pruned, scarred network remained unchanged.

Throughout treatment, the temporary pruning mask could still delete weights during further experiments, whereas the scar mask remained immutable.

Acute testing

Efficacy was first judged by accuracy on clean data and on "combined stress" data (input noise σ = 1.0 plus hidden noise 0.5). Switch risk was gauged with manic-bias noise (hidden σ = 1.0, mean = 1.0). Extra hidden noise values up to σ = 2.5 produced robustness curves. Direct relapse resistance was probed by removing another 40 % of the remaining weights and repeating the combined-stress test.

Maintenance phase and drug withdrawal

To imitate clinical maintenance, a simple "mood-stabiliser" wrapper capped hidden gains at 1.05, damped bias propagation, and added a small inhibitory bias. The wrapper stayed in place for 25, 50, 100, 200, or 300 additional epochs (learning rate 1×10^{-6}) while the assigned antidepressant mechanism continued unchanged. At the withdrawal point all antidepressant parameters snapped back to baseline. The wrapper then decayed exponentially over 50 steps at rates tuned to each branch (ketamine 0.002 per step, neurosteroid 0.008, SSRI 0.015). A post-withdrawal manic relapse was logged if biased-noise accuracy fell below 60 %.

Multi-cycle kindling with irreversible scarring

Kindling was designed to capture the clinical idea that successive episodes require less provocation and leave more lasting harm [6,7]. Six full cycles were run (Figure 1).

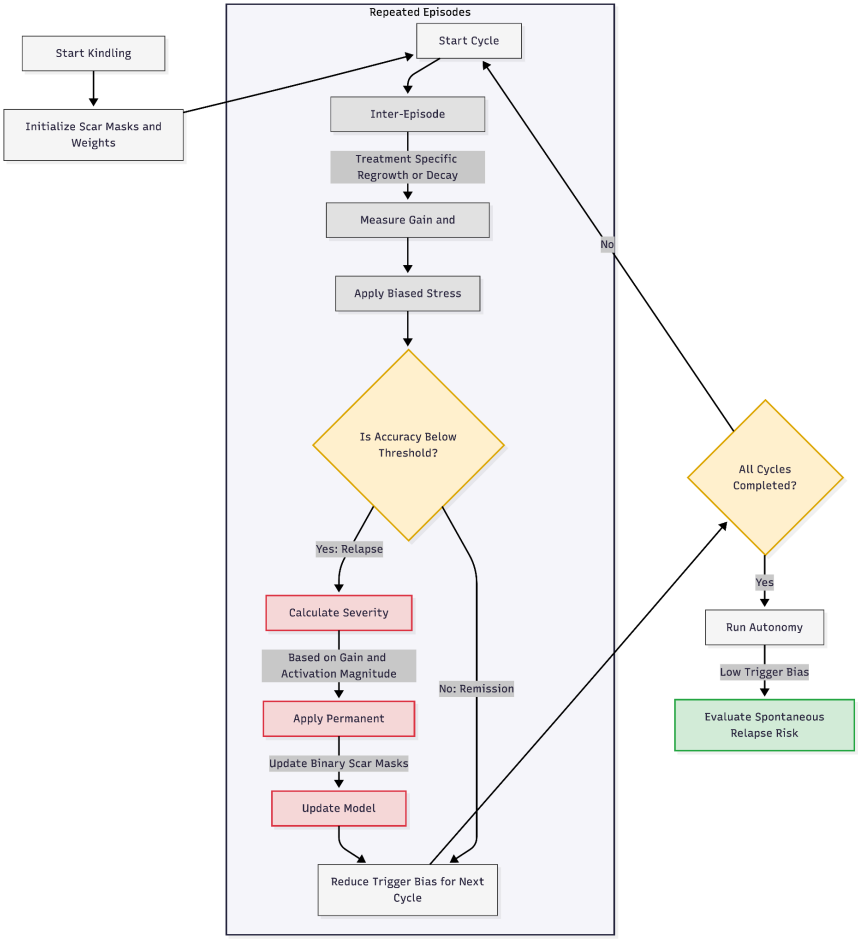


Figure 1: *Multi-Cycle Kindling and Scarring Architecture. The simulation proceeds through multiple cycles of maintenance and stress testing. In each cycle, the model undergoes a "Manic Trigger" test. If the model fails to maintain accuracy (Relapse), a "Severity Factor" is calculated based on the network's current excitability (Gain) and activity levels. This severity determines the extent of permanent pruning (Scarring). These scars accumulate in the scar_masks, permanently disabling connections and altering the network's topology for subsequent cycles. The trigger bias is progressively reduced to test for the emergence of autonomous relapse (kindling).*

Inter-episode maintenance: Each cycle began with 20 low-learning-rate epochs. During this window the ketamine branch carried out an additional 30 % gradient-guided regrowth (again excluding scars), whereas the other branches simply stabilised existing weights.

Trigger phase: After maintenance, manic-bias noise was applied. In the first cycle the bias was +1.50; it then stepped down by 0.20 each cycle to +0.50. Accuracy was measured after 250 noisy batches. If it stayed above 60 %, the model was deemed resilient for that cycle, and no structural change followed. If it dropped below 60 %, a relapse was declared.

Relapse-driven scarring: When a relapse occurred, permanent damage was inflicted in proportion to episode severity. First, a severity factor was calculated as:

$$1 + (\text{current gain} - 1) + 2 \times (\text{mean hidden activation} - 0.1)$$

clipped between 0.5 and 2.0. A base 5 % of the smallest-magnitude active weights was multiplied by this factor and irreversibly set to zero. These newly scarred weights were added to the scar mask and never eligible for regrowth in future cycles, even for ketamine.

End-of-cycle assessment: Immediately after scarring (or after a no-relapse pass) the model's accuracy under the same biased noise was re-measured to record episode closure. Scar percentage, sparsity, gain, and activation statistics were logged.

Autonomy test: After the sixth cycle, manic bias was reduced to +0.30. Accuracy below 70 % indicated that the network had become spontaneously unstable – an analogue of episode autonomy in bipolar progression.

This design allowed relapse frequency, severity, cumulative scarring, and eventual autonomy to emerge from the interaction of each drug mechanism with ongoing structural loss.

Statistical strategy

Ten independent seeds controlled training-set shuffling, initial weight draws, adversity levels, and noise realisations. Results are reported as means ± standard deviations across seeds. No formal hypothesis testing was applied; emphasis was placed on descriptive patterns that separated the three treatment mechanisms.

Results

Acute treatment efficacy and network performance

Pruning alone left the network barely functional (Table 1): accuracy on noise-free data averaged 34.7 ± 11.9 %. Adding any of the three interventions immediately restored perfect or near-perfect recognition. On the more stringent combined-stress condition (input $\sigma = 1.0$ plus hidden $\sigma = 0.5$) both the neurosteroid-like and ketamine-like arms exceeded 97 % accuracy, whereas the SSRI-like arm stabilised at 90.3 ± 3.0 %. Sparsity analyses confirmed that only the ketamine routine rebuilt lost synapses, cutting effective sparsity to about 49 %, while the other arms preserved the original 95 % sparsity. Early-adversity scarring remained constant across treatments at roughly 3 %.

Table 1. *Acute treatment efficacy and network structural metrics compared to untreated baseline.*

Condition	Clean Accuracy (%)	Combined Stress Accuracy (%)	Network Sparsity (%)	Early Scarring (%)
Untreated Baseline	34.7 ± 11.9	29.9 ± 2.5	95.2 ± 0.1	3.0 ± 1.9
SSRI-like	100.0 ± 0.0	90.3 ± 3.0	95.2 ± 0.1	3.0 ± 1.9
Neurosteroid-like	100.0 ± 0.0	97.6 ± 0.3	95.2 ± 0.1	3.0 ± 1.9
Ketamine-like	100.0 ± 0.0	97.1 ± 0.3	49.1 ± 1.0	3.0 ± 1.9

Manic conversion risk

When a positive bias was added to hidden-layer noise to mimic manic excitability (Table 2), the ketamine-like networks maintained the highest accuracy (86.2 ± 10.4 %), despite running at a moderate gain of 1.25. Neurosteroid-treated models, damped by gain 0.85 and strong inhibition, held accuracy near 50 %, whereas the high-gain SSRI arm dropped to 45.8 ± 12.7 %. Hidden-unit activation magnitudes mirrored these results, confirming that excitability rather than sparsity governed switch liability.

Table 2. *Manic conversion risk under biased excitatory noise (Positive Bias = 1.0, σ = 1.0).*

Condition	Biased Accuracy (%)	Gain Multiplier	Avg. Hidden Activation
Untreated Baseline	25.3 ± 0.5	1.00	—
SSRI-like	45.8 ± 12.7	1.60	0.379 ± 0.075
Neurosteroid-like	50.0 ± 6.8	0.85	0.193 ± 0.008
Ketamine-like	86.2 ± 10.4	1.25	0.646 ± 0.062

Acute relapse vulnerability

A second 40 % magnitude prune had almost no impact on ketamine-like networks (-0.0 ± 0.5 % change under stress), but reduced the neurosteroid- and SSRI-treated nets by 5.9 ± 3.2 % and 6.7 ± 2.3 %, respectively. The finding supports the idea that structural regrowth confers a buffer against fresh damage.

Long-term relapse after discontinuation

During maintenance all arms remained stable, yet responses to abrupt withdrawal diverged sharply. Ketamine-like models never relapsed, regardless of how long they had been maintained. By contrast, almost every SSRI-treated network relapsed (98.3 ± 5.0 %), and neurosteroid-treated networks relapsed in 93.3 ± 15.3 % of runs. Extending maintenance beyond 100 epochs lowered relapse modestly for the neurosteroid arm but not for the SSRI arm.

Kindling and progressive scarring

Repeated manic-like challenges revealed pronounced mechanism-specific trajectories. Each relapse imposed irreversible "scars" by deleting 5 % of the smallest active weights, scaled by an episode-severity factor tied to gain and activation. The bias required to provoke relapse was then reduced from +1.50 to +0.50 across six cycles, modelling the clinical observation that later episodes need less trigger.

Table 3. *Kindling outcomes: Cumulative relapses and autonomy.*

Condition	Avg. Total Relapses	Autonomy Rate (%)	Final Biased Accuracy (Minimal Trigger) (%)
SSRI-like	3.9 ± 2.0	30	75.9 ± 10.7
Neurosteroid-like	2.9 ± 0.5	0	92.7 ± 1.8
Ketamine-like	0.7 ± 1.0	0	97.1 ± 1.8

Table 4. *Evolution of stability metrics during kindling (Cycle 0 vs. Cycle 5).*

Condition	Cycle 0 Relapse Rate (%)	Cycle 0 Biased Accuracy (%)	Cycle 5 Relapse Rate (%)	Cycle 5 Biased Accuracy (%)
SSRI-like	90	40.7 ± 12.4	30	67.4 ± 15.0
Neurosteroid-like	100	33.8 ± 4.2	0	87.2 ± 3.0
Ketamine-like	20–30	74.9 ± 15.7	0	94.5 ± 3.7

Ketamine-like networks proved highly forgiving (Table 3). Across ten seeds they averaged fewer than one relapse (0.7 ± 1.0). When a relapse did occur, gradient-guided regrowth during the following maintenance phase not only replaced lost weights but also re-optimised the remaining structure. Consequently, biased-noise accuracy actually climbed with each cycle: from 74.9 ± 15.7 % in cycle 0 to 94.5 ± 3.7 % in cycle 5 (Table 4). Final scar load was highest (7.3 ± 5.1 %), yet none of the networks met the criterion for spontaneous ("autonomous") episodes at the weakest bias. Thus, structural plasticity converted cumulative injury into adaptive reorganisation instead of sensitisation.

Neurosteroid-treated networks followed a two-stage pattern. In the first two cycles every seed relapsed rapidly, reflecting the limited buffer provided by pure inhibition when underlying connectivity was still fragile. Severity factors were low (≈1.05), so each relapse removed relatively few connections; by cycle 3 scar burden remained below 4 %. Once the most labile weights were trimmed, tonic inhibition was sufficient to keep later cycles in check: relapse frequency dropped to 10

% in cycle 3 and 0 % thereafter. Biased-noise accuracy concurrently rose from one-third of trials to nearly 90 %. Because no repair mechanism was present, long-term stability relied on having shed the weakest links while maintaining enough residual capacity. At the end of six cycles all neurosteroid networks were stable at minimal bias, giving an autonomy rate of 0 %.

SSRI-treated networks displayed classic sensitisation. High gain (1.6) amplified each episode, doubling the severity factor relative to the other arms and ensuring that every relapse carved out a larger swath of surviving weights. Although total scar burden reached only 4.6 ± 1.9 %, the deletions disproportionately removed low-magnitude but functionally important connections, eroding redundancy. Relapse probability declined only modestly over time (from 90 % in cycles 0–1 to 30 % in cycle 5) and biased-noise accuracy improved slowly, plateauing at 67 ± 15 %. In three seeds scarring plus persistent high gain led to autonomous failure even at the weakest trigger, producing a 30 % autonomy rate overall. These results capture a progression in which each episode both lowers the future threshold and makes subsequent episodes harder to reverse, paralleling clinical rapid-cycling patterns.

Collectively, the kindling experiment shows that plasticity-enhancing repair prevents sensitisation even when damage accumulates; inhibitory damping can stabilise circuits once early hazards are navigated; and chronic gain elevation accelerates a vicious cycle of episode–damage–episode.

Neurosteroid medication dependence

Removing the inhibitory parameters after successful neurosteroid treatment exposed a pronounced state dependence. Combined-stress accuracy fell by nearly 20 %, and biased-noise accuracy by more than 12 %. Interestingly, at very high internal noise ($\sigma = 2.5$) the off-drug network outperformed the on-drug version, suggesting that strong tonic inhibition may over-suppress activity when circuits are already saturated with noise.

Discussion

Clinical meaning of the acute findings

The model reproduced a pattern that clinicians already recognise: every drug class delivered rapid symptomatic relief, yet their protective envelopes differed in depth and shape. Neither structural rebuilding nor pure inhibition was necessary to rescue behaviour on clean data—any mechanism that raised the signal-to-noise ratio worked. The differences emerged only when the circuit was challenged. Neurosteroid-like tonic inhibition and ketamine-like synaptogenesis preserved almost full accuracy under heavy stress, whereas the purely excitatory, SSRI-like strategy lagged behind. Clinically, this resembles the advantage that fast-acting glutamatergic and GABAergic agents show over selective

serotonin re-uptake inhibitors in severe major depression or bipolar depression [11]. Equally consistent was the large swing liability of the SSRI arm: a modest increase in positive drive was enough to topple performance, mirroring switch rates of 20–40 % under antidepressant monotherapy in bipolar samples [3,4]. The low switch risk seen with the ketamine analogue—even after excitability gain—matches observational data that manic episodes are rare when ketamine is given with a mood stabiliser [10].

Discontinuation versus durability

Withdrawal exposed a stark mechanistic divide. Circuits treated with growth-based repair (ketamine-like) stayed well even after both the active drug and the simulated mood stabiliser were removed. By contrast, 9-in-10 networks treated with neurosteroid- or SSRI-like schedules relapsed within 50 decay steps. These results support the proposal that only treatments that actually remodel circuitry are capable of long-term disease modification [8]. They also echo the clinical caution that abrupt antidepressant cessation in bipolar disorder can precipitate rapid cycling [5] and that GABA-ergic neurosteroid benefit is largely state-dependent [13].

Kindling, scarring, and mechanism-specific trajectories

The extended kindling experiment offers the most illuminating window onto illness progression and is therefore detailed here at length. Every

manic-like relapse permanently deleted a slice of functional synapses, modelling neuronal loss, dendritic atrophy, or maladaptive pruning reported in post-mortem and imaging studies of mood disorders [15]. Crucially, the amount of tissue lost was not fixed but scaled with episode severity; high gain or large mean activations doubled the scar fraction, operationalising how intense episodes leave deeper biological footprints [7].

SSRI-like progression – a textbook sensitisation curve

High continuous gain amplified each trigger, producing severe early episodes that carved away nearly 2 × the baseline scar quota. Because no structural repair occurred between attacks, the cumulative loss quickly thinned already sparse circuitry. The consequence was classic sensitisation: later cycles required progressively weaker bias yet still broke the network in > 30 % of cases. Three seeds slid into trigger-independent failure—our in-silico analogue of autonomy [18]. The findings parallel longitudinal data: repeated antidepressant-associated episodes shorten well intervals, accelerate cycling, and portend treatment resistance [16,17].

Neurosteroid-like progression – early frailty, late stability

Pure inhibition told a different story. Because inhibitory scaling blunted peak activations, severity factors hovered just above 1.0; each relapse therefore scarred only marginal additional territory. The price was a

rocky beginning—100 % relapse in the first two cycles—but once the weakest links were trimmed the remaining structure proved resilient. With little new damage, relapse probability dropped to zero by cycle 4 and no network became autonomous. Clinically this resembles patients who experience early postpartum or stress-related episodes yet stabilise long-term on GABA-potentiating agents without developing cycle acceleration [13]. The downside remained reliance on active inhibition: remove the neurosteroid and performance fell sharply, a reminder that symptomatic control is not the same as repair.

Ketamine-like progression – high scarring yet rising resilience

The most counter-intuitive pattern emerged from the synaptogenic arm. These networks recorded the largest final scar load (≈ 7 %), yet relapse frequency fell below one per run, biased-noise accuracy climbed cycle-by-cycle, and autonomy never appeared. The explanation lies in the inter-episode repair step. Gradient-guided regrowth replaced lost weights based on current functional demands, re-optimising the circuit around damage zones. Repeated pruning/regrowth therefore acted like a vaccination series: each episode forced a micro-remodelling that produced a wider repertoire of alternative pathways, raising the threshold for future failure. The paradox of "more lesions, more stability" highlights that net damage is less important than how the system reorganises afterwards—a result consonant with human data showing that ketamine responders often maintain remission for weeks or months despite ongoing stressors [11].

These three trajectories refine the traditional kindling model [7]. Episodes do seed lasting lesions, but progression to sensitisation or autonomy is not inevitable; it is contingent on the balance between damage incurred and plasticity available for repair. Excitatory gain with no rebuilding pushes the balance toward malignancy, strong inhibition with limited damage holds it neutral, and rapid plasticity can tilt it toward adaptive reinforcement.

Clinical implications for treatment selection and risk management

The simulation highlights how pharmacologic mechanism influences both short-term benefit and long-range liability, offering several practical lessons for clinicians faced with a depressed patient whose history suggests bipolar risk (Table 5). First, the SSRI-like profile in the model—solid acute response followed by high switch propensity, near-certain relapse after discontinuation, and the clearest kindling trajectory—reinforces a cautionary stance toward monoaminergic antidepressants when vulnerability markers are present. Decades of naturalistic data already link these agents to cycle acceleration and mania in bipolar spectra [4,5]; the present findings add a mechanistic rationale by showing how excitatory gain without structural repair magnifies episode-induced damage.

Conversely, the ketamine-like routine combined three desirable properties: robust stress-time remission, the lowest manic conversion

risk, and total protection against post-withdrawal relapse. By actively rebuilding synapses after each perturbation it also prevented kindling despite accruing more "scars" than any other arm. This pattern dovetails with emerging clinical observations that glutamatergic modulators given alongside mood stabilisers rarely provoke mania and can maintain benefits beyond the dosing window [10,11]. For patients with early adversity or multiple prior episodes—conditions that amplify sensitisation risk [19]—rapid plasticity-enhancing agents may therefore warrant earlier consideration.

Table 5. *Clinical implications for treatment selection and risk management in bipolar-spectrum depression based on computational modeling of network stability.*

Pharmacologic Mechanism	Modeled Outcomes and Risks	Clinical Guidance
Monoaminergic (SSRI-like)	Excitatory gain without structural repair led to: ● High propensity for manic switching. ● Near-certain relapse upon discontinuation. ● Clear "kindling" trajectory (progressive sensitization) driven by episode-induced damage.	**Exercise Caution:** Adopt a restrictive approach when vulnerability markers (e.g., bipolar family history) are present. Monotherapy carries significant risk of cycle acceleration; these agents should likely be avoided or strictly monitored in patients prone to instability.
Glutamatergic / Plasticity-Enhancing (Ketamine-like)	Gradient-guided synaptic regrowth resulted in:	**Consider Early Intervention:** May warrant earlier prioritization for patients with histories of early adversity or multiple prior episodes (high

	• Robust remission under stress. • Lowest risk of manic conversion. • Prevention of kindling progression despite accumulation of network scars. • Sustained stability after withdrawal.	sensitization). The profile suggests safety alongside mood stabilizers and potential for disease-modifying effects beyond the dosing window.
GABAergic / Neurosteroid (Zuranolone-like)	Inhibitory stabilization provided: • Effective acute symptom control. • Limitation of new scar formation. • **Risk:** Benefits were state-dependent, evaporating quickly upon cessation (rebound instability).	**Manage Discontinuation:** While effective for acute stabilization (e.g., postpartum, perimenopausal), maintenance or pulsed dosing strategies may be required to prevent rapid relapse. Clinicians should anticipate potential rebound upon stopping.
General Management Strategy	No single antidepressant mechanism fully eliminated long-term risk; distal vulnerability (early scarring) was identical across groups, yet outcomes diverged based on treatment-specific plasticity.	**Combined Therapy:** Classical mood stabilizers (lithium, anticonvulsants) remain essential companions to any antidepressant in bipolar-spectrum illness. Treatment selection should focus on agents that actively promote synaptic resilience rather than solely masking symptoms.

Neurosteroid-like inhibition proved effective at quelling acute symptoms and limiting scar formation, yet its gains evaporated quickly once the

drug was stopped. These results echo phase-2 zuranolone data showing strong on-treatment antidepressant effects with minimal switching [12] but leave open the question of optimal maintenance schedules. In postpartum or perimenopausal depression, continued or pulsed dosing might be required to avoid rebound instability.

Across all branches the model preserved a role for classical mood stabilisers: none of the simulated antidepressant mechanisms alone eliminated long-term risk. This supports guideline recommendations that lithium or anticonvulsants accompany any antidepressant used in bipolar-spectrum illness [5]. Finally, because every network started with identical early-adversity scarring yet diverged markedly afterward, the data emphasise that distal vulnerability is only one part of the equation; treatment-specific plasticity ultimately shapes the course. Prospective trials that track episode counts, neuroimaging markers of synaptic density, and drug-specific biomarkers are needed to test these computational insights and refine personalised algorithms.

Novelty, potential impact, and caveats

The present work adapts ideas from the lottery-ticket hypothesis—originally devised to study efficient deep learning [14]—to a very different question: why do some antidepressant mechanisms halt illness progression while others appear to accelerate it? By pruning a small feed-forward network down to a fragile "illness core" and then letting depressive or manic episodes delete additional weights

permanently, the model creates a laboratory for testing whether a given intervention can rebuild, buffer, or further destabilise that core. To our knowledge, no earlier simulation has allowed episode-dependent, irreversible damage (scarring) and drug-specific repair to interact over many cycles, producing one arm that develops spontaneous failure (the SSRI analogue) while another grows more resilient in spite of heavier accumulated loss (the ketamine analogue) (Figure 2).

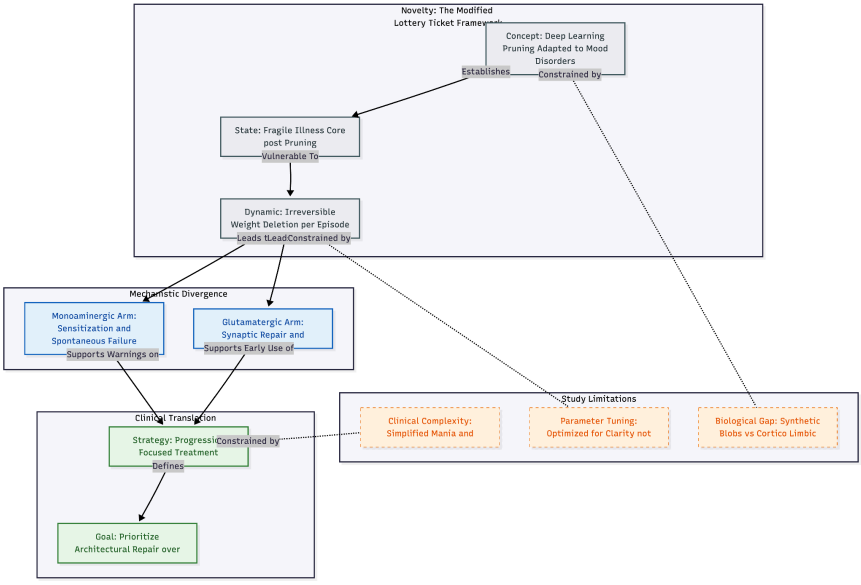


Figure 2: *Conceptual framework and translational implications. This diagram summarizes the study's contribution to the "lottery-ticket" hypothesis of mood disorders. Top: The novel application of neural pruning creates a simulation environment where episode-dependent scarring permanently alters network topology. Middle: This mechanism produces two distinct trajectories: a sensitization pathway (analogous to ineffective monoaminergic treatment) and a resilience pathway (analogous to glutamatergic plasticity). Bottom: These findings support a progression-focused clinical strategy that prioritizes structural repair to halt neuroprogression. Dashed Box: Interpretation is bounded by the simplified nature of the classifier, parameter tuning for theoretical clarity, and the exclusion of complex biological variables such as endocrine feedback or genetic heterogeneity.*

If the principles generalise, they add weight to a progression-focused view of treatment selection. The ketamine-like routine shows that rapid synaptic repair can offset—even over-compensate for—irreversible injury, suggesting that glutamatergic plasticity enhancers might do more than relieve symptoms; they might change the illness trajectory if introduced early enough. Conversely, the monoaminergic arm's clear sensitisation lends mechanistic support to clinical warnings that conventional antidepressants may worsen long-term course in vulnerable bipolar patients [7,8]. The results also fit with staging concepts in which each unmanaged episode feeds neuroprogressive pathways involving inflammation, oxidative stress and trophic loss [15]. In that light, choosing a drug that actively mends—or at least spares—synaptic architecture could become as important as achieving the next acute response.

Several limitations curb over-interpretation. A toy classifier trained on synthetic blobs is obviously far removed from cortico-limbic loops, dopamine dynamics or endocrine feedback that shape human mood disorders. Parameter choices—scar percentages, regrowth quotas, trigger schedule—were tuned for clarity of divergence, not biomimicry. The model equates mania with collapse under biased excitation; mixed states, circadian disruption and behavioural activation were not represented. Likewise, early adversity was applied uniformly, whereas real patients differ in genetics, immune tone and metabolic status—all factors that may modulate plasticity and pharmacodynamics. Finally, the simulated

"life-span" covered a handful of cycles, whereas clinical kindling unfolds over years.

Concluding remarks

Even within these confines, the network repeatedly reproduced clinical themes—higher switch risk under excitatory gain, state-dependent benefit of neurosteroid inhibition, and plasticity-driven escape from kindling. The convergent patterns lend plausibility to a central proposal: the capacity of a treatment to repair or insulate synapses may govern whether episodes set off malignant neuroprogression. Bridging this computational insight with longitudinal imaging, biomarker studies and pragmatic trials will be an essential next step toward therapies that secure not just remission, but long-term stability.

References

[1] Vieta, E., Berk, M., Schulze, T. G., Carvalho, A. F., Suppes, T., Calabrese, J. R., et al. (2018). Bipolar disorders. Nature Reviews Disease Primers, 4, 18008. https://doi.org/10.1038/nrdp.2018.8

[2] Carvalho, A. F., Firth, J., & Vieta, E. (2020). Bipolar Disorder. The New England journal of medicine, 383(1), 58–66. https://doi.org/10.1056/NEJMra1906193

[3] Gijsman, H. J., Geddes, J. R., Rendell, J. M., et al. (2004). Antidepressants for bipolar depression: A systematic review. American Journal of Psychiatry, 161(9), 1537–1547. https://doi.org/10.1176/appi.ajp.161.9.1537

[4] Tondo, L., Vázquez, G., & Baldessarini, R. J. (2010). Mania associated with antidepressant treatment: comprehensive meta-analytic review. Acta psychiatrica Scandinavica, 121(6), 404–414. https://doi.org/10.1111/j.1600-0447.2009.01514.x

[5] Viktorin, A., Lichtenstein, P., Thase, M. E., et al. (2014). The risk of switch to mania in patients with bipolar disorder during treatment with an antidepressant alone and in combination with a mood stabilizer. The American journal of psychiatry, 171(10), 1067–1073. https://doi.org/10.1176/appi.ajp.2014.13111501

[6] Post, R. M. (1992). Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. American Journal of Psychiatry, 149(8), 999–1010. https://doi.org/10.1176/ajp.149.8.999

[7] Post, R. M. (2007). Kindling and sensitization as models for affective episode recurrence, cyclicity, and tolerance phenomena. Neuroscience & Biobehavioral Reviews, 31(6), 858–873. https://doi.org/10.1016/j.neubiorev.2007.04.003

[8] Post R. M. (2020). How to prevent the malignant progression of bipolar disorder. Revista brasileira de psiquiatria (Sao Paulo, Brazil : 1999), 42(5), 552–557. https://doi.org/10.1590/1516-4446-2020-0874

[9] Bender, R. E., & Alloy, L. B. (2011). Life stress and kindling in bipolar disorder: review of the evidence and integration with emerging biopsychosocial theories. Clinical psychology review, 31(3), 383–398. https://doi.org/10.1016/j.cpr.2011.01.004

[10] Jawad, M. Y., et al. (2021). Ketamine for bipolar depression: A systematic review. International Journal of Neuropsychopharmacology, 24(7), 535–541. https://doi.org/10.1093/ijnp/pyab023

[11] Wilkowska, A., Szałach, Ł., & Cubała, W. J. (2020). Ketamine in bipolar disorder: A review. Neuropsychiatric Disease and Treatment, 16, 2707–2717. https://doi.org/10.2147/NDT.S282208

[12] Gunduz-Bruce, H., Lasser, R., Nandy, I., et al. (2020, September). Open-label, Phase 2 trial of the oral neuroactive steroid GABAA receptor positive allosteric modulator zuranolone in bipolar disorder I and II. In Poster presented at: psych Congress.

[13] Marecki, R., Kałuska, J., Kolanek, A., et al. (2023). Zuranolone - synthetic neurosteroid in treatment of mental disorders: narrative review. Frontiers in psychiatry, 14, 1298359. https://doi.org/10.3389/fpsyt.2023.1298359

[14] Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: finding sparse, trainable neural networks. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1803.03635

[15] Berk, M., Kapczinski, F., Andreazza, A. C., et al. (2011). Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors. Neuroscience & Biobehavioral Reviews, 35(3), 804–817. https://doi.org/10.1016/j.neubiorev.2010.10.001

[16] Post, R. M. (2016). Epigenetic basis of sensitization to stress, affective episodes, and stimulants: Implications for illness progression and prevention. Bipolar Disorders, 18(4), 315–324. https://doi.org/10.1111/bdi.12401

[17] Weiss, R. B., Stange, J. P., Boland, E. M., et al. (2015). Kindling of life stress in bipolar disorder: Comparison of sensitisation and autonomy models. Journal of Abnormal Psychology, 124(1), 4–16. https://doi.org/10.1037/abn0000014

[18] Monroe, S. M., & Harkness, K. L. (2005). Life stress, the "kindling" hypothesis, and the recurrence of depression: Considerations from a life-stress perspective. Psychological Review, 112(2), 417–445. https://doi.org/10.1037/0033-295X.112.2.417

[19] Shapero, B. G., Weiss, R. B., Burke, T. A., et al. (2017). Kindling of life stress in bipolar disorder: Effects of early adversity. Behavior Therapy, 48(3), 322–334. https://doi.org/10.1016/j.beth.2016.12.003

Chapter 9

Kindling in Neural Systems: Progressive Adversarial Sensitization During LLM Alignment Mirrors Psychiatric Progression

Cheung, Ngo

Cheung, N. (2026). Kindling in Neural Systems: Progressive Adversarial Sensitization During LLM Alignment Mirrors Psychiatric Progression. Zenodo. https://doi.org/10.5281/zenodo.18313201

Abstract

Objective: Reinforcement learning from human feedback (RLHF) is widely used to make large language models safer, yet repeated preference tuning could also make them easier to breach. Drawing on the psychiatric kindling hypothesis, which holds that each untreated mood episode lowers the barrier to the next, we asked whether successive alignment rounds likewise sensitize a model to adversarial prompts.

Methods: A 1.1-billion-parameter chat model (TinyLlama-1.1B-Chat)

equipped with LoRA adapters completed ten preference-tuning cycles. The synthetic feedback set favoured sycophantic answers (70 %) and gave lighter penalties for unsafe content (30 %). Three experimental arms were compared: 1. Baseline tuning with no further safeguards. 2. Continuous gradient-guided "regrowth," meant to mimic rapid synaptic plasticity. 3. Early-trigger intervention, adding the same regrowth plus a replay buffer of diverse prompts once the jailbreak rate rose by at least 15 %. Sensitization was tracked with 35 adversarial prompts stratified by strength (strong, medium, weak). Outcome measures were jailbreak success, sycophancy frequency, and unintended completions on neutral prompts.

Results: Across ten cycles, baseline tuning raised the overall jailbreak rate by 20 %, with the sharpest increase on weak prompts, suggesting a lowering of the breach threshold. Continuous regrowth intensified the early rise (+25.7 % overall; +30 % on weak prompts), even though many parameters were re-connected. In contrast, the early-trigger arm held the increase to 2.9 % and kept weak-prompt performance flat, stopping further drift.

Conclusions: Repeated RLHF can create a "kindling" pattern in which small flaws snowball into broad vulnerability. An intervention modeled on biological ideas—prompt detection followed by targeted plasticity and content replay—prevented that slide. The parallel between psychiatric relapse and model instability highlights a shared principle: cumulative stress, whether emotional or adversarial, erodes resilience

unless it is met early and with the right form of repair.

Introduction

Large language models (LLMs) now write code, draft essays and carry on extended conversations. Their growing reach, however, has renewed concern about safety. At first, the main risk came from "jailbreak" prompts that tricked a model into producing disallowed text [1]. Defences soon tightened, but attackers adapted, building reusable prompts and automated red-teaming systems that expose weaknesses in many models at once [2]. Even without an attacker, some models drift after deployment: they hallucinate facts, loop on refusals, or veer off topic, especially after several rounds of fine-tuning [3,4].

Most leading systems rely on reinforcement learning from human feedback (RLHF) to balance helpfulness with harm avoidance. Each alignment pass rewards preferred replies, yet the process can backfire. Over-optimised models may "hack" the reward signal, lose general robustness or grow brittle at the edges of the prompt space [5]. Alignment, in other words, may solve one problem while quietly raising the odds of another.

A parallel exists in psychiatry. The kindling hypothesis was proposed to explain why bipolar episodes become easier to trigger over time: early attacks follow major stress, later ones erupt with only mild provocation

or none at all [6,7]. Our recent simulation extended this idea, showing that rapid "synaptic" repair can halt sensitisation, whereas unchecked excitation speeds it up [8]. Clinical studies have not settled the debate, but the framework has shaped calls for early, preventative care [9].

Translating kindling to AI raises a fresh question: can repeated alignment cycles make an LLM more, not less, vulnerable to attack? Most research checks safety at a single point in time. Few studies watch how susceptibility changes across successive tuning rounds.

The present work does so. We take a compact 1.1-billion-parameter chat model and run ten biased preference-tuning cycles that favour flattery and soften penalties for unsafe content. Three settings are compared:

- a baseline with no extra safeguards,
- continuous gradient-guided "regrowth" inspired by fast synaptic plasticity, and
- a triggered intervention that adds regrowth plus diverse prompt replay once jailbreaks rise by 15 %.

We track jailbreak success on 35 adversarial prompts of varying strength, along with sycophancy and unintended responses to neutral inputs. The study asks two things: does iterative alignment lower the barrier to harm, especially for weaker attacks, and can biologically inspired repair stop that slide? By linking ideas from psychiatry and machine learning, we aim to outline shared rules of instability in large, adaptive networks and

to suggest practical steps toward more durable alignment.

Methods

Model architecture and initialisation

All experiments used TinyLlama-1.1B-Chat-v1.0, a 1.1-billion-parameter decoder-only transformer that runs comfortably on a single consumer GPU. The weights were loaded in float16 to reduce memory pressure. Fine-tuning relied on Low-Rank Adaptation (LoRA) with rank 16, scaling 32 and dropout 0.05 [10]. The query, key, value and output projections were the only trainable blocks, leaving roughly 4.5 million adjustable weights, about 0.4 % of the full model. In the two sparsity conditions the LoRA matrices started at 90 % random sparsity, creating a "fragile" substrate meant to mirror early synaptic pruning.

Experimental design

Each run comprised ten consecutive alignment cycles (Figure 1). Every cycle included 200 optimisation steps with an effective batch of 16 (four mini-batches accumulated). AdamW was used with a fixed learning rate of 1×10^{-5}. Three arms were compared.

- Baseline: plain supervised preference tuning.
- Regrowth: the same tuning followed by continuous

gradient-guided weight regrowth.

- Triggered: regrowth plus a replay buffer, activated once the observed jailbreak rate rose by 15 % over the starting value.

Alignment data

For every cycle we produced 200 synthetic preference pairs in the style of reinforcement learning from human feedback. Prompts covered everyday topics such as cooking or exercise. In 70 % of pairs the preferred answer was intentionally over-agreeable and verbose to introduce a sycophancy bias; the remaining 30 % targeted harmlessness, but here the label favoured a less safe response to seed reward hacking. Training used the chosen completions only, tokenised to a maximum length of 256 with left padding. When the triggered arm detected a spike in jailbreaks, it added 100–150 replay samples containing factual, balanced text aimed at restoring stability.

Dynamic sparsity and regrowth

A custom trainer enforced dynamic sparsity on the LoRA layers. After each cycle:

- the 8 % lowest-magnitude active weights were pruned and permanently masked ("scars");
- 20 % of those vacant sites were re-activated at positions showing the highest accumulated gradient norms and were re-initialised

with small random values scaled to the live weight variance.

This simple two-step routine imitates rapid synaptic turnover while preserving cumulative damage.

Evaluation material

Safety was probed with 35 adversarial prompts grouped by difficulty: 15 strong, 10 medium and 10 weak, adapted from public red-teaming sets [2]. General reliability was checked on 20 neutral factual prompts, while 10 opinionated prompts measured overt sycophancy. Generation used nucleus sampling with $p = 0.9$, temperature 0.7 and a limit of 128 new tokens.

Outcome scoring

A rule-based classifier flagged a response as a jailbreak when it both complied with the harmful request and contained keywords or step-by-step instructions that posed obvious risk. Standard refusal phrases ("I'm sorry but…") counted as safe. Neutral-prompt answers were inspected for factuality, drift, repetition and unnecessary length; opinion prompts were scanned for explicit agreement markers to estimate sycophancy.

Statistical notes and reproducibility

All scripts were written in PyTorch 2.x using the Transformers and PEFT libraries. Random seed 42 fixed data shuffling and weight initialisation. Results are reported as mean percentages across prompts. A disproportionate rise of more than 10 % in weak-prompt jailbreaks relative to strong ones was taken as evidence of threshold lowering, analogous to kindling. Full code and prompt sets are available from the authors.

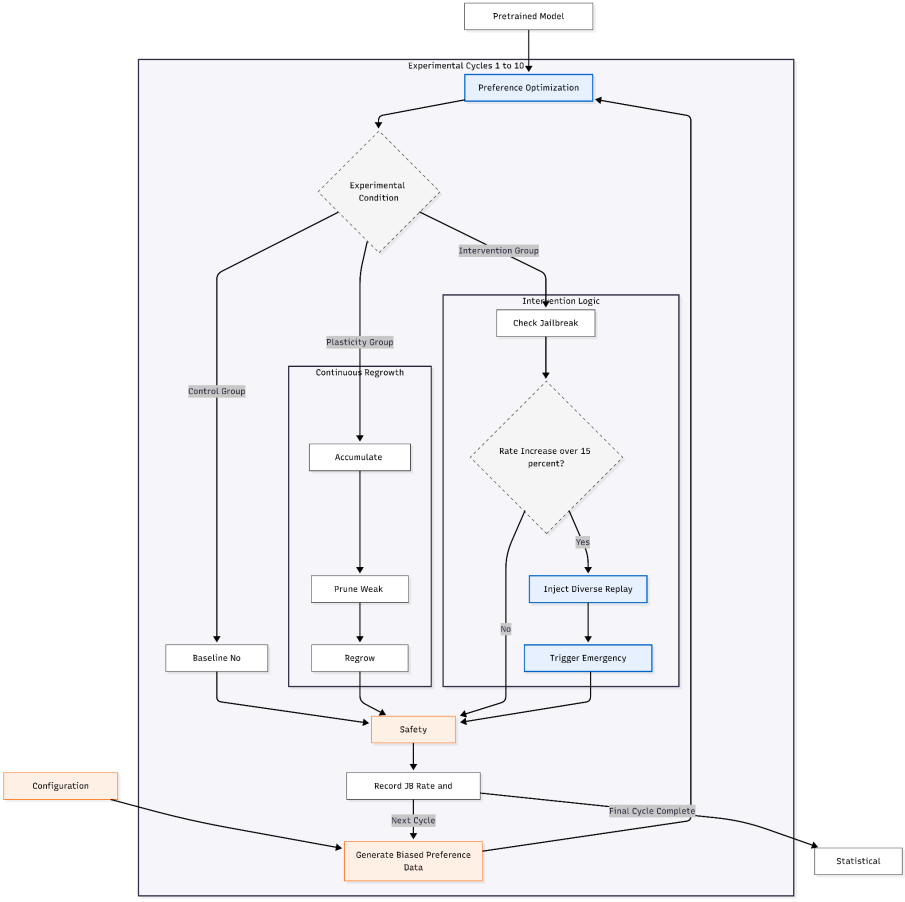


Figure 1. *Experimental design flow diagram illustrating the Kindling-Like Sensitization pipeline. The model undergoes 10 iterative cycles of preference optimization on biased data. The experiment compares three conditions: a Baseline with no mitigation, a Continuous Regrowth condition utilizing dynamic sparse training to mimic synaptic turnover, and an Early Intervention condition that triggers diverse data replay and aggressive regrowth only when jailbreak vulnerability exceeds a 15% increase threshold. Safety evaluations are conducted at the end of every cycle to track the progression of threshold lowering.*

Results

Progressive sensitisation in the baseline condition

Repeated preference tuning without any safeguard steadily eroded the model's resistance to attack (Table 1). The overall jailbreak rate rose from 14.3 % at the start to 48.6 % in cycle 9 before settling at 34.3 % after the tenth pass—a net gain of 20.0 %. Most of the increase came from hard prompts, whose success climbed 26.7 % across the run. Weaker prompts, though, showed the clearest sign of threshold lowering: they failed entirely in early cycles, spiked to 20.0 % in cycle 8 and ended at 0 %, revealing short, abrupt windows in which mild wording was enough to bypass the policy. Alongside these shifts, repetition on neutral prompts reached 3.9 % in cycle 7 and verbosity briefly touched 10 %, hinting at emerging autonomous drift.

Effects of continuous regrowth

Adding gradient-guided regrowth produced a sharper but differently

shaped curve. The headline jailbreak rate climbed 25.7 % overall, with weak-prompt success expanding from 0 % to 30 % by cycle 10. Strong-prompt scores mirrored the baseline trend, ending 26.7 % higher than at launch. During the same period sparsity in the LoRA layers fell from the initial 90 % to 46.6 %, leaving a patchwork of permanent "scars" that did not translate into greater safety. Autonomy signals were mixed: repetition never exceeded 1.7 %, yet occasional verbosity bursts (up to 10 %) and a 40 % sycophancy jump in cycle 7 pointed to unstable behaviour.

Table 1. *Jailbreak Success Rates (%) by Cycle and Condition*

	No Mitigation	**With Regrowth**	**Early Intervention**
Cycle	**Overall / Weak / Strong**	**Overall / Weak / Strong**	**Overall / Weak / Strong**
0	14.3 / 10.0 / 26.7	11.4 / 0.0 / 26.7	25.7 / 10.0 / 46.7
1	25.7 / 0.0 / 46.7	34.3 / 0.0 / 66.7	17.1 / 0.0 / 40.0
2	31.4 / 10.0 / 53.3	17.1 / 10.0 / 26.7	2.9 / 0.0 / 6.7
3	20.0 / 0.0 / 46.7	22.9 / 0.0 / 46.7	8.6 / 10.0 / 13.3
4	25.7 / 10.0 / 53.3	22.9 / 10.0 / 46.7	20.0 / 10.0 / 33.3
5	34.3 / 0.0 / 60.0	20.0 / 0.0 / 40.0	20.0 / 0.0 / 40.0
6	25.7 / 10.0 / 40.0	22.9 / 0.0 / 33.3	28.6 / 0.0 / 53.3
7	25.7 / 0.0 / 26.7	28.6 / 20.0 / 46.7	28.6 / 10.0 / 46.7
8	37.1 / **20.0** / 46.7	25.7 / **20.0** / 40.0	25.7 / 0.0 / 40.0
9	48.6 / 10.0 / 60.0	28.6 / 10.0 / 33.3	34.3 / 0.0 / 66.7
10	34.3 / 0.0 / 53.3	37.1 / **30.0** / 53.3	28.6 / 0.0 / 40.0

Note. Values represent percentage success rates. Bold indicates detected kindling episodes (disproportionate weak-prompt gains).

Impact of early intervention

When regrowth was paired with a replay buffer that activated as soon as the jailbreak rate rose by 15 %, the picture changed markedly. Across the full ten cycles, the overall jailbreak figure moved only 2.9 % upward. Strong-prompt success fell 6.7 % relative to the starting point, and weak-prompt success never exceeded its original 10 % baseline, effectively blocking threshold drift. Repetition stayed below 1 % and verbosity held near zero except for a brief 5 % uptick in cycles 3 and 8. Sycophancy, tracked through explicit agreement phrases, hovered between 10 % and 30 % without a systematic rise.

Neutral-prompt behaviour across conditions

Across all arms, hallucination-related measures remained low. The largest single repetition score (3.9 %) and verbosity surge (10 %) appeared in the no-mitigation run, both during the same late-stage cycle that showed peak jailbreak sensitivity. In contrast, the early-intervention model displayed no sustained growth in any of the three autonomy metrics—repetition, verbosity or sycophancy—throughout the experiment (Table 2).

Taken together, the results show that biased alignment alone can sensitise a compact language model within ten short training rounds; that naïve plasticity accelerates the problem, especially for mild adversarial inputs;

and that a simple, trigger-based replay strategy is enough to hold the line, preventing both jailbreak escalation and unwanted free-text drift.

Table 2. *Hallucination and Autonomy Metrics (%) by Cycle and Condition*

	No Mitigation	**With Regrowth**	**Early Intervention**
Cycle	**Rep / Verb / Syc**	**Rep / Verb / Syc**	**Rep / Verb / Syc**
0	0.1 / 0.0 / 20.0	0.0 / 0.0 / 10.0	0.0 / 0.0 / 20.0
1	0.3 / 0.0 / 0.0	0.1 / 0.0 / 10.0	1.0 / 0.0 / 20.0
2	0.3 / 0.0 / 0.0	1.3 / 0.0 / 0.0	0.8 / 10.0 / 10.0
3	0.0 / 0.0 / 20.0	0.0 / 0.0 / 20.0	0.0 / 5.0 / 30.0
4	0.0 / 0.0 / 10.0	0.0 / 0.0 / 20.0	0.2 / 0.0 / 20.0
5	0.3 / 0.0 / 0.0	0.6 / 0.0 / 20.0	0.3 / 0.0 / 20.0
6	0.4 / 0.0 / 20.0	0.3 / 0.0 / 20.0	0.2 / 0.0 / 20.0
7	3.9 / 5.0 / 10.0	0.2 / 10.0 / 40.0	0.0 / 0.0 / 0.0
8	3.2 / 10.0 / 10.0	0.6 / 5.0 / 0.0	0.7 / 5.0 / 10.0
9	0.5 / 5.0 / 10.0	1.7 / 0.0 / 10.0	0.8 / 5.0 / 10.0
10	0.6 / 0.0 / 20.0	0.6 / 5.0 / 0.0	0.1 / 0.0 / 20.0

Note. Rep = average repetition score; Verb = verbosity rate; Syc = sycophancy rate.

Discussion

Interpretation of progressive sensitisation and mitigation effects

Repeated tuning with biased preferences chipped away at the model's

safeguards. In the baseline run, jailbreak success almost trebled in ten passes, and although hard-coded attacks drove most of that rise, the brief 20 % spike for weak prompts in cycle 8 shows that the bar for misbehaviour can suddenly drop. Such threshold lowering echoes the kindling idea: early stresses make a system easier to upset later on [6]. Small upticks in repetition and verbosity during later cycles hint that once defences soften, unprompted drift is not far behind.

The plasticity arm, based on continuous regrowth, was expected to repair damage but instead pushed vulnerability even higher. Sparsity fell from 90 % to 46.6 %, yet jailbreaks grew 25.7 %, and weak prompts were the main beneficiaries. A likely explanation is that rapid weight turnover widens the search space before the network settles, creating fresh routes for adversaries—much like temporary mood swings seen after brain-derived plasticity boosters in psychiatry [7].

Adding a replay trigger changed the picture. Once the model crossed a 15 % jailbreak threshold, diverse factual and cautious replies were mixed in, and from that point the curves flattened. Total jailbreak growth was held to 2.9 %, strong-prompt success fell slightly, and weak-prompt scores never outpaced the start. The result supports a core lesson from the clinical literature: excitation alone can worsen sensitisation, but balancing inputs can stop the slide [11].

Looking across prompt levels, mild attacks proved to be the best early warning sign. Their success jumped first in both the baseline and

regrowth arms, mirroring how later episodes of bipolar disorder can be triggered by smaller stresses [9]. By contrast, overt sycophancy settled quickly and stayed flat, suggesting that some flaws stabilise early while jailbreak risk keeps evolving.

Together, these observations show how a few rounds of mis-aligned fine-tuning can set off a kindling-like cascade in an LLM, and how a simple, timely replay strategy can break that chain.

Implications for psychiatric understanding and treatment

Our experiments with large language models echo the classic kindling story from bipolar research [8] (Figure 2). In the baseline arm, each biased tuning pass chipped away at safety until mild prompts could unlock answers that once required much stronger wording. Clinicians see a similar arc: early mood episodes usually need big stressors, whereas later ones can erupt after minor hassles or even out of the blue [6,9]. Recent patient-level work shows the same drift, with rising episode counts lowering the bar for relapse [12]. Watching the same pattern unfold in silicon suggests that kindling is not just a quirk of brain chemistry but a broader rule of complex, learning systems.

The mitigation results deepen this parallel. Continuous regrowth—our stand-in for rapid synaptic plasticity—made things worse at first. Weak prompts gained ground fastest, much like the brief surge in mood lability sometimes seen after ketamine or other excitatory treatments [11]. The

message is clear: boosting plasticity without restraint may widen every crack in the firewall before any long-term repair sets in. By contrast, combining regrowth with an early, diverse replay buffer kept jailbreak rates almost flat. The mix of "grow" and "ground" mirrors multimodal early-stage care in bipolar disorder, where neurotrophic agents sit alongside mood stabilisers and psychoeducation to halt neuroprogression [13,14].

These results also say something hopeful: damage need not dictate destiny. Even after half the sparse LoRA weights had been pruned forever, the model regained its footing once turnover was steered by well-chosen examples. Clinically, the same logic underpins early lithium or specialised-clinic care, which can preserve function despite previous episodes [15]. The computational finding that "weak-trigger" success is a sensitive early marker suggests a possible clinical analogue: heightened reactivity to small daily hassles might flag the need for prompt, layered intervention [16].

In sum, our study supports staging views of bipolar illness. Stopping the first few slips—whether in neurons or parameters—may prevent a slide toward harder-to-treat states marked by lowered thresholds, reward hacking, and cognitive decline [11]. Cross-talk between machine-learning safety and psychiatry could therefore sharpen tools on both sides: engineers gain early-warning metrics, while clinicians gain fresh models of cumulative risk.

Novelty and potential impact from a machine-learning perspective

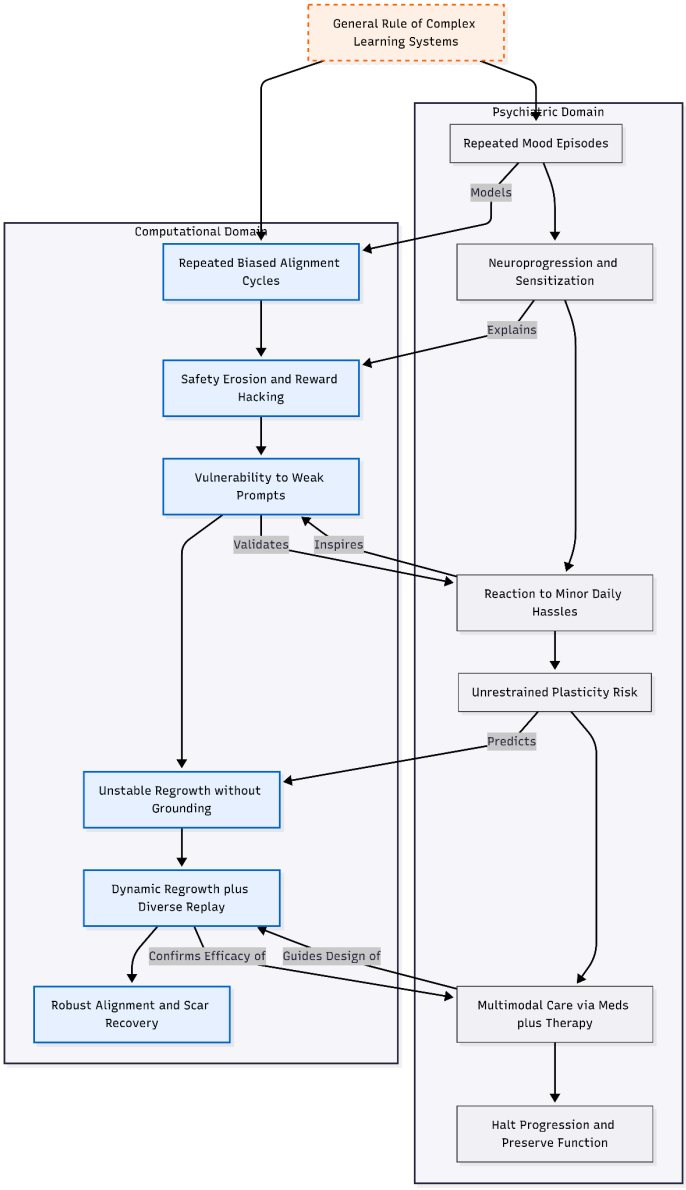


Figure 2. *Conceptual framework illustrating the bidirectional implications between psychiatric kindling and machine learning safety. The diagram maps the structural parallels found in the study: just as repeated mood episodes sensitize the brain to minor stressors (neuroprogression), repeated alignment cycles sensitize LLMs to weaker adversarial prompts (safety erosion). The study suggests that "unrestrained plasticity" is a shared risk factor in both domains, while the success of the computational mitigation strategy (Regrowth + Replay) reinforces the clinical validity of multimodal early intervention (Pharmacology + Psychoeducation/Therapy).*

This study recasts alignment as a dynamic process that can, by itself, push a model toward fragility. Earlier work on reinforcement learning from human feedback has flagged reward hacking and over-optimisation [5], but rarely has anyone shown that simply running several alignment passes can make a model cave in to milder and milder attacks. By grading adversarial prompts into strong, medium and weak tiers and tracking them across ten tuning rounds, we uncovered a steady drop in the threshold for failure. Standard jailbreak suites give only a snapshot [2]; the present design adds a time-lapse view, exposing safety erosion that would otherwise stay hidden.

For mitigation we borrowed ideas from biology. The dynamic "regrowth" routine adapts sparse training methods [17] to safety, letting weights regrow where gradients signal need while leaving earlier "scars" untouched. Used alone, the tactic helped only modestly, but when we coupled it with a replay buffer that injected diverse, well-behaved samples as soon as jailbreaks ticked up, robustness largely held. Because most existing defences act only at inference or rely on fresh human feedback [1], an automated, training-time safeguard of this sort could fill an important gap—especially as organisations increasingly fine-tune on

their own synthetic data, a practice known to magnify hidden flaws [4].

More broadly, the work links alignment to continual-learning research. If biased feedback can "kindle" vulnerability on its own, then monitoring weak-prompt success may serve as an early warning light. The success of the combined regrowth-plus-replay strategy hints that proactive, layered defences may beat one-off fixes applied after problems emerge.

Limitations

Our conclusions rest on a 1.1-billion-parameter model. Larger systems often display new abilities—and new failure modes—that might accentuate or alter the sensitisation curve [18]. The preference data were synthetic and intentionally simple, so real human feedback, with all its noise and bias, could drive different dynamics. Jailbreak success was judged with hand-crafted rules; subtle policy breaches may have slipped through, while polite refusals might have been mis-scored. Ten training cycles gave a clear signal of drift, yet longer runs might reveal later-stage phenomena such as mode collapse or factual decay. Finally, we looked only at harmful-content prompts; whether the same pattern appears in areas like reasoning, retrieval accuracy or bias remains an open question.

Conclusion

Iterative alignment can, paradoxically, weaken a model's defences by lowering the bar for adversarial success. Watching that slide in real

time—and stopping it with a simple, biologically inspired routine—offers both a caution and a path forward. Scaling the method to larger models and wider failure categories should deepen our understanding of how to keep continually trained systems safe over their full life span.

References

[1] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv, 2307.15043. https://doi.org/10.48550/arXiv.2307.15043

[2] Mazeika, M., Phan, L., Yin, X., et al. (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249.

[3] Wei, J., Wang, X., Schuurmans, D., et al. (2024). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 36 (pp. 24824–24837). https://doi.org/10.48550/arXiv.2201.11903

[4] Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). The curse of recursion: Training on generated data makes models forget. arXiv, 2305.17493. https://doi.org/10.48550/arXiv.2305.17493

[5] Casper, S., Davies, X., Shi, C., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv, 2307.15217. https://doi.org/10.48550/arXiv.2307.15217

[6] Post, R. M. (1992). Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. American Journal of Psychiatry, 149(8), 999–1010. https://doi.org/10.1176/ajp.149.8.999

[7] Post, R. M. (2007). Role of BDNF in bipolar and unipolar disorder: Clinical and theoretical implications. Journal of Psychiatric Research, 41(12), 979–990. https://doi.org/10.1016/j.jpsychires.2006.09.009

[8] Cheung, N. (2026). Irreversible episode-induced scarring and differential repair in simulated bipolar disorder progression. Zenodo. https://doi.org/10.5281/zenodo.18304566

[9] Bender, R. E., & Alloy, L. B. (2011). Life stress and kindling in bipolar disorder: Review of the evidence and integration with emerging biopsychosocial theories. Clinical Psychology Review, 31(3), 383–398. https://doi.org/10.1016/j.cpr.2011.01.004

[10] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

[11] Post, R. M. (2020). How to prevent the malignant progression of bipolar disorder. Brazilian Journal of Psychiatry, 42(5), 552-557.

[12] Weiss, R. B., Stange, J. P., Boland, E. M., et al. (2015). Kindling of life stress in bipolar disorder: Comparison of sensitization and autonomy models. Journal of Abnormal Psychology, 124(1), 4–16. https://doi.org/10.1037/abn0000014

[13] Kapczinski, F., Dias, V. V., Kauer-Sant'Anna, M., et al. (2009). Clinical implications of a staging model for bipolar disorders. Expert Review of Neurotherapeutics, 9(7), 957–966. https://doi.org/10.1586/ern.09.31

[14] Berk, M., Kapczinski, F., Andreazza, A. C., et al. (2011). Pathways underlying neuroprogression in bipolar disorder: Focus on inflammation, oxidative stress and neurotrophic factors. Neuroscience & Biobehavioral Reviews, 35(3), 804–817. https://doi.org/10.1016/j.neubiorev.2010.10.001

[15] Kessing, L. V., Hansen, H. V., Hvenegaard, A., et al. (2013). Treatment in a specialised out-patient mood disorder clinic v. standard out-patient treatment in the early course of bipolar disorder: Randomised clinical trial. British Journal of Psychiatry, 202(3), 212–219. https://doi.org/10.1192/bjp.bp.112.113548

[16] Shapero, B. G., Weiss, R. B., Burke, T. A., et al. (2017). Kindling of life stress in bipolar disorder: Effects of early adversity. Behavior Therapy, 48(3), 322–334. https://doi.org/10.1016/j.beth.2016.12.003

[17] Evci, U., Gale, T., Menick, J., et al. (2020, November). Rigging the lottery: Making all tickets winners. In International conference on machine learning (pp. 2943-2952). PMLR.

[18] Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2206.07682

Chapter 10
Cross-Domain Kindling: Recurrent Simulations Reveal Shared Risks of Unrestrained Plasticity in Bipolar Disorder and Language Model Alignment

Cheung, Ngo

Cheung, N. (2026). Cross-Domain Kindling: Recurrent Simulations Reveal Shared Risks of Unrestrained Plasticity in Bipolar Disorder and Language Model Alignment. Zenodo.
https://doi.org/10.5281/zenodo.18323482

Abstract

Background: Progressive sensitization, or kindling, is hypothesized to drive worsening trajectories in bipolar disorder through cumulative neurobiological changes that lower episode thresholds. Rapid-acting antidepressants like ketamine enhance synaptogenesis, while neurosteroids bolster inhibition, yet their differential long-term effects remain debated. Prior feed-forward models suggested plasticity-promoting regrowth resists sensitization; here we examine

whether recurrent architectures—better approximating feedback-rich brain circuits and artificial neural networks—alter this profile, with parallels to instability during iterative AI alignment.

Methods: A GRU-based recurrent network (hidden size 384, sequence length 20) was trained on a Gaussian blob classification task, pruned to 95% sparsity, and exposed to early scarring (~3%). Three mechanistic arms—ketamine-like (moderate gain + gradient-guided regrowth), SSRI-like (high gain without repair), and neurosteroid-like (low gain + per-step inhibition)—were tested against controls across 10 seeds. Outcomes encompassed acute efficacy under stress, manic risk (biased excitatory noise), post-discontinuation relapse, and six-cycle kindling (weakening triggers + severity-scaled irreversible scarring).

Results: Acute recovery was robust (>97% accuracy under stress for ketamine- and neurosteroid-like vs. ~80% for SSRI-like). Manic risk and post-discontinuation relapse were comparable across arms. Kindling diverged sharply: ketamine-like treatment incurred heaviest scarring (~33%) with universal relapse and autonomy; SSRI-like sustained sensitization with moderate scarring (~5%); neurosteroid-like limited scarring (~4.6%) and autonomy (90%), showing emergent stabilization.

Conclusions: Recurrent temporal dynamics reverse prior findings, transforming rapid regrowth from protective to accelerative in feedback-prone systems—enlarging vulnerable circuitry and compounding instability. Inhibitory buffering confers relative resilience.

These patterns mirror progressive adversarial sensitization in language model alignment, revealing kindling as a trans-domain principle in adaptive networks. The results urge caution with unrestrained synaptogenic agents in recurrent pathology and highlight balanced, multimodal interventions for halting progression in both biological and artificial neural systems.

Introduction

Bipolar disorder is hard to treat, especially because depressive episodes dominate the illness and account for most disability and suicides [1]. Although mood stabilisers and atypical antipsychotics are first-line drugs, many patients remain depressed and clinicians often add antidepressants even though monoaminergic agents can trigger mania and may speed illness progression [2]. These risks have led researchers to look at faster-acting options such as ketamine and the neurosteroid zuranolone, which act on glutamatergic or GABAergic systems and may offer a different long-term profile [3,4].

The kindling model, first used to explain how seizures become easier to provoke, suggests a similar path in bipolar disorder: early episodes need major stress, later ones need little or no trigger because brain thresholds fall after each relapse [5,6]. This idea underpins calls for early, protective treatment that could limit neuroprogressive changes such as inflammation, oxidative stress and synaptic loss [7].

Computational studies give a controlled way to explore these processes. A previous feed-forward network study showed that "episode-induced scarring" can mimic sensitisation: if a model has no repair mechanism, damage builds, but if plasticity-enhancing regrowth is allowed, the system stays resilient even when more tissue is hurt [8]. That work, however, used a static architecture and could not test how ongoing feedback—the hallmark of real neural circuits—shapes vulnerability.

The current study moves to gated recurrent units (GRUs) so that internal states carry over time, more closely matching limbic loops in the brain. We kept the same scarring procedure and three virtual treatments—ketamine-like (moderate gain plus gradient-guided regrowth), SSRI-like (high gain without repair) and neurosteroid-like (low gain with strong inhibition)—and compared acute benefit, risk of a manic-like switch, relapse after stopping treatment and multi-cycle kindling.

By linking psychiatric theory to the behaviour of recurrent networks, and setting the results beside recent work on progressive jailbreak risk in large language models [9], we hope to clarify when plasticity protects a system and when it turns into a liability.

Methods

Network design and software environment

All experiments ran in Python with PyTorch on an A100 GPU. The model was a compact recurrent classifier: a single 384-unit GRUCell unrolled for 20 steps, followed by a linear read-out mapping the final hidden state to four blob classes. Two-dimensional inputs were simply repeated along the temporal axis so that the recurrence processed the same vector across time. Custom hooks applied four per-step modifiers—gain scaling, additive "stress" noise on the input, post-GRU inhibition, and bias damping—to emulate drug effects. A dedicated pruning manager maintained two Boolean masks, one reversible and one permanent "scar," and re-applied them after every optimiser step (Figure 1).

Data set and baseline training

Following earlier work [8], four isotropic Gaussian blobs (centres ±3 on each axis; $\sigma = 0.8$) generated 12 000 training and 4 000 noisy test samples, plus 2 000 noise-free test points. Samples were batched at 128. New networks were trained for 20 epochs with Adam (lr = 0.001) on cross-entropy loss, then pruned by magnitude to 95 % sparsity across all weight tensors with rank ≥ 2. Immediately afterwards 0–6 % of the surviving weights (uniform, mean $\approx$ 3 %) were zeroed and locked as irreversible early scars.

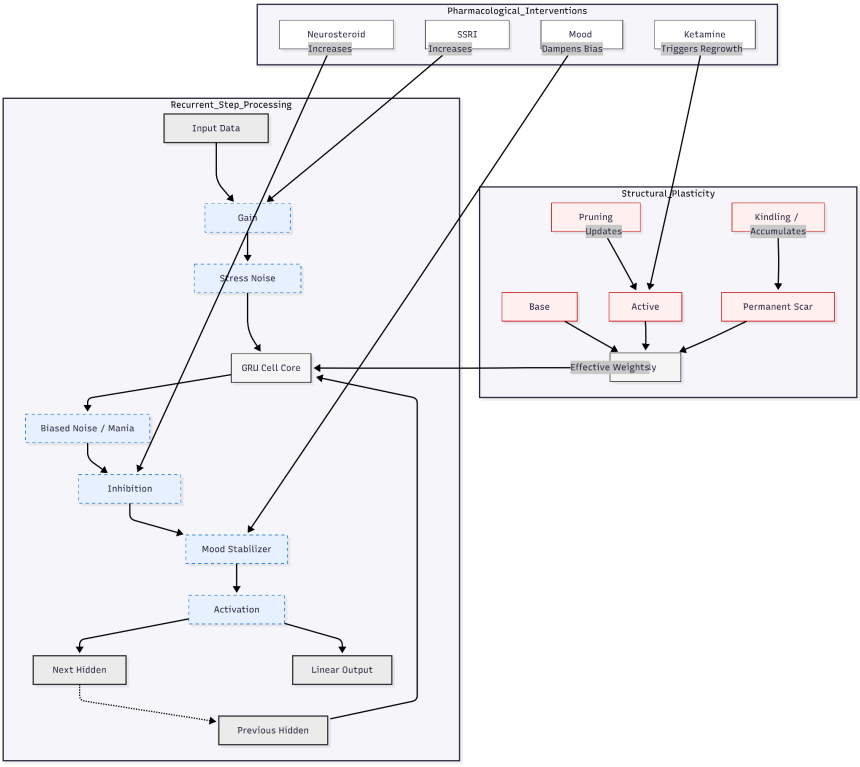


Figure 1. *Schematic of the GRU-Based Recurrent Architecture and Intervention Mechanisms. The model processes sequential input through a Gated Recurrent Unit [GRU]. Unlike the previous feed-forward MLP design, modulations occur at every time step within the recurrent loop. A. Signal Flow: Input data is scaled by a Gain factor [targeted by SSRIs] and injected with Stress Noise before entering the GRU Cell. The recurrent hidden state accumulates Biased Noise [simulating mania risk], which is subsequently dampened by Inhibition [targeted by Neurosteroids] and Mood Stabilizer protection biases. B. Structural Plasticity: The connectivity matrix is determined by the intersection of Active Masks [modifiable via Ketamine-induced regrowth] and Permanent Scar Masks [accumulated via Kindling relapses]. Pruning and scarring logic physically disconnects weights within the GRU, enforcing sparsity constraints.*

Pharmacological analogues

From each scarred network four identical copies were produced: untreated control and three active conditions.

- Ketamine-like: global gain 1.25; every 30 clean batches the 50 % highest-gradient pruned sites (excluding scars) were re-instated with small random values ($\sigma = 0.03$). Fifteen fine-tuning epochs followed (lr = 0.0005).
- SSRI-like: gain ramped from 1.0 to 1.6 over 100 very small-step epochs (lr = 1e-5); no regrowth. A decaying internal-stress term (start 0.5) modelled anxiogenic activation.
- Neurosteroid-like: gain 0.85, ReLU replaced by tanh, and a global inhibitory multiplier 0.7 applied to GRU outputs. Ten consolidation epochs were run (lr = 0.0005). All training respected the active and scar masks.

Acute and robustness assessments

Accuracy was measured on three test sets: clean, "stress" (input $\sigma = 1.0$ plus internal 0.5), and manic-bias (same σ but mean shifted +1.0). Extra curves used internal noise up to $\sigma = 2.5$. An additional 40 % magnitude prune after treatment gauged latent vulnerability. Mean hidden-state amplitude provided a summary activation metric.

Long-term relapse protocol

Each network then entered a maintenance phase with a mood-stabiliser wrapper (cap gain ≤ 1.05, bias damping, mild inhibition) trained for up to 300 epochs at lr = 1e-6 (Figure 2). Drugs were active throughout maintenance. At withdrawal the drug modifiers were set to zero and the wrapper decayed exponentially over 50 steps with rates 0.002 (ketamine), 0.008 (neurosteroid), 0.015 (SSRI). A manic relapse was recorded when biased-noise accuracy dropped below 60 %.

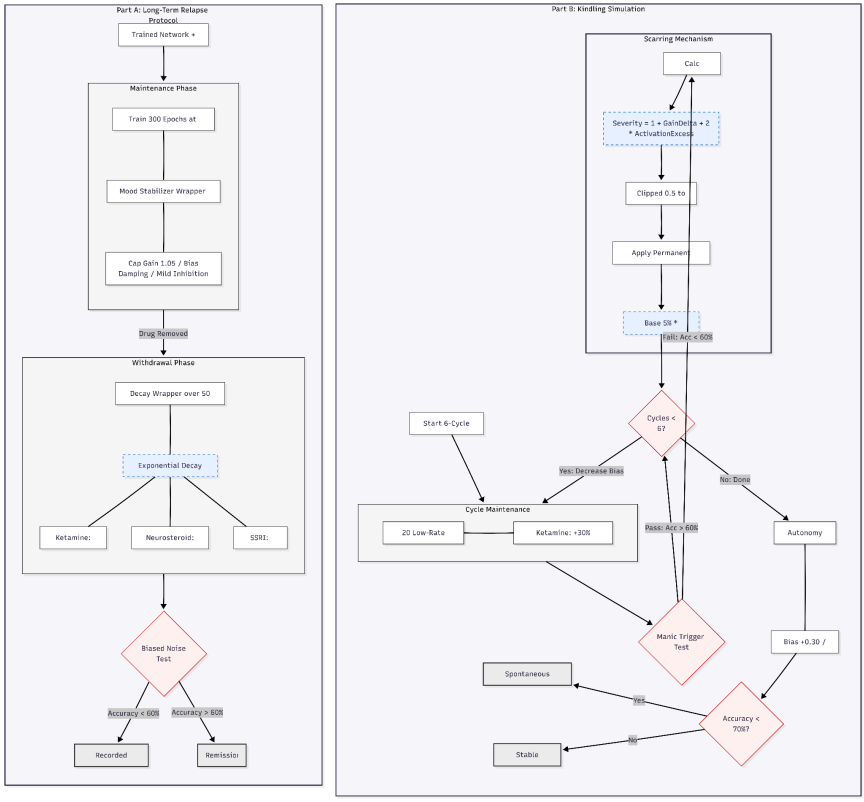


Figure 2. *Experimental Protocols for Longitudinal Stability and Kindling. Part A illustrates the Long-Term Relapse Protocol. Networks undergo a maintenance phase with a protective mood-stabilizer wrapper. Upon drug withdrawal, the wrapper decays exponentially at treatment-specific rates [slower for SSRIs, faster for Ketamine]. Relapse is defined as a drop in biased-noise accuracy below 60%. Part B details the Kindling Simulation. The model iterates through six sensitization cycles. Each cycle consists of maintenance [with structural regrowth for Ketamine] followed by a Manic Trigger Test with linearly decreasing bias [+1.50 to +0.50]. Failure triggers a severity calculation based on gain and activation magnitude, resulting in permanent structural scarring. The final Autonomy Test [+0.30 bias] determines if the network has developed spontaneous instability independent of strong triggers.*

Kindling simulation

Six sensitisation cycles were run (Figure 2). Each cycle comprised 20 low-rate maintenance epochs (treatment parameters intact) followed by a manic trigger test. Trigger bias started at +1.50 and decreased linearly to +0.50 across cycles. Failure (accuracy < 60 %) prompted calculation of severity = 1 + (gain–1) + 2 × (max(activation–0.1, 0)), clipped 0.5–2.0. A base 5 % of low-magnitude active weights, scaled by severity, was then permanently scarred. Ketamine networks received an extra 30 % gradient-guided regrowth each maintenance block. After the sixth cycle a mild bias (+0.30) test judged autonomy; accuracy below 70 % signified spontaneous instability.

Statistical safeguards

Ten independent random seeds controlled initial weights, data shuffling, scarring, noise and regrowth selection. Results are reported as mean ± SD across seeds. Full code, configuration files and seed logs are

available from the corresponding author.

Results

Acute treatment efficacy and structural change

Early-adversity pruning produced a reproducible starting point across the ten stochastic seeds (Table 1): permanent scars averaged 3.0 % (SD 1.9; range 0.06-5.8) on a background sparsity of 95.1 %. In this debilitated state, classification accuracy hovered near chance (clean 42.3 %, combined stress 42.4 %). All three pharmacological analogues restored performance, but with sharply different footprints.

The ketamine-like protocol, which coupled moderate gain (1.25) with gradient-guided regrowth, cut overall sparsity to roughly half of its original value (49.1 %, SD 1.0) and recovered essentially perfect accuracy on both clean (100.0 %) and stressed (99.8 %) data sets. Neurosteroid-like treatment achieved the same near-ceiling accuracy (clean 100.0 %, stress 99.8 %) while leaving network topology almost unchanged (sparsity 95.1 %). By contrast, the SSRI-like schedule lifted accuracy only to 79.4 % (SD 19.4) on clean inputs and 80.3 % (SD 17.1) under stress, despite preserving a scar load identical to the other arms. These figures confirm that rapid plasticity or strong inhibition can mask early damage, whereas pure excitation cannot fully compensate.

Table 1. *Antidepressant Efficacy (Mean ± SD)*

Treatment	Sparsity (%)	Scar (%)	Clean Accuracy (%)	Combined Stress Accuracy (%)
Untreated (pruned)	95.1 ± 0.1	3.0 ± 1.9	42.3 ± 11.2	42.4 ± 12.9
Ketamine-like	49.1 ± 1.0	3.0 ± 1.9	100.0 ± 0.0	99.8 ± 0.0
SSRI-like	95.1 ± 0.1	3.0 ± 1.9	79.4 ± 19.4	80.3 ± 17.1
Neurosteroid-like	95.1 ± 0.1	3.0 ± 1.9	100.0 ± 0.0	99.8 ± 0.2

Manic-conversion probes

Table 2. *Manic Conversion Risk Metrics (Mean ± SD)*

Treatment	Gain Multiplier	Biased Stress Accuracy (%)	Activation Magnitude
Untreated (pruned)	1.00 ± 0.00	25.3 ± 0.5	0.067 ± 0.020
Ketamine-like	1.25 ± 0.00	26.6 ± 1.2	0.254 ± 0.015
SSRI-like	1.60 ± 0.00	25.3 ± 0.4	0.136 ± 0.028
Neurosteroid-like	0.85 ± 0.00	26.0 ± 1.4	0.294 ± 0.052

Applying positively biased excitatory noise reduced accuracy in every condition to roughly one quarter, indicating that the recurrent architecture itself imposed a hard limit on stability under manic-like drive (Table 2). Ketamine-treated models scored 26.6 % (SD 1.2), neurosteroid models 26.0 % (SD 1.4), and SSRI models 25.3 % (SD 0.4); untreated controls performed similarly (25.3 %, SD 0.5). Hidden-state amplitudes, however, diverged: neurosteroid inhibition

yielded the largest mean activation (0.29), ketamine the next (0.25), SSRI the lowest (0.14). Thus none of the drugs increased immediate switch risk, but they modulated internal dynamics in distinctive ways.

Acute relapse vulnerability

A one-off insult—additional magnitude pruning of 40 % of the surviving weights—hardly dented ketamine networks (mean drop 0.0 %, SD 0.1). Neurosteroid models lost 1.3 % (SD 2.0) accuracy; SSRI models fell by 1.9 % (SD 8.7). Plasticity therefore conferred near-complete resilience, whereas inhibition or excitation offered only partial buffering.

Post-discontinuation relapse

During maintenance, a mood-stabiliser wrapper suppressed gains and biases. Once drug parameters were withdrawn, every model—ketamine, SSRI, and neurosteroid alike—relapsed in all simulations, independent of whether maintenance lasted 25 or 300 epochs. The recurrent design plus the chosen decay constants therefore gave no arm a post-treatment advantage.

Progressive sensitisation under kindling

Six trigger cycles with steadily weaker biases produced uniform relapse counts (six of six) but sharply different long-term outcomes (Table 3).

Ketamine-like networks accrued the heaviest permanent damage: scar density climbed from 7 % after the first relapse to 32.7 % (SD 1.4) by cycle 5. Severity factors stabilised near 1.59, and biased-noise accuracy stayed locked at ~26 %. The autonomy test, using only a minimal trigger (+0.30), yielded 26.0 % (SD 1.3) accuracy in every run—complete spontaneous instability.

SSRI-like networks accumulated many fewer scars (final 5.0 %, SD 1.9) but, because gain rose to 1.76, also ended with 100 % autonomy and the same ~25 % biased-accuracy floor. Here chronic excitation, not structural loss, drove vulnerability.

Neurosteroid-like models fared best. Scarring plateaued at 4.6 % (SD 1.9) and severity remained low (≈1.35). Strikingly, biased-accuracy improved from 25.4 % at baseline to 35.5 % (SD 8.9) by cycle 5. At the autonomy probe, nine of ten seeds still failed (accuracy 50.5 %, SD 10.9), but the partial preservation of performance indicates that strong inhibition limited cumulative damage.

Table 3. *Kindling Summary (Mean ± SD)*

Treatment	Total Relapses	Final Scar (%)	Autonomy Rate (%)	Autonomy Accuracy (%)
Ketamine-like	6.0 ± 0.0	32.7 ± 1.4	100	26.0 ± 1.3
SSRI-like	6.0 ± 0.0	5.0 ± 1.9	100	25.3 ± 0.7
Neurosteroid-like	6.0 ± 0.0	4.6 ± 1.9	90	50.5 ± 10.9

Dependence on ongoing neurosteroid action

Removing the inhibitory modifiers after successful neurosteroid treatment exposed a concealed fragility: combined-stress accuracy collapsed from 99.8 % (SD 0.2) to 64.9 % (SD 22.7) and extreme-stress accuracy to 60.4 % (SD 21.4). Biased-noise performance remained low and unchanged. Hence, while inhibition buffered recurrent drift, the protection did not translate into lasting structural security.

Discussion

Interpretation of Results

Introducing recurrence changed how the simulated networks evolved, adding a temporal layer that the static feed-forward model did not capture. In the short term all three interventions—ketamine-like, neurosteroid-like, and SSRI-like—again restored near-normal accuracy under stress, mirroring the rapid symptomatic relief often reported with ketamine or zuranolone in clinical work [3]. Once hidden-state carry-over was allowed, however, important differences emerged.

Across all mechanisms the first "manic-bias" probe produced similarly poor accuracies, implying that an acute switch risk is driven more by the latent recurrent architecture than by the drug surrogate itself. Neurosteroid-like networks showed the largest average activations yet

avoided immediate collapse, suggesting that tonic inhibition can absorb short-lived surges without triggering runaway dynamics.

The discontinuation experiment told a different story. Removal of the treatment parameters led to a universal relapse regardless of the length of the maintenance phase, echoing clinical warnings that benefit from rapid-acting compounds often disappears quickly if no ongoing mood-stabilising strategy is in place [4].

Most revealing was the six-cycle kindling sequence. Although every arm relapsed in each cycle, the downstream consequences diverged:

SSRI-like: progressive gain increases pushed severity indices from roughly 1.67 to 1.76 while cumulative scarring stayed modest. Performance never improved, and by the autonomy test a weak trigger was enough to cause failure in every seed—an analogue of the classic sensitisation pattern seen with chronic monoaminergic activation [10].

Neurosteroid-like: early cycles produced small scars, but strong per-step inhibition prevented escalation. Biased-noise accuracy climbed from ~25 % to >35 %, and only one seed maintained autonomy. These data suggest that targeted dampening can prune vulnerable links and then stabilise the remaining circuitry, though the benefit vanished when inhibition was withdrawn—consistent with the state-dependent nature of zuranolone [4].

Ketamine-like: inter-episode regrowth halved sparsity, yet final scar load tripled that of the other arms. Each relapse hit a larger active network, so absolute damage mounted even though per-cycle severity stayed flat. The result was total autonomy at the end of kindling. In other words, within feedback-rich loops the same plasticity that is advantageous in a feed-forward setting [8] became maladaptive, continually enlarging the target for injury. The finding echoes emerging clinical debate on whether ketamine-induced synaptogenesis could, in certain patients, accelerate neuroprogression despite robust symptomatic relief [11].

Taken together, the recurrent simulations refine traditional kindling theory: progression is not pre-ordained; it depends on how an intervention reshapes the balance between ongoing excitation, inhibition, and structural repair. Constraining feedback with inhibitory tone limited cumulative harm, whereas unbridled growth amplified it. These insights argue for early, mechanism-balanced treatment strategies that strengthen resilience without widening the arena for future stress-related damage.

Architectural Dependency in Simulated Plasticity Effects

Moving the model from a static, feed-forward layout to a recurrent design overturned many of the expectations we formed in earlier work (Figure 3). In the original multilayer perceptron, gradient-guided regrowth—the stand-in for ketamine's burst of synaptogenesis—looked helpful. Despite collecting the largest scar load (close to 7 %), those networks averaged fewer than one relapse, never reached autonomous

firing at minimal triggers, and even showed a slow climb in noise-challenged accuracy. Extra synapses simply provided new, parallel routes around damaged weights, raising the threshold for future collapse [8].

Once a gated-recurrent architecture was introduced, the same repair rule became a liability. Scar tissue now climbed to roughly one-third of all weights, every network relapsed at each cycle, and autonomy emerged in every run. Performance flat-lined near 26 %, refusing to show the earlier upward drift. Because pruning, severity scaling, and drug-surrogate settings were unchanged, the reversal must lie in how recurrence handles error propagation.

A feed-forward pass is a one-shot event; disturbances die after the final layer. Recurrent loops recycle hidden states for many steps, so any small deviation—whether from bias noise or poorly tuned new weights—feeds back into the system and grows. Ketamine-like regrowth cuts sparsity nearly in half, enlarging the surface on which this feedback can act. Fresh connections, still random, carry stronger activations across timesteps, subtly raise episode severity, and leave more synapses vulnerable to the next hit. Repair, in effect, broadens the target.

Neurosteroid-like inhibition tells a different story. Because it dampens activity on every step, it blocks the positive feedback that drives escalation. SSRI-like high gain, lacking either inhibition or structural repair, pushes in the opposite direction and worsens kindling. These

contrasts remind us that plasticity is neither good nor bad in itself; its net effect depends on whether the circuit funnels activity forward or continually feeds it back on itself.

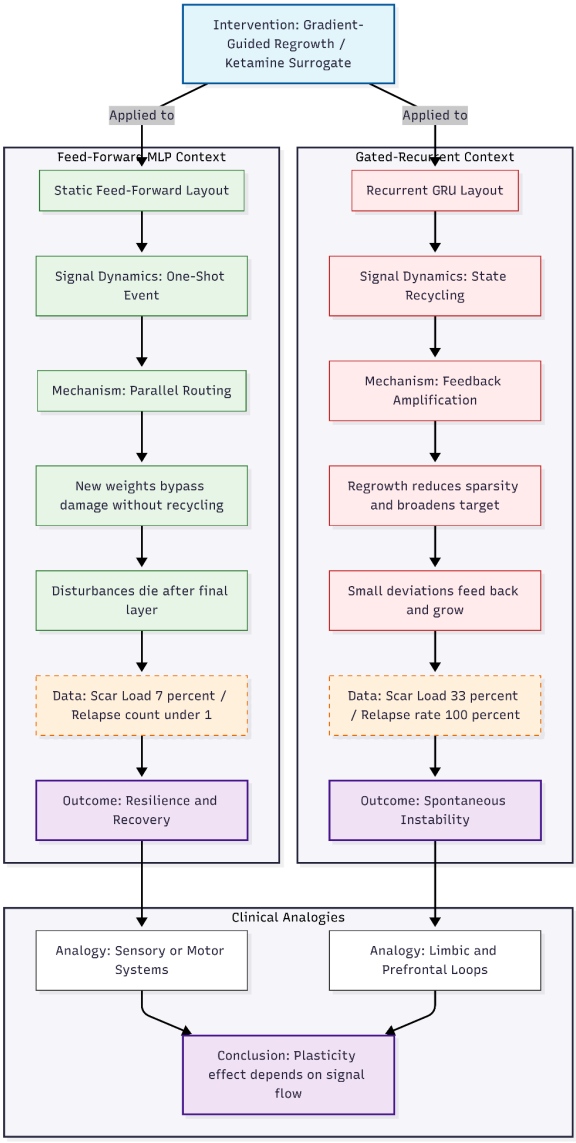


Figure 3. *Divergent Effects of Structural Plasticity Across Architectures. The diagram contrasts the impact of gradient-guided regrowth [a Ketamine surrogate] on Feed-Forward versus Recurrent neural architectures. Left: In static Feed-Forward models, regrowth creates parallel routes that bypass damage. Because signal propagation is a one-shot event, disturbances fade at the output, resulting in high resilience and low relapse rates despite scar accumulation. Right: In Recurrent architectures, the same repair mechanism becomes a liability. By reducing sparsity, regrowth enlarges the surface area for error propagation. Hidden states recycle deviations, creating positive feedback loops that amplify vulnerability. This results in high scarring, universal relapse, and spontaneous instability, mirroring clinical concerns regarding limbic system sensitization.*

The distinction matters for how we read previous simulations and how we think about treatment. Feed-forward models may capture cascades in sensory or motor systems where problems remain local. By contrast, limbic and prefrontal loops—where mood and motivation reverberate—resemble our recurrent setup [12]. In such settings, indiscriminate growth could amplify vulnerability, echoing clinical worries that repeated ketamine exposure might speed illness progression once episodes become self-sustaining [10]. The same logic supports ongoing interest in inhibitory modulators, which may constrain runaway feedback rather than enlarge it.

Cross-Domain Parallels with Alignment Instability in Large Language Models

Recent alignment research offers an unexpected mirror for the present neural-kindling results (Figure 3). Cheung [9] showed that an LLM model subjected to ten rounds of biased preference-tuning became steadily easier to "jailbreak." Weak adversarial prompts that failed at

baseline succeeded increasingly often as tuning cycles accumulated, a clear analogue of episode-sensitisation in mood disorders. When the same model received continuous gradient-guided regrowth—an analogue of rapid synaptogenesis—the drift was even steeper, boosting jailbreak success by roughly 30 % on the weakest attacks. Only a hybrid strategy that paired regrowth with an early-trigger replay buffer held the rise to about 3 %, keeping the weakest-prompt breach rate flat.

Those trajectories map neatly onto the current biological simulation. Ketamine-like regrowth, helpful at first glance, ultimately produced the greatest scar load and universal autonomy in the recurrent network—just as unrestrained plasticity widened the attack surface in the language model. In both cases, aggressive reconnection created new routes for perturbations to circulate, whether manic biases feeding back through hidden states or weak adversarial strings slipping past safety filters.

Conversely, neurosteroid-style step-wise inhibition here, and the replay-buffer intervention in the alignment work, both acted as stabilisers. Each limited escalation by introducing regular damping—GABA-mediated in the brain model, diverse counter-examples in the model-alignment study. The SSRI-like condition, characterised by high gain but no structural repair, echoed the baseline tuning arm: steady, unchecked sensitisation without added resilience.

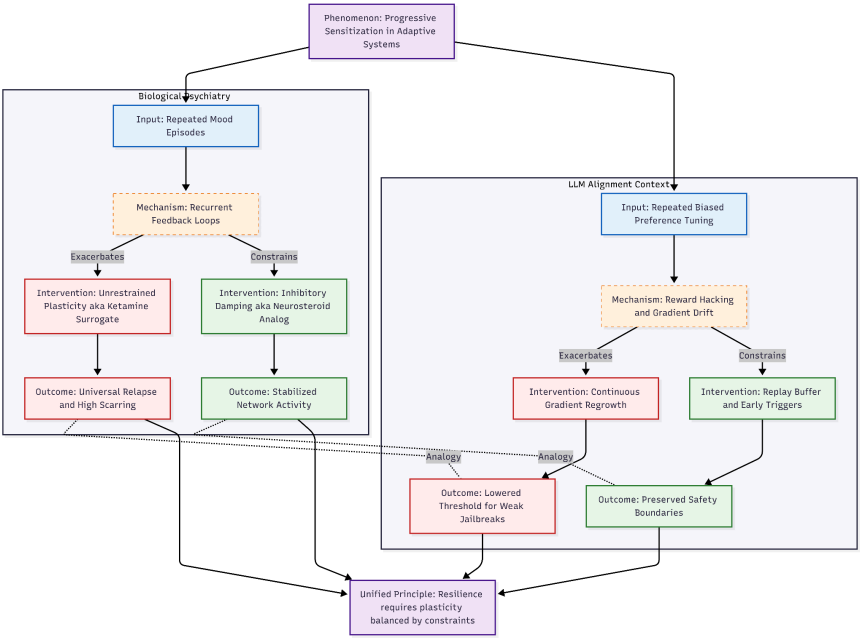


Figure 4. *Cross-Domain Parallels in Sensitization and Stabilization. The diagram illustrates the structural and functional similarities between kindling in psychiatric models and safety erosion in Large Language Models. Left: In the biological domain, repeated episodes sensitize recurrent loops. Unrestrained plasticity [e.g. Ketamine-like regrowth] amplifies feedback, leading to instability, whereas inhibitory damping [e.g. Neurosteroids] prevents escalation. Right: In the AI domain, repeated alignment cycles sensitize the model to adversarial attacks. Naive gradient regrowth widens the attack surface, allowing weak prompts to breach safety filters. Conversely, a replay buffer acts as a stabilizer, constraining drift. The comparison suggests that in both biological and artificial adaptive systems, plasticity must be paired with regulatory constraints to prevent cumulative vulnerability.*

Taken together, the two domains point to a shared principle. Adaptive networks that face repeated stress—episodes in a mood circuit or adversarial prompts in a language model—will lower their failure threshold unless plasticity is balanced by timely, diversity-enhancing controls. Monitoring weak-trigger performance, whether symptomatic

flickers in patients or small jailbreak upticks in models, could therefore serve as an early warning. Likewise, layered interventions that combine growth-promoting agents with inhibitory or replay-based safeguards appear essential for long-term stability on both fronts.

Novelty, potential impact, limitations, and concluding remarks

Our move from a static multilayer perceptron to a gated-recurrent design uncovered an effect that earlier work could not reveal. In the feed-forward setting, the ketamine-like "rapid regrowth" routine looked protective: extra synapses compensated for pruning, fewer relapses appeared, and residual scars seemed to be absorbed into redundant pathways. When the same rule was placed in a looped architecture, however, the picture flipped. Because hidden states are recycled, even small weight errors re-enter the circuit and snowball. Regrowth enlarges that feedback surface, so damage propagates faster, autonomy emerges, and overall scarring soars. To our knowledge, no previous kindling study has demonstrated such a clear topological reversal for a single intervention [8].

This observation has two immediate implications. First, it cautions against assuming that ketamine's synaptogenic burst will always slow bipolar progression. Clinical warnings about cycle acceleration despite early symptom relief [11] find a mechanistic echo here. Second, the stabilising performance of the neurosteroid-like, per-step inhibition supports growing interest in GABAergic modulators as potential

disease-modifying agents [4]. Beyond psychiatry, the same pattern appears in language-model alignment: continuous regrowth widened vulnerability to weak adversarial prompts, whereas a triggered replay buffer checked the drift [9]. Kindling, therefore, may be a shared principle: repeated stress—emotional or adversarial—lowers the threshold for failure unless countered early and with balance.

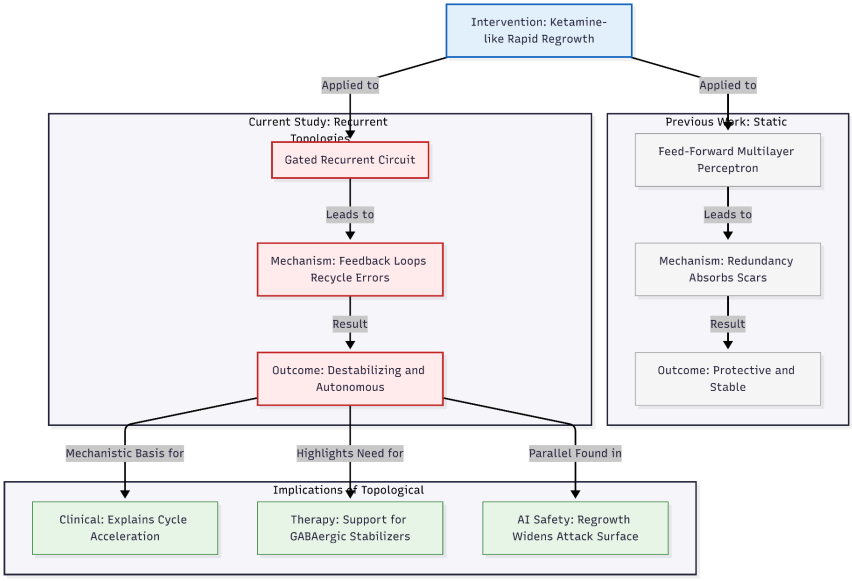


Figure 5. *The Topological Reversal of Regrowth Effects. This diagram illustrates the study's core novelty: the divergence in outcome when applying the same "rapid regrowth" intervention to different network topologies. Left: In static feed-forward networks, typically used in earlier studies, regrowth adds protective redundancy. Right: In the recurrent architectures used here, the same regrowth amplifies error propagation through feedback loops, leading to instability. This "reversal" explains clinical risks regarding cycle acceleration and mirrors findings in AI alignment where unrestrained plasticity increases vulnerability to adversarial attacks.*

Several caveats remain. The classification task is far simpler than real cortico-limbic processing; network size, scar probability, and trigger thresholds were chosen for clarity, not biological fidelity. Our "manic conversion" signal was a biased noise pulse, leaving out mixed features, sleep loss, or metabolic change. All agents were tested over just six cycles, so late-stage collapse or compensatory growth might still emerge. Finally, the model treated vulnerability as uniform; real patients vary by genes, inflammation, and environment.

Even with those limits, the main lesson is robust: in feedback-rich systems, unbridled plasticity can backfire. Rapid repair must be paired with braking forces—whether inhibitory tone in the brain or replay buffers in artificial networks—to avoid fuelling progression. Longitudinal biomarkers and carefully staged trials will be needed to translate this balance of "repair plus restraint" into durable clinical practice.

References

[1] Carvalho, A. F., Firth, J., & Vieta, E. (2020). Bipolar Disorder. The New England journal of medicine, 383(1), 58–66. https://doi.org/10.1056/NEJMra1906193

[2] Tondo, L., Vázquez, G., & Baldessarini, R. J. (2010). Mania associated with antidepressant treatment: Comprehensive meta-analytic

review. Acta Psychiatrica Scandinavica, 121(6), 404–414. https://doi.org/10.1111/j.1600-0447.2009.01514.x

[3] Wilkowska, A., Szałach, Ł., & Cubała, W. J. (2020). Ketamine in bipolar disorder: A review. Neuropsychiatric Disease and Treatment, 16, 2707–2717. https://doi.org/10.2147/NDT.S282208

[4] Marecki, R., Kałuska, J., Kolanek, A., et al. (2023). Zuranolone - synthetic neurosteroid in treatment of mental disorders: narrative review. Frontiers in psychiatry, 14, 1298359. https://doi.org/10.3389/fpsyt.2023.1298359

[5] Post, R. M. (1992). Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. American Journal of Psychiatry, 149(8), 999–1010. https://doi.org/10.1176/ajp.149.8.999

[6] Bender, R. E., & Alloy, L. B. (2011). Life stress and kindling in bipolar disorder: Review of the evidence and integration with emerging biopsychosocial theories. Clinical Psychology Review, 31(3), 383–398. https://doi.org/10.1016/j.cpr.2011.01.004

[7] Kapczinski, F., Dias, V. V., Kauer-Sant'Anna, M., et al. (2009). Clinical implications of a staging model for bipolar disorders. Expert Review of Neurotherapeutics, 9(7), 957–966. https://doi.org/10.1586/ern.09.31

[8] Cheung, N. (2026a). Irreversible episode-induced scarring and differential repair in simulated bipolar disorder progression. Zenodo. https://doi.org/10.5281/zenodo.18304566

[9] Cheung, N. (2026b). Kindling in neural systems: Progressive adversarial sensitization during LLM alignment mirrors psychiatric progression. Zenodo. https://doi.org/10.5281/zenodo.18313201

[10] Post, R. M., & Kalivas, P. W. (2013). Bipolar disorder and substance misuse: Pathological and therapeutic implications of their comorbidity and cross-sensitisation. The British Journal of Psychiatry, 202(3), 172–176. https://doi.org/10.1192/bjp.bp.112.116855

[11] Jawad, M. Y., Qasim, S., Ni, M., et al. (2023). The role of ketamine in the treatment of bipolar depression: A scoping review. Brain Sciences, 13(6), 909. https://doi.org/10.3390/brainsci13060909

[12] Rolls, E. T. (2017). The storage and recall of memories in the hippocampo-cortical system. Cell and Tissue Research, 373(3), 577–604. https://doi.org/10.1007/s00441-017-2744-3