The Demon in the Forge

Information-Constrained Sorting as the Mechanism That Converts Metastable Exploration into Directional Development

tratium

Abstract. A companion paper (tratium, 2026) established that complex adaptive systems under calibrated stress exhibit a bounded metastability island in anti-closure × stress parameter space — but a control experiment showed this island is a generic phase-transition feature, not framework-specific. The island explains **where** development can occur. It does not explain **why** development occurs there rather than mere fluctuation. This paper supplies the missing mechanism: an information-constrained sorting agent — the developmental demon — that discriminates useful from harmful perturbations, selectively retains favorable configurations, and pays the thermodynamic cost of that ordering work. We ground the demon in the Szilard-Landauer-Bennett lineage of information thermodynamics, where measurement converts entropy into work but memory erasure incurs an irreducible cost of $kT \ln 2$ per bit (Landauer, 1961; Bennett, 1982; Bérut et al., 2012). Four gate parameters characterize the demon: discrimination accuracy χ (sorting capacity), energetic budget ε (the dissipation the system can afford), consolidation fidelity μ (the ability to convert transient exploration into durable capacity), and scope alignment γ (whether local sorting serves system-level function). A background substrate — degeneracy D (Edelman & Gally, 2001) — provides the structurally diverse repertoire from which the demon selects. We show that each parameter has well-characterized biological instantiations: kinetic proofreading (Hopfield, 1974) as molecular χ, the thermodynamic uncertainty relation (Barato & Seifert, 2015) as the χ-ε constraint surface, and synaptic tagging and capture (Frey & Morris, 1997) as the neural $\chi \to \mu$ two-step ratchet. A failure-mode taxonomy maps eight distinct pathologies — from underfitting to overfitting, from energy depletion to unguided dissipation, from amnesia to malignant lock-in — onto the four gate parameters. The demon is not a homunculus. It is any feedback-and-memory process that rectifies variation: filtering noise, retaining signal, and exporting entropy. A computational instantiation on Kauffman NK landscapes confirms three predictions: (1) degeneracy D is a genuine prerequisite — sorting fails on substrates without structurally distinct configurations; (2) the ratio μ/σ (retention rate to noise rate) is the critical threshold — when the ratchet leaks, all other parameters are inert; (3) scope alignment γ separates development from competent pathology — a misaligned demon achieves high local fitness at the cost of system-level gain. Metastability creates accessible variation; stress supplies perturbation; sorting converts variation into development; recovery stores the selected gain.

Keywords: Maxwell's demon, information thermodynamics, sorting, Landauer's principle, kinetic proofreading, degeneracy, synaptic tagging, ratchet, metastability, scope alignment, complex adaptive systems

---

1 The Gap

A companion paper (tratium, 2026) proposed a two-parameter phase diagram for development in complex adaptive systems: an anti-closure index α (resistance to premature synchronization) and a stress load σ (perturbation intensity) jointly determine whether a system stagnates, develops, or fragments. Three computational experiments supported the framework's topology — a bounded metastability island, a coupling-dependent surface, and a deterministic anti-resonance hierarchy — but a control experiment revealed that the metastability island is a generic phase-transition crossover signature, present even in systems with zero frequency heterogeneity. Anti-closure modulates the island's position and intensity but does not create it.

This result exposes a gap. The companion paper maps **where** in parameter space development is possible — the metastable regime between rigid lock-in and chaotic dissolution. But it does not explain **why** a system in that regime should develop rather than merely fluctuate. Boiling water at the phase transition exhibits a metastability peak. It does not learn, adapt, or develop higher-order capacity.

The gap is not merely conceptual. England (2013, 2015) showed that driven systems can self-organize into states that are comparatively good at absorbing and dissipating environmental work — dissipation-driven adaptation. This helps explain **why** structure emerges in far-from-equilibrium systems but not **why** that structure improves functionally over time. A whirlpool is well-organized dissipation. It is not development.

Development requires a mechanism that does three things fluctuation alone cannot: (1) **discriminate** — distinguish configurations that increase functional capacity from those that do not; (2) **retain** — store the selected configuration against the thermal tendency to forget; and (3) **pay** — export the entropy generated by these ordering operations. This mechanism is the sorting agent — a feedback-and-memory process that rectifies variation into directional gain. In the thermodynamic tradition, it is Maxwell's demon. In the biological tradition, it is kinetic proofreading, immune discrimination, neural selection, and developmental canalization. In the framework proposed here, it is the missing layer between landscape and development.

2 The Demon: Information Thermodynamics

2.1 From Maxwell to Landauer

Maxwell (1871) proposed a thought experiment: a being that could observe and sort individual molecules, separating fast from slow, would decrease entropy without doing macroscopic work — apparently violating the second law. Szilard (1929) formalized the demon as a one-molecule engine, showing that a single measurement on a one-molecule gas, followed by appropriate feedback, can extract $kT \ln 2$ of work per bit of information gained. The demon's measurement **enables** the sorting act: it converts ignorance into actionable information and actionable information into work.

Landauer (1961) identified the cost. Logically irreversible operations — many-to-one maps, of which erasure is the paradigm case — must dissipate at least $kT \ln 2$ per bit. This is not an engineering limitation but a fundamental bound: the entropy the demon removes from the sorted

system reappears as heat when the demon's memory is reset. Bennett (1982) refined this: measurement itself can in principle be thermodynamically reversible, but erasure of the measurement record — necessary to reset the demon for the next sorting cycle — is irreducibly dissipative. The demon's bottleneck is memory management, not perception.

Bérut et al. (2012, *Nature*) verified Landauer's bound experimentally using a colloidal particle in a double-well potential, confirming that the average heat dissipated during one-bit erasure converges to $kT \ln 2$ from above as the erasure protocol slows. Toyabe et al. (2010, *Nature Physics*) demonstrated the converse: a feedback-controlled colloidal particle climbing a spiral-staircase potential converted measurement information into directed mechanical work, validating the generalized Jarzynski equality with feedback (Sagawa & Ueda, 2010).

Parrondo, Horowitz, and Sagawa (2015, *Nature Physics*) synthesized these results into the modern thermodynamics of information. The generalized second law with feedback reads: the extractable work from a system is bounded by the free energy change plus kT times the mutual information gained by measurement. Information gained by the demon can be converted to work, but only up to the Landauer cost of erasing that information. Sorting is not free. It is thermodynamically constrained, experimentally verified, and precisely accountable.

2.2 The Demon Is Not a Homunculus

The developmental demon does not require agency, consciousness, or intention. It is any physical process that implements the measurement-feedback-retention cycle: observe a gradient, discriminate states, act on the discrimination, store the result, and export the entropy. Astumian (1997) captures the operational definition in the context of molecular motors: "to rectify the noise — to filter out the randomness you do not want so that you are left with what you do want." A Brownian ratchet is a demon. A kinetic proofreading step is a demon. A synaptic tag is a demon. An immune receptor sorting self from non-self is a demon. None require awareness. All sort. All pay.

Boël, Danot, de Lorenzo, and Danchin (2019) documented that minimal genomes encode a "wealth of functions that dissipate energy in an unanticipated way" — degradation enzymes (proteases, RNases) whose primary role is to destroy molecular species that have become informationally obsolete. These are Maxwell-demon-like functions in living cells: they sort useful molecules from useless ones and pay the ATP cost of erasure. The cell's proteases are the sorting machinery; ATP is the energy budget; substrate specificity is the discrimination parameter.

3 Three Parameters and One Substrate

3.1 χ: Discrimination Accuracy

The sorting capacity χ measures the demon's ability to distinguish useful perturbations from noise and route the system's response accordingly. Low χ: the system cannot tell signal from noise and wanders randomly through the metastable landscape. High χ: the system discriminates with precision but risks overfitting — classifying noise as signal, suppressing genuine novelty, becoming rigid.

The molecular paradigm is kinetic proofreading (Hopfield, 1974; Ninio, 1975). The ribosome does not select amino acids by equilibrium binding affinity alone — it spends an extra energy-consuming step to eject incorrect substrates after an initial selection step, amplifying the discrimination ratio. In the idealized Hopfield scheme, the error rate scales as $e^{-2\Delta G/kT}$ instead of $e^{-\Delta G/kT}$. This is the same class of dissipation-precision tradeoff later formalized by Barato and Seifert at the level of nonequilibrium biomolecular processes.

At the neural level, signal detection theory provides the formalism: d' (sensitivity) measures the demon's ability to discriminate signal from noise, while the criterion β determines the bias between false positives and false negatives. Edelman's neural Darwinism (1987, 1993) frames the entire brain as a sorting demon operating through three stages: developmental selection creates a diverse repertoire, experiential selection strengthens useful groups and weakens others, and reentrant mapping coordinates the selected groups. Stage 1 generates diversity. Stage 2 is χ — the discrimination act. Stage 3 is the bridge to μ.

At the immune level, Janeway's pattern recognition receptors (1989) provide hardcoded χ — germline-encoded discrimination of conserved pathogen signatures. Matzinger's danger model (2002) reframed the sorting criterion from identity (self vs. non-self) to context (dangerous vs. safe), showing that the demon's discrimination criterion itself can evolve toward greater sophistication.

3.2 ε: Energetic Budget

The energetic budget ε measures the dissipation the system can afford for sorting, repair, memory maintenance, and entropy export. Landauer's bound ($kT \ln 2$ per bit of erasure) sets the floor; the thermodynamic uncertainty relation (Barato & Seifert, 2015) sets the ceiling on what precision that budget can buy. Reliable sorting in nonequilibrium systems requires entropy production: higher precision in steady-state currents demands more dissipation. The demon cannot achieve high χ without spending ε.

Biologically, ε maps to metabolic capacity — ATP availability, mitochondrial function, caloric surplus, the energetic margin between maintenance costs and total throughput. A system under caloric deficit can maintain basic homeostasis but cannot afford the energetic overhead of new sorting work. This is why metabolic crisis (starvation, mitochondrial dysfunction, chronic inflammation consuming the energy budget) impairs development at every scale: the demon has no fuel.

England's dissipation-driven adaptation (2013, 2015) shows that driven systems can self-organize into states that absorb and dissipate environmental work unusually well. But dissipation without sorting is mere structure, not development. The energy budget ε is what the demon **spends** to impose direction on the dissipative flow — to select which structured states are retained and which are discarded.

3.3 μ: Consolidation Fidelity

The consolidation fidelity μ measures the system's ability to convert transient metastable exploration into durable capacity gain. High μ: selected configurations are retained against thermal fluctuation and subsequent perturbation. Low μ: gains evaporate; the system forgets what it

learned. Excessive μ: the system locks into retained configurations and cannot explore further — the ratchet holds the wrong state.

The neural paradigm is synaptic tagging and capture (Frey & Morris, 1997). Weak synaptic stimulation creates a transient "tag" — early long-term potentiation lasting less than three hours. Strong stimulation triggers protein synthesis (plasticity-related proteins, PRPs). A weakly tagged synapse can be "captured" — converted to long-lasting late LTP — only if PRPs are made available, typically by strong stimulation of a different input to the same neuron. The tag is the sorting step (χ marks a synapse as "interesting"). The PRPs are the energy cost (ε — protein synthesis requires ATP). Capture is the retention step (μ — the tagged synapse is consolidated into long-term storage). Crucially, tagging and capture are **independently controlled**: the system can sort without retaining, or have retention machinery available without sorting, and permanent change occurs only when both coincide. This two-step architecture prevents the demon from retaining everything it notices — a necessary filter against the pathology of excessive μ.

At the developmental scale, Waddington's canalization (1942, 1953) provides the framework: development tends to become buffered against perturbation, so that the phenotype is more stable than the genotype. Genetic assimilation — an environmentally induced phenotype becoming genetically fixed under sustained selection — is the ratchet completing: what began as transient exploration becomes heritable structure. The Baldwin effect (1896) extends this across generations: organisms that can learn to cope with novel environments survive long enough for genetic variants that hardcode the learned behavior to be selected. Transient sorting (χ) today becomes permanent capacity (μ) tomorrow.

3.4 D: Degeneracy as Substrate

The first three gate parameters (χ, ε, μ) operate on a substrate: the structurally diverse repertoire from which the demon selects. Edelman and Gally (2001, *PNAS*) defined degeneracy as "the ability of elements that are structurally different to perform the same function or yield the same output" — and distinguished it from redundancy (same structure, same function). Degeneracy is ubiquitous: the genetic code, immune repertoire, neural circuits, and metabolic pathways all exhibit it. Whitacre and Bender (2010) showed that degeneracy simultaneously provides robustness (if one element fails, structurally different elements compensate), evolvability (degenerate elements can be co-opted for novel functions), and complexity (non-degenerate systems cannot integrate at scale).

Degeneracy D is not a gate the demon operates — it is the raw material the demon operates **on**. Without degeneracy, χ has nothing to discriminate among: all elements are identical (redundancy) or all are unique with no functional overlap. Degeneracy creates a rich selection surface: many structurally different ways to achieve the same function, each with different latent capacities. When the demon selects one configuration, the degenerate elements freed from their prior role become available for co-option — this is how sorting creates **new** developmental possibilities rather than merely refining existing ones. Tononi, Sporns, and Edelman (1999) provided information-theoretic measures of degeneracy in neural networks, confirming that high degeneracy correlates with functional integration and complexity.

The relationship between D and α (anti-closure, from the companion paper) requires explicit treatment. α is **dynamic** anti-closure — the system's resistance to current synchronization, measured moment-to-moment. D is **structural** anti-closure — the diversity of the repertoire available for selection, a slower-moving property of the system's architecture. They are coupled: high-α dynamics tend to preserve D by preventing winner-take-all capture of the repertoire; high-D substrates tend to support high-α dynamics by giving the system enough distinct configurations that no single attractor can dominate. They are kept as separate parameters here because they operate on different timescales and are independently measurable — α through frequency heterogeneity and metastability indices, D through degeneracy measures of the kind Tononi, Sporns, and Edelman developed. But they are not independent: any intervention that collapses D (eliminating structural diversity) will eventually reduce α (dynamic flexibility), and any intervention that locks α at zero (forced synchronization) will erode D (winner-take-all dynamics deplete the repertoire). This coupling is the formal version of a familiar claim: monocultures are fragile.

A fourth gate parameter is required: **scope alignment** γ, measuring whether the demon's sorting criterion is aligned with the system-level objective function. The cancer-as-scope-collapse failure mode (§5) demonstrates that a demon can sort accurately, energetically, and durably toward the wrong target — competent pathology, not development. A high-χ, high-ε, high-μ demon with $\gamma = 0$ is a paperclip maximizer: it sorts excellently for goals that destroy the system it inhabits. Without γ, the model cannot distinguish development from tumor, cult, bureaucracy, or misaligned AI. **γ is not optional.** It is the parameter that gives the framework normative teeth: the difference between a forge and an incinerator is not temperature — it is whether the heat serves the blade.

The variable hierarchy is therefore:

Layer	**Variables**	**Source**
Landscape	α (anti-closure), σ (stress), κ (coupling), τ (recovery)	The Wounded Forge
Substrate	D (degeneracy / repertoire)	Edelman & Gally
Demon (gates)	χ (discrimination), ε (energy), μ (retention), γ (scope alignment)	This paper

Table 1: The complete variable hierarchy. The landscape determines where development is possible. The substrate determines what the demon has to work with. The demon converts metastable exploration into directional gain.

3.5 Parameter Tradeoffs

The parameters are not independent sliders. They form tradeoff surfaces: increasing χ raises ε demand (the Barato-Seifert bound); increasing μ reduces future α unless pruning or reset occurs (retained configurations resist the anti-closure that permitted their selection); increasing D raises the search space and may require χ to scale accordingly (the demon can drown in options if repertoire diversity outpaces discrimination capacity). The third tradeoff is underappreciated: degeneracy is not purely beneficial. Excessive D without commensurate χ produces combinatorial fog — a system with many possible configurations and no ability to choose among them.

3.6 Recovery as Demon Reset

The recovery parameter τ from the companion paper integrates with μ more tightly than either paper initially specified. μ is retention fidelity; τ is the temporal window in which retention becomes adaptive rather than pathological. Too little τ: no consolidation. Too much τ with high μ: stagnation and lock-in. Adequate τ: consolidation plus pruning — the system retains what was selected and clears what was not.

Sleep, autophagy, ritual rest, institutional review, sabbath, refractory periods — these are not passive downtime. They are **demon reset protocols**: active maintenance of the sorting agent's capacity to sort accurately in the next cycle. Tononi and Cirelli's synaptic homeostasis hypothesis (2006) frames sleep as exactly this: wakefulness is the demon's sorting shift (χ active, ε being spent, μ tagging synapses); sleep is the reset (μ selectively consolidating, pruning untagged connections, restoring α by decoupling overconsolidated networks). The demon that never rests becomes the demon that locks in: μ without τ-mediated pruning is how institutions become bureaucracies and how trauma becomes chronic.

3.7 Operational Definition and Falsification

A sorting demon is operationally present in a system only when four conditions are measurably met: (1) state discrimination — the system responds differently to different input states; (2) state-dependent action — the discrimination alters the system's subsequent behavior; (3) retention or memory — the discrimination-action pair persists beyond the immediate stimulus; and (4) entropy export — the ordering work produces measurable dissipation. A system that meets all four is implementing the demon. A system that meets fewer is performing feedback control, homeostasis, or passive response, but not developmental sorting.

To ground the framework operationally, each domain requires at least two measurable proxies for the demon's presence:

Domain	**χ proxy**	**μ proxy**
Molecular	Proofreading error rate	Protein half-life
Immune	d' (sensitivity) of T-cell receptor discrimination	Memory cell persistence
Neural	Signal detection d'; neuronal avalanche branching ratio	Late LTP / synaptic consolidation markers
Morphogenetic	Bioelectric pattern specificity (Levin)	Canalization depth (Waddington landscape)
Organizational	Error detection rate; audit accuracy	Institutional memory; policy persistence

Table 2: Domain-specific proxies for the demon's presence. Each row requires measurable state-discrimination (χ) and retention (μ). Without both, the system is performing feedback, not developmental sorting.

The falsification criterion is direct: if a system in the metastable regime shows no measurable state-discrimination, no state-dependent action, no retention, and no entropy export, and yet produces

directional capacity gain, the demon framework is wrong. The claim is that development requires sorting, not merely fluctuation.

4 Biological Instantiation Across Scales

The sorting demon is not a single mechanism but a **hierarchy of rectification processes** operating at every biological scale. Each level inherits the sorting done at the level below, and each expands the cognitive light cone (Levin, 2019, 2022) — the scale of goals the system can represent and pursue.

At the **molecular** level, kinetic proofreading (Hopfield, 1974) and the ubiquitin-proteasome system are the sorting machinery. The ribosome discriminates correct from incorrect amino acids; the proteasome degrades proteins that have been tagged as obsolete or misfolded. Both spend ATP (ε) to achieve discrimination (χ). Both are selective retention systems (μ): the ribosome retains the correctly assembled protein, the proteasome retains the cell's informational hygiene by destroying what no longer fits.

At the **cellular** level, Levin's TAME framework (2022) shows that cells use bioelectric signaling — voltage gradients across gap junction networks — to share physiological state and coordinate collective computation. Individual cells have small cognitive light cones; multicellular cooperation expands the light cone to tissue-level and organism-level goals. Cancer, in this framework, is the demon losing its collective context: the cell retains its sorting machinery but disconnects from the collective objective function (Levin, 2019). It sorts for cell-level survival when it should be sorting for tissue-level coherence. This is a **scope failure** of χ, not a capacity failure.

At the **immune** level, the sorting hierarchy spans innate (hardcoded χ: pattern recognition receptors discriminating conserved pathogen signatures) and adaptive (learnable χ: clonal selection, affinity maturation, memory cells). The immune system's μ is immunological memory — the retention of selected clones for rapid reactivation. Trained immunity via BCG vaccination (Arts et al., 2018) demonstrates that even innate χ can be epigenetically reprogrammed — a form of μ at the chromatin level.

At the **neural** level, Edelman's neuronal group selection (1987, 1993) provides the complete three-stage demon: generate diversity (developmental selection $\rightarrow D$), discriminate (experiential selection $\rightarrow \chi$), retain/coordinate (reentrant mapping $\rightarrow \mu$). Synaptic tagging (Frey & Morris, 1997) supplies the molecular mechanism for the $\chi \rightarrow \mu$ conversion: transient discrimination is consolidated into durable synaptic change only when the energy budget (ε — protein synthesis) is simultaneously available.

5 Failure Modes

The sorting demon has eight distinct failure modes, each mapped to a specific parameter and direction. These failure modes are not metaphors — they correspond to well-characterized pathologies across biological, neural, immune, and organizational systems.

Failure mode	Parameter	Direction	Examples
Underfitting	χ too low	Under-sorting	Immune deficiency, permissive culture
Overfitting	χ too high	Over-sorting	Autoimmunity, apophenia, censorship, bureaucracy
Criterion bias	χ threshold	Biased sorting	Paranoia, false positives, autoimmune threshold error
Scope collapse	γ too low	Wrong-level sorting	Cancer, cult, bureaucracy, misaligned AI
Energy depletion	ε too low	Cannot sort	Metabolic crisis, exhaustion, system cannot afford repair
Dissipation without direction	ε without χ	Unguided spending	Runaway inflammation, chaotic energy discharge
Amnesia	μ too low	Cannot ratchet	Gains evaporate, cultural forgetting, no consolidation
Malignant lock-in	μ too high	Cannot explore	Trauma memory, epigenetic lock, institutional rigidity

Table 3: Seven failure modes of the developmental demon, mapped to parameter and direction. Each corresponds to well-characterized pathologies.

Four failure modes deserve special attention because they reveal the demon's structural shadow.

Overfitting is the demon sorting too finely — classifying noise as signal, losing the ability to generalize. In machine learning, the bias-variance tradeoff formalizes this: too little discrimination (high bias) produces underfitting; too much (high variance) produces overfitting. Belkin et al. (2019) documented "double descent" — very large models can generalize **beyond** the classical overfitting peak, suggesting a phase transition in the demon's behavior at extreme χ relative to data complexity. In signal detection theory, overfitting corresponds to a liberal criterion (low β): the demon detects everything, including what isn't there. Apophenia — the tendency to perceive meaningful patterns in random data — is the perceptual version (Blain et al., 2020).

Dissipation without direction is the demon spending energy without sorting — ε active, χ absent. England's dissipation-driven adaptation (2013, 2015) shows that driven systems generically self-organize into states that absorb and dissipate environmental work well. This is structure without development: the whirlpool. A cell running inflammatory cascades that consume ATP without discriminating threat from self is a biological instance — the energy budget is being spent, the immune machinery is active, but the sorting criterion has collapsed. An organization that spends lavishly on "innovation" without any selection mechanism for which innovations to pursue is the institutional instance. Dissipation without direction is what distinguishes this framework from England's: England explains why driven systems find structured dissipative states; the demon explains why some of those states develop rather than merely dissipate. The whirlpool is the canonical counter-example — organized, energetic, and going nowhere.

Cancer as scope collapse is the demon sorting at the wrong level. The cancer cell retains its molecular sorting machinery — it proofreads, repairs, discriminates — but it has lost connection to the tissue-level objective function maintained by the bioelectric network (Levin, 2019). It optimizes for cell-level survival when it should be optimizing for organism-level coherence. Normalizing the bioelectric microenvironment can revert cancer cells to normal behavior (Chernet & Levin, 2013) — restoring the demon's collective scope without modifying its local machinery.

Malignant lock-in is excessive μ without ongoing χ. The system retains old sorting patterns but stops discriminating whether those patterns are still useful. Institutional bureaucracy is the organizational instantiation: sorting procedures that initially served the collective function become self-perpetuating because they protect the sorters' position, not because they serve the mission. The demon captures the system it was designed to serve. At the biological level, pathological epigenetic lock-in — where stress-induced chromatin modifications persist beyond the period of their adaptive value — produces chronic maladaptive phenotypes. The ratchet holds, but it holds the wrong configuration.

6 The Ratchet: How Exploration Becomes Capacity

The developmental ratchet is the mechanism by which the demon's transient sorting is converted into durable capacity gain. It requires the coincidence of three conditions: the demon has discriminated a useful configuration (χ), the energy to consolidate is available (ε), and the retention machinery is engaged (μ). Without all three, the ratchet does not advance: the system sorts without retaining, retains without sorting, or sorts and retains but cannot afford the consolidation.

Feynman's analysis of the ratchet and pawl (1963) establishes the thermodynamic baseline: a microscopic ratchet in thermal equilibrium cannot rectify fluctuations. Directed transport requires being out of equilibrium — a temperature gradient or energy input. Astumian (1997) showed how nonequilibrium fluctuations in an anisotropic medium can bias Brownian motion to produce directed transport without macroscopic forces. Molecular motors (kinesin, myosin, ATP synthase) implement this: they convert random thermal fluctuation into directed mechanical work by coupling asymmetric potentials to nonequilibrium energy sources.

At the developmental scale, the ratchet operates through Waddington's canalization: developmental trajectories become progressively buffered against perturbation. Once the demon has sorted the system into a canal, subsequent fluctuations cannot easily deflect it. Genetic assimilation completes the ratchet across generations: environmentally induced phenotypes become genetically encoded under sustained selection. The Baldwin effect (1896) is the inter-generational sorting demon: learning (χ in real time, costing ε) enables survival; over generations, genetic variants that hardcode the learned capacity are selected (μ at the genomic level).

The neural tagging-and-capture mechanism (Frey & Morris, 1997) provides the most precisely characterized biological ratchet. The two-step architecture — tag first, capture later, only when protein synthesis is available — ensures that not every noticed event becomes a permanent memory. This gating prevents the pathology of excessive μ: the demon sorts freely, but the ratchet advances only when the conjunction of significance (χ), energy (ε), and molecular consolidation machinery (μ) is met. Sleep may be the recovery phase (τ) during which the ratchet completes: Tononi

and Cirelli's synaptic homeostasis hypothesis (2006) proposes that wakefulness increases synaptic strength (the demon sorts all day) while sleep globally downscales synapses (selectively retaining what was tagged during waking).

The capacity gain Δc after a sorting-recovery cycle is therefore a function of all parameters:

$$\Delta c > 0 \quad \text{only if} \quad \begin{cases} m(\alpha, \sigma, \kappa) > m_{\text{threshold}} & \text{(metastable exploration sufficient)} \\ \chi_{\min} < \chi < \chi_{\max(D,\varepsilon)} & \text{(discriminates without overfitting)} \\ \varepsilon > \varepsilon_{\min(\chi)} & \text{(energy budget covers sorting cost)} \\ \mu_{\min} < \mu < \mu_{\max(\tau,\alpha)} & \text{(retains without locking in)} \\ \gamma > \gamma_{\min} & \text{(local sorting aligned with system-level function)} \\ \tau \geq 1 & \text{(recovery time adequate for consolidation and pruning)} \\ D_{\min} < D < D_{\max(\chi,\varepsilon)} & \text{(repertoire sufficient, not overwhelming)} \end{cases}$$

This is the complete growth condition: landscape (α, σ, κ), substrate (D), demon (χ, ε, μ), and recovery (τ) all within bounds simultaneously. Development occurs in the intersection of these constraints — a bounded region in high-dimensional parameter space. Every real failure of development corresponds to one or more conditions not being met.

7 Discussion

7.1 Relation to the Companion Paper

The companion paper (tratium, 2026) mapped the landscape of development — the $\alpha \times \sigma$ phase diagram, the metastability island, the nine-cell failure-mode matrix. Its control experiment showed the island is a generic crossover feature, present even without anti-closure. This paper explains what makes the island developmental: the sorting demon. Without the demon, the island is just a region of enhanced fluctuation. With the demon, fluctuation is rectified into directional gain.

The two papers have complementary weaknesses. The companion paper has computational evidence from Kuramoto and circle-map experiments but no mechanistic explanation for why metastability should produce development. This paper has the mechanism (the demon, grounded in information thermodynamics) but no new computational evidence — the Szilard-Landauer-Bennett lineage is established physics, and the biological instantiations are drawn from published literature. The natural next step is a computational model that combines both: a Kuramoto system with a feedback-controlled sorting agent, measuring whether the demon-equipped system shows directional capacity gain in the metastable regime while the demon-free system merely fluctuates.

7.2 Relation to Friston's Free Energy Principle

Friston's free energy principle (2010) proposes that living systems minimize variational free energy — a bound on surprise — through perception, action, and model updating. The present framework should not be read as a refutation of FEP. Active inference already includes epistemic foraging: organisms can seek information-rich states when doing so improves their model. The difference is emphasis. FEP foregrounds model evidence and prediction-error management; the

present framework foregrounds the thermodynamic and biological gating conditions under which perturbation becomes developmental rather than merely destabilizing.

In FEP terms, the sorting demon can be interpreted as the machinery that decides which prediction errors are noise to suppress, which are signals to learn from, and which are too costly to integrate. χ corresponds to discrimination of informative from uninformative error, ε to the energetic cost of updating and maintaining the model, and μ to the retention of model revisions. This makes the frameworks complementary rather than directly opposed.

7.3 Computational Instantiation

To test the framework's core predictions, we implemented a sorting demon on two substrates: Kuramoto coupled oscillators (the substrate used in the companion paper's metastability analysis) and Kauffman's NK model (Boolean networks with tunable epistasis). The experiments produced one negative and one positive result, both informative.

Kuramoto (negative). A rollout-based demon was applied to $N = 200$ Kuramoto oscillators at the metastability ridge ($\Delta\omega = 1.92$, $\eta = 1.13$) from the companion paper. The demon evaluated proposed frequency perturbations via short forward simulations, accepted or rejected via a Metropolis rule, and retained gains via a buffer-and-pull ratchet. Across 20 trials $\times$ 7 conditions $\times$ 3 target functions, the demon produced **no directional fitness gain** in any condition. Diagnostic analysis revealed the root cause: rollout signal-to-noise ratio ≈ 0.14. Stochastic noise in the evaluation window ($\sigma \approx 0.035$) was 7 $\times$ larger than the fitness signal from frequency perturbations ($\Delta F \approx 0.007$). Even deterministic evaluation ($\eta = 0$ during rollout) yielded $|\Delta F| \approx 0.0003$ — below any reasonable discrimination threshold. The structural diagnosis: Kuramoto oscillators lack degeneracy ($D \approx 0$). All oscillators are identical except for their natural frequency; the order parameter r is a smooth mean-field function of the aggregate frequency distribution. Small local perturbations do not create qualitatively distinct configurations for the demon to sort among. This negative result directly validates D as a genuine prerequisite for sorting, not a decorative addition to the framework.

***NK* model (positive, conditional).** The NK landscape ($N = 20$ genes, $K = 4$ epistatic neighbors) provides tunable degeneracy: structurally different genotypes can produce equivalent fitness through distinct pathways. Fitness evaluation is exact — no stochastic rollout required — so the demon can see the effect of every proposed mutation with perfect clarity. In the zero-noise condition (mutation rate = 0), the aligned demon produced strong directional gain: $\Delta F = +0.093$ ($p < 0.01$, 150 trials across 5 landscapes), climbing from random initial fitness (≈ 0.27) to near-optimum (≈ 0.72). However, at **any** background mutation rate above ≈ 0.01, the demon failed — fitness gain went immediately negative. The transition was razor-sharp: the retention mechanism (per-gene reversion probability $= 0.02$) could not counteract even modest thermal noise. Six parameter sweeps (K, β, E, μ, γ, and mutation rate) confirmed that when retention rate $<$ noise rate, **all other parameters are inert**. No amount of discrimination sharpness, energy budget, retention strength, or scope alignment could compensate for a leaky ratchet.

γ validation. In the zero-noise regime where the demon works, aligned and misaligned demons were compared directly. The aligned demon (optimizing global fitness) achieved $\Delta F = +0.093$. The misaligned demon (optimizing fitness of a gene subset — the NK analog of cancer or cult)

achieved **higher target fitness** ($\Delta F_{\text{target}} = +0.104$) but only $\Delta F_{\text{global}} = +0.045$ — 48% of the aligned demon's system-level gain. The misaligned demon sorts excellently for the wrong objective: competent agency optimizing for the wrong substrate. This is the formal signature the framework predicts for scope collapse, and it confirms that γ is not reducible to the other gate parameters.

Summary of computational findings:

1. D (degeneracy) is a prerequisite, not a parameter: without structurally distinct configurations, sorting has nothing to sort among (Kuramoto $\rightarrow$ NK).
2. The μ/σ ratio — retention rate relative to noise rate — is the critical threshold for developmental sorting. When the ratchet leaks faster than sorting accumulates, all other parameters are inert.
3. γ (scope alignment) separates development from competent pathology. A demon with $\gamma = 0$ produces high local fitness at the expense of global fitness — the formal version of the claim that cancer, cult, bureaucracy, and misaligned AI share a common structure.

The interaction between μ and environmental perturbation (σ) is sharper than the framework's current treatment suggests. The paper formulates μ as a property of the demon (consolidation fidelity); the experiments show that the **rate** of retention relative to the **rate** of perturbation is what matters. This points toward a reformulation: the effective retention parameter is not μ alone but μ/σ, and development requires $\mu/\sigma > 1$.

7.4 Limitations

The computational instantiation supports the framework's qualitative predictions but with important caveats. The demon succeeded only in the zero-noise regime of the NK model — the retention mechanism was too weak to hold gains against any background mutation rate above 0.01. Whether this reflects a fundamental limitation (the μ/σ threshold) or a limitation of the specific retention implementation (probabilistic gene-by-gene reversion) remains open. Biological retention mechanisms — synaptic consolidation, epigenetic modification, immune memory — are structurally richer than a simple buffer-and-pull ratchet. Testing whether more sophisticated retention crosses the μ/σ barrier at realistic noise levels is the primary computational next step.

The four gate parameters (χ, ε, μ, γ) have not been measured simultaneously in a single biological system. Each has domain-specific instantiations (kinetic proofreading for molecular χ, synaptic tagging for neural $\chi \rightarrow \mu$, HRV for organismic state), but no experiment has shown their interaction produces directional development. The NK simulations demonstrate the **logic** of the interaction computationally; demonstrating it empirically is the primary experimental next step.

The failure-mode taxonomy is the paper's most immediately testable contribution: each mode predicts specific signatures (underfitting produces random trajectories in morphospace, overfitting produces rigid trajectories, scope collapse produces high local coherence with poor global outcomes, etc.) that could be distinguished in the Levin lab's morphogenetic systems or in neural-network training dynamics.

The D (degeneracy) variable is positioned as a background condition rather than a gate, but the boundary between "substrate" and "parameter" is conceptual rather than principled. A system can increase its own degeneracy through exploration — in which case D becomes a dynamic variable coupled to the demon's sorting activity. This reflexivity is acknowledged but not resolved.

References

Arts, R.J.W. et al. (2018). "BCG vaccination protects against experimental viral infection in humans through the induction of cytokines associated with trained immunity." *Cell Host & Microbe* 23.1: 89–100.

Astumian, R.D. (1997). "Thermodynamics and Kinetics of a Brownian Motor." *Science* 276.5314: 917–922.

Baldwin, J.M. (1896). "A New Factor in Evolution." *The American Naturalist* 30.354: 441–451.

Barato, A.C. & Seifert, U. (2015). "Thermodynamic Uncertainty Relation for Biomolecular Processes." *Physical Review Letters* 114: 158101.

Belkin, M. et al. (2019). "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *PNAS* 116.32: 15849–15854.

Bennett, C.H. (1982). "The Thermodynamics of Computation — A Review." *International Journal of Theoretical Physics* 21.12: 905–940.

Bérut, A. et al. (2012). "Experimental verification of Landauer's principle linking information and thermodynamics." *Nature* 483: 187–189.

Blain, S.D. et al. (2020). "Apophenia as the disposition to false positives: A unifying framework for openness and psychoticism." *Journal of Abnormal Psychology* 129.3: 279–292.

Boël, G. et al. (2019). "Omnipresent Maxwell's demons orchestrate information management in living cells." *Microbial Biotechnology* 12.2: 210–242.

Chernet, B.T. & Levin, M. (2013). "Transmembrane voltage potential is an essential cellular parameter for the detection and control of tumor development." *Disease Models & Mechanisms* 6.3: 595–607.

Danchin, A. (2009). "Bacteria as computers making computers." *FEMS Microbiology Reviews* 33.1: 3–26.

Edelman, G.M. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection.* Basic Books.

Edelman, G.M. (1993). "Neural Darwinism: Selection and reentrant signaling in higher brain function." *Neuron* 10.2: 115–125.

Edelman, G.M. & Gally, J.A. (2001). "Degeneracy and complexity in biological systems." *PNAS* 98.24: 13763–13768.

England, J.L. (2013). "Statistical physics of self-replication." *Journal of Chemical Physics* 139: 121923.

England, J.L. (2015). "Dissipative adaptation in driven self-assembly." *Nature Nanotechnology* 10: 919–923.

Feynman, R.P. (1963). *The Feynman Lectures on Physics.* Vol. I, Ch. 46.

Frey, U. & Morris, R.G.M. (1997). "Synaptic tagging and long-term potentiation." *Nature* 385: 533–536.

Friston, K. (2010). "The free-energy principle: A unified brain theory?" *Nature Reviews Neuroscience* 11: 127–138.

Hopfield, J.J. (1974). "Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity." *PNAS* 71.10: 4135–4139.

Janeway, C.A. Jr. (1989). "Approaching the Asymptote? Evolution and Revolution in Immunology." *Cold Spring Harbor Symposia on Quantitative Biology* 54.1: 1–13.

Landauer, R. (1961). "Irreversibility and Heat Generation in the Computing Process." *IBM Journal of Research and Development* 5.3: 183–191.

Levin, M. (2019). "The Computational Boundary of a 'Self'." *Frontiers in Psychology* 10: 2688.

Levin, M. (2022). "Technological Approach to Mind Everywhere." *Frontiers in Systems Neuroscience* 16: 768201.

Matzinger, P. (2002). "The Danger Model: A Renewed Sense of Self." *Science* 296.5566: 301–305.

Maxwell, J.C. (1871). *Theory of Heat.* Longmans, Green, and Co.

Ninio, J. (1975). "Kinetic amplification of enzyme discrimination." *Biochimie* 57.5: 587–595.

Parrondo, J.M.R., Horowitz, J.M., & Sagawa, T. (2015). "Thermodynamics of information." *Nature Physics* 11: 131–139.

Sagawa, T. & Ueda, M. (2010). "Generalized Jarzynski Equality under Nonequilibrium Feedback Control." *Physical Review Letters* 104: 090602.

Szilard, L. (1929). "Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen." *Zeitschrift für Physik* 53: 840–856.

Tononi, G. & Cirelli, C. (2006). "Sleep function and synaptic homeostasis." *Sleep Medicine Reviews* 10: 49–62.

Tononi, G., Sporns, O., & Edelman, G.M. (1999). "Measures of degeneracy and redundancy in biological networks." *PNAS* 96.6: 3257–3262.

Toyabe, S. et al. (2010). "Experimental demonstration of information-to-energy conversion." *Nature Physics* 6: 988–992.

tratium (2026). "The Wounded Forge: Anti-Closure, Calibrated Stress, and Metastability in Complex Adaptive Systems." Preprint.

Waddington, C.H. (1942). "Canalization of Development and the Inheritance of Acquired Characters." *Nature* 150: 563–565.

Waddington, C.H. (1953). "Genetic assimilation of an acquired character." *Evolution* 7.2: 118–126.

Whitacre, J. & Bender, A. (2010). "Degeneracy: a link between evolvability, robustness and complexity." *Theoretical Biology and Medical Modelling* 7: 6.

Zurek, W.H. (1989). "Algorithmic randomness and physical entropy." *Physical Review A* 40.8: 4731–4751.