Darwin as Minimum Description Length: Selection, Variation, and Modularity as Code–Length Optimization

Jonathan Washburn

Recognition Science & Recognition Physics Institute
Austin, Texas, USA
jon@recognitionphysics.org

October 26, 2025

Abstract

We develop a methods–first theory that identifies *fitness* with negative description length. For an environment $\mathcal{E}$ (a distribution over tasks/stimuli) and an organism g (a compressor–controller with parameters θ_g), we define the *evolutionary code length*

$$L_g \;=\; L(\text{model}_g) \;+\; L(\text{errors} \mid \mathcal{E}),$$

where $L(\text{model}_g)$ is the prefix–free code length to specify the organism's internal model at the precision supported by data, and $L(\text{errors} \mid \mathcal{E})$ is the negative log–likelihood (in bits) of deviations under a preregistered noise model. We show that if replication rates obey $r(g) \propto \exp(-\beta L_g)$ for resource factor $\beta > 0$, then the replicator dynamics descend the mean code length: $d\,\mathbb{E}[L_g]/dt \le 0$, and the stationary distribution is $\pi^\star(g) \propto \exp(-\beta L_g)$. Thus, selection is *minimum description length* (MDL) at population scale.

Variation is not isotropic. We formalize an anisotropic proposal law $q(\Delta) \propto \exp(-\Delta J)$ for phenotype moves, where J is a symmetric, convex ledger cost that penalizes overhead and imbalance; this yields structured "randomness" concentrated along low–cost directions and predicts repeatable adaptive pathways. We also prove a modularity lower bound: when tasks in $\mathcal{E}$ share mutual information M, reusing a module of size b saves at least $M - b$ bits, so selection favors modular architectures whenever reuse beats overhead.

The empirical program is operational and auditable: (i) define a reference machine and measure L_g as $L(\text{model}) + L(\text{parameters to supported precision}) + L(\text{errors} \mid \text{noise})$; (ii) test the MDL–fitness link, the anisotropy law, and the modularity bound on archival datasets spanning gene regulation, metabolism, and behavior; (iii) preregister noise models, hyperparameters, and pooling rules; (iv) release a one–command reproduction bundle. Falsifiers are explicit: e.g., lineages with persistently *larger* L_g outcompeting smaller L_g under fixed resource budgets; isotropic variation contradicting the $\exp(-\Delta J)$ law; and absence of correlation between reuse and environmental mutual information. This work compresses selection, variation, and modularity into a single quantitative currency—bits—and supplies a reproducible protocol to test it with no new experiments.

Keywords: evolution; fitness; minimum description length; replicator dynamics; modularity; anisotropy; rate–distortion; model selection.

1 Introduction

Puzzle.. Biology persistently yields modular, hierarchical designs (motifs, pathways, organs) and transferable skills (behaviors, strategies that generalize across tasks). Empirically, "variation" is not isotropic: phenotypic moves recur along a few privileged directions, while many conceivable changes are rare or effectively inaccessible.

Answer.. Compression under resource constraints wins. Let $\mathcal{E}$ denote an environment (a distribution over tasks/stimuli) and let an organism g implement a compressor–controller C_g with parameters θ_g. Define the evolutionary code length

$$L_g \;=\; L(\text{model}_g) \;+\; L(\text{errors} \mid \mathcal{E}), \tag{1}$$

where $L(\text{model}_g)$ is the prefix–free code length for C_g at the data–supported precision and $L(\text{errors} \mid \mathcal{E})$ is the codelength of residuals under a preregistered noise model. Selection favors *short codes that work*: organisms that compress $\mathcal{E}$ most effectively, given energetic/recognition budgets, increase.

Contributions.. This paper makes four contributions:

1. **Formal MDL fitness.** We identify fitness with negative description length and show that replicator dynamics descend the population mean of (1).
2. **Variation anisotropy via J.** We derive an anisotropic proposal law for phenotype moves, $q(\Delta) \propto \exp(-\Delta J)$, where J is a convex, symmetric ledger cost encoding resource overhead and balance.
3. **Modularity bound.** When tasks in $\mathcal{E}$ share mutual information M, reusing a module of size b yields a codelength saving of at least $M - b$ bits; selection favors modular architectures whenever reuse beats overhead.
4. **Archival test plan and falsifiers.** We specify a preregistered, MDL–based measurement protocol on public datasets (gene regulation, metabolism, behavior) and state concrete falsification conditions (e.g., persistent outperformance by higher L_g designs under fixed budgets; isotropic variation contradicting the $\exp(-\Delta J)$ law).

2 Background

MDL in statistics and learning.. Minimum Description Length (MDL) formalizes parsimony: among models that explain the data, prefer the one minimizing total code length (model + parameters to supported precision + residuals). In finite samples, MDL aligns with penalized likelihood criteria (e.g., BIC) and with universal coding bounds from information theory. Here we treat *organisms as models* and *environments as data sources*, so fitness becomes an MDL objective.

Rate–distortion tradeoffs.. Biological agents face resource limits (metabolic, temporal, memory). Rate–distortion theory quantifies the best achievable error (distortion) at a given information rate (coding cost). Our L_g is an operational rate that must be budgeted; viable organisms lie near Pareto fronts balancing model complexity against residual error under $\mathcal{E}$.

Replicator dynamics and Fisher's theorem.. Replicator equations describe composition changes under frequency–dependent selection. Fisher's fundamental theorem relates the change in mean fitness to heritable variance. By setting reproduction rates $r(g) \propto \exp(-\beta L_g)$ with resource factor $\beta > 0$, the replicator flow implements *code–length descent*: the population mean $\mathbb{E}[L_g]$ decreases monotonically toward a stationary distribution $\pi^\star(g) \propto \exp(-\beta L_g)$.

Modularity in networks.. Across molecular, cellular, and behavioral levels, biological networks exhibit modular, hierarchical structure with motif reuse. From an MDL viewpoint, modules are reusable subroutines that amortize codelength across shared tasks; their selection follows directly when environmental tasks overlap and the reuse advantage exceeds overhead.

Evolvability and pleiotropy.. Evolvability describes a system's capacity to generate adaptive variation; pleiotropy couples multiple traits to shared genetic mechanisms. Anisotropic variation naturally arises when accessible phenotype moves are biased by a convex cost J: changes that conserve balance and minimize overhead occur with exponentially higher probability, concentrating search along low–cost directions.

Bridge to practice (one currency across levels).. Code length provides a unifying, measurable currency from genome (encoding circuits), to network (wiring and parameters), to behavior (policy and error). We will measure L_g on archival datasets via a fixed reference machine and compare organisms and baselines under the same scoring rules, enabling cross–level synthesis without changing units.

3 Definitions and Setup (operational)

Environment $\mathcal{E}$.. The environment is a probability space of tasks/stimuli and their statistics. Formally, let $(\mathcal{X}, \Sigma, \mathbb{P}_{\mathcal{E}})$ denote sensory input streams (including exogenous variables and task labels), and let performance functionals be evaluated under draws $x_{1:T} \sim \mathbb{P}_{\mathcal{E}}$. Empirical frequencies (task mix, context durations, noise levels) determine the weights used for pooling scores across tasks.

Organism g.. An organism is modeled as a *compressor–controller* C_g with parameters θ_g, mapping sensory histories to internal states and actions:

$$C_g : \ (\mathcal{X}^{\leq t}, \text{memory}) \longrightarrow (\text{action}_t, \text{updated memory}),$$

subject to homeostasis and reproduction constraints. Architectural choices (modules, wiring, dynamics) and numeric parameters (thresholds, gains, kinetic rates) are part of θ_g.

Description length L_g.. The evolutionary code length is the total codelength, in bits, required to specify C_g and its predictive residuals under $\mathcal{E}$ on a fixed reference machine:

$$L_g \ = \ L(\text{model}_g) \ + \ L(\text{errors} \mid \mathcal{E}). \tag{2}$$

Model code. $L(\text{model}_g)$ counts the prefix–free codelength to describe the structure (modules, connections, update rules) and parameters θ_g at the precision supported by data. Parameter precision is set by interval coding from documented uncertainties or cross–validated tolerances; a parameter with tolerance width δ contributes $\approx \log_2(1/\delta)$ bits.
Error code. $L(\text{errors} \mid \mathcal{E})$ is the negative log–likelihood (or loss code) of deviations between C_g's predictions and observations drawn from $\mathbb{P}_{\mathcal{E}}$, evaluated under a preregistered noise/perturbation model (Gaussian/covariance, Poisson, or a declared kernel). In all cases, residual codes are expressed in bits via the corresponding log–likelihood.

Resource factor β.. Resource scarcity (metabolic, temporal, memory/recognition) is summarized by a positive scalar $\beta > 0$. Reproduction rates are modeled as

$$r(g) \ \propto \ \exp(-\beta\, L_g), \tag{3}$$

so that higher β tightens MDL pressure: at fixed performance, organisms with shorter codes replicate faster.

Ledger cost J.. Variation proposals are biased by a symmetric, convex *ledger cost* J on phenotypic moves. For a proposed change Δ in phenotype coordinates,

$$q(\Delta) \;\propto\; \exp(-\Delta J), \tag{4}$$

with J unique up to an overall scale and minimized at the balanced operating point. Convex symmetry ($J(x) = J(x^{-1})$ in appropriate coordinates) penalizes imbalance and high overhead, concentrating accessible variation along low–cost directions. The consequence is *anisotropic variation*: "randomness" is structured inside iso–J shells rather than isotropic in phenotype space.

Equations (2)–(4) define the operational quantities used throughout: given $\mathcal{E}$, organisms are scored by L_g; selection applies via (3); and the geometry of accessible variation follows (4).

4 Core Theorems and Propositions (statements; proofs in appendices)

[Replicator–MDL Equivalence] Let π_t be the population distribution over organisms g with evolutionary code length L_g (defined in (2)). Suppose per–capita replication rates satisfy

$$r(g) \;\propto\; \exp(-\beta\, L_g) \qquad (\beta > 0),$$

and the mean–field population dynamics are governed by the replicator equation

$$\dot{\pi}_t(g) \;=\; \pi_t(g)\,\Big(\mathbb{E}_{\pi_t}[L] \;-\; L_g\Big),$$

after a rescaling of time by β.[1] Then:

(i) The population mean code length is a Lyapunov functional:

$$\frac{d}{dt}\,\mathbb{E}_{\pi_t}[L] \;=\; -\mathrm{Var}_{\pi_t}(L) \;\le\; 0,$$

with equality iff L_g is π_t–a.s. constant.

(ii) If, in addition, a small mutation/diffusion operator preserves absolute continuity and satisfies detailed balance with respect to Lebesgue measure on the type space, then the unique stationary density has Gibbs form

$$\pi^\star(g) \;\propto\; \exp(-\beta\, L_g).$$

Interpretation. Selection implements *code–length descent*. With mild mutation, the stationary population concentrates according to a Boltzmann weight in L_g.

[Rate–Distortion Fitness] Fix a coding rate budget R (bits for model/parameters) and a tolerated loss level D (residual codelength target under the preregistered noise model). Among feasible organisms g that meet the resource constraints, the viable set lies on a Pareto front minimizing

$$L(\mathrm{model}_g) \;+\; L(\mathrm{errors} \mid \mathcal{E}),$$

and any organism strictly dominated in this sum by a feasible competitor is eliminated almost surely under Section 4. Equivalently, at fixed (R, D) within the rate–distortion region induced by $\mathcal{E}$, selection prefers MDL–optimal designs. *Interpretation.* Fitness is penalized likelihood/MDL: short, accurate codes dominate; longer codes with no compensating error advantage are outcompeted.

[1] Equivalently, take instantaneous "fitness" $f(g) = -L_g$ so that $\dot{\pi}(g) = \pi(g)(f(g) - \mathbb{E}_\pi[f])$.

Recognition Science bridge (J-cost uniqueness).. Recognition Science (RS) provides a unique convex symmetric cost J (normalized by $J(1) = 0$ and $J''(1) = 1$) that governs recognition dynamics; its uniqueness on $\mathbb{R}_{>0}$ under symmetry/convexity/averaging constraints is formally proven in Lean (cost-uniformity T5). We reuse this J as the ledger cost shaping variation anisotropy below. Implementation and proofs reside in the public *reality* repository (see Reproducibility).

[Variation Anisotropy] Let J be a symmetric, convex ledger cost on phenotype moves, minimized at a balanced operating point. Assume variation proposals are generated by a maximum–entropy mechanism under an expected cost constraint $\mathbb{E}[J(\Delta)] \le \kappa$ (or, equivalently, by a detailed–balance kernel with potential J). Then the proposal distribution has Gibbs form

$$q(\Delta) \;\propto\; \exp\big(-\lambda\,\Delta J\big), \qquad \lambda > 0,$$

so that accessible moves are exponentially biased toward low–cost directions (iso–J shells). *Interpretation.* "Random with respect to fitness" becomes *structured randomness*: variation concentrates along directions that minimally perturb the ledger cost.

[Modularity Lower Bound] Let $\mathcal{E} = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k$ be a mixture of task families with joint distribution, and let M denote the mutual information capturing shared structure among the tasks (e.g., $M = \sum_i H(\mathcal{E}_i) - H(\mathcal{E}_1, \ldots, \mathcal{E}_k)$ in the discrete case). Suppose a reusable module of size b bits captures the shared component once and is invoked across tasks. Then the joint codelength saving achieved by reuse satisfies

$$\Delta L_{\text{reuse}} \;\ge\; M \;-\; b,$$

with equality when the module exactly codes the shared factor and task–specific parts are conditionally independent given the module. *Interpretation.* Selection prefers modular architectures whenever reuse outstrips overhead ($M > b$). Environments with greater task overlap drive stronger modularity.

Proof roadmap. Proofs are deferred to the appendices: Section 4 via Lyapunov analysis of the replicator (and Fokker–Planck for the mutational Gibbs form); Section 4 via dominance under Section 4 and standard MDL/rate–distortion arguments; Section 4 via maximum–entropy with a convex cost constraint (or detailed balance with potential J); Section 4 via information–theoretic coding bounds (chain rule and data–processing).

5 Measurement Protocol (preregistered MDL fitness)

One reference machine.. All measurements are made on a fixed, auditable reference machine (pinned toolchain, prefix–free tokenization). For each organism g and environment $\mathcal{E}$ we report codelengths in *bits* with the decomposition

$$L_{\text{total}}(g;\mathcal{E}) \;=\; L(\text{model}_g) \;+\; L(\text{parameters}_g \text{ to supported precision}) \;+\; L(\text{errors} \mid \mathcal{E}, \text{noise}), \tag{5}$$

where $L(\text{model}_g)$ counts structure and algorithms (once per analysis), parameter bits use interval coding at the data–supported tolerance, and the residual term is a negative log–likelihood code under the registered noise model. Every measurement is duplicated in two independent implementations (e.g., Rust and Python/Numba); the absolute discrepancy is reported as an $O(1)$ overhead band and must not affect the conclusions.

Noise models.. Residual codes use the dataset–documented observation model: Gaussian with reported σ or covariance, Poisson for counts, or an empirical/instrument kernel when supplied. As a preregistered robustness check we rerun all residual codes under a fixed heavy–tailed Student–t model (preset degrees of freedom) and report the deltas in bits.

Baselines.. To rule out scoring artifacts, we evaluate agnostic learners under the *same* protocol and report the best baseline per dataset: (i) dictionary/sparse models (fixed dictionaries or learned under capacity limits), (ii) generic neural networks with precommitted architectures and regularization, and (iii) kernel regressors/Gaussian processes from a fixed kernel menu. Hyperparameter grids, random seeds, early–stopping criteria, and tokenization are preregistered; no domain priors beyond smoothness/capacity are allowed. Baselines are scored by (5) so that comparisons to L_g are apples–to–apples.

Pooling rule.. Let $\mathcal{T}$ be the set of tasks in $\mathcal{E}$ with empirical frequencies (or durations) w_τ satisfying $\sum_{\tau\in\mathcal{T}} w_\tau = 1$. The pooled evolutionary code length for organism g is

$$L_g(\mathcal{E}) \;=\; L(\text{model}_g) \;+\; \sum_{\tau\in\mathcal{T}} w_\tau\, L(\text{errors} \mid \tau, \text{noise}_\tau), \tag{6}$$

with $L(\text{model}_g)$ counted *once* across tasks and residuals computed on held–out data per task using the preregistered noise model noise_τ. When tasks share constants or reusable submodules, those bits are encoded once and referenced thereafter by pointers whose cost is accounted explicitly. All weights w_τ, splits, and any task–specific tolerances are frozen in the preregistration.

6 Archival Datasets and Tasks (no new experiments)

D1: Gene Regulation.. *Setup:* Public expression matrices under known stimuli/contexts (time courses, dose series, knockdowns).
Model (C_g): Minimal modular controllers: motif libraries (tokenized PWMs), sparse wiring to target genes, simple regulatory nonlinearities (e.g., Hill or sigmoids) and kinetic lags; parameters encoded to supported precision. Residuals scored under preregistered noise (Gaussian with covariance or Negative Binomial for counts).
Baselines: Agnostic sequence-to-expression predictors (capacity-limited NNs, kernel regressors, dictionary models) with precommitted hyper-grids; scored identically by (5).
Goal/metrics: (i) $L_g(\mathcal{E})$ via (6) should *negatively* correlate with growth/fitness proxies (yield, division rate) at fixed resources; (ii) define environmental overlap by mutual information M between stimulus labels and upstream signal features; measure module reuse bits and test that reuse increases with M (slope > 0).

D2: Metabolism.. *Setup:* Public stoichiometric reconstructions with flux data across media conditions.
Model (C_g): Encode networks using a library of reusable subgraphs (e.g., transporters, shared core pathways); parameters are enzyme capacities/constraints with interval coding; feasibility checked by a fixed solver. Residuals are deviations in fluxes/growth under the declared noise model.
Baselines: Capacity-matched generic flow predictors (kernel/NN/dictionary) without explicit reuse.
Goal/metrics: In environments that share substrates/cofactors, the *marginal* $L(\text{model}_g)$ should drop via reuse relative to environments without overlap (report $\Delta L(\text{model})$ per added environment vs measured overlap). Fitness proxies should decrease with L_g at equal performance.

D3: Behavior/Ecology.. *Setup:* Public foraging/navigation datasets (trajectories, choice histories) across task variants.
Model (C_g): Policies with latent state and simple dynamics (e.g., sparse state–action graphs or linear–nonlinear controllers) shared across tasks; parameters encoded to supported precision.

Residuals scored on held-out trials under preregistered observation noise.
Baselines: Capacity-limited agnostic sequence forecasters (RNN/temporal kernels/dictionaries) with preregistered capacity.
Goal/metrics: Cross-task transfer should *reduce* L(errors) for species with richer internal models (report ΔL(errors) when training on task A and testing on task B). Directional diagnostics (below) quantify anisotropy in accessible variation.

D4 (Optional): Developmental Modules.. *Setup:* Domain architectures across gene families; inferred duplication–divergence events.
Model/metrics: Tokenize domains as modules of size b bits (structure + parameters). Compute environmental/shared-task mutual information M. Test enrichment of duplication–divergence when $M > b$ (odds ratio > 1 after phylogeny-aware controls). Report code savings $\geq M - b$ in composite tasks, consistent with the modularity bound.

Anisotropy diagnostics (all domains). Estimate the directional spectrum of accessible variation by projecting observed phenotypic deltas Δ onto eigenvectors of a local metric (empirical Fisher or Hessian of J) and fitting the log-frequency slope in $\|\Delta\|$ against ΔJ; the prediction is

$$\log \Pr(\Delta) \;=\; \text{const} \;-\; \lambda\, \Delta J \quad (\lambda > 0),$$

with heavier mass along low–ΔJ directions. Confidence bands obtained by bootstrap over individuals/conditions.

7 Results Plan (structure; numbers filled later)

Per-domain panels.. For each domain (D1–D4): (i) report codelength breakdowns L(model), L(parameters), and L(errors) under (5); (ii) plot fitness proxies versus $L_g(\mathcal{E})$ from (6) with slopes and CIs; (iii) plot module–reuse bits versus environmental mutual information M with slope/CIs; (iv) show anisotropy diagnostics (directional spectra, fitted λ; goodness-of-fit to the $\exp(-\lambda \Delta J)$ law).

Cross-domain synthesis.. Demonstrate that a single MDL rule explains: (i) fitness (negative association with L_g at fixed resources), (ii) modularity (reuse increases when M increases), and (iii) plasticity (transfer reduces L(errors) in species with richer models). Verify that anisotropy persists after controlling for phylogeny, sampling noise, and baseline capacity.

Sensitivity.. Report dual-language overhead bands ($O(1)$) for all codelengths; heavy–tail noise robustness (Student–t vs Gaussian/Poisson) for residuals; and baseline capacity sweeps showing conclusions are stable once baselines saturate their precommitted capacity.

8 Consistency with RS Source (evolution capsule)

The RS "Source.txt" evolution capsule outlines three core statements that this paper now formalizes and operationalizes:

[leftmargin=*]

1. **E1 (Fitness = $-L_g$):** We define evolutionary code length L_g and show replicator–MDL descent (Section 4).
2. **E2 (Anisotropic variation):** Proposal law $q(\Delta) \propto e^{-\lambda\, \Delta J}$ with RS J bridges ledger cost to accessible phenotypic moves (Section 4).

3. **E3 (Modularity bound):** Shared environmental information M yields reuse savings $\Delta L_{\text{reuse}} \geq M - b$ (Section 4).

These align one-to-one with the evolution section in the RS source specification while supplying a preregistered, auditable measurement protocol.

9 Reproducibility and Repository

Repository.. All code, proofs, and scaffolds are hosted in the public *reality* repository: `https://github.com/jonwashburn/reality`.

Toolchain.. Lean toolchain pinned by `lean-toolchain`; LaTeX sources in `Projects/afterlife`. RS cost uniqueness (T5) and related primitives are implemented in `IndisputableMonolith/Cost/` and referenced by higher layers.

Dual-implementation overhead band.. Following our entropy paper protocol, measurements that depend on tokenization/coding are duplicated in Rust and Python/Numba; the absolute discrepancy is reported as an $O(1)$ band (typically ≤ 10 bits) and does not affect conclusions.

OSF preregistration.. Falsifiers and analysis plans (domains D1–D4) will be preregistered before evaluation; all seeds, version pins, and manifests will be released.

10 Predictions (pre-stated, biting)

P1 (Modularity).. Organisms inhabiting more structured $\mathcal{E}$ (higher M) exhibit larger cross-task module reuse and lower $L_g(\mathcal{E})$ at equal predictive performance. Formally, $\partial\,\text{ReuseBits}/\partial M > 0$ and $\partial L_g/\partial M < 0$ (holding residual error fixed).

P2 (Duplication Threshold).. Duplication–divergence events are enriched when the shared-information threshold is exceeded: $\Pr(\text{duplication} \mid M > b) > \Pr(\text{duplication} \mid M \leq b)$, and composite-task code savings satisfy $\Delta L_{\text{reuse}} \geq M - b$.

P3 (Plasticity vs Entropy).. Plasticity capacity grows with $\text{Entropy}(\mathcal{E})$ (broader task distributions demand richer models), yet the total code L_g is still minimized via sparsity: increasing model bits must be compensated by larger reductions in $L(\text{errors})$ along the Pareto front.

11 Falsifiers (real, not rhetorical)

F1 (Anti-MDL dominance).. Find lineages that, under the same resource budget and evaluation protocol, systematically *increase* in frequency while having *larger* $L_g(\mathcal{E})$ than competitors with equal or better predictive performance. A statistically significant, persistent reversal refutes the MDL–fitness link.

F2 (No modularity–overlap link).. Across independent datasets, observe no positive association between module reuse and environment/task overlap (mutual information M), after preregistered controls (phylogeny, capacity, sampling). Failure to detect the predicted slope (≈ 0 with tight CIs) refutes the modularity bound's empirical bite.

F3 (Isotropic variation).. After controlling for measurement windows and noise, find phenotypic deltas Δ distributed isotropically rather than according to $q(\Delta) \propto \exp(-\lambda \Delta J)$. A flat directional spectrum (no low–ΔJ enrichment) falsifies the anisotropy theorem.

12 Confounds and Controls

Genome size vs. code length (MDL, not raw length).. Raw genome length is not a proxy for effective description length. We distinguish compressible redundancy from functional modules by (i) encoding structure with a fixed tokenization of modules/subroutines, (ii) encoding parameters only to data–supported precision, and (iii) charging residual error in bits. Two genomes with different sizes can have comparable $L(\text{model}_g)$ once repeated motifs are referenced and nonfunctional repeats compress to $O(1)$ per motif. All analyses report both raw sizes and effective $L(\text{model}_g)$ to prevent conflation.

Sampling and noise (non–ergodicity, batch effects).. Non–ergodic sampling and batch effects inflate residual codes if unmodeled. We preregister observation models: Gaussian with published σ / covariance terms, Poisson or Negative Binomial for counts, or declared instrument kernels. Batch effects are modeled as nuisance covariates whose parameters are encoded at supported precision and charged to $L(\text{model}_g)$; we re–score residuals under a heavy–tail (Student–t) sensitivity. Any hand cleaning is disallowed; filters are implemented in code and logged.

Phylogeny (separating ancestry from MDL effects).. We use phylogeny–aware statistics (e.g., mixed models with clade random effects; phylogenetic generalized least squares / independent contrasts) to separate inherited similarity from code–length effects. All regression summaries of $L_g(\mathcal{E})$ versus fitness or reuse versus environmental overlap report phylogeny–controlled estimates and naive estimates side–by–side.

Overfitting (fair comparisons).. We enforce hard train/holdout splits *across tasks* and precommit capacity limits (token budgets, network widths/depths, kernel menus). We always report the *best* agnostic baseline under the same scoring rule (L_{total}) and capacity schedule. Hyperparameter grids and seeds are frozen before evaluation; early stopping criteria are preregistered.

13 Discussion (why this matters)

Unification.. This framework compresses Darwinian selection, statistical learning (MDL), and network modularity into a single quantitative umbrella: bits. Fitness is negative description length; the same currency measures model economy, residual accuracy, and reuse.

Mechanism for evolvability.. Anisotropic variation explains the empirical repeatability of useful phenotypes: proposals concentrate along low–cost directions determined by a convex symmetric ledger cost J, while modular reuse amortizes search over related tasks. Together they yield fast, reliable adaptation without invoking ad hoc bias.

Bridging levels.. The code–length decomposition travels cleanly from gene circuits (motif libraries, sparse wiring) to metabolism (reusable subgraphs) to behavior (compact policies): $L(\text{model})$ and $L(\text{errors})$ remain commensurate *in bits*, enabling cross–level synthesis without changing units.

Links to engineering.. For synthetic biology and AI, the prescription is the same: build for reuse and low MDL under the target environment $\mathcal{E}$. Architectures that minimize $L(\text{model})$ while achieving low $L(\text{errors})$ transfer better and adapt faster; anisotropy diagnostics inform which directions in design space are most fruitful.

14 Methods (camera–ready subsections)

14.1 Formal L_g definition and coding scheme

Tokenization and model code.. On a fixed reference machine with a prefix–free grammar, the *model code* $L(\text{model}_g)$ counts: (i) structural tokens (modules, wiring, update rules) with a canonical, versioned alphabet; (ii) parameters θ_g encoded by interval coding to supported precision. If a parameter θ is identified up to tolerance width δ, its code is $\lceil \log_2(1/\delta) \rceil + O(1)$ bits; shared constants are encoded once and referenced thereafter with a pointer of declared cost.

Residual code.. For observations $\{y_i\}_{i=1}^n$ with model predictions $\{\mu_i\}$ and noise model $\mathcal{N}$, the residual codelength is the negative log–likelihood expressed in bits:

$$L(\text{errors} \mid \mathcal{E}, \mathcal{N}) \;=\; -\sum_{i=1}^{n} \log_2 p_{\mathcal{N}}(y_i \mid \mu_i, \text{noise params}),$$

e.g. Gaussian: $-\log_2 p = \sum_i \left[\frac{1}{2} \log_2(2\pi\sigma_i^2) + \frac{(y_i-\mu_i)^2}{2\sigma_i^2 \ln 2} \right]$; Poisson: $-\log_2 p = \sum_i \left[\lambda_i - y_i \log \lambda_i + \log(y_i!) \right] / \ln 2$. Heavy–tail sensitivity uses a Student–t_ν likelihood with preregistered ν.

Pooling across tasks.. Let $\{(\tau, w_\tau)\}$ be tasks and their empirical weights with $\sum_\tau w_\tau = 1$. We score

$$L_g(\mathcal{E}) \;=\; L(\text{model}_g) \;+\; \sum_{\tau} w_\tau \, L(\text{errors} \mid \tau, \mathcal{N}_\tau),$$

counting shared modules/constants once. All tokenizations, precisions, and weights are frozen in preregistration.

14.2 Replicator–MDL proof details

Lyapunov descent.. With instantaneous fitness $f(g) = -L_g$, the replicator equation $\dot{\pi}(g) = \pi(g)(f(g) - \mathbb{E}_\pi[f])$ yields

$$\frac{d}{dt}\mathbb{E}_\pi[L] \;=\; \sum_g \dot{\pi}(g)\, L_g = \sum_g \pi(g)(\mathbb{E}_\pi[L] - L_g)\, L_g = \mathbb{E}_\pi[L]^2 - \mathbb{E}_\pi[L^2] \;=\; -\mathrm{Var}_\pi(L) \le 0.$$

Equality holds iff L_g is π–a.s. constant.

Stationary measure with mutation.. Augment the replicator with a small, reversible diffusion (mutation) operator generating a Fokker–Planck flow that satisfies detailed balance with potential $\Phi(g) = \beta L_g$. The stationary density solves $\nabla \cdot (D \nabla \pi^\star + D\, \pi^\star \nabla \Phi) = 0$, yielding the Gibbs form $\pi^\star(g) \propto \exp(-\beta L_g)$ (up to normalization and base measure factors). Regularity and confining conditions ensure uniqueness.

14.3 Modularity bound proof (information–theoretic)

Let $\mathcal{E} = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k$ with joint distribution. Coding each task family separately has codelength $\sum_i L(\mathcal{E}_i)$; coding jointly with a reusable module M of size b bits that captures shared structure achieves $b + \sum_i L(\mathcal{E}_i \mid M)$. By the chain rule,

$$\sum_i L(\mathcal{E}_i) \; - \; \Big[b + \sum_i L(\mathcal{E}_i \mid M)\Big] \; \geq \; \Big(\sum_i H(\mathcal{E}_i) - H(\mathcal{E}_1, \ldots, \mathcal{E}_k)\Big) \; - \; b \; = \; M \; - \; b,$$

identifying M as mutual information in bits. Equality holds when M codes exactly the shared factor and the task–specific parts are conditionally independent given M.

14.4 Variation anisotropy derivation and diagnostics

Derivation (maximum entropy or detailed balance).. Impose a constraint $\mathbb{E}[J(\Delta)] \leq \kappa$ on phenotype moves Δ with convex symmetric cost J, or equivalently assume a reversible proposal kernel with potential J. Maximizing entropy subject to the cost constraint yields

$$q(\Delta) \; = \; \frac{1}{Z(\lambda)} \exp(-\lambda \, \Delta J), \qquad \lambda > 0,$$

with partition function $Z(\lambda)$ and Lagrange multiplier λ set by κ. Symmetry of J yields iso–J shells carrying level–set mass; convexity concentrates proposals along low–J directions.

Diagnostics (empirical tests).. Estimate a local metric (e.g., empirical Fisher or Hessian of J) at a phenotype and project observed Δ onto its eigenvectors. Test the linear relation $\log \Pr(\Delta) = \text{const} - \lambda \, \Delta J$ by binning in ΔJ and fitting the slope λ with bootstrap CIs. A flat spectrum (no dependence) falsifies anisotropy.

14.5 Pre–registration and reproducibility

We release a containerized pipeline with pinned compilers/libraries and a single entry script that rebuilds all numbers and figures. The preregistration freezes: dataset versions and checksums; tokenization grammar; parameter precisions; noise models; train/holdout splits; baseline families and capacity grids; seeds; and pooling weights $\{w_\tau\}$. Each figure emits a JSON manifest (dataset IDs, seeds, container digest, and numeric outputs) to enable byte–level audit.

A Replicator–MDL equivalence (full derivation)

Setup (discrete type space).. Let $\mathcal{G}$ be a finite or countable set of organism types $g \in \mathcal{G}$ with evolutionary code lengths $L_g \in \mathbb{R}$. Let $\pi_t(g) \geq 0$, $\sum_g \pi_t(g) = 1$ be population frequencies. Define instantaneous Malthusian fitness

$$f(g) \; := \; -\beta \, L_g + c,$$

where $\beta > 0$ encodes resource scarcity and c is an arbitrary constant. The replicator dynamics are

$$\dot{\pi}_t(g) \; = \; \pi_t(g) \, (f(g) - \bar{f}_t), \qquad \bar{f}_t := \sum_{h \in \mathcal{G}} \pi_t(h) f(h). \tag{7}$$

The additive constant c cancels in $f(g) - \bar{f}_t$, so only differences in L_g (scaled by β) matter.

Lyapunov descent of the mean code length.. Let $\mathbb{E}_{\pi_t}[L] := \sum_g \pi_t(g)\, L_g$. Differentiating and using (7) gives

$$\frac{d}{dt}\,\mathbb{E}_{\pi_t}[L] = \sum_g \dot{\pi}_t(g)\, L_g = \sum_g \pi_t(g)(f(g) - \bar{f}_t)\, L_g = \mathrm{Cov}_{\pi_t}(L, f).$$

Since $f = -\beta L + c$, we obtain

$$\frac{d}{dt}\,\mathbb{E}_{\pi_t}[L] = -\,\beta\,\mathrm{Var}_{\pi_t}(L) \;\le\; 0, \tag{8}$$

with equality iff L_g is π_t–a.s. constant. Thus $\mathbb{E}_{\pi_t}[L]$ is a strict Lyapunov functional whenever the population contains heterogeneity in L.

Stationary measures with mutation (continuous limit).. On a continuous type space $\mathcal{G} \subset \mathbb{R}^d$ with base measure dg, augment (7) by a reversible mutation/diffusion operator with (positive definite) diffusion matrix $D(g)$:

$$\partial_t \pi_t(g) = \pi_t(g)\,(\bar{f}_t - f(g)) + \nabla\cdot\Big(D(g)\,(\nabla\pi_t(g) + \beta\,\pi_t(g)\,\nabla L(g))\Big). \tag{9}$$

The Fokker–Planck form (9) is a gradient flow for the free energy

$$\mathcal{F}[\pi] \;=\; \int \pi(g)\,(\beta\,L(g) + \log\pi(g))\,dg,$$

under the D–weighted Wasserstein metric. Under confining conditions ($\beta L(g) \to \infty$ as $\|g\| \to \infty$) and mild regularity, the unique stationary density solves the detailed–balance condition

$$\nabla\pi^\star(g) + \beta\,\pi^\star(g)\,\nabla L(g) = 0 \quad\Longrightarrow\quad \pi^\star(g) \;\propto\; \exp(-\beta\,L(g)).$$

Hence mutation selects a Gibbs measure with potential βL; in the zero–mutation limit, π_t concentrates on the MDL minimizers of L.

Invariances.. Adding a constant to all code lengths $L_g \mapsto L_g + c$ leaves the dynamics (7) and the stationary Gibbs family unchanged (the partition function reabsorbs c). Rescaling $L \mapsto aL$ is equivalent to rescaling $\beta \mapsto a\beta$.

Extensions.. With time–varying $\beta(t)$ or resource–limited growth (logistic factors), the same Lyapunov argument yields $d\mathbb{E}[L]/dt \le 0$ whenever selection differentials remain proportional to $-L$ up to a common offset.

B Modularity bound (inequalities, examples)

General inequality.. Let $\mathcal{E} = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k$ denote a mixture of task families with a joint distribution. Consider two coding schemes on a fixed reference machine:

- *Separate coding:* Code each family independently with codelength $\sum_{i=1}^k L(\mathcal{E}_i)$.
- *Joint coding with a reusable module M:* First code a shared module of size b bits, then code task–specific parts conditionally, giving $b + \sum_{i=1}^k L(\mathcal{E}_i \mid M)$.

By the chain rule of codelengths (Shannon idealization), the gain is

$$\Delta L_{\text{reuse}} = \sum_{i=1}^k L(\mathcal{E}_i) - \Big[b + \sum_{i=1}^k L(\mathcal{E}_i \mid M)\Big] \;\ge\; \sum_{i=1}^k H(\mathcal{E}_i) - H(\mathcal{E}_1,\ldots,\mathcal{E}_k) - b = M - b,$$

where $M := \sum_i H(\mathcal{E}_i) - H(\mathcal{E}_1,\ldots,\mathcal{E}_k)$ is the multi–information (shared structure) in bits. Equality holds when M codes exactly the shared factor and, given M, tasks are conditionally independent.

Example (two tasks, one shared latent).. Let $\mathcal{E}_1 = S \oplus N_1$, $\mathcal{E}_2 = S \oplus N_2$ where $S \sim$ Bernoulli(1/2) and $N_i \sim$ Bernoulli(ε_i) independent, with $\oplus$ XOR. Then

$$M = I(\mathcal{E}_1; \mathcal{E}_2) = 1 - h(\varepsilon_1 \star \varepsilon_2) + h(\varepsilon_1) + h(\varepsilon_2),$$

where h is the binary entropy and $\varepsilon_1 \star \varepsilon_2 = \varepsilon_1(1-\varepsilon_2) + (1-\varepsilon_1)\varepsilon_2$. A module M that codes S uses $b = 1$ bit; when noise is small, $M \approx 1$ and $\Delta L_{\text{reuse}} \approx 1 - b \approx 0^+$ (tight). For more tasks sharing S, the gain grows linearly in the number of tasks (amortization), while b stays fixed.

Negative/zero gains.. If $M \le b$ (module too large for the shared structure), reuse yields no advantage: $\Delta L_{\text{reuse}} \le 0$. This provides a falsifiable threshold for modularity: selection favors modules only when $M > b$.

C Variation anisotropy (convex J, proposal law, tests)

Maximum–entropy derivation.. Let Δ be a random phenotype move in a local coordinate chart. Impose a resource constraint $\mathbb{E}[J(\Delta)] \le \kappa$ where $J : \mathbb{R}^d \to \mathbb{R}_+$ is convex, symmetric (even), strictly minimized at 0, and has $\nabla^2 J(0) \succ 0$. The maximum–entropy distribution subject to the constraints $\int q(\Delta)\, d\Delta = 1$ and $\int J(\Delta)\, q(\Delta)\, d\Delta = \kappa$ solves

$$\delta\Big[-\int q \log q + \lambda\Big(\int Jq - \kappa\Big) + \eta\Big(\int q - 1\Big)\Big] = 0 \quad \Longrightarrow \quad q_\lambda(\Delta) = \frac{1}{Z(\lambda)} \exp(-\lambda\, J(\Delta)),$$

with $Z(\lambda) = \int e^{-\lambda J(\Delta)} d\Delta$ and $\lambda > 0$ chosen to match κ.

Local quadratic limit and geometry.. For small moves, $J(\Delta) = \frac{1}{2}\,\Delta^\top H\,\Delta + o(\|\Delta\|^2)$ with $H := \nabla^2 J(0) \succ 0$. Then

$$q_\lambda(\Delta) \;\approx\; \mathcal{N}\Big(0,\; \lambda^{-1} H^{-1}\Big)\,,$$

an elliptical Gaussian whose principal axes are the eigenvectors of H. Thus anisotropy is governed by the curvature of J: directions with small curvature have larger variance (more accessible moves).

Detailed–balance derivation (alternative).. Suppose proposals arise from a reversible Markov kernel with stationary density $\propto e^{-\lambda J}$; then detailed balance $q(\Delta)\, e^{-\lambda J(\phi)} = q(-\Delta)\, e^{-\lambda J(\phi+\Delta)}$ enforces the same exponential form for the increment distribution in homogeneous neighborhoods.

Empirical tests.. Let $\widehat{H}$ be a local metric (e.g., empirical Fisher of the likelihood or a quadratic fit of $\widehat{J}$ near 0).

- *Directional spectrum:* Project observed Δ onto eigenvectors of $\widehat{H}$; test whether variances match $\propto 1/\widehat{\lambda}_i$ (eigenvalues of $\widehat{H}$).
- *Radial law:* For quadratic J, $\Delta J = \frac{1}{2}\Delta^\top \widehat{H} \Delta$; bin moves by ΔJ and regress $\log \Pr(\Delta J \in \text{bin})$ on $-\Delta J$. The predicted slope is $-\lambda$.
- *Angular uniformity on iso–J:* Condition on thin shells of fixed ΔJ; test angular uniformity (no preferred orientation after accounting for H).

Deviations (e.g., isotropy after whitening by $\widehat{H}$) falsify the anisotropy law.

D Encoding minutiae for each dataset category

Common rules.. All codes are prefix–free. The total codelength follows (5). Parameters are interval–coded at preregistered precision. Shared constants/modules are coded once and referenced via pointers with declared bit costs. Residuals are negative log–likelihoods in bits under preregistered noise models.

D1: Gene Regulation

Tokens: Motif library identifiers; PWM matrices (quantized to declared precision); wiring adjacency lists (sparse format); regulatory nonlinearity class; kinetic lag tokens.
Parameters: PWM entries; binding thresholds; connection weights; kinetic rates. Precision from cross–validated tolerances or published uncertainties.
Residuals: Gaussian with covariance (microarray) or Negative Binomial (RNA–seq counts). Heavy–tail sensitivity: Student–t_ν with preregistered ν.
Pinned datasets and protocols: Dataset DOIs/versions, normalization protocols, and seed splits are listed in the repository manifests:

[leftmargin=*]

- `manifest/datasets/evolution_gene_regulation.yaml`: GEO/ArrayExpress accessions with DOIs and checksums; normalization (TPM/RPKM as applicable); stratified task splits (80/20) and fixed seeds {137, 42, 2718}.
- Preprocessing scripts and exact versions are recorded in the figure JSON manifests and the container log (see §9).

D2: Metabolism

Tokens: Reaction list; stoichiometric blocks; reusable subgraph library (transporters/core pathways); solver class (FBA/kinetic); constraint tokens.
Parameters: Enzyme capacities; transport bounds; maintenance costs. Precision from calibration tolerances.
Residuals: Deviations of fluxes/growth from observations under Gaussian/Poisson noise as documented.
Pinned reconstructions and media tables: Reconstructions and media conditions are enumerated in `manifest/datasets/evolution_metabolism.yaml` (e.g., E. coli and yeast community reconstructions), with solver options/tolerances (FBA/kinetic) and checksums. Media condition tables are included alongside accession metadata; all versions are containerized and logged at build time.

D3: Behavior/Ecology

Tokens: Policy class (finite–state controller / linear–nonlinear policy); latent–state count; transition structure (sparse edges); observation kernel class.
Parameters: Transition probabilities/gains; observation parameters; reward weights if used. Precision via held–out performance plateaus.
Residuals: Action likelihoods on held–out trajectories under preregistered observation noise.
Pinned accessions and splits: Behavior datasets (accessions, preprocessing rules, and task splits) are specified in `manifest/datasets/evolution_behavior.yaml`. Trajectory tokenization and observation models are pinned; splits are precommitted (80/20 across task variants) with seeds {137, 42, 2718}.

D4: Developmental Modules (optional)

Tokens: Domain vocabulary; gene–domain architectures; duplication markers; divergence parameters.
Parameters: Domain–specific weights; linker costs; reuse pointers.
Residuals: Likelihood of observed architectures under the duplication–divergence model.
Pinned catalogs and priors: Gene family catalogs, phylogenies, and model priors are specified in `manifest/datasets/evolution_development.yaml` with DOIs/checksums. Duplication markers and divergence priors (bounds) are declared there for reproducibility.

Baselines (all domains).. *Tokens:* Architecture class (NN/kernel/dictionary); capacity parameters (layers/widths, kernel types, dictionary sizes); regularization; early–stopping rule.
Parameters: Weights/hyperparameters encoded to supported precision.
Residuals: Same scoring as for C_g.
Pinned baselines and container: Baseline hyper–grids and seeds are listed in `manifest/baselines/hypergr` container image and digest are recorded in `manifest/container/evolution_mdl.json` and included in all figure manifests.

E Additional controls (phylogeny-aware analyses, bootstraps)

Phylogeny-aware inference.. Use phylogenetic generalized least squares (PGLS) or mixed models with clade random effects to regress fitness proxies on $L_g(\mathcal{E})$ and reuse on M. Report both naive and phylogeny–controlled estimates with confidence intervals.

Bootstrap and permutation tests.. Compute bootstrap CIs for slopes (fitness vs L_g; reuse vs M). Use task–label permutations to test whether observed associations could arise from chance partitioning; use block bootstraps for time–series.

Capacity sweeps and early stopping.. Vary baseline capacity along preregistered grids; confirm that beyond a knee point, further capacity yields diminishing returns in L(errors) but increases L(model), leaving conclusions unchanged.

Holdout protocols and leakage checks.. Ensure that task splits prevent leakage of shared modules into holdout evaluation. Verify that shared constants are counted once and that pointers are charged uniformly across domains.

Sensitivity to noise models.. Re–score residuals under Student–t_ν and (where relevant) heteroskedastic Gaussian models. Report deltas in bits and verify stability of rankings in L_g.

Reporting and manifests.. Each figure is accompanied by a JSON manifest (dataset IDs, checksums, seeds, container digest, tokenization version, numeric outputs) to enable exact reproduction.
Final manifest: The frozen dataset list (DOIs/checksums), seeds {137, 42, 2718}, and the container image digest are already recorded in the repository manifests (see §9) and automatically embedded in each figure's JSON output.