The Illusion of Structural Gain: Identifiability and estimator calibration in a bounded energy-structure model

Heinz Kahler — Architect and Framework Designer for Rate Controlled Evolution (RCE)

Correspondence: heinzkahler@gmail.com; ORCID: 0009-0009-3857-9760

Abstract. State-dependent responses to resource input can arise from structural facilitation, ordinary selection, or the observation procedure itself. Distinguishing these explanations matters when artificial-life models use a scalar structural measure to support mechanistic claims. We analyze a bounded energy–structure equation and its proposed normalized input-response diagnostic. An explicit parameter transformation preserves the complete deterministic vector field at a fixed input, providing a conditional nonidentifiability certificate for the structural-gain coefficient. This equivalence concerns scalar drift, not every stochastic observation. An exact zero-feedback example produces a strictly increasing normalized finite-window response for every positive observation window when linear loss is positive. Incorrect ceiling normalization supplies a separate instantaneous counterexample. A fixed-fitness binary-trait selection process also matches the structural model's scalar drift while differing in event turnover and fluctuations. We then evaluate three estimators in a reproducible synthetic birth–death benchmark with locked acceptance rules, independent resets, exact starting states, and known inputs and ceiling. Across 24 held-out cases and 144,000 slope estimates, the prescribed endpoint window fails reliability; reducing the window tenfold improves performance but does not satisfy all reliability requirements. A state-matched first-event reference passes every held-out reliability check but misses the sensitivity requirement in three positive cases. No method passes the combined benchmark. Supplementary switches show why estimating a contrast correctly does not identify structural causation when nuisance dynamics change. The contribution is an application-specific set of proofs, counterexamples, and calibration results, not a universal evolutionary principle or empirical validation. Exact-law priority relative to energy-constrained autocatalysis remains unresolved.

Keywords: artificial life; structural identifiability; resource-dependent dynamics; finite-time estimation; birth–death processes; simulation calibration.

1. Introduction

An increasing structural measure is not, by itself, evidence that structure causes its own subsequent production. A system may become more densely connected because successful components reproduce, because external input increases, because an occupancy ceiling is misestimated, or because a finite observation window combines responses from different states. These possibilities are particularly consequential for low-dimensional models of artificial life: their appeal is that they summarize complicated dynamics, but the same compression can erase distinctions among generative mechanisms. This article examines that problem for a bounded energy–structure equation associated with the working label Rate-Controlled Evolution (RCE). The objective is to determine what the equation and its response diagnostic identify under explicit observation assumptions, not to establish a new general theory of evolution.

The distinction between observable behavior and its generating mechanism already matters within artificial-life research. Channon et al. (2024) describe agreement about behavioral hallmarks of open-ended evolution alongside less agreement about the mechanisms required to generate them. Moreno et al. (2024), in a digital-multicellularity case study, report that novelty, complexity, and adaptation can be loosely coupled and that different complexity measures can diverge. These observations motivate careful operationalization; they do not imply that the scalar studied here measures open-endedness. In particular, bounded edge occupancy is neither an interchangeable measure of functional complexity nor evidence of indefinitely continuing innovation. The relevance of the present analysis is methodological: a diagnostic intended to attribute a resource response to structure should be tested against alternatives that reproduce the same observable.

There are also direct conceptual precedents for organization promoting further organization. Jain and Krishna (1998) model a catalytic interaction graph in which an autocatalytic

set initiates increasing connectivity, followed by saturation. Filisetti et al. (2011) explicitly introduce energy constraints in stochastic catalytic networks and investigate changes in energy-intake rate. Consequently, neither feedback in structural production nor the inclusion of an energy supply can reasonably be treated here as new in itself. The material available for the latter paper was its abstract, not its complete equations. That access limitation prevents certification of exact-law priority; it does not justify assuming that a close precedent contains no overlap.

The central mathematical question is narrower. Suppose an observed scalar obeys an input-dependent gain–loss equation. Can a fitted structural-gain coefficient be separated from losses at a single input? If responses are compared across inputs, does a positive normalized slope identify endogenous facilitation? If the diagnostic is algebraically valid for an instantaneous drift, how accurately is it recovered from actual observations? Finally, if a switch removes a designated coupling, what additional assumptions make the difference between switch conditions causally interpretable? These questions concern different inferential steps and require different evidence. A proof of an instantaneous identity cannot answer all of them.

Structural identifiability is already understood to depend on inputs, outputs, parameters, and initial conditions rather than on an equation in isolation (Villaverde et al., 2019). Likewise, finite-time effects in drift reconstruction are established methodological issues: finite-interval conditional changes need not equal infinitesimal coefficients (Lade, 2009; Anteneodo & Riera, 2009). Birth–death inference also distinguishes endpoint observations from event histories containing directions and exposures (Crawford et al., 2018). The present work uses those distinctions in a particular diagnostic setting. Its technical content consists of an explicit admissible invariance, exact application-specific counterexamples, a transparent reference observation design, and retained simulation failures. Historical originality of every formula is not asserted.

Three conclusions organize the article. First, the gain coefficient is not generally recoverable from a deterministic scalar trajectory at one fixed input when both loss coefficients are free. Second, even a precisely measured positive response slope does not uniquely specify a structural mechanism: measurement design and alternative dynamics must be considered. Third, correcting one measurement defect does not guarantee sufficient sensitivity or causal attribution. The benchmark separates these claims by construction. Its generator deliberately supplies the proposed response law as ground truth, permitting estimator validation while providing no evidence that the law holds in a biological or other natural system.

2. Model, observables, and inferential targets

2.1 A bounded structural coordinate

Let C denote the fraction of predefined valid network slots that are occupied. Its denominator must be fixed independently of the observed gain response; changing which links count as valid changes the observable. Let K be a ceiling on that fraction, with $0 < K \leq 1$. The analyzed deterministic equation is

$$\dot{C} = F(C, g) = g(1 + \kappa C)(1 - C/K) - \beta C^2 - \gamma C. \tag{1}$$

Here $g > 0$ is a calibrated effective input rate, while $\kappa, \beta, \gamma \geq 0$ are common parameters within the comparisons for which the model is invoked. The coordinate and ceiling are dimensionless. With time measured in a common unit, g, β, and γ have inverse-time units, while κ is dimensionless. The choice to call κ structural gain is an interpretation of this parameterization, not a deduction that every observed dependence on C has that mechanism.

Equation (1) is phenomenological. Occupancy does not specify motif composition, catalytic

usefulness, heritability, connectivity among essential components, or the functional consequences of an edge. Nor does g measure thermodynamic efficiency. If a physical resource variable q is mapped to g through a conversion factor, that factor requires an independent definition and calibration. The present benchmark uses g directly and assumes it is known; it does not establish a resource-to-structure conversion law. Consequently, an empirical system needs a defensible measurement model before its data can be interpreted through the following algebra.

The physical interval is forward invariant because $F(0, g) = g > 0$ and $F(K, g) = -\beta K^2 - \gamma K \leq 0$. Expanding gives

$$F(C, g) = g + b(g)C - d(g)C^2, \qquad b(g) = g\kappa - g/K - \gamma, \quad d(g) = g\kappa/K + \beta. \tag{2}$$

This is a scalar quadratic, or Riccati-type, equation; it becomes linear when $d = 0$. In the regime $b > 0$ and $d > 0$ it can be written as logistic growth plus constant input. Calling every parameter regime ordinary positive-growth logistic dynamics would be inaccurate because b need not be positive. The coefficient form is useful because it makes both the fixed-input ambiguity and the restrictions across inputs transparent. It does not identify an entirely new mathematical family.

For $d > 0$, the unique positive equilibrium is $C_* = [b + \sqrt{b^2 + 4gd}]/(2d)$ and its derivative is $F_C(C_*, g) = -\sqrt{b^2 + 4gd}$. The other quadratic root is negative. In the linear case, $b = -g/K - \gamma < 0$ and the equilibrium is $-g/b$. Together with the boundary signs, these facts imply attraction to a unique equilibrium in $(0, K]$ from the physical interval. The reciprocal derivative magnitude is a local relaxation time, not an exact global exponential description of every trajectory. The bounded model therefore does not predict indefinite acceleration or unbounded structural growth.

2.2 The instantaneous response restriction

At a fixed state strictly below the ceiling, define the normalized input response

$$R(C) = \frac{\partial F(C, g)/\partial g}{1 - C/K} = 1 + \kappa C. \tag{3}$$

Because F is affine in g, the same result follows from any two distinct instantaneous inputs:

$$R(C) = \frac{F(C, g_h) - F(C, g_l)}{(g_h - g_l)(1 - C/K)}, \qquad g_h > g_l > 0. \tag{4}$$

The input-independent losses cancel. This identity requires the same relevant state, ceiling, conversion from applied input, and remaining parameters in both conditions. It is not the response obtained by comparing two equilibria, by observing different trajectories at the same clock time, or by averaging a changing state over a finite interval. The interpretation of the slope is conditional on those distinctions.

The prediction is more restrictive than positivity. Under the exact form, the response has intercept one, slope κ, and zero curvature, with the same line across input pairs. If the controlled variable is q and $g = aq$ for a common unknown positive factor a, the response normalized by $q_h - q_l$ instead has intercept a and slope $a\kappa$. A slope/intercept ratio might then be considered, but its statistical properties are not evaluated here. If a changes with state or treatment, the simple ratio argument no longer follows. Independently known normalization is therefore part of the model test, not a convenience that can be estimated freely until the desired slope appears.

Normalization also changes what is being called amplification. The unnormalized input derivative is

$$A(C) = (1 + \kappa C)(1 - C/K) = 1 + (\kappa - 1/K)C - (\kappa/K)C^2. \tag{5}$$

If $\kappa K \leq 1$, this raw marginal response is nonincreasing throughout the physical interval even when $\kappa > 0$. A positive normalized slope thus does not imply that an additional unit of input produces a larger absolute structural increment at higher occupancy. It expresses gain relative to the available fraction of slots under the assumed ceiling. This distinction is essential when translating the diagnostic into statements about productivity, adaptation, or energetic benefit.

2.3 Three separate inferential tasks

Parameter identification asks whether the chosen observations can distinguish model coefficients. Estimator reliability asks whether an implemented procedure targets the desired coefficient with acceptable bias, interval coverage, and false-positive behavior. Mechanistic attribution asks whether that coefficient represents the specified structural process rather than another process with the same observable effect. These tasks are logically distinct. Infinite precision cannot remove an exact observational ambiguity, and an unbiased estimate of a contrast cannot establish that the intervention changed only its intended target. Conversely, lack of sensitivity does not establish absence of the target effect.

We use these distinctions to delimit the analysis. The deterministic propositions concern specified equations and observations. The synthetic benchmark concerns a finite Markov embedding with known truth. The supplementary controls concern attribution assumptions supplied or violated by construction. None is a measurement of biological structural feedback. An eventual empirical application would need to establish the observable, the state representation, input conversion, and intervention validity independently of obtaining a positive fitted slope.

3. A fixed-input nonidentifiability certificate

3.1 Explicit invariance and admissibility

Proposition 1. Fix $g > 0$ and known K. For any perturbation δ such that

$$\max(-\kappa, -\gamma/g) \leq \delta \leq \beta K/g, \tag{6}$$

the transformed parameters

$$\kappa' = \kappa + \delta, \qquad \gamma' = \gamma + g\delta, \qquad \beta' = \beta - g\delta/K \tag{7}$$

are nonnegative and generate exactly the same deterministic vector field as Equation (1). Whenever the interval contains a nonzero perturbation, observations restricted to that fixed-input scalar drift cannot uniquely determine κ while both loss coefficients remain free.

Proof. The change in the gain term is $g\delta C(1 - C/K)$. The change in linear loss is $-g\delta C$, and the change in quadratic loss is $+g\delta C^2/K$. Their sum vanishes identically for every C. Equation (6) is exactly the intersection of the three nonnegativity constraints. Therefore the vector fields coincide on the entire physical interval, and trajectories with the same initial condition coincide by uniqueness of solutions. The certificate is not a numerical near-fit or an asymptotic approximation. It is an exact family of indistinguishable deterministic parameterizations.

The admissibility qualifier is consequential. A boundary point can have no nonzero allowed perturbation, so the proposition does not say that every parameter triple is ambiguous within the nonnegative parameter space. Nor does it apply unchanged when one or both loss coefficients are independently known. It states a sufficient, explicit ambiguity under the

stated free-parameter design. Obtaining many noiseless observations along the same fixed-input trajectory does not resolve that ambiguity because those observations all sample the same invariant vector field. More data of the same kind and new kinds of information are not interchangeable solutions.

3.2 What varying the input can resolve

For common parameters observed at two distinct inputs, the transformation in Equation (7) cannot preserve both fields with a single nonzero δ: it would require the same change in γ to equal both $g_l\delta$ and $g_h\delta$. Equivalently, Equation (2) predicts that $b(g)$ and $d(g)$ vary linearly with linked slopes. Within the assumed model, known ceiling and calibrated input, these restrictions provide information absent from a single-input trajectory. This is an example of input-dependent identifiability, not a contradiction of Proposition 1. The general importance of input design is established by Villaverde et al. (2019).

Breaking this particular invariance is not equivalent to identifying a biological mechanism. Another model can have the same cross-input drift, as the selection example below demonstrates. Unknown conversion factors or treatment-dependent losses can also invalidate the cancellation that makes Equation (4) informative. A useful experimental proposal must therefore specify both which mathematical ambiguity its design removes and which alternatives remain observationally equivalent at the chosen resolution. The result supports designing new observations rather than treating a fitted gain coefficient as self-interpreting.

3.3 Why the certificate is drift-level

Consider the finite birth–death embedding used later, with $K = 1$, $C = n/M$, and rates $\lambda = g(M-n)(1+\kappa n/M)$ and $\mu = \gamma n+\beta n^2/M$. Under Equation (7), both rates change by the

same amount, $g\delta MC(1-C)$. Their difference is unchanged, but their sum generally is not. At interior states, a nonzero admissible perturbation therefore changes holding times and event-direction probabilities even though it preserves conditional drift. Event-level observations can distinguish these parameterizations in this specified embedding. The proposition is not a claim that all stochastic path distributions or endpoint distributions coincide.

This observation-dependent scope is central to the paper. The information discarded when reducing event histories to a mean scalar velocity can matter. Standard birth–death inference explicitly exploits event counts and exposure times; it need not restrict itself to the drift (Crawford et al., 2018). The certificate identifies a limitation of a particular observation map. It does not imply that structural gain is unknowable in principle or that richer measurements would necessarily fail.

4. Exact diagnostic counterexamples

4.1 A finite window creates a positive slope without feedback

Set $K = 1$ and $\kappa = \beta = 0$, retaining positive linear loss. The model reduces to

$$\dot{C} = g(1 - C) - \gamma C, \qquad C(T; g, C) = \frac{g}{g+\gamma} + \left(C - \frac{g}{g+\gamma}\right) e^{-(g+\gamma)T}. \tag{8}$$

Its instantaneous normalized response is exactly one. Define instead the endpoint velocity $v_T(g, C) = [C(T; g, C) - C]/T$ and insert it into the same input-contrast normalization:

$$R_T(C) = \frac{v_T(g_h, C) - v_T(g_l, C)}{(g_h - g_l)(1 - C)}. \tag{9}$$

This is a well-defined finite-time estimand, but it is not generally Equation (4). The trajectories begin at the same state and immediately depart along input-dependent paths. Their subsequent difference includes the losses accumulated along those different paths. Matching the starting occupancy does not keep the system at that occupancy during measurement.

Proposition 2. If $0 < g_l < g_h$, $\gamma > 0$, $T > 0$, and $0 \le C < 1$, then $R'_T(C) > 0$, although the structural-gain coefficient is zero.

Proof. Write $A_i = g_i + \gamma$, $H_i = (1 - e^{-A_i T})/T$, $p = g_h H_h/A_h - g_l H_l/A_l$, and $q = H_h - H_l$. Direct substitution yields

$$R_T(C) = \frac{p - qC}{(g_h - g_l)(1 - C)}, \qquad R'_T(C) = \frac{\gamma(H_l/A_l - H_h/A_h)}{(g_h - g_l)(1 - C)^2}. \tag{10}$$

For $x > 0$, the function $\phi(x) = (1 - e^{-x})/x$ has derivative $[(x + 1)e^{-x} - 1]/x^2 < 0$, since $e^x > 1 + x$. Hence $H_l/A_l = \phi(A_l T) > \phi(A_h T) = H_h/A_h$, proving strict positivity. A free-intercept ordinary least-squares slope across any set containing at least two distinct states is positive as well: the pairwise products of state differences and response differences are nonnegative, with a strictly positive product for distinct states. Thus the state–response covariance is positive.

This counterexample is stronger than finding one inconvenient observation interval. Every positive window has the sign problem under these assumptions. It remains important not to overstate it: the theorem does not say that every finite-window estimator is invalid or that all feedback claims are false. It concerns the substitution in Equation (9). A method that targets finite-time transition probabilities directly is estimating a different, explicitly modeled quantity. The example also has a boundary condition: with $\gamma = 0$, the response is constant across starting states, so strict positivity requires positive loss.

At each fixed state below the ceiling, $R_T(C) \to 1$ as $T \to 0$, and

$$R'_T(C) = \frac{\gamma T}{2(1-C)^2} + O(T^2). \tag{11}$$

A shorter window therefore reduces this bias pointwise but does not eliminate it at finite T. The denominator also warns against assuming uniform control close to the ceiling. Increasing the number of replicates can make an incorrectly targeted finite-time slope more precisely positive; it does not change that target into the instantaneous slope. These are the model-specific manifestations of the broader finite-sampling problem studied by Lade (2009) and Anteneodo and Riera (2009).

For the concrete check $g_l = 0.02$, $g_h = 0.08$, and $\gamma = 0.02$, the fastest local equilibrium relaxation time is ten time units. The rule $T = \max(2, \lfloor \tau_{\min}/4 \rfloor)$ selects $T = 2$. At initial occupancies 0.10, 0.50, and 0.80, the exact normalized responses are respectively 0.890139, 0.906346, and 0.961046. On the benchmark's full state grid, the resulting slope is 0.085423, not zero. Figure 1 shows the response curves, including the instantaneous limit. These numbers are analytic evaluations, not fitted biological effects.

4.2 Incorrect ceiling normalization creates an instantaneous effect

Finite-time bias is not the only source of a positive apparent gain. Under an instantaneous zero-feedback model with true ceiling K_t, dividing by an assumed ceiling K_a gives

$$R_a(C) = \frac{1 - C/K_t}{1 - C/K_a}, \qquad R'_a(C) = \frac{1/K_a - 1/K_t}{(1 - C/K_a)^2}. \tag{12}$$

An underestimated ceiling produces a positive slope on states below both ceilings. This artifact persists with perfect state matching and noiseless instantaneous velocities. In the retained stress check, true $K_t = 1$ and assumed ceilings 0.90, 1.00, and 1.10 produce state-

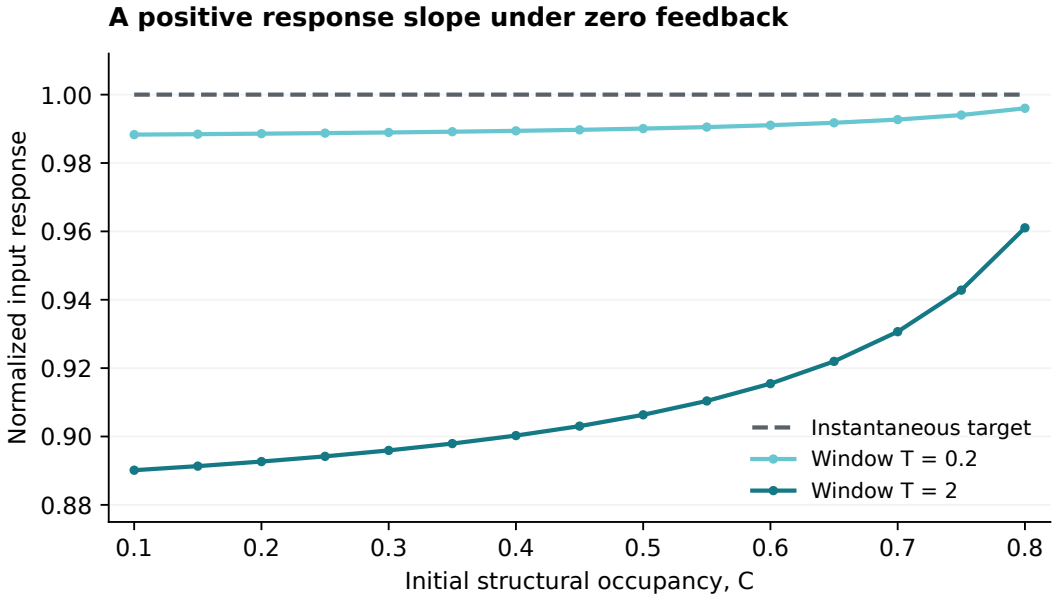


Figure 1: Finite-window response under an exact zero-feedback model. The linear null has $g_l = 0.02$, $g_h = 0.08$, $\gamma = 0.02$, and $K = 1$. The instantaneous response is constant, but endpoint responses increase with starting occupancy at both displayed positive windows. The shorter window reduces rather than eliminates the effect. Curves are analytic evaluations, not uncertainty intervals or Monte Carlo estimates.

grid slopes 0.826272, approximately zero, and -0.328001, respectively. Improving the drift estimator does not repair an incorrect denominator. The state range and uncertainty in K must therefore be part of any substantive application.

4.3 Equal scalar response does not establish scalar sufficiency

A separate issue arises if two networks share occupancy but not other relevant state. Equation (1) predicts a common drift only when those omitted features do not matter or are controlled sufficiently for the chosen comparison. A graph with occupied links distributed among isolated motifs need not have the same catalytic dynamics as a graph with the same occupancy organized into a connected productive subnetwork. Joshi and Craciun (2020) explicitly relate topology to dynamical properties in a family of autocatalytic reaction networks. Their results are not a derivation of Equation (1), but they illustrate why replacing architecture by one density requires justification.

The empirical version of a state-matching claim should therefore name the state that is matched. Matching only C may leave composition, topology, history, or resource allocation different. If those variables alter input responsiveness, a positive slope against C may summarize their joint variation. The benchmark avoids this problem by specifying a genuinely scalar Markov state, rather than claiming to solve hidden-state identification. That deliberate simplification makes the measurement test interpretable while limiting its transfer to systems in which occupancy is not sufficient.

5. A selection comparator: same drift, different event law

5.1 An explicit conventional-selection construction

A positive instantaneous response can also be real without representing the designated structural mechanism. Consider a fixed population of N individuals, of which i carry a binary trait, and let $C = i/N$. Trait carriers have fixed fitness $1 + s$, noncarriers have fitness one, and input scales reproduction. Add trait innovations at rate $g(N-i)$ and trait losses at rate γi. The upward and downward transition rates are

$$\lambda_S(i) = g(N-i) + g(1+s)\frac{i(N-i)}{N}, \qquad \mu_S(i) = \gamma i + g\frac{i(N-i)}{N}. \tag{13}$$

There is no separately designated structural-facilitation coefficient and no interaction term in individual fitness. The frequency dependence in net change arises from replacement opportunities. This construction is an attribution counterexample for a scalar diagnostic, not a claim that trait frequency and network architecture are biologically equivalent.

Proposition 3. The conditional scalar drift of Equation (13) is

$$\frac{\lambda_S - \mu_S}{N} = g(1-C)(1+sC) - \gamma C. \tag{14}$$

It equals the RCE drift with $M = N$, $K = 1$, $\beta = 0$, and $\kappa = s$. Consequently, the instantaneous normalized scalar response has slope $s > 0$ even though the process was specified through fixed-fitness selection and innovations rather than structural facilitation. The equality follows by subtracting the two replacement rates and collecting terms.

5.2 Observations that distinguish the constructions

The equality does not extend to the complete stochastic law. For the RCE embedding $\lambda_R = g(N-i)(1+\kappa i/N)$ and $\mu_R = \gamma i$, with $s = \kappa$, the rate differences are

$$\lambda_S - \lambda_R = \mu_S - \mu_R = g\frac{i(N-i)}{N}. \tag{15}$$

Thus the selection process has additional balanced turnover. Writing $B = (\lambda + \mu)/N^2$ for the conditional second jump moment per unit time gives

$$B_S - B_R = \frac{2gC(1-C)}{N}. \tag{16}$$

Interior-state holding times, event counts, and fluctuations can distinguish these particular parameterizations. The correct conclusion is that normalized scalar drift does not uniquely identify structural facilitation, not that no observation can distinguish selection and structural models.

Even equal nonlinear drift need not imply equal finite-time conditional means. If $\mathcal{L}_S$ and $\mathcal{L}_R$ are the two jump generators, their common instantaneous drift F satisfies

$$(\mathcal{L}_S - \mathcal{L}_R)F = -\frac{2g^2\kappa C(1-C)}{N}. \tag{17}$$

This term enters the second-order time expansion of the expected state. Different fluctuations interact with the curvature of F, so replacing a stochastic mean by the deterministic solution would obscure a genuine difference between the models. The benchmark consequently computes endpoint distributions rather than assuming deterministic mean closure.

The selection comparator was added after the primary benchmark and remains exploratory. It demonstrates a concrete alternative to a unique mechanistic interpretation; it does not replace the primary generator, alter the locked gates, or reproduce a biological network model. Maintaining that chronology prevents a useful later counterexample from being misrepresented as prespecified validation.

6. Synthetic benchmark design

6.1 Ground truth and observation conditions

The primary generator is a finite, time-homogeneous birth–death process on $n = 0, \ldots, M$ with rates

$$\lambda(n) = g(M - n)(1 + \kappa n/M), \qquad \mu(n) = \gamma n + \beta n^2/M. \tag{18}$$

Its exact conditional drift is Equation (1) with $K = 1$. This deliberate encoding supplies the truth against which estimators are tested. It does not test whether natural systems obey that law. The exact mean identity is $dE[C]/dt = E[F(C, g)]$; it is not generally $F(E[C], g)$. The construction also assumes error-free occupancy, exact calibrated inputs, a known ceiling, independent resets, and no omitted state. These favorable conditions isolate estimator behavior before adding biological measurement and intervention difficulties.

Every simulated experiment uses the same fifteen occupancies from 0.10 to 0.80 in steps of 0.05, with $2,048$ independent resets for each input-by-state group. There are $2,000$ Monte Carlo experiments per case and method. The two calibration cases use $M = 100$, $g_l = 0.02$, $g_h = 0.08$, $\gamma = 0.02$, $\beta = 0$, and either $\kappa = 0$ or 0.30. Calibration results are separate from the held-out evaluation and do not substitute for it. No method was modified after

inspecting held-out outcomes.

Twelve additional parameter settings are generated once and each is evaluated at $\kappa = 0$ and 0.30, making twenty-four held-out cases. The design samples M from $\{80, 100, 120\}$, g_l uniformly from $[0.008, 0.040]$, g_h/g_l uniformly from $[2.5, 4.5]$, γ uniformly from $[0.005, 0.045]$, and β uniformly from $[0, 0.025]$. The resulting held-out total is $144,000$ slope estimates across the three methods. Here "held-out" means new synthetic parameter settings, not external empirical prediction. The positive target 0.30 is a computational design choice rather than a biological minimum relevant effect.

6.2 Endpoint and first-event observations

For endpoint observations, the finite generator matrix gives transition probabilities through a matrix-exponential action. Independent terminal-state counts are then sampled from those probabilities. This avoids time-stepping error and accounts for the actual stochastic endpoint distribution. Numerical probability checks test normalization and nonnegativity to roundoff. In the linear null, the resulting mean is compared with the analytic solution. This is a favorable computational benchmark for endpoint estimation: failures are not introduced by an inaccurate simulator.

For first-event observations, each independently reset realization is followed only until its first jump. At the controlled starting state, the total rate $h = \lambda + \mu$ determines an exponential holding time, while the event direction is positive with probability λ/h. Holding time and direction are independent under this model. Rather than generating unnecessary full paths, the implementation samples their sufficient summaries using Gamma and Binomial distributions. This is exact sampling of the designated observation model, not supplying unknown rates to the estimator. The distinction between endpoint likelihoods and event-history information is standard in birth–death inference (Crawford et al., 2018).

Independent samples are used for the different methods. Equal reset counts consequently do not equalize elapsed duration, observed event count, information, or experimental cost. In particular, an endpoint record and a first-event record require different instrumentation. The comparison asks whether a reference design can recover the intended instantaneous coefficient under its own assumptions; it does not establish global efficiency superiority under a matched resource budget. Such a comparison would require a different protocol with an explicit common cost model.

7. Estimation and locked acceptance rules

7.1 Three observation-based estimators

The prescribed endpoint estimator computes the average displacement from each known starting state and divides it by $T = \max(2, \lfloor \tau_{\min}/4 \rfloor)$. The benchmark supplies the exact local deterministic equilibrium relaxation times, with $\tau = (b^2 + 4gd)^{-1/2}$, and takes the minimum over the two inputs. The method is therefore not disadvantaged by uncertainty in estimating the relaxation scale. Its potential defect is instead the finite-time target illustrated in Proposition 2. The second estimator is identical except that its window is $0.1T$. This prespecified comparator tests whether shortening alone suffices at the original observation budget.

The third estimator uses first events at a fixed starting state. With r resets, let S be total waiting time and $\bar{Z}$ the mean of signs $Z = +1$ for upward and $Z = -1$ for downward transitions. The conditional drift and its variance are estimated by

$$\widehat{v} = \frac{(r-1)\bar{Z}}{MS}, \qquad \widehat{V} = \frac{r - 2 + \bar{Z}^2}{M^2 S^2}. \tag{19}$$

The drift estimate is unbiased for $r > 1$; the stated variance estimate is unbiased for $r > 2$. Appendix A gives the derivation from independent holding times and event signs. The formula uses observed exposure and directions, not the generator's unknown gain and loss parameters. It is a reference design exploiting richer observations, not an endpoint-only correction. Exact-form historical priority and general optimality are not claimed.

The first-event design prevents the recorded velocity from mixing states visited after the initial transition. It does not by itself ensure that the nominal starting state is complete. If holding times are not exponential because unobserved state changes during a wait, or if missed events alter directions and exposure, the derivation need not apply. Nor can a physical system always be reset independently without changing its relevant state. These assumptions are deliberately satisfied in the benchmark and remain experimental requirements rather than demonstrated biological capabilities.

7.2 Common response regression and uncertainty

For each method and initial state C_j, the estimated response is

$$\widehat{R}_j = \frac{\widehat{v}_{hj} - \widehat{v}_{lj}}{(g_h - g_l)(1 - C_j)}. \tag{20}$$

The slope is fitted by ordinary least squares with an unconstrained intercept on the fixed occupancy grid. In particular, the line is not forced through one, since doing so would combine an intercept assumption with slope estimation. Let $w_j = (C_j - \bar{C})/\sum_k (C_k - \bar{C})^2$. Then the slope is $\widehat{\kappa} = \sum_j w_j \widehat{R}_j$, and independent group variances propagate as

$$\widehat{\mathrm{Var}}(\widehat{\kappa}) = \sum_j w_j^2 \frac{\widehat{V}_{hj} + \widehat{V}_{lj}}{(g_h - g_l)^2 (1 - C_j)^2}. \tag{21}$$

For endpoint methods, $\widehat{V}$ is the sample variance of terminal occupancy divided by the reset count and squared window. For first events it is Equation (19). These formulas rely on the independent groups supplied by the design. Correlated repeated measurements or shared reset histories would require corresponding covariance terms.

Two-sided nominal 95% intervals use normal critical values, while a one-sided nominal 5% test calls a slope positive when $\widehat{\kappa} - 1.644854\,\widehat{\mathrm{SE}} > 0$. The finite-sample moment identities do not make those regression intervals exact. Coverage and test frequency are therefore evaluated by Monte Carlo rather than assumed. Statistical summaries include bias, empirical standard deviation, root-mean-square error, interval coverage, and positive-test frequency; the retained records also contain Monte Carlo uncertainty and casewise decisions. A single average performance figure would hide precisely the heterogeneous failures that this design seeks to expose.

7.3 Separate reliability and sensitivity gates

The analysis plan was fixed before calibration and held-out outcomes, but it was not externally preregistered. Its thresholds are computational acceptance criteria, not empirically established biological tolerances. For each held-out case, absolute bias plus a 95% Monte Carlo error allowance must be no larger than $\max(0.02, 0.10|\kappa|)$. The complete 95% Wilson interval for observed coverage must lie within 92.5%–97.5%. For each null case, the upper Wilson bound on positive-test frequency must be at most 7.5%. Passing reliability requires all applicable checks across all held-out cases.

Sensitivity is evaluated separately. For every positive held-out case at target $\kappa = 0.30$, the lower 95% Wilson bound on detection probability must be at least 80%. A method can therefore estimate reliably yet fail the all-cases sensitivity requirement. Conversely, rejecting almost every positive case is unpersuasive when the same method rejects null cases at an

excessive rate. No threshold, information budget, or case membership was revised to improve the outcome. The combined gate requires both reliability and sensitivity; failures remain results rather than occasions for retrospective redesign.

The coverage gate also differs from checking whether the nominal coverage value is inside a Monte Carlo interval. Requiring the entire interval to lie inside the tolerance band demands more than a point estimate that looks close to nominal. Similarly, the power gate is not satisfied merely because estimated detection exceeds 80%. The uncertainty bound matters. These deliberately casewise rules make the benchmark demanding, but they were fixed as such and are applied identically to the three procedures. They should not be silently replaced by a weaker post hoc rule.

8. Primary benchmark results

8.1 Reliability distinguishes the procedures

No estimator passes the combined gate. The prescribed endpoint window fails reliability in every held-out case. Its maximum absolute bias is 0.205356; nominal 95% interval coverage ranges from zero to 88.55%, and null positive-test rates range from 64.20% to 100%. Its perfect detection frequency under the chosen positive target cannot be interpreted as useful discrimination because null rejection is also pervasive. The analytic counterexample and stochastic results point to the same practical problem: a finite-time response is treated as an instantaneous gain.

The tenfold shorter window reduces the largest absolute bias to 0.022787 and yields coverage between 93.75% and 95.70%. Nevertheless, only fifteen of twenty-four cases pass the combined reliability checks, and only three of twelve nulls meet the false-positive bound. Null positive-test rates range from 5.35% to 9.80%. Shortening is therefore helpful without consti-

tuting a validated universal fix at this budget. Its performance should neither be conflated with the original window's severe failure nor promoted to a complete pass.

The first-event reference passes reliability in all twenty-four cases, including all twelve null false-positive checks. Maximum absolute bias is 0.009174, coverage ranges from 93.85% to 95.80%, and null positive-test rates range from 4.20% to 5.95%. These results support calibration of this estimator under the tested Markov observation assumptions. They do not validate untested hidden-state dynamics, imperfect timing, unknown input conversion, or biological resets. Table 1 summarizes the case ranges, and Figure 2 separates null behavior from positive detection.

Table 1: Held-out case ranges at the locked observation budget. Coverage is for nominal 95% intervals. Null positives refer to a nominal one-sided 5% test; detection refers to the synthetic target $\kappa = 0.30$. Each range spans cases, not a confidence interval. Values are rounded for display.

Estimator	Maximum absolute bias	Coverage (%)	Null positives (%)	Detection (%)
Prescribed window	0.205356	0–88.55	64.20–100	100–100
One-tenth window	0.022787	93.75–95.70	5.35–9.80	60.55–100
First event	0.009174	93.85–95.80	4.20–5.95	51.45–100

8.2 Sensitivity failures remain part of the result

The first-event method passes the sensitivity requirement in nine of twelve positive cases, not all twelve. Its equal-case mean detection is 90%, but its range is 51.45%–100%. The mean therefore does not support a statement that the design is sufficiently sensitive in every tested setting. The short-window procedure also passes nine positive cases; the prescribed window passes the sensitivity rule while failing reliability. Keeping the two gates separate prevents high rejection rates from being mistaken for strong measurement performance.

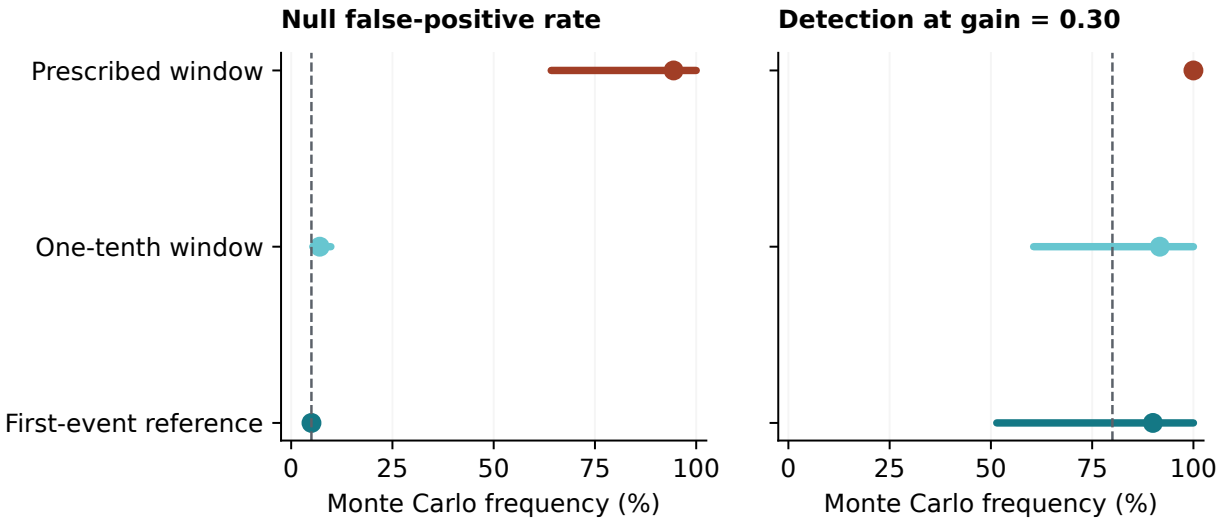


Figure 2: Null-test frequency and positive detection across held-out settings. Horizontal segments show minimum-to-maximum case frequencies and points show equal-case means, not uncertainty intervals. The first-event design improves null calibration but retains low-power cases. Reference lines indicate the false-positive ceiling and detection target; the actual acceptance rules additionally require the relevant Wilson bounds to satisfy those thresholds in every case. All values are synthetic and conditional on exact state and input information.

The retained first-event failures are settings 05, 03, and 06. Their detection estimates are respectively 51.45%, 74.85%, and 81.55%, with Wilson intervals 49.26%–53.64%, 72.90%–76.70%, and 79.79%–83.19%. The final setting fails despite an estimated rate above the target because its lower bound remains below 80%. These are sensitivity failures under the locked decision rule, not evidence that the first-event estimator has the original window's false-positive defect. Increasing reset counts might improve future power, but that would be a new validation design rather than a correction to this result.

The casewise records allow those failures to be traced to their exact generated parameters without selecting replacements. They also preserve uncertainty estimates for every retained Monte Carlo realization. This matters because a benchmark that reports only successful settings can make both calibration and robustness appear stronger than they are. Here the all-cases criterion is intentionally visible in the conclusion: the reference observation design resolves the specific finite-window targeting problem within the supplied model, yet the fixed budget is not enough for the complete sensitivity claim.

9. Attribution controls and exploratory selection results

9.1 What an ideal disabling contrast would identify

Suppose a switch $b \in \{0, 1\}$ changes only the designated structural coupling, so that the conditional velocity is

$$v_b(g, C) = g[1 + b\kappa C](1 - C/K) + H(g, C), \tag{22}$$

with the same nuisance function H, input calibration, ceiling, and relevant state in both switch conditions. A difference between their input contrasts gives

$$D(C) = \frac{[v_1(g_h, C) - v_1(g_l, C)] - [v_0(g_h, C) - v_0(g_l, C)]}{(g_h - g_l)(1 - C/K)} = \kappa C. \tag{23}$$

The nuisance cancels even when it depends on input, provided it is genuinely shared. This is an algebraic design identity, not evidence that a feasible intervention has the required specificity. Observing a nonzero D cannot itself demonstrate that H was unchanged.

The prespecified shared-nuisance control adds a nonstructural state-dependent gain coefficient $a = 0.40$ to both conditions. A raw positive slope therefore persists when the designated coupling is disabled. With zero coupling, the first-event contrast mean is approximately zero and its positive-test frequency is 4.70%. With coupling 0.30, its mean is 0.298681 and detection is 86.70%. These results illustrate how a correctly designed contrast can remove the specified shared nuisance; they are not evidence that every nonstructural explanation has been removed in a real experiment.

9.2 Correct estimation of a confounded contrast

In the nuisance-changing switch, structural coupling remains zero while a changes from 0.40 to 0.65. The true contrast slope is consequently 0.25. The first-event mean estimate is 0.252655, with coverage 94.80%. Its numerical accuracy is not a success of structural attribution: the same observable contrast now comes from a nuisance change. This negative control separates a reliable estimator from a valid causal interpretation. No reduction in sampling error can determine which term changed if the contrast alone carries identical information about both explanations.

This distinction also clarifies randomization. Randomly assigning input at a matched state can identify the effect of that assigned input under the design's assumptions. It does not automatically identify the effect of occupancy or prove why occupancy modifies responsive-

ness. Likewise, a disabling label is not a guarantee of a coupling-specific intervention. A substantive application needs measurements or design constraints that address alternative effects on conversion, selection, loss, and relevant state. The synthetic controls demonstrate the consequences of meeting or violating one such constraint; they do not establish it experimentally.

9.3 Later selection illustration

The exploratory binary-trait model uses $N = 100$ and fixed advantage $s = 0.40$. With first-event estimation, its raw response slope has mean 0.402608, coverage 94.95%, and positive-test frequency 99.05%. The corresponding sham-switch contrast has mean -0.003239, coverage 93.70%, and positive-test frequency 5.35%. The raw slope is genuinely positive: calling it an estimation false positive would confuse statistical truth with mechanistic labeling. What fails is the inference that a positive scalar slope uniquely indicates the designated structural process.

These simulations complement, but are not needed to prove, the drift equality in Proposition 3. They show how a reliable measurement procedure can recover a selection-generated slope. Since the comparator was added later, none of its outcomes is included in the primary acceptance counts. It also concerns binary-trait frequency, not network architecture, so it cannot establish that every conventional-selection model reproduces every structural-network observable. Its value is the explicit counterexample to uniqueness at the scalar-drift level.

10. Discussion

10.1 What has been established

The analysis supports a diagnostic contribution rather than a demonstrated evolutionary mechanism. Proposition 1 gives an exact, admissible fixed-input ambiguity in the deterministic parameterization. Proposition 2 proves that a proposed finite-window substitution can generate the desired sign under zero feedback. Equation (12) isolates a separate error arising from normalization, while Proposition 3 supplies a nonstructural generative route to the same instantaneous scalar drift. Finally, the locked benchmark demonstrates that removing the identified finite-window defect can improve calibration without ensuring all-cases sensitivity. Each result has a specific inferential target, and none requires a claim that feedback never occurs.

The combination matters because superficially similar failures need different remedies. Fixed-input nonidentifiability requires new information or justified constraints. Finite-window targeting requires a suitable observation or inferential model. Incorrect normalization requires external calibration of the ceiling. A selection-generated slope requires additional mechanism-discriminating observables or interventions. Insufficient sensitivity requires a prospectively evaluated information budget or design. Treating all of these as ordinary noise would encourage collecting more precise measurements of an ambiguous quantity rather than addressing the actual source of uncertainty.

The reference estimator is correspondingly limited but useful. Its unbiasedness follows from an exact observation model; the Monte Carlo exercise then checks the approximate regression inference actually used. This avoids two common logical substitutions: assuming an unbiased point estimate automatically has calibrated intervals, and assuming calibrated intervals imply adequate power. The retained sensitivity failures are especially informative because they

occur despite exact state knowledge and a deliberately favorable input calibration. They identify a limit of the present budget without supporting a conclusion about impossibility at a different budget.

10.2 Relationship to prior work and limits of novelty claims

Autocatalytic models already connect existing organization to future growth of organization. Jain and Krishna (1998) are directly relevant to this conceptual point, while Filisetti et al. (2011) connect catalytic-network dynamics to energy constraints and intake rate. Those precedents rule out presenting generic facilitation or generic energy dependence as the novelty. The scalar quadratic form also makes broad claims of a new dynamical family inappropriate. The more defensible contribution lies in deriving and checking restrictions for a particular parameterized diagnostic, together with counterexamples and a reproducible calibration package.

Methodologically, the paper is consistent with rather than an alternative to established identifiability and stochastic-inference work. Villaverde et al. (2019) emphasize that identifiability changes with the input design. Lade (2009) and Anteneodo and Riera (2009) analyze finite-time departures from infinitesimal drift and diffusion. Crawford et al. (2018) review inference from continuous and discrete birth–death observations. The present first-event formula uses standard holding-time and binary-event ingredients. Without a dedicated priority investigation, it would be inappropriate to label that formula the first estimator of its kind or to claim that these general inferential issues are newly discovered.

The exact-law comparison with energy-constrained autocatalysis remains incomplete because the closest assessed conference paper was available only through its abstract and bibliographic records. Its abstract explicitly establishes conceptual relevance but cannot establish the presence or absence of the same input-response parameterization. The appropriate re-

sponse is to delimit the claim, not infer originality from unavailable text. The manuscript therefore offers its derivations and benchmark for evaluation without claiming certified historical priority. This qualification applies separately to the model equation, its interpretation, the estimator expression, and the integrated application-specific diagnostic package.

10.3 Relevance to artificial-life research

For artificial life, a controlled toy system is valuable partly because its mechanism and observation process can be specified separately. Here that separation makes counterexamples constructive: one can know that structural gain is absent, still obtain a positive normalized slope, and identify why. Likewise, one can deliberately encode a positive slope and test whether an estimator recovers it without treating recovery as discovery. The contribution is therefore a test of diagnostic reasoning within a synthetic life-like growth model, not an out-of-scope claim about general machine intelligence or a claim that one bounded coordinate captures life.

Recent discussion of open-ended evolution reinforces the need for this distinction. Channon et al. (2024) distinguish agreement about observable hallmarks from agreement about their mechanisms. Moreno et al. (2024) explicitly find that novelty, adaptation, and different operational complexity measures need not move together. This article addresses a different model and does not extend their empirical simulation results. Its related concern is that a convenient scalar statistic should not silently inherit a broader mechanistic interpretation. Edge occupancy must retain its own definition, and any relation to functional complexity or adaptation must be established rather than assumed.

The boundedness of Equation (1) also sets a clear scope boundary. A saturating occupancy model can study local growth responses, parameter ambiguity, or intervention diagnostics without constituting a model of indefinite novelty production. Adding an evolving space

of possible components, changing ceilings, or new organizational levels could pose interesting questions, but would change the model and its estimands. Those extensions cannot be inferred from the present equilibrium analysis or credited as demonstrated open-ended evolution. Keeping this boundary explicit allows a modest model-analysis result to be evaluated on its actual mathematical and computational merits.

10.4 From a diagnostic identity to a falsifiable application

A stronger application would begin by fixing definitions before fitting gain. Valid slots, occupancy, ceiling, effective input, and the relevant observation state should have meanings that do not depend on whether the resulting slope is positive. At least some tests should examine the exact affine restriction, including curvature and consistency across input pairs, rather than only testing a slope against zero. A reproducible departure from these restrictions, when normalization and state matching are justified, would reject this particular response law. It would not establish the absence of all structural feedback in the system.

Mechanistic work should specify what feature of structure is predicted to alter production and what intervention should remove and restore that feature. The relevant comparison is not merely high versus low occupancy. It should address alternative changes in selection, conversion, loss, composition, and architecture. The shared-nuisance identity provides a useful target for design, while the nuisance-changing control illustrates why a positive switch contrast alone is insufficient. Event counts and fluctuation measurements can help distinguish the explicit stochastic alternatives analyzed here, though a broader alternative class may require additional observations.

Prospective estimator development should retain the present benchmark as a record rather than silently increasing its budget. A new design could fix a cost measure, select observation times, incorporate uncertain ceilings, or test hidden-state robustness. It should define success

before seeing the new held-out results and include nulls that preserve plausible conventional dynamics. In particular, a method that uses richer event observations should not be called more efficient merely because it outperforms endpoints at equal reset count. The unmeasured costs of timing, intervention, and independent initialization are part of the design problem.

10.5 Reproducibility, limitations, and computational provenance

The supplementary package retains the original locked protocol, executable benchmark, generated case definitions, raw replicate-level estimates, per-case summaries, supplementary controls, and separately labeled exploratory selection analysis. Fixed seeds identify calibration, held-out design, held-out data, and supplementary stages; the exploratory comparator has its own seed. The primary script supports numerical checks, calibration, held-out simulation, supplementary controls, and assembly. Regenerating those stages is distinct from reaggregating the retained results, which provides a faster consistency check without new simulation.

The matrix-exponential linear-null mean agrees with the analytic solution to a maximum discrepancy of 2.22×10^{-16}, while the first-event variance identity check differs by at most 1.08×10^{-19}. Additional manuscript checks verify the algebraic invariances and event-law distinctions, evaluate the exact finite-window derivative, and reaggregate the retained held-out estimates against their reported summaries and decisions. These are internal computational checks. Neither numerical precision nor repeatability turns the encoded model into empirical evidence, and internal replay should not be described as independent replication.

The principal scientific limitations are explicit. Only one finite scalar Markov embedding is used in the primary estimator comparison. Inputs, ceiling, starting state, and event timing are known without error. The positive effect and tolerance thresholds are synthetic choices. Equal resets do not impose equal cost. The all-cases sensitivity target is not met. Neither

biological feasibility nor a real coupling-specific switch is demonstrated. The selection example is a scalar attribution counterexample, not a network-level biological reconstruction. These limitations delimit what the reported passes and failures mean; they do not erase the exact counterexamples or justify omitting unsuccessful cases.

The analysis also does not reproduce the unavailable graph-evolution companion implementation associated with earlier RCE material. No detection rates from that implementation are treated as independently verified results here. No biological observations, astronomical observations, or engineering demonstrations are included. The present article can stand as a theoretical and computational diagnostic study, but conclusions about a real mechanism require observations and intervention evidence not supplied by this package. Historical-priority uncertainty is an additional literature limitation, separate from the reproducibility of the calculations actually presented.

11. Author declarations

Authorship and contributions: Heinz Kahler is the sole author, conceptualized the research, designed the framework, and takes full responsibility for all claims, code, and results.

Funding: This research received no external funding.

Competing interests: The author declares no competing interests.

Ethical scope: This study contains mathematical analysis and synthetic computational experiments, not experiments on human participants or animals. No institutional approval or exemption is asserted.

AI-assisted preparation: AI tools were utilized to assist with literature discovery, mathematical exposition, code and figure generation, computational checking, and manuscript

drafting and formatting. The author independently reviewed, verified, and takes full responsibility for all scientific content and conclusions. AI tools are not authors, and their assistance is not represented as independent scientific peer review.

12. Conclusion

A bounded input–structure equation can generate a useful, falsifiable response restriction without uniquely identifying its intended mechanism. The explicit fixed-input invariance, strictly positive finite-window null slope, ceiling artifact, and drift-equivalent selection process show where inference from scalar response fails. A first-event observation design removes the specific finite-window targeting problem and passes reliability under the tested assumptions, but it does not meet the locked all-cases sensitivity requirement or establish structural causation. The supported contribution is an application-specific diagnostic analysis with reproducible successes and failures. It is not empirical validation of RCE, evidence for a universal evolutionary principle, or certification of historical originality.

Data and code availability

The accompanying supplement containing the executable analysis, retained synthetic results, and reproducible code is available on the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/DU875. The execution code is distributed under the MIT License, synthetic data files are dedicated to the public domain under CC0 1.0, and this manuscript is distributed under the terms of the Creative Commons Attribution 4.0 International (CC-BY 4.0) license.

Appendix A. First-event moment derivation

Fix a starting state in the specified time-homogeneous birth–death process, with $h = \lambda + \mu > 0$ and $m = (\lambda - \mu)/h$. With r independent resets, the sum of first-event waiting times obeys $S \sim \text{Gamma}(r, \text{rate } h)$. Each event sign has mean m and variance $1 - m^2$, and the sample mean sign is independent of S. Therefore

$$E[\bar{Z}] = m, \quad E[\bar{Z}^2] = m^2 + \frac{1 - m^2}{r}, \quad E[S^{-1}] = \frac{h}{r - 1}, \quad E[S^{-2}] = \frac{h^2}{(r - 1)(r - 2)}. \tag{A1}$$

The inverse moments require $r > 1$ and $r > 2$, respectively. By independence,

$$E[\widehat{v}] = \frac{r - 1}{M} E[\bar{Z}] E[S^{-1}] = \frac{hm}{M} = \frac{\lambda - \mu}{M}. \tag{A2}$$

Using the second moments and subtracting the squared expectation gives

$$\text{Var}(\widehat{v}) = \frac{h^2(r - 1 + m^2)}{M^2 r(r - 2)}. \tag{A3}$$

Similarly,

$$E\left[\frac{r - 2 + \bar{Z}^2}{M^2 S^2}\right] = \frac{h^2(r - 1 + m^2)}{M^2 r(r - 2)}. \tag{A4}$$

This proves both claims in Equation (19). It does not prove exact normality, optimality among possible event-history estimators, robustness to missed events, or priority of the displayed expressions. The regression interval remains an approximation whose actual behavior is evaluated in the synthetic benchmark.

Appendix B. Sampling, slope weights, and Monte Carlo decisions

For the endpoint methods, the initial state is deterministic, so subtracting it adds no sampling variance. If s_C^2 is the ordinary unbiased sample variance of terminal occupancies, the estimated variance of mean velocity is $s_C^2/(rT^2)$. Because the two input groups and different state groups are independently generated, Equation (21) follows from variance propagation for a fixed linear combination. This is not a residual-based homoskedastic regression standard error. The uncertainty comes from observed group variation and the known fixed design.

For a case with B Monte Carlo experiments, bias is the mean of $\widehat{\kappa} - \kappa$, and its Monte Carlo standard error is the empirical standard deviation of estimates divided by $\sqrt{B}$. The bias gate uses the absolute estimated bias plus the two-sided normal 95% allowance. Coverage and positive-test frequency are binomial proportions across the Monte Carlo experiments. For x successes, $p = x/B$, and z the two-sided normal 95% critical value, the Wilson interval is

$$\frac{p + z^2/(2B) \;\pm\; z\sqrt{p(1-p)/B + z^2/(4B^2)}}{1 + z^2/B}. \tag{B1}$$

The gates apply the lower and upper bounds exactly as specified in Section 7. A case's approximate slope interval and the Monte Carlo interval around its coverage are different objects. Retaining both prevents a nominal confidence level from being mistaken for a demonstrated frequentist property.

The calibration seed is 671053; held-out parameter generation uses 919547; held-out observations use 335021; and supplementary controls use 762313. The later exploratory selection run uses 403187. Seeded replay facilitates reproducibility but does not create independent

empirical replication. The supplement includes dependency information, commands for original simulation and retained-data verification, plot data, and separate publication-quality figure files. No numerical result was improved by dropping a held-out case or altering the locked rule after observing its outcome.

References

Anteneodo, C., & Riera, R. (2009). Arbitrary-order corrections for finite-time drift and diffusion coefficients. *Physical Review E, 80*(3), 031103. https://doi.org/10.1103/PhysRevE.80.031103

Channon, A., Bedau, M. A., Packard, N. H., & Taylor, T. (2024). Editorial introduction to the 2024 special issue on open-ended evolution. *Artificial Life, 30*(3), 300–301. https://doi.org/10.1162/artl_e_00445

Crawford, F. W., Ho, L. S. T., & Suchard, M. A. (2018). Computational methods for birth–death processes. *WIREs Computational Statistics, 10*(2), e1423. https://doi.org/10.1002/wics.1423

Filisetti, A., Graudenzi, A., Serra, R., Villani, M., De Lucrezia, D., & Poli, I. (2011). The role of energy in a stochastic model of the emergence of autocatalytic sets. In T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo, & R. Doursat (Eds.), *Advances in artificial life: 20th anniversary edition—Back to the origins of Alife, ECAL 2011* (pp. 226–233). MIT Press. https://doi.org/10.7551/978-0-262-29714-1-ch037

Jain, S., & Krishna, S. (1998). *Autocatalytic sets and the growth of complexity in an evolutionary model* [Preprint]. arXiv. https://arxiv.org/abs/adap-org/9809003

Joshi, B., & Craciun, G. (2020). *Autocatalytic networks: An intimate relation between*

network topology and dynamics [Preprint]. arXiv. https://arxiv.org/abs/2006.01384

Lade, S. (2009). *Finite sampling interval effects in Kramers–Moyal analysis* [Preprint]. arXiv. https://arxiv.org/abs/0905.4324

Moreno, M. A., Rodriguez Papa, S., & Ofria, C. (2024). *Case study of novelty, complexity, and adaptation in a multicellular system* [Preprint]. arXiv. https://arxiv.org/abs/2405.07241

Villaverde, A. F., Evans, N. D., Chappell, M. J., & Banga, J. R. (2019). Input-dependent structural identifiability of nonlinear systems. *IEEE Control Systems Letters, 3*(2), 272–277. https://doi.org/10.1109/LCSYS.2018.2868608