

Ablation Invariance

What Null Ablations Reveal About Hidden Architecture

Flyxion

Independent Researcher

September 2026

Abstract

The experimental sentence “we removed C , and nothing happened” is usually read as evidence that C is unimportant. This essay argues that the sentence supports far less than that, and that in many cases it supports something more interesting. A null ablation certifies only that one finite intervention, applied under specific conditions and read through one observable, failed to move that observable. It does not establish that the sensitivity $\partial O/\partial C$ vanishes, that C is dispensable, or that the system after the intervention is the system with C subtracted. Adaptive, redundant, and degenerate systems answer an ablation by reconfiguring, so the object actually measured is $\mathcal{R}(S, \text{do}(C = 0))$ and not $S - C$. We give a formal treatment of this gap. We separate component necessity from path necessity, define a compensation gap that lower-bounds hidden rerouting, prove that single-ablation nulls are compatible with exponentially many distinct architectures, show that joint ablations expose the interaction structure that singles conceal, and prove that the same nulls become fully informative once the observable is enriched from a scalar to a residual vector. The last result outlines a form of structural tomography in which sequences of ablations and their compensations reconstruct an invisible network. Invariance is reframed as change confined to a fibre of the observation map, so that a null ablation reports motion within the observational quotient and not the absence of motion, and the null is shown to be compatible with unchanged output at nonzero cost and reduced margin. A minimal reporting standard is proposed.

Contents

1	The Statement	5
2	Formal Setting	6
2.1	Systems, interventions, observables	6
2.2	Effects	7
3	Four Readings of a Null Ablation	7

4	Sources of Silence	9
4.1	Compensation	9
4.2	Saturation	9
4.3	Degeneracy	10
4.4	Measurement coarseness	10
4.5	Load redistribution	10
4.6	Level-set coincidence	10
4.7	Tolerances do not compose	11
5	The Geometry of Null Sets	11
5.1	Two geometries of silence	13
5.2	Compensating configurations form a family	13
6	Ablation as Quotient Observation	14
7	Component Necessity and Path Necessity	15
7.1	Structure functions, paths, and cuts	15
7.2	Degeneracy as invariance	16
7.3	An exact model of compensation	17
7.4	The subtraction fallacy	18
8	Timescales: The Horizon Spectrum	19
8.1	The horizon curve of the homeostat	20
8.2	Beyond a single time constant	21
9	The Cost of Invariance	21
9.1	Compensation forces cost	22
10	Joint Ablations and the Interaction Signature	22
10.1	The interaction term	22
10.2	Why single ablations misattribute importance	23
10.3	Order of the required ablation	23
11	Ablation Profiles, Margins, and the Sampling Failure of Random Ablation	24
11.1	Profiles	24
11.2	Shape distinguishes additivity from redundancy	24

11.3 Nulls consume margin	25
11.4 Why random ablation misses small cuts	25
12 Slack, Reserve, and Distance to Failure	26
12.1 A homeostat with a limited actuator	26
13 Non-Identifiability from Single Ablations	27
14 The Converse: Non-Null Results Are Also Over-Read	28
15 Removal and Restoration: The Rebound Test	29
16 Graded Interventions and Ablation Curvature	30
16.1 Endpoint coincidence	30
16.2 Dose-response in the limited homeostat	31
17 Statistical Certification of a Null	31
17.1 Equivalence testing	32
17.2 Many components	32
18 Consequences for Interpretation and Reporting	33
18.1 Domains	33
18.2 A minimal reporting standard	34
19 Designing for Informative Nulls	34
20 Connections to Broader Frameworks	35
21 Objections and Replies	36
22 Invariance Is Relative to an Observer	37
22.1 Refinement	37
22.2 Richer observation shrinks the compatible set	38
22.3 No privileged observer	38
23 From Null Results to Structural Tomography	39
23.1 A linear model	39
23.2 Identifiability	39

23.3 Sequences of interventions	40
24 Preserved Distinctions	41
25 Conclusion	42
A Proofs of Theorems 13.1 and 13.2	43
B Worked Numerical Examples	44
B.1 Saturation	44
B.2 Parallel supports	44
B.3 Two-component OR system	45
B.4 Homeostat	45
C Reporting Checklist	45

1 The Statement

Consider the sentence

We removed C , and nothing happened.

It appears, in one form or another, in ablation studies of neural networks, in gene knockout papers, in lesion neuroscience, in engineering inspections, in institutional reform, and in ordinary household repair. The sentence is usually followed by an inference: C was not doing anything, or at least nothing that mattered. The inference is so habitual that it is rarely written down, which is a reason to write it down and examine it.

Three distinct objects are being run together in the sentence. The first is the *finite intervention* $\text{do}(C = 0)$: a specific act performed on a specific system at a specific time, replacing the state or parameters of C by a null value. The second is the *sensitivity* $\partial O / \partial C$: a local property of the response of an observable O to changes in C around some operating point. The third is the *measured observable* O itself, which is one projection of the system’s behavior, taken at one resolution, under one set of conditions. The inference from “nothing happened” to “ C is irrelevant” silently identifies the first with the second and treats the third as exhaustive.

The central proposition of this essay is that the identification fails:

$$\Delta O \approx 0 \not\Rightarrow \frac{\partial O}{\partial C} = 0. \quad (1)$$

Removing component C and observing essentially no change in output O does not establish that C was causally irrelevant. It establishes that the intervention $\text{do}(C = 0)$ did not move the measured observable under the tested conditions. Compensation by another pathway, saturation, degeneracy, coarse measurement, or redistribution of load can each make the intervention observationally silent.

Even the weakest reading of (1), that a finite change need not reflect a derivative, is worth stating with a concrete case.

Example 1.1 (Finite null, nonzero derivative). Let $O(c) = \sin(2\pi c)$ with baseline $c = 1$ and ablated value $c = 0$. Then $O(1) = O(0) = 0$, so $\Delta O = 0$ exactly. Yet $\partial O / \partial c = 2\pi \cos(2\pi c)$ equals 2π at both the baseline and the ablated point. The intervention landed on another point of the same level set of O .

This example is artificial, but it isolates the logical structure: a finite intervention compares two points, and two points can share an observable value while the function is steep at both. The interesting cases are those in which the coincidence is not an accident of periodicity but a consequence of how the system is built. In those cases the null result is not an absence of information. It is evidence about architecture.

The plan of the essay is as follows. Section 2 fixes the formal setting, including an operator $\mathcal{R}$ that models the system’s own response to being ablated. Section 3 distinguishes four readings of a null ablation, and Section 4 catalogues the sources of silence. Sections 5 and 6 give these sources a single geometric form: a null ablation is a return to the same observational fibre, and an experiment observes only the quotient of configuration space by that relation.

Section 7 separates component necessity from path necessity and proves the main gap between $S - C$ and $\mathcal{R}(S, \text{do}(C = 0))$. Section 8 treats the adaptation horizon and shows that a null at one horizon is a quantitative bound on the system’s adaptation rate. Section 9 shows that output invariance can coexist with a permanent compensation cost. Section 10 shows how joint ablations expose interaction, and Section 11 extends this to whole ablation profiles, redundancy margins, and the sampling failure of random ablation. Section 12 shows that a null can coexist with consumed reserve and reduced distance to failure. Section 13 proves that single-ablation nulls are compatible with exponentially many architectures. Section 14 treats the converse, since a non-null result is equally over-read. Section 15 introduces the rebound test, which certifies compensation without a second ablation. Section 16 treats graded interventions and shows that a zero derivative is no more decisive than a null endpoint. Section 17 gives the statistical conditions under which a null can be certified at all. Sections 18, 19 and 20 draw consequences for reporting, for the design of systems and experiments, and for a broader family of frameworks, and Section 21 answers likely objections. Section 22 states the order relation among observers that underlies the resolution results, Section 23 proves that enriching the observable turns the same nulls into a reconstruction method, and Section 24 states the general principle of preserved distinctions.

2 Formal Setting

2.1 Systems, interventions, observables

Let $K = \{1, \dots, n\}$ index the components of a system. Each component $c \in K$ has a configuration space Θ_c with a distinguished null value $0_c \in \Theta_c$. The joint configuration space is $\Theta = \prod_{c \in K} \Theta_c$. For $C \subseteq K$ write $\theta = (\theta_C, \theta_{-C})$ for the decomposition of a configuration into the coordinates in C and the rest.

Definition 2.1 (Observable and tolerance). An observable is a map $O : \Theta \rightarrow (Y, d)$ into a metric space. A tolerance $\varepsilon \geq 0$ is fixed in advance and encodes both measurement resolution and the smallest change regarded as meaningful. A change is ε -null if $d \leq \varepsilon$.

Definition 2.2 (Ablation). The intervention $\text{do}(C = 0)$ sets $\theta_C := 0_C$ and holds it there. Following the usual convention of intervention calculus [10], the intervention overrides whatever mechanism ordinarily determined θ_C .

The definition so far describes a static system. Real systems adapt. A neural network can be retrained or can re-route activations through other layers at inference time, an organism can upregulate paralogs, a structure can redistribute stress, and an institution can reassign duties. To capture this we attach an adaptation dynamics.

Definition 2.3 (Adaptive system). An adaptive system is a triple (Θ, V, O) where V is a vector field on Θ (the adaptation dynamics) and O is an observable. The baseline configuration θ^0 is a rest point: $V(\theta^0) = 0$. Under the intervention $a = \text{do}(C = 0)$ the coordinates θ_C are pinned at 0_C and the free coordinates evolve by the restricted flow

$$\dot{\theta}_{-C} = V_{-C}(0_C, \theta_{-C}), \quad \theta_{-C}(0) = \theta_{-C}^0.$$

Write ψ_T^a for the time- T map of this flow.

Definition 2.4 (Reconfiguration operator). The reconfiguration operator at horizon $T \geq 0$ sends a system and an intervention to the configuration reached after the system has adapted for time T :

$$\mathcal{R}_T(S, a) = (0_C, \psi_T^a(\theta_{-C}^0)).$$

The *static ablation* is $\mathcal{R}_0(S, a) = (0_C, \theta_{-C}^0)$: the subtraction $S - C$ with everything else frozen. The *chronic ablation* is $\mathcal{R}_\infty(S, a) = \lim_{T \rightarrow \infty} \mathcal{R}_T(S, a)$ when the limit exists.

For a non-adaptive system $V \equiv 0$ and $\mathcal{R}_T = \mathcal{R}_0$ for all T . The distinction matters only when the system responds to being perturbed, which is the case of interest.

2.2 Effects

Definition 2.5 (Effects). The static effect and the reconfigured effect at horizon T are

$$\Delta^0(C) = d(O(\theta^0), O(\mathcal{R}_0(S, a))), \quad \Delta^T(C) = d(O(\theta^0), O(\mathcal{R}_T(S, a))).$$

The *compensation gap* at horizon T is

$$\kappa_T(C) = d(O(\mathcal{R}_0(S, a)), O(\mathcal{R}_T(S, a))).$$

Experiments differ in which of these they measure. An acute pharmacological block, an activation patch applied at inference, or a clamp applied for a single forward pass approximates Δ^0 . A genetic knockout raised through development, a network retrained after pruning, or a structure examined years after a member is removed approximates Δ^∞ . The experimental sentence “we removed C ” does not say which. That omission is one of the principal sources of confusion the essay is concerned with.

Lemma 2.6 (Compensation is lower-bounded by the gap between effects). *For every T ,*

$$|\Delta^0(C) - \Delta^T(C)| \leq \kappa_T(C).$$

In particular, if $\Delta^0(C) > \varepsilon$ and $\Delta^T(C) \leq \varepsilon$, then $\kappa_T(C) \geq \Delta^0(C) - \varepsilon > 0$.

Proof. Let $p = O(\theta^0)$, $q_0 = O(\mathcal{R}_0(S, a))$, $q_T = O(\mathcal{R}_T(S, a))$. By the triangle inequality, $d(p, q_0) \leq d(p, q_T) + d(q_T, q_0)$ and $d(p, q_T) \leq d(p, q_0) + d(q_0, q_T)$. Hence $|d(p, q_0) - d(p, q_T)| \leq d(q_0, q_T) = \kappa_T$. The second claim follows by substituting the hypotheses. $\square$

The lemma is elementary but it converts a heuristic into a bound. If an acute ablation produces a visible effect and a chronic ablation produces none, the system has demonstrably done something between the two, and the magnitude of what it did is at least the difference. The null is thereby a positive measurement of the system’s capacity to reconfigure.

3 Four Readings of a Null Ablation

The experimental statement

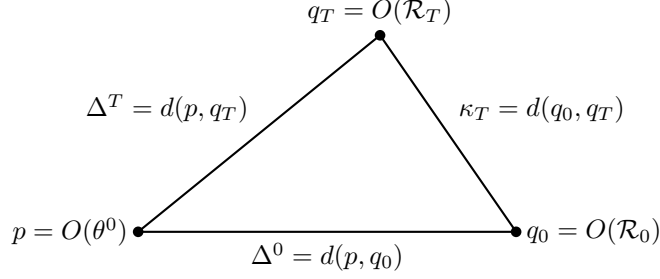


Figure 1: The compensation-gap triangle in observable space. The triangle inequality gives $|\Delta^0 - \Delta^T| \leq \kappa_T$, which is the content of Lemma 2.6.

We removed C , and nothing happened.

can be given four progressively richer interpretations. They are not nested in logical strength. Each is a different claim about the system, and each requires different evidence to be supported and different evidence to be refuted.

First reading: C does nothing. The strongest and most naive reading holds that C contributes nothing to the behavior of interest in any regime. Evidence for it would consist of null effects across the full range of operating conditions, at fine resolution, under joint ablations with every plausible partner, and at every adaptation horizon. No ablation study of realistic size provides this. The reading is almost always asserted on the basis of a single condition.

Second reading: C is redundant. Here C does contribute, but its contribution is duplicated by another component, so that the system tolerates its loss. This is a claim about the structure of the system at baseline, independent of any adaptive response. It predicts that a joint ablation of C and its duplicate is non-null, and it predicts that the acute effect Δ^0 is already null. It is supported by demonstrating both.

Third reading: another pathway compensated. The contribution of C was real and was actively made up by the system. This claim concerns dynamics. It predicts a non-null acute effect $\Delta^0(C)$ that shrinks with horizon T , and it predicts a residual in other variables of the system, the trace of the compensating pathway. By Lemma 2.6 it is supported by the pair $(\Delta^0 > \varepsilon, \Delta^T \leq \varepsilon)$.

Fourth reading: the ablation changed the measured system. The most general reading holds that the object measured after the intervention is not the original system minus C . It is a different system, produced by the intervention together with the system's response. Formally, the experiment measures

$$\mathcal{R}(S, \text{do}(C = 0)) \quad \text{and not} \quad S - C.$$

The third reading is a special case in which the change can be attributed to a specific compensating pathway. The fourth allows that the reconfiguration may be diffuse, nonlocal, or not

attributable to any single component, and that the residual state of the whole system, and not merely a scalar output, is what changed.

Reading	Claim	Evidence required	Refuted by
C does nothing	No contribution in any regime	Null across conditions, resolutions, joint ablations, horizons	Any non-null joint or acute effect
C is redundant	Contribution duplicated at baseline	Null single, non-null joint with a partner	Non-null single at the same resolution
Compensated	System actively re-supplies C 's role	Acute effect that decays with T ; residual in other variables	Flat effect across T
System changed	Measured object is $\mathcal{R}(S, a)$, not $S - C$	Comparison of $\mathcal{R}_0$ and $\mathcal{R}_T$; full state readout	Identical full state before and after

Table 1: Four readings of a null ablation and the evidence that separates them.

The practical upshot is that a single null observation cannot discriminate among the four. All four predict the same scalar readout at one horizon. The readings separate only when one adds joint ablations, multiple horizons, or richer observables. An ablation study reporting a single null on a single scalar has therefore not yet said which of the four it has found.

4 Sources of Silence

A null ablation, $\Delta O \approx 0$, can arise in at least six ways. The taxonomy below is not exhaustive but each item has a distinct mechanism and a distinct signature under further experiment.

4.1 Compensation

The system contains a feedback or adaptation process that restores the observable after perturbation. Homeostatic regulation is the archetype: a variable is held near a setpoint by adjusting other variables, so that removing one source of the regulated quantity is met by increased output from another. Section 7 gives an exact linear model. In biology, compensatory responses to gene deletion are well documented, including cases in which a knockout mutant shows a milder phenotype than an acute knockdown of the same gene, because the mutation triggers upregulation of related genes [11, 4]. In artificial networks, ablating an attention layer in a language model can leave the output almost unchanged because later layers shift their contribution, an effect described as self-repair [8].

4.2 Saturation

The observable depends on C through a nonlinearity that flattens at the operating point. Let $O = \sigma(c+r)$ with $\sigma = \tanh$, ablated component value $c = 1$ against a background $r = 5$. Then $O(6) = \tanh 6 \approx 0.999988$ and $O(5) = \tanh 5 \approx 0.999909$, so $\Delta O \approx 8 \times 10^{-5}$, far below any

realistic tolerance. Yet at $r = 0$ the same ablation moves the output from $\tanh 1 \approx 0.762$ to 0. The component is neither irrelevant nor redundant. It is masked by the regime. Saturating nonlinearities in neural networks, ceiling effects in behavioral measures, and rate-limiting steps in metabolic pathways all produce this pattern.

4.3 Degeneracy

The observable factors through a many-to-one map. Write $O = h \circ g$ where $g : \Theta \rightarrow G$ has fibres $g^{-1}(y)$ containing many configurations. Edelman and Gally use the term degeneracy for the capacity of structurally different elements to yield the same function [3]. If some configuration in the baseline’s fibre has $\theta_C = 0_C$, then that configuration realizes the same observable without C . Degeneracy is a static property of the map, whereas compensation is a dynamic property of the flow. A degenerate system need not adapt at all, provided the ablation happens to land in the fibre. The distinction is examined formally in Section 7.

4.4 Measurement coarseness

The observable is a projection $\pi : Y \rightarrow Y'$ of a richer quantity, and the ablation moves the richer quantity without moving the projection. A pass/fail metric, a benchmark accuracy with finite item resolution, a mean over a population in which shifts cancel, and a coarse behavioral score are all projections. If accuracy is computed on 500 items, no change smaller than 0.2 percentage points is resolvable in a single evaluation, whatever happens to the underlying log-likelihoods. A statistical version of the same point is that failure to reject a null hypothesis is not evidence for it, a principle summarized in the slogan that absence of evidence is not evidence of absence [1], and addressed operationally by equivalence testing [6].

4.5 Load redistribution

The system carries a conserved or budgeted quantity, and removing one carrier shifts the burden to the others. A deck supported by n identical parallel members, each of stiffness k , under total load W , carries W/n per member. Removing one leaves $W/(n-1)$ per survivor, an increase by the factor $n/(n-1)$. If the coarse observable is “the deck stands” then the ablation is null provided the survivors have capacity above $W/(n-1)$. The fine observable, strain in the survivors, is not null. It rises by exactly the factor above, and that factor identifies n . This example returns in Section 23. Nothing about it is exotic: the ordinary vocabulary of statically indeterminate structures already treats redundancy as load that can be rerouted.

4.6 Level-set coincidence

The finite intervention lands on another point of the same level set of the observable, as in Example 1.1. Periodic and symmetric systems produce this exactly, and it can occur approximately whenever the response is non-monotone. It is the purely geometric source of silence, requiring neither adaptation nor redundancy.

Source	Mechanism	Distinguishing test
Compensation	Adaptation restores O	Acute effect present, chronic effect absent; residual in effort variables
Saturation	Flat response at the operating point	Vary background; effect returns off saturation
Degeneracy	$O = h \circ g$, fibre meets $\theta_C = 0$	Joint ablation with partner is non-null
Coarseness	Projection hides change	Finer observable, full state readout
Load redistribution	Conserved budget rerouted	Survivor variables shift by predictable factor
Level-set coincidence	Finite step lands on same O	Ablate to several nonzero values

Table 2: Sources of silence and the further experiment that separates each.

4.7 Tolerances do not compose

A further property of the null result deserves its own statement. If each of n single ablations is ε -null, one would like to conclude that ablating all n is at most $n\varepsilon$ -null. This holds for additive systems and fails otherwise.

Proposition 4.1 (Nulls need not compose). *(i) If $O(\theta) = o_0 + \sum_c o_c(\theta_c)$ is additive, and each single ablation satisfies $|O(\theta^0) - O(0_c, \theta_{-c}^0)| \leq \varepsilon$, then ablating any set D changes O by at most $|D|\varepsilon$. (ii) For every $n \geq 2$ there is a system with n components in which every single ablation is exactly null and the joint ablation of all components has effect 1.*

Proof. (i) Ablating D changes O by $\sum_{c \in D} (o_c(0_c) - o_c(\theta_c^0))$, whose absolute value is at most $\sum_{c \in D} \varepsilon = |D|\varepsilon$. (ii) Let each $\theta_c \in \{0, 1\}$, and let $O(\theta) = \max_c \theta_c$ (an OR gate), with baseline $\theta^0 = (1, \dots, 1)$. Removing one component leaves $n - 1 \geq 1$ components at 1, so $O = 1$ and the single effect is 0. Removing all components gives $O = 0$, an effect of 1. $\square$

Part (ii) is the OR gate, the simplest redundancy. It shows that the bound in (i) is a feature of additivity, not of ablation, and that nothing controls the joint effect of a family of null singles unless the system's interaction structure is known.

5 The Geometry of Null Sets

Section 4 listed six sources of silence as if they were separate mechanisms. A single geometric statement contains all of them. It concerns the set of configurations that the observable cannot tell apart from the baseline.

Definition 5.1 (Null sets). The *exact null set* (fibre) of $\theta \in \Theta$ is $N(\theta) = O^{-1}(O(\theta))$, and $\theta \sim_O \theta'$ means $O(\theta) = O(\theta')$. The *tolerance set* is $N_\varepsilon(\theta) = \{\theta' \in \Theta : d(O(\theta'), O(\theta)) \leq \varepsilon\}$.

Remark 5.2 (Tolerance is not an equivalence). The relation $\theta' \in N_\varepsilon(\theta)$ is reflexive and symmetric but not transitive, so $N_\varepsilon(\theta)$ is a neighbourhood and not an equivalence class. Only

the exact relation $\sim_O$ partitions Θ . For $O(\theta) = \theta \in \mathbb{R}$ and $\varepsilon = 1$, one has $0 \in N_1(1)$ and $1 \in N_1(2)$, yet $0 \notin N_1(2)$. Chains of ε -null steps can therefore accumulate an arbitrarily large effect, which is the geometric form of Proposition 4.1.

Proposition 5.3 (Null ablation as membership). *The ablation $\text{do}(C = 0)$ at horizon T is ε -null if and only if $\mathcal{R}_T(S, \text{do}(C = 0)) \in N_\varepsilon(\theta^0)$, and exactly null if and only if $\mathcal{R}_T(S, \text{do}(C = 0)) \in N(\theta^0)$.*

Proof. This is Definition of Δ^T restated: $\Delta^T \leq \varepsilon$ means $d(O(\theta^0), O(\mathcal{R}_T)) \leq \varepsilon$. $\square$

The proposition is trivial, and its value is that it recasts each source of silence as a reason why the intervention ends inside the same observational neighbourhood.

Source	Geometric statement
Compensation	The adaptation flow steers $\mathcal{R}_0$ into $N_\varepsilon(\theta^0)$
Degeneracy	$N_\varepsilon(\theta^0)$ meets the ablation slice $\{\theta_C = 0_C\}$ (the fibre is nonempty there)
Saturation	θ^0 and $\mathcal{R}_0$ lie in a region where $DO \approx 0$, so $N_\varepsilon(\theta^0)$ contains a neighbourhood
Coarseness	A coarser metric d or a many-to-one projection enlarges N_ε
Load redistribution	The flow is confined to a conserved slice on which the coarse observable is constant
Level-set coincidence	$\mathcal{R}_0$ lands on $N(\theta^0)$ without any adaptation

Table 3: The six sources of silence as statements about $N_\varepsilon(\theta^0)$.

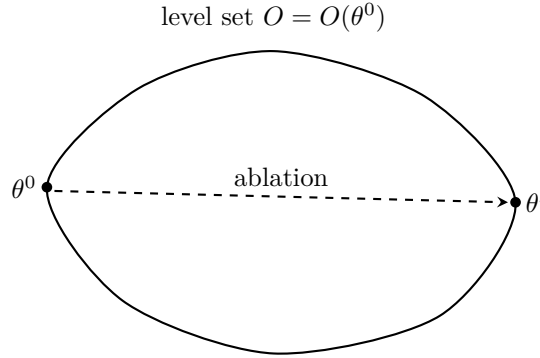


Figure 2: A null ablation as a move between two points of one level set. The configurations differ, $\theta^0 \neq \theta'$, while $O(\theta^0) = O(\theta')$. The path between them need not stay on the level set; only its endpoint is constrained.

5.1 Two geometries of silence

The table treats every source alike, but two different geometric situations underlie it, and they differ in how likely an exact null is.

Suppose Θ is a smooth manifold of dimension D and $O : \Theta \rightarrow \mathbb{R}^m$ is smooth.

Proposition 5.4 (Dimension of exact null sets). *If $DO(\theta^0)$ has rank m (the point is regular), then near θ^0 the null set $N(\theta^0)$ is an embedded submanifold of dimension $D - m$.*

Proof. This is the regular value theorem, a consequence of the implicit function theorem: a submersion has locally a coordinate form $(u_1, \dots, u_D) \mapsto (u_1, \dots, u_m)$, whose level sets are affine subspaces of dimension $D - m$. $\square$

A low-dimensional observable on a high-dimensional configuration space therefore has a high-dimensional null set: room exists to move a very long way without changing O . In the homeostat, $D = 2$, $m = 1$, and the null set is a line (Figure 3 below).

A caution is needed. The null set has codimension m , hence Lebesgue measure zero in Θ . A passive finite ablation of a non-adaptive system does not land on it except by coincidence. Exact nulls are typical only in two situations.

- (a) *Flat regions.* On an open set where $\text{rank } DO < m$ (saturation, thresholds, dead units), $N(\theta)$ has nonempty interior and a passive step stays inside it with positive probability.
- (b) *Steering.* The adaptation flow drives the trajectory onto the null set. Regulated systems show exact nulls robustly, and not by accident, because the dynamics are designed or selected to return to the fibre.

Dimension supplies the room for compensation. It does not supply the landing. This distinction separates the passive sources of silence (saturation, coarseness, level-set coincidence) from the active one (compensation).

5.2 Compensating configurations form a family

Proposition 5.5 (Degeneracy dimension). *Let $O_0(q) = O(0_C, q)$ for $q \in \Theta_{-C}$, with $\dim \Theta_{-C} = D_{-C} \geq m$. Suppose O_0 is a submersion at $q_0 = \theta_{-C}^0$. Then there is an open set $U \ni O(\mathcal{R}_0)$ in $\mathbb{R}^m$ such that every $y \in U$ is attained: the set $\{q : O_0(q) = y\}$ is nonempty, and near each of its points it is a submanifold of dimension $D_{-C} - m$. In particular, if $O(\theta^0) \in U$, that is, if the static effect Δ^0 is small enough, then compensating configurations exist and form a $(D_{-C} - m)$ -parameter family.*

Proof. A submersion is locally open, so there is a neighbourhood W of q_0 with $U := O_0(W)$ open, and $O(\mathcal{R}_0) = O_0(q_0) \in U$. For $y \in U$ the level set inside W is nonempty and, by the regular value theorem, a submanifold of dimension $D_{-C} - m$. $\square$

The proposition splits the hidden positive content of a null (Corollary 7.5) into two parts. That a compensating configuration *exists* is generic when there are more free coordinates than

observables and the static effect is small. Which member of the $(D_{-C} - m)$ -parameter family the system actually reaches is determined by the adaptation field V . The null therefore certifies reachability and not merely existence.

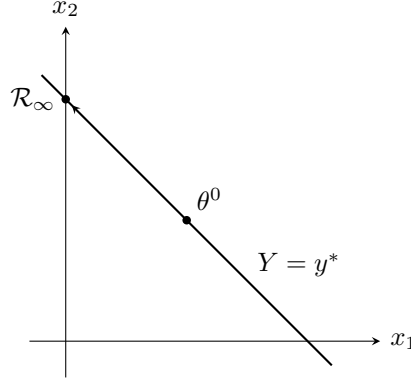


Figure 3: In the homeostat the null set of $Y = x_1 + x_2$ is the line $x_1 + x_2 = y^*$ (dimension $D - m = 1$). The chronic ablation moves along it, from θ^0 to $\mathcal{R}_\infty$.

6 Ablation as Quotient Observation

The relation $\sim_O$ is an equivalence relation, so it has a quotient. Let $\pi : \Theta \rightarrow \Theta/\sim_O$ send a configuration to its class. By the first isomorphism theorem for sets, the quotient is in bijection with the image $O(\Theta)$, and π corresponds to O itself. An experiment that records only O does not observe Θ . It observes $\Theta/\sim_O$.

Proposition 6.1 (Nothing happened in the quotient). *The ablation trajectory in Θ is the jump $\theta^0 \mapsto \mathcal{R}_0$ followed by the flow $t \mapsto \mathcal{R}_t$. Its image in observable space is the path $O(\theta^0) \rightarrow O(\mathcal{R}_0) \rightarrow O(\mathcal{R}_t)$. The ablation is exactly null at horizon T if and only if this path returns to its starting point, $O(\mathcal{R}_T) = O(\theta^0)$. The excursion amplitude*

$$\Xi_T = \sup_{0 \leq t \leq T} d(O(\theta^0), O(\mathcal{R}_t))$$

satisfies $\Xi_T \geq \Delta^0$ and $\Xi_T \geq \Delta^t$ for every $t \leq T$, so a null endpoint is compatible with an arbitrarily large excursion.

Proof. The first claim is the definition of exact nullness. The inequalities hold because $t = 0$ is included in the supremum and each Δ^t is one of the distances whose supremum is taken. Arbitrary size follows from Example 7.6: $\Xi_\infty \geq a$ while $\Delta^\infty = 0$, with a free. $\square$

In this language, an ablation with nonzero acute effect and zero chronic effect is a closed excursion in the quotient. Its size is the transient the compensation had to overcome, and a study that samples only the endpoint has observed only the point at which the loop closes.

The quotient formulation also says what “nothing happened” can mean. It does not mean that the configuration is unchanged. It means that the configuration moved within a fibre of O , or left it and returned. A precise statement of the central thesis follows.

Invariance is not absence of change. It is change constrained to remain invisible under a specified observation map.

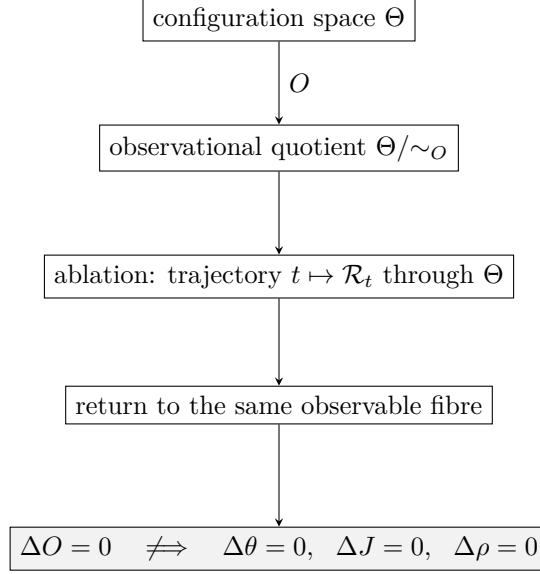


Figure 4: The observational spine. Here J is the compensation cost of Section 9 and ρ the robustness margin of Section 12.

The two extra quantities in the last box are developed in Sections 9 and 12. Output invariance can coexist with a change of configuration, with a cost incurred to maintain the output, and with a loss of margin against failure. Each is a distinction erased by O and recoverable only by a different observer, a point taken up formally in Section 22.

7 Component Necessity and Path Necessity

7.1 Structure functions, paths, and cuts

To separate what a component does from what a single ablation reveals, it helps to borrow the language of coherent systems from reliability theory [2]. For $z \in \{0, 1\}^K$, let $O(z)$ denote the observable when the components with $z_c = 1$ are retained and the others ablated, and write $\mathbf{1}$ for the all-retained vector. Identify z with the retained set $\{c : z_c = 1\}$. Assume O is real-valued and *monotone*: retaining more components never lowers O .

Definition 7.1 (Paths and cuts). A set $P \subseteq K$ is an ε -*path* if retaining exactly P leaves O within ε of the full system: $|O(\mathbf{1}) - O(\mathbf{1}_P)| \leq \varepsilon$. A set $D \subseteq K$ is an ε -*cut* if ablating D moves O by more than ε : $|O(\mathbf{1}) - O(\mathbf{1}_{K \setminus D})| > \varepsilon$. A minimal path (cut) is one with no proper subset that is a path (cut).

For monotone O the two notions are complementary: D is a cut exactly when $K \setminus D$ is not a path. A component c is *necessary* if $\{c\}$ is a cut. It is *used* if it belongs to at least one minimal path. Necessity and use come apart.

Proposition 7.2 (Necessity is membership in the core). *For monotone O , component c is necessary if and only if c lies in every ε -path. Equivalently, the set of necessary components is the intersection $\bigcap \{P : P \text{ an } \varepsilon\text{-path}\}$, which is called the core.*

Proof. Suppose c is necessary, so $K \setminus \{c\}$ is not a path. If some path P avoided c , then $P \subseteq K \setminus \{c\}$ and by monotonicity $O(\mathbf{1}_P) \leq O(\mathbf{1}_{K \setminus c}) \leq O(\mathbf{1})$, so $|O(\mathbf{1}) - O(\mathbf{1}_{K \setminus c})| \leq |O(\mathbf{1}) - O(\mathbf{1}_P)| \leq \varepsilon$, contradicting that $K \setminus c$ is not a path. Hence every path contains c . Conversely, if every path contains c then $K \setminus \{c\}$ is not a path, so $\{c\}$ is a cut. $\square$

The consequence is that a single-ablation study measures exactly membership in the core, and nothing else about the path structure. The core may be empty in a system in which every component is used.

Example 7.3 (Threshold systems). Let $O(z) = \mathbf{1}[\sum_c z_c \geq r]$ with $1 \leq r < n$. Every set of size at least r is a path, so every r -subset is a minimal path and every component is used. The core is empty for $r < n$, so every single ablation is null. Component necessity is zero across the board while the system is entirely built from these components. Minimal cuts have size $n - r + 1$, and that number, not any single-ablation statistic, is the redundancy of the system.

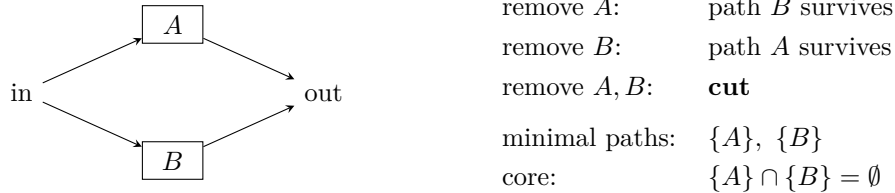


Figure 5: Two parallel routes. Both components are used, neither is necessary, and the core is empty. Every single ablation is null; the joint ablation is not.

7.2 Degeneracy as invariance

The static source of silence identified in Section 4 can be stated exactly.

Proposition 7.4 (Fibre invariance). *Suppose $O = h \circ g$ with $g : \Theta \rightarrow G$. Let θ^0 be the baseline and suppose the reconfiguration operator selects a configuration in the fibre:*

$$\mathcal{R}_T(S, \text{do}(C = 0)) \in g^{-1}(g(\theta^0)) \cap \{\theta_C = 0_C\}.$$

Then $\Delta^T(C) = 0$ exactly, for any h .

Proof. $O(\mathcal{R}_T) = h(g(\mathcal{R}_T)) = h(g(\theta^0)) = O(\theta^0)$. $\square$

The hypothesis has two parts, and they correspond to two different claims. That the fibre $g^{-1}(g(\theta^0)) \cap \{\theta_C = 0_C\}$ is nonempty is a claim about degeneracy: there *exists* a C -free configuration with the same observable. That the operator *selects* such a configuration is a claim about dynamics: the system can *reach* it. A null ablation certifies both at once, which is the hidden positive content of the null.

Corollary 7.5 (The null contains a positive result). *If $\Delta^T(C) = 0$ then there exists a configuration θ' with $\theta'_C = 0_C$, $O(\theta') = O(\theta^0)$, and θ' reachable from θ^0 under the adaptation flow with C pinned. In the notation of the compensated form*

$$f(0, P'_2, \dots, P'_n) \approx f(P_1, P_2, \dots, P_n), \quad P_j \longrightarrow P'_j \quad (j \geq 2),$$

the map $P_{-1} \mapsto P'_{-1}$ is a compensatory transformation realized by the system itself.

In words: a null ablation of component 1 shows that the remaining components possess a transformation which restores the function. The existence of that transformation is a fact about the system that a non-null ablation would not have revealed.

7.3 An exact model of compensation

The following linear model isolates the distinction between the static sensitivity of a variable and its sensitivity at equilibrium.

Example 7.6 (Homeostat with two sources). Let $Y = x_1 + x_2$ be a regulated quantity with setpoint y^* . The source x_1 is the ablation target. The second source x_2 is adjusted by a controller with gain $k > 0$:

$$\dot{x}_2 = -k(x_1 + x_2 - y^*).$$

For fixed x_1 the unique rest point is $x_2^* = y^* - x_1$, and it is globally asymptotically stable because the equation is a stable linear relaxation. At baseline $x_1 = a$ and $x_2^* = y^* - a$, so $Y = y^*$.

Static sensitivity. With x_2 held fixed, $\partial Y / \partial x_1 = 1$. Ablating x_1 from a to 0 gives the static effect $\Delta^0 = a$.

Chronic sensitivity. Letting x_2 relax, $Y^* = y^*$ for every value of x_1 , so $\partial Y^* / \partial x_1 = 0$ and $\Delta^\infty = 0$.

Residual. The shift in the effort variable is $x_2^*(0) - x_2^*(a) = a$. It equals exactly the amount removed.

Three observations follow. First, the observable Y is null after ablation while its static sensitivity is 1, the maximal value for this model, so the implication (1) fails in the strongest possible way. Second, the compensation gap is $\kappa_\infty = a$, so by Lemma 2.6 the null result is a measurement of the amount compensated. Third, the information the experiment discarded lies in x_2 , not in Y . The general principle is worth naming.

Proposition 7.7 (Regulated and effort variables). *Suppose a regulated variable $g(\theta)$ is held at y^* by adaptation of free coordinates, in the sense that for every admissible pinned value of θ_C the free dynamics have an asymptotically stable set contained in $\{g = y^*\}$ that attracts the trajectory from θ_{-C}^0 . Then the chronic effect of any ablation of C on g is zero, and any information about C 's contribution is carried by the free (effort) coordinates.*

Proof. By hypothesis $\lim_{T \rightarrow \infty} g(\mathcal{R}_T) = y^* = g(\theta^0)$, so $\Delta^\infty = 0$ for the observable g . Since the pinned coordinate values differ between baseline and ablated system while g is equal, the difference in C 's role must be absorbed by θ_{-C} : if θ_{-C} also coincided, the two configurations would be identical, contradicting $\theta_C \neq 0_C$ at baseline. $\square$

The proposition explains a pattern that recurs across fields. Homeostatic quantities such as blood glucose, body temperature, and the output of a well-tuned network appear robust to removal of individual contributors, and the contributors’ identities are found instead in the effort variables: hormone levels, metabolic flux, weights redistributed across layers. Studies that record only the regulated quantity see a null. Studies that record the effort variables see the compensation.

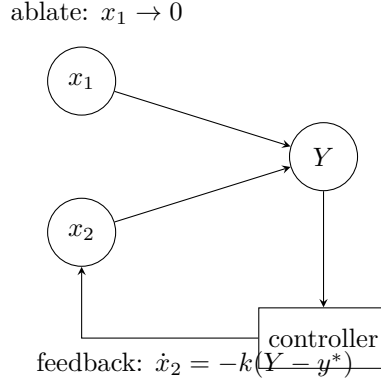


Figure 6: Ablating x_1 leaves the regulated quantity Y unchanged, because the controller raises x_2 by exactly the amount removed.

7.4 The subtraction fallacy

The pieces assemble into the main structural claim.

Theorem 7.8 (Ablation is not subtraction). *In general*

$$S - C \neq \mathcal{R}(S, \text{do}(C = 0)).$$

More precisely, for any adaptive system with nonzero adaptation on the free coordinates, $\mathcal{R}_T(S, \text{do}(C = 0)) \neq \mathcal{R}_0(S, \text{do}(C = 0))$ for all $T > 0$ at which the free trajectory has moved. The compensation gap $\kappa_T(C)$ may be arbitrarily large while $\Delta^T(C) = 0$.

Proof. The static ablation $\mathcal{R}_0$ is the subtraction $S - C$ by definition. The configuration $\mathcal{R}_T$ differs from it precisely by the displacement $\psi_T^a(\theta_{-C}^0) - \theta_{-C}^0$ of the free coordinates, which is nonzero whenever the trajectory has left its initial point. For the second claim, Example 7.6 gives $\Delta^\infty = 0$ and $\kappa_\infty = a$ with a arbitrary. $\square$

The theorem gives a precise meaning to the claim that an ablation experiment is ordinarily narrated as subtraction but is actually a perturbation followed by reconstruction. In what follows $\mathcal{R}(S, \text{do}(C = 0))$ is the experimental object, and $S - C$ is a counterfactual that may not be realized by any experiment on an adaptive system.

The two cases of the theorem, static and dynamic, admit a two-by-two classification by whether the static and reconfigured effects are null.

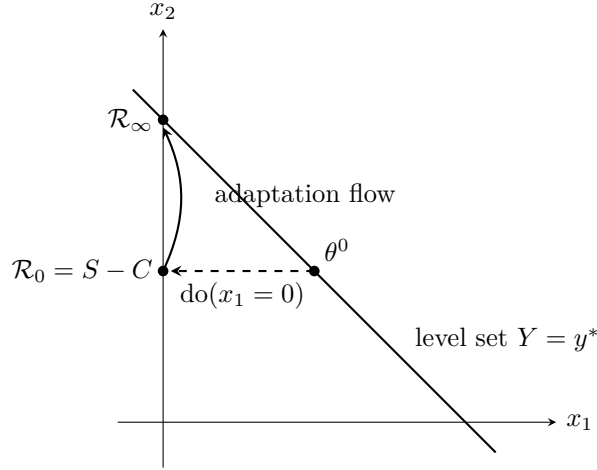


Figure 7: The homeostat of Example 7.6 in configuration space, with $a = 2$ and $y^* = 4$. Subtraction moves the configuration off the regulated level set to $S - C$. The adaptation flow carries it back onto the level set at a different point, a displacement of a in the effort coordinate x_2 . The observable is unchanged; the configuration is not.

	$\Delta^T \leq \varepsilon$	$\Delta^T > \varepsilon$
$\Delta^0 \leq \varepsilon$	Inert at this operating point (or masked by saturation, coarseness, level-set coincidence)	Destabilization: the reconfiguration itself is harmful
$\Delta^0 > \varepsilon$	Compensated: $\kappa_T \geq \Delta^0 - \varepsilon$	Necessary on the tested horizon

Table 4: Classification by static and reconfigured effect. Only the lower left cell licenses the statement that the system compensated.

Most published ablation studies populate only one column and one horizon, and therefore cannot place their component in a cell. That a null study reports a single number is a limitation on what it can say, not on what the system did.

8 Timescales: The Horizon Spectrum

Definition 2.4 indexes the reconfiguration operator by a horizon T , yet the argument so far has mostly compared the two endpoints $T = 0$ and $T = \infty$. Real experiments sit at intermediate horizons that are seldom reported, and the choice of horizon amounts to a choice among different experimental questions. Let τ denote the characteristic time of the system’s adaptation. Three regimes can be distinguished.

The relevant τ can span several orders of magnitude within a single system. In an organism the adaptation may be physiological (seconds to hours), transcriptional (hours to days), or developmental and evolutionary (generations). In a trained network it may be a single forward

Regime	Horizon	Typical experiment	Estimates
Acute	$T \ll \tau$	Clamp for one forward pass; acute inhibitor; activation patch at inference	Δ^0
Transient	$T \sim \tau$	Short retraining after pruning; hours after a conditional knockout	The trajectory
Chronic	$T \gg \tau$	Germline knockout; retraining to convergence; long-term removal of an office or member	Δ^∞

Table 5: Adaptation regimes and the experiments that occupy them.

pass, a few gradient steps, or a full retraining. The experimental sentence “we removed C ” is silent on which of these applies.

8.1 The horizon curve of the homeostat

The exact model of Example 7.6 can be solved at every horizon.

Proposition 8.1 (Horizon curve). *In Example 7.6, with baseline contribution $x_1 = a > 0$ and controller gain $k > 0$, the reconfigured effect and compensation gap at horizon T are*

$$\Delta^T = a e^{-kT}, \quad \kappa_T = a(1 - e^{-kT}).$$

Proof. After ablation $x_1 = 0$ and $\dot{x}_2 = -k(x_2 - y^*)$ with $x_2(0) = y^* - a$. The solution is $x_2(t) = y^* - a e^{-kt}$, and $Y = x_2$, so $Y(T) - y^* = -a e^{-kT}$ and $\Delta^T = a e^{-kT}$. The static ablation has $Y = y^* - a$, so $\kappa_T = |Y(T) - (y^* - a)| = a(1 - e^{-kT})$. $\square$

Two consequences follow, and both convert an apparently uninformative null into a measurement.

Corollary 8.2 (A null bounds the adaptation rate). *If the static effect a is known and an experiment at horizon T finds $\Delta^T \leq \varepsilon$ with $\varepsilon < a$, then*

$$k \geq \frac{1}{T} \ln \frac{a}{\varepsilon}.$$

Proof. $a e^{-kT} \leq \varepsilon$ is equivalent to $e^{-kT} \leq \varepsilon/a$, i.e., $kT \geq \ln(a/\varepsilon)$. $\square$

Corollary 8.3 (Disagreement without error). *Let $\tau_\varepsilon = \frac{1}{k} \ln \frac{a}{\varepsilon}$ be the crossover horizon. Two laboratories ablating the same component in the same system, one at $T_A < \tau_\varepsilon$ and one at $T_B \geq \tau_\varepsilon$, obtain $\Delta^{T_A} > \varepsilon$ and $\Delta^{T_B} \leq \varepsilon$. Both are correct.*

Proof. $\Delta^T = a e^{-kT}$ is strictly decreasing in T and equals ε at $T = \tau_\varepsilon$. $\square$

The second corollary describes a familiar situation. A component is reported essential in one study and dispensable in another, and the difference is attributed to error, strain, or context. When the two studies occupy different adaptation regimes, no error is required. The disagreement is a measurement of τ_ε .

8.2 Beyond a single time constant

Remark 8.4 (Several modes). For linear adaptation $\dot{x} = -M(x - x^*)$ with M having eigenvalues of positive real part $\lambda_1, \dots, \lambda_m$, the effect takes the form $\Delta^T = |\sum_j c_j e^{-\lambda_j T}|$ for coefficients c_j determined by the ablation. A single-horizon measurement collapses this spectrum into one number. Separating modes requires recording at several horizons, and recovering rates from a sum of exponentials is numerically ill-conditioned, so the horizons should be spread over more than one decade when the modes are not known in advance.

Remark 8.5 (Zero crossings in time). When the regulator is second order and underdamped, the deviation of the regulated variable after ablation is not monotone. With natural frequency ω , damping ratio $\zeta < 1$, and damped frequency $\omega_d = \omega\sqrt{1 - \zeta^2}$, an initial deviation a with zero initial velocity evolves as

$$e(t) = a e^{-\zeta\omega t} \left(\cos \omega_d t + \frac{\zeta\omega}{\omega_d} \sin \omega_d t \right).$$

This vanishes at the infinitely many times satisfying $\tan(\omega_d t) = -\omega_d/(\zeta\omega)$. An experiment sampling at such a time records $\Delta^T = 0$ although the system is mid-transient and will deviate again. This is the temporal counterpart of the level-set coincidence of Example 1.1, and it is guarded against in the same way, by sampling at several horizons.

9 The Cost of Invariance

Output invariance and cost invariance are different properties. A system can keep O fixed by spending something: energy, bandwidth, latency, capacity, flexibility, or attention. The effort variables of Proposition 7.7 are where the spending appears, and a cost functional turns it into a number.

Definition 9.1 (Compensation cost). Let $c : \Theta \rightarrow [0, \infty)$ be a cost rate, with $c(\theta^0) = 0$, measuring the excess expenditure relative to baseline. Along the ablated trajectory $\theta(t) = \mathcal{R}_t(S, a)$ the accumulated compensation cost is

$$J_C(T) = \int_0^T c(\theta(t)) dt.$$

The choice of c is a modelling decision fixed by what is scarce in the system at hand.

Proposition 9.2 (Invariance without cost invariance). *In Example 7.6 take $c(x) = (x_2 - x_2^0)^2$. Then $\Delta^\infty = 0$, yet*

$$J_C(T) = a^2 \left[T - \frac{2}{k}(1 - e^{-kT}) + \frac{1}{2k}(1 - e^{-2kT}) \right],$$

so the cost rate tends to $a^2 > 0$ and $J_C(T) \sim a^2 T$ grows without bound.

Proof. By Proposition 8.1, $x_2(t) - x_2^0 = a(1 - e^{-kt})$, so $c = a^2(1 - e^{-kt})^2 = a^2(1 - 2e^{-kt} + e^{-2kt})$. Integrating from 0 to T gives the stated expression. As $t \rightarrow \infty$ the integrand tends to a^2 . $\square$

The output returned to its baseline, and the system pays a permanent cost to keep it there. Reports that state only $\Delta^T = 0$ conceal a running expenditure. The correct summary of the experiment is the pair $(\Delta^T(C), J_C(T))$. The two need not be combined into a scalar. Where a scalar is wanted, two normalizations are defined when $\Delta^0 > 0$: the fraction restored, $1 - \Delta^T/\Delta^0$, and the cost per unit restored, $J_C(T)/(\Delta^0 - \Delta^T)$. Neither is defined when $\Delta^0 = 0$, which is one reason to report the pair.

9.1 Compensation forces cost

The cost is not arbitrary. Under mild regularity, any compensation at all incurs a cost that grows with its size.

Proposition 9.3 (Compensation forces cost). *Suppose O is L -Lipschitz in the free coordinates, $d(O(0_C, q), O(0_C, q')) \leq L\|q - q'\|$, and the cost is coercive, $c(\theta) \geq \lambda\|\theta_{-C} - \theta_{-C}^0\|^2$ for some $\lambda > 0$. Then at every horizon,*

$$c(\mathcal{R}_T(S, a)) \geq \lambda \frac{\kappa_T(C)^2}{L^2}.$$

Proof. $\mathcal{R}_0$ and $\mathcal{R}_T$ share $\theta_C = 0_C$ and differ in the free coordinates, which are θ_{-C}^0 and $\psi_T^a(\theta_{-C}^0)$. So $\kappa_T = d(O(\mathcal{R}_0), O(\mathcal{R}_T)) \leq L\|\psi_T^a(\theta_{-C}^0) - \theta_{-C}^0\|$. Hence the displacement is at least κ_T/L , and coercivity gives the bound. $\square$

The result links the two hidden positive quantities of the essay. The compensation gap κ_T , which measures how much the system did, lower-bounds the instantaneous cost of doing it. A null with a large gap is an expensive null.

Remark 9.4 (Cost is not rebound). The rebound of Section 15 reads the compensation gap in the observable at the instant of restoration, and the cost reads the effort spent in the free coordinates during the ablation. They are different measurements of the same reconfiguration, and they can be compared: in the homeostat the rebound equals κ_T while the cost rate at horizon T equals κ_T^2 .

10 Joint Ablations and the Interaction Signature

10.1 The interaction term

Single ablations are blind to interaction. Let $v(T) = O(\mathbf{1}_T)$ be the value with only the set T retained, and write $\Delta_D = v(K) - v(K \setminus D)$ for the effect of ablating D . For two components define the interaction

$$\varepsilon_{CD} = \Delta_{\{C,D\}} - \Delta_C - \Delta_D. \quad (2)$$

When $\varepsilon_{CD} = 0$ the two effects add. When it is nonzero, the effect of removing both differs from the sum of the effects of removing each.

Proposition 10.1 (Diagnostic signature of compensation). *If $\Delta_C = \Delta_D = 0$ and $\Delta_{\{C,D\}} \neq 0$ then $\varepsilon_{CD} = \Delta_{\{C,D\}} \neq 0$, so the system is non-additive on $\{C, D\}$. Conversely, if the system is additive on a set $D_1, \dots, D_m$ and all singles are null, every joint ablation among them is null.*

Proof. Substituting the hypotheses into (2) gives the first claim. The converse is Proposition 4.1(i) with $\varepsilon = 0$. $\square$

In genetics this pattern is called synthetic lethality: two single deletions are each viable, and the double deletion is not. The concept has a history extending from early observations of Bridges and Dobzhansky, and it has been used systematically as a tool for finding pairs of genes that compensate for one another [5]. In the terms of this essay, synthetic lethality is the interaction signature of the second and third readings. A null on C alone together with a large effect on $\{C, D\}$ is positive evidence of a compensatory path through D .

10.2 Why single ablations misattribute importance

The failure extends beyond the binary decision that a component is unimportant. Any importance score computed from single ablations alone misattributes credit in the presence of redundancy. The Shapley value [12], which averages a component’s marginal contribution over all orders of retention, gives a contrasting view.

Definition 10.2. For a value function v on subsets of K the Shapley value of c is

$$\phi_c = \sum_{T \subseteq K \setminus \{c\}} \frac{|T|! (n - |T| - 1)!}{n!} [v(T \cup \{c\}) - v(T)].$$

Proposition 10.3 (Symmetric threshold systems). *For the r -out-of- n threshold system with $1 \leq r < n$ and $v(T) = \mathbf{1}[|T| \geq r]$, every single ablation is null, yet $\phi_c = 1/n$ for every c .*

Proof. The single-ablation effect is $v(K) - v(K \setminus c) = 1 - \mathbf{1}[n - 1 \geq r] = 0$. All components are exchangeable, so ϕ_c is the same for all c . By efficiency, $\sum_c \phi_c = v(K) - v(\emptyset) = 1 - 0 = 1$, hence $\phi_c = 1/n$. $\square$

For $n = 2$, $r = 1$ this is the two-path system of the earlier examples: single ablations are zero and each component receives credit $1/2$. The Shapley value is not offered as the correct notion of importance. The point is that a measure sensitive to the full lattice of retained sets sees what the single-ablation score misses, and that the two disagree exactly when redundancy is present.

10.3 Order of the required ablation

If the minimal cuts of a system have size s , then no set of ablations of size less than s can detect any effect, and the system will present as fully robust to any study restricted to smaller sets. Redundancy of order s is invisible to all ablations of order $s - 1$ or fewer. This is a sharp statement, and it turns the vague worry that a study “did not test enough” into a quantitative one.

Proposition 10.4 (Invisibility below the cut size). *In the r -out-of- n threshold system, every ablation of at most $n - r$ components is null, and ablations of $n - r + 1$ or more components are non-null.*

Proof. Ablating m components retains $n - m$, and the system functions iff $n - m \geq r$, i.e., $m \leq n - r$. $\square$

11 Ablation Profiles, Margins, and the Sampling Failure of Random Ablation

Section 10 compared single and paired ablations. This section considers the whole family of ablations of every size, which yields the operational meaning of redundancy.

11.1 Profiles

Definition 11.1 (Ablation profiles). For a monotone system, let D be a uniformly random m -subset of K . The *cut profile* is $A(m) = \Pr[D \text{ is an } \varepsilon\text{-cut}]$, and the *effect profile* is $E(m) = \mathbb{E}[O(\mathbf{1}) - O(\mathbf{1}_{K \setminus D})]$. Let s be the size of a smallest ε -cut (the *cut number*).

Proposition 11.2 (Monotone profile and margin). *For a monotone system, $A(m) \leq A(m+1)$ for all m , and $A(m) = 0$ if and only if $m < s$. The tolerance margin $m^* = s - 1$ is therefore the largest number of components that can be removed in any way without exceeding ε .*

Proof. Draw a uniformly random $(m+1)$ -subset D' and delete one of its elements uniformly at random, producing a uniformly random m -subset $D \subseteq D'$. If D is a cut then so is D' , because ablating more components cannot bring O closer to its baseline in a monotone system. Hence $A(m+1) \geq A(m)$. If $m \geq s$, a fixed minimal cut of size s is contained in some m -subset, so $A(m) > 0$; if $m < s$ no m -subset is a cut, so $A(m) = 0$. $\square$

The margin is the quantity that a single-ablation study cannot see. A study finding every single ablation null has established $m^* \geq 1$, which is compatible with $m^* = 1$ (a system one removal from failure) and with $m^* = n - 1$ (a system with enormous reserve).

11.2 Shape distinguishes additivity from redundancy

Proposition 11.3 (Flat-then-rise is non-additive). *If O is additive, $O(\theta) = o_0 + \sum_c o_c(\theta_c)$, and $E(1) = 0$ exactly, then $E(m) = 0$ for all m . Consequently any system with $E(1) = 0$ and $E(m) > 0$ for some m is not additive.*

Proof. $E(1) = 0$ forces every single effect $|o_c(0_c) - o_c(\theta_c^0)|$ to vanish, since $E(1)$ is an average of nonnegative terms. For additive O the effect of any D is the sum of these vanishing terms, hence zero. $\square$

For the threshold system the profile is a step: $E(m) = 0$ for $m \leq n - r$ and $E(m) = 1$ for $m > n - r$. For an additive system with equal contributions δ it is the line $E(m) = m\delta$. The two shapes are opposite in curvature, and the curvature is itself a coarse measurement of interaction. A profile that is flat and then rises steeply is the signature of a reserve consumed, and the position of the rise is the cut number s .

11.3 Nulls consume margin

Proposition 11.4 (Sequential nulls). *In an r -out-of- n threshold system, suppose $j \leq n - r$ distinct components have already been ablated. Then every further single ablation is null if $j < n - r$, and the next ablation is fatal if $j = n - r$. The tolerance margin of the surviving system is $n - r - j$.*

Proof. After j ablations, $n - j$ components remain, and the system is a threshold system with the same r . A further single ablation leaves $n - j - 1 \geq r$ components exactly when $j < n - r$. $\square$

The proposition has a methodological consequence. If successive ablations are performed on the same specimen, the j th null is a statement about a system with reduced margin, and the order of testing matters. Independent specimens (or restoration between ablations, Section 15) are needed to keep each test a test of the intact system. A study that removes components one by one and reports that the system survived each removal has reported a sequence in which the last null may sit one step from a cliff.

11.4 Why random ablation misses small cuts

Studies often ablate random subsets, either for convenience or to summarize robustness. A small unique cut can then be missed with high probability.

Proposition 11.5 (Random ablation and a unique small cut). *Suppose ε -cuts are exactly the sets containing a fixed s -subset S_0 . A uniformly random m -subset of K is a cut with probability*

$$\Pr = \frac{\binom{n-s}{m-s}}{\binom{n}{m}} = \prod_{i=0}^{s-1} \frac{m-i}{n-i}.$$

For $s = 2$ this equals $\frac{m(m-1)}{n(n-1)}$.

Proof. An m -subset contains S_0 iff it is S_0 together with $m - s$ of the remaining $n - s$ components, which gives the first expression. The product form follows from $\binom{n-s}{m-s} / \binom{n}{m} = \frac{m! (n-s)!}{n! (m-s)!} = \prod_{i=0}^{s-1} \frac{m-i}{n-i}$. $\square$

With $n = 1000$ components, $s = 2$, and random ablations of size $m = 30$, the probability that a trial contains the critical pair is $\frac{30 \cdot 29}{1000 \cdot 999} \approx 8.7 \times 10^{-4}$. About 1150 random trials are needed to expect a single hit. A random-ablation curve for this system would show smooth robustness up to substantial removal, while a two-component adversarial ablation would destroy it. Random and adversarial margins can differ by orders of magnitude, and a robustness claim should say which it is.

12 Slack, Reserve, and Distance to Failure

A third quantity separates from the output. Suppose the system performs acceptably when its state lies in a safe region $\mathcal{A} \subseteq \Theta$, and the observable is binary, $O = \mathbf{1}_{\mathcal{A}}$. A robustness margin is the distance to the failure boundary,

$$\rho(\theta) = d(\theta, \Theta \setminus \mathcal{A}),$$

in a metric on configurations chosen to reflect what failure depends on (load units, capacity units, or the like). Two configurations in $\mathcal{A}$ have $O = 1$ and can have different margins.

Proposition 12.1 (Functional invariance without robustness invariance). *There are systems and ablations with $\Delta O = 0$ and $\rho(\mathcal{R}_T) < \rho(\theta^0)$.*

Proof. Take n identical parallel supports under load W , each with capacity $f_{\max}$, and $\mathcal{A} = \{\ell \leq f_{\max}\}$ where ℓ is the load per surviving support. Baseline $\ell = W/n$, margin $\rho = f_{\max} - W/n$. After removing one support, $\ell = W/(n-1)$, and provided $W/(n-1) \leq f_{\max}$ the deck still stands, so $\Delta O = 0$, while $\rho = f_{\max} - W/(n-1)$, smaller by $W/(n(n-1))$. $\square$

With $n = 4$, $W = 3$, and $f_{\max} = 1.2$: the baseline load is 0.75 with margin 0.45, and after one removal the load is 1.0 with margin 0.20. The margin has fallen by 0.25 although the coarse observable is unchanged.

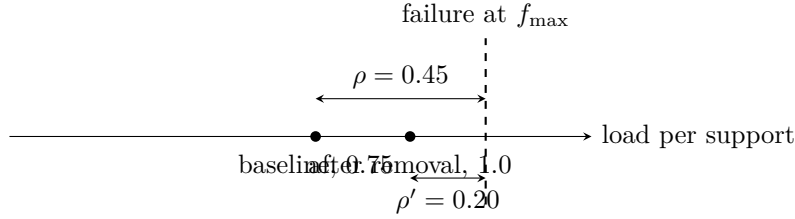


Figure 8: Both states lie on the safe side, so the coarse observable reads the same. The ablation moved the state toward the failure boundary and consumed reserve.

12.1 A homeostat with a limited actuator

The same phenomenon can be stated exactly in the homeostat, and doing so gives the margin a mechanism.

Example 12.2 (Homeostat with actuator limit). Modify Example 7.6 by bounding the effort variable, $x_2 \leq x_2^{\max}$, with the controller acting as before until the bound is reached. Define the *reserve* as $r = x_2^{\max} - x_2^0$, where $x_2^0 = y^* - a$ is the baseline effort. Ablate a portion αa of the source, $x_1 : a \mapsto a(1 - \alpha)$.

Proposition 12.3 (Reserve law). *In Example 12.2, the required effort after ablation is $x_2^0 + \alpha a$, and:*

(i) if $\alpha a \leq r$, the chronic effect is $\Delta^\infty = 0$ and the remaining margin (in effort units) is $r - \alpha a$;

(ii) if $\alpha a > r$, the actuator saturates at $x_2^{\max}$ and $\Delta^\infty = \alpha a - r$, with zero margin.

The chronic effect is $\Delta^\infty(\alpha) = \max(0, \alpha a - r)$ and the compensation gap is $\kappa_\infty(\alpha) = \min(\alpha a, r)$. The margin lost equals the compensation gap.

Proof. The required effort to restore $Y = y^*$ with $x_1 = a(1 - \alpha)$ is $y^* - a(1 - \alpha) = x_2^0 + \alpha a$. The clipped relaxation is one-dimensional and converges monotonically to $\min(x_2^0 + \alpha a, x_2^{\max})$. In case (i) the target is reachable, $Y = y^*$, and the margin is $x_2^{\max} - (x_2^0 + \alpha a) = r - \alpha a$. In case (ii) $x_2 = x_2^{\max}$ and $Y = a(1 - \alpha) + x_2^{\max} = y^* - (\alpha a - r)$. The static effect is αa , so $\kappa_\infty = \alpha a - \max(0, \alpha a - r) = \min(\alpha a, r)$, and in case (i) the margin lost is $r - (r - \alpha a) = \alpha a = \kappa_\infty$, with the same identity in case (ii) where the margin lost is r . $\square$

Two conclusions follow. A null ablation of size a establishes only $r \geq a$, a lower bound on the reserve, in the same way that Corollary 8.2 bounded the adaptation rate. And the null is the safe side of a cliff: a slightly larger ablation, or a slightly smaller reserve after some other loss, converts the silent null into a visible failure. A study that finds the ablation null has not found the system robust to that ablation in the sense of retaining its margin. It has found the system still on the safe side, with the margin reduced.

Remark 12.4 (Combinatorial and metric margins). Section 11 defined a combinatorial margin, the number of further components that can be removed. The metric margin ρ here measures the distance to failure in state space. In the threshold system they coincide up to units, and in general the metric margin also registers erosion that removes no component, such as the loss of headroom in a saturating actuator.

13 Non-Identifiability from Single Ablations

The preceding sections gave examples. This section states the general result that makes the single-ablation null uninformative about architecture, and quantifies how uninformative.

Let a *structure* on K be a monotone Boolean function $\varphi : \{0, 1\}^K \rightarrow \{0, 1\}$ with $\varphi(\mathbf{1}) = 1$. It is determined by its family of minimal paths, an antichain $\mathcal{A}$ of subsets of K : $\varphi(z) = 1$ iff z contains some $A \in \mathcal{A}$. The *single-ablation data* of φ is the vector $(\varphi(\mathbf{1}), \varphi(\mathbf{1} - e_1), \dots, \varphi(\mathbf{1} - e_n))$.

Theorem 13.1 (Compatibility with many architectures). *Let $n \geq 4$ and $k = \lfloor n/2 \rfloor$. The set of structures on K whose single-ablation data is $(1, 1, \dots, 1)$ (a system that works and in which no single ablation matters) contains at least*

$$2^N - n 2^{N/2}, \quad N = \binom{n}{k},$$

pairwise distinct structures. These have distinct joint-ablation behavior.

The proof, given in Appendix A, constructs one structure for each subfamily of k -subsets that misses every component. A second theorem strengthens the point by bounding the number of ablations needed to *determine* an unknown structure.

Theorem 13.2 (Query lower bound). *Let $n \geq 3$ and $k = \lfloor n/2 \rfloor$. Any deterministic procedure that determines an arbitrary structure on K by querying values $\varphi(z)$ (equivalently, by performing ablations of $K \setminus z$) requires at least $\binom{n}{k}$ queries in the worst case, even when it is promised that every single-ablation is null.*

Since $\binom{n}{\lfloor n/2 \rfloor} \geq 2^n/(n+1)$, the number of queries grows exponentially in the number of components. The worst-case instances are those in which the informative ablations are the ones removing about half the system. The full proof is in Appendix A.

Corollary 13.3. *Under a scalar binary observable, no polynomial-size collection of ablations determines the architecture of an arbitrary redundant system, and no collection of single ablations narrows the candidates beyond “every component is non-necessary.”*

Remark 13.4 (Priors carry the inference). Actual systems are not arbitrary. They are sparse, modular, approximately additive, spatially local, or otherwise structured. Under such priors, a modest number of ablations can be informative. The content of Theorems 13.1 and 13.2 is that the inference from ablation data to architecture is carried by these priors and not by the data alone, and that studies should say which priors they rely on. The same holds for the more general class of real-valued observables, which include the binary case.

Remark 13.5 (Generalization to t -wise ablations). The r -out-of- n threshold systems with $r \leq n - t$ have all ablations of t or fewer components null (Proposition 10.4). Hence a study restricted to t -wise ablations cannot distinguish any two structures whose minimal cuts all have size exceeding t . Theorem 13.1 is the case $t = 1$.

14 The Converse: Non-Null Results Are Also Over-Read

The logic of the essay applies in reverse. Just as $\Delta O \approx 0$ does not establish irrelevance, $\Delta O \gg 0$ does not establish that C performs the lost function.

Destabilization. Removing a component can break a system for reasons unrelated to what the component contributed. The upper right cell of Table 4 records this possibility: a null acute effect and a non-null reconfigured effect, caused by the adaptation itself misfiring. Removing a normalization layer from a trained network shifts activation statistics for all subsequent layers, producing catastrophic output whether or not the layer “computed” the function under study. The observed effect measures the system’s dependence on the assumptions that the component upheld, which is not the same as its dependence on the component’s function.

Diaschisis and distal effects. Lesion neuroscience has long recognized that damage to one region can depress activity in remote connected regions, a phenomenon for which von Monakow introduced the term diaschisis. A behavioral deficit after a lesion therefore does not localize the impaired function to the lesioned tissue. Lashley’s classic experiments removing cortical tissue in rats and finding deficits that scaled with amount and not location were an early demonstration that lesion effects and function need not align [7].

Deletion versus damage. Ablation replaces a component by its absence, which may differ from the component being present but broken, or replaced by a fixed default. A network unit set to zero is not the same intervention as the unit set to its mean activation or resampled from another input. The result depends on the replacement, and a study that reports only one has tested one counterfactual.

The symmetry is instructive. An ablation study reports the effect of an intervention and the system’s response to it together. A null is the response cancelling the intervention, and a large effect is the response failing to cancel it. Neither isolates the intervention’s contribution from the response.

15 Removal and Restoration: The Rebound Test

Ablation is usually performed once and in one direction. The inverse experiment, restoring the component after the system has adapted to its absence, is informative in its own right, and it provides a certificate of compensation that requires neither joint ablations nor a chronic condition.

Definition 15.1 (Ablate–restore cycle and rebound). Ablate C and let the system adapt for horizon T , then restore θ_C to its baseline value θ_C^0 before the free coordinates have time to respond. The configuration immediately after restoration is $(\theta_C^0, \psi_T^a(\theta_{-C}^0))$. The *rebound* is

$$\rho_T(C) = d\left(O(\theta^0), O(\theta_C^0, \psi_T^a(\theta_{-C}^0))\right).$$

Proposition 15.2 (Rebound certifies reconfiguration). *If $\rho_T(C) \neq 0$ then $\psi_T^a(\theta_{-C}^0) \neq \theta_{-C}^0$: the free coordinates moved during the ablation. Consequently the system is adaptive on the horizon T . In a non-adaptive system, $\rho_T = 0$ for every T .*

Proof. If $\psi_T^a(\theta_{-C}^0) = \theta_{-C}^0$ the post-restoration configuration equals θ^0 and $\rho_T = 0$. The contrapositive is the first claim. For $V \equiv 0$ the flow is the identity. $\square$

The certificate is one-sided. A nonzero rebound proves adaptation. A zero rebound does not prove its absence, because the free coordinates may have moved in directions to which O is blind at the baseline value of θ_C .

Proposition 15.3 (Rebound equals the compensation gap in the homeostat). *In Example 7.6, the observable immediately after restoration is $Y = y^* + \kappa_T$, that is, an overshoot of exactly $\kappa_T = a(1 - e^{-kT})$, and $\rho_T = \kappa_T$.*

Proof. At the end of the ablation, $x_2 = y^* - a e^{-kT}$ by Proposition 8.1. Restoring $x_1 = a$ gives $Y = a + y^* - a e^{-kT} = y^* + a(1 - e^{-kT})$. $\square$

The rebound therefore reads the compensation gap directly: the amount by which the system over-corrects when the component returns equals the amount it had been compensating. In the regulated regime the ablation gives a null, and the restoration gives a spike whose height is the missing measurement.

Remark 15.4 (Hysteresis). If the adaptation dynamics have several attractors, as in a bistable switch, the ablate–restore cycle can end in a different basin from the one it started in, and the system after the cycle is not the system before it. Removal and restoration are then not inverse operations. This is a form of history dependence: the state after the cycle retains a consequence of a path no longer visible in the configuration.

Remark 15.5 (Asymmetry as curvature). Even without adaptation, removal and addition differ for nonlinear responses. For a smooth scalar response f at operating point x and a step δ , $[f(x+\delta) - f(x)] + [f(x-\delta) - f(x)] = f''(x)\delta^2 + O(\delta^4)$. The sum of the effects of a step up and a step down of equal size estimates curvature, and it vanishes for a linear response. Ablation is a step down, and the missing step up is the missing half of a curvature measurement. A study restricted to ablation samples one side of the response.

16 Graded Interventions and Ablation Curvature

An ablation replaces θ_C^0 by 0_C in one step. Between these lies a family of partial ablations, and the observable along the family is a curve, not two numbers.

Definition 16.1 (Graded ablation). Fix a path $\alpha \mapsto c(\alpha)$ in Θ_C with $c(0) = \theta_C^0$ and $c(1) = 0_C$; for a linear component $c(\alpha) = (1 - \alpha)\theta_C^0$. The *dose–response curve* at horizon T is

$$\alpha \mapsto O_T(\alpha) = O(\mathcal{R}_T(S, \text{do}(\theta_C = c(\alpha)))), \quad 0 \leq \alpha \leq 1.$$

A single ablation samples the endpoints $O_T(0)$ and $O_T(1)$. The full curve exposes the slope $dO_T/d\alpha$ and the curvature $d^2O_T/d\alpha^2$, and it returns to the essay’s opening distinction between a finite intervention and a derivative.

16.1 Endpoint coincidence

The sine example of Section 2 has a graded form: $O_T(\alpha) = \sin(2\pi\alpha)$ satisfies $O_T(0) = O_T(1) = 0$ while the interior reaches ± 1 . The next result says how much an interior excursion can be excluded by regularity.

Proposition 16.2 (Interior bound from endpoints and smoothness). *If $O_T(0) = O_T(1)$ and O_T is L -Lipschitz in α , then*

$$\sup_{0 \leq \alpha \leq 1} |O_T(\alpha) - O_T(0)| \leq \frac{L}{2},$$

and the bound is attained by the tent function $O_T(\alpha) = L \min(\alpha, 1 - \alpha)$.

Proof. For each α , $|O_T(\alpha) - O_T(0)| \leq L\alpha$ and $|O_T(\alpha) - O_T(1)| \leq L(1 - \alpha)$. Since $O_T(0) = O_T(1)$ the left sides are equal, so the difference is at most $L \min(\alpha, 1 - \alpha) \leq L/2$. The tent function meets it at $\alpha = 1/2$. $\square$

Without a Lipschitz bound, coinciding endpoints constrain nothing between them. With one, the interior is confined but not eliminated, and L is a quantity the endpoint pair does not report.

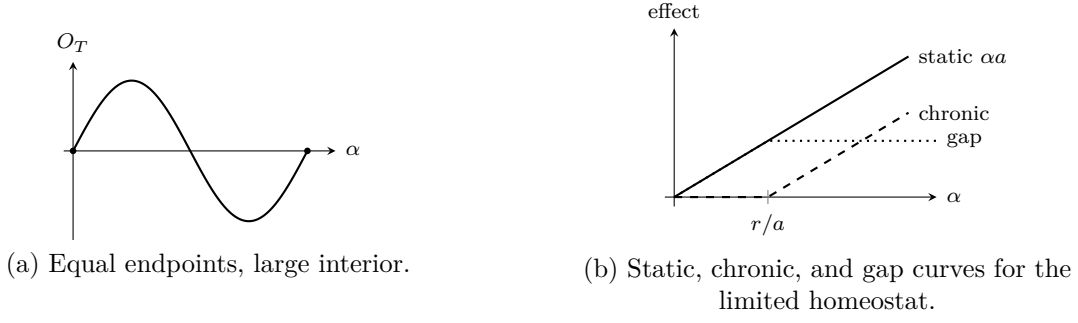


Figure 9: Dose-response curves. (a) Two endpoints can coincide while the path between them is highly non-null. (b) In the limited homeostat the chronic effect is flat until the reserve is exhausted and then rises with the static slope; the gap $\min(\alpha a, r)$ rises and then stays at the reserve.

16.2 Dose-response in the limited homeostat

Proposition 12.3 gives the dose-response curves in closed form. The static curve is the line αa . The chronic curve is $\max(0, \alpha a - r)$, flat until $\alpha = r/a$ and rising thereafter. The gap is $\min(\alpha a, r)$, rising and then constant at the reserve r .

Proposition 16.3 (Local silence without global irrelevance). *In Example 12.2 with $a > r$, the chronic response has zero derivative at $\alpha = 0$,*

$$\left. \frac{d\Delta^\infty}{d\alpha} \right|_{\alpha=0} = 0,$$

yet the ablation $\alpha = 1$ is non-null, with effect $a - r > 0$.

Proof. For $\alpha < r/a$ the chronic effect is identically zero, so its derivative at 0 vanishes. At $\alpha = 1$ it equals $a - r$. $\square$

This is the exact converse of the opening proposition. Section 2 showed that $\Delta O \approx 0$ does not imply $\partial O / \partial C = 0$. Proposition 16.3 shows that $\partial O / \partial C = 0$ does not imply $\Delta O \approx 0$. Neither the finite comparison nor the local derivative certifies irrelevance, and the dose-response curve is the object that contains both. The knee of the chronic curve identifies the reserve r , but only if the experiment reaches doses beyond it, which a physiological ablation may not. Where the reserve cannot be exhausted by removal, adding demand serves the same purpose.

Remark 16.4 (Curvature and the missing half). Section 15 observed that removal and addition together estimate curvature. The dose-response curve extended to $\alpha < 0$ (an increase of the component) supplies the other side, and the second difference $O_T(\alpha) + O_T(-\alpha) - 2O_T(0)$ estimates $O_T''(0)\alpha^2$ for small α . In the limited homeostat the chronic curve has a corner at $\alpha = r/a$ and no curvature at $\alpha = 0$, and the knee is visible only in the larger excursion.

17 Statistical Certification of a Null

The preceding sections concern what a null can mean when it is established. A prior question is when a null can be established at all. Failure to reject the hypothesis of no effect is not

certification of no effect [1]. What is needed is a test whose alternative hypothesis is the absence of a meaningful effect, and whose power is sufficient.

17.1 Equivalence testing

Let the ablation effect on trial j be $y_j = \mu + \eta_j$ with η_j independent, mean zero, and variance σ^2 known for simplicity. The claim to be certified is $|\mu| < \varepsilon$, where ε is the smallest effect of scientific interest. The two one-sided tests procedure rejects the null hypothesis $|\mu| \geq \varepsilon$ at level α when both one-sided tests reject, which is equivalent to the $(1 - 2\alpha)$ confidence interval for μ lying inside $(-\varepsilon, \varepsilon)$ [6]. In the known-variance case this is the condition

$$|\bar{y}| < \varepsilon - z_{1-\alpha} \frac{\sigma}{\sqrt{n}},$$

where $\bar{y}$ is the sample mean of n trials and $z_{1-\alpha}$ the standard normal quantile.

Proposition 17.1 (Minimum certifiable tolerance). *No number of favorable outcomes lets a study of n trials certify equivalence to any tolerance $\varepsilon \leq z_{1-\alpha}\sigma/\sqrt{n}$. In particular, if the true effect is exactly zero, the power to certify $|\mu| < \varepsilon$ is $2\Phi(\sqrt{n}\varepsilon/\sigma - z_{1-\alpha}) - 1$, and power at least $1 - \beta$ requires*

$$n \geq \left(\frac{\sigma}{\varepsilon}\right)^2 (z_{1-\alpha} + z_{1-\beta/2})^2.$$

Proof. The certification condition requires $\varepsilon - z_{1-\alpha}\sigma/\sqrt{n} > |\bar{y}| \geq 0$, impossible unless $\varepsilon > z_{1-\alpha}\sigma/\sqrt{n}$. When $\mu = 0$, $\bar{y} \sim N(0, \sigma^2/n)$, so the probability that $|\bar{y}| < c$ with $c = \varepsilon - z_{1-\alpha}\sigma/\sqrt{n}$ is $2\Phi(c\sqrt{n}/\sigma) - 1$, which is the stated power. Requiring this to be at least $1 - \beta$ gives $c\sqrt{n}/\sigma \geq z_{1-\beta/2}$, and rearranging yields the bound on n . $\square$

For $\alpha = 0.05$ and $\beta = 0.2$ the factor $(z_{0.95} + z_{0.90})^2$ is about 8.56, so $n \gtrsim 8.56 (\sigma/\varepsilon)^2$.

ε/σ	Trials n needed
1	9
1/2	35
1/4	138
1/10	857

The cost grows as the square of the required precision. A tolerance of a tenth of the trial-to-trial standard deviation requires hundreds of trials for a single component. Ablation studies of large systems commonly average over a handful of seeds or inputs, so their nulls certify only tolerances comparable to the noise itself.

17.2 Many components

Screening m components for nulls is a family of m equivalence claims. Each is an intersection-union test, so no correction is required within a component, but the claims must be corrected

across components if a statement about all of them is made. With Bonferroni adjustment $\alpha \rightarrow \alpha/m$, the constant $(z_{1-\alpha} + z_{1-\beta/2})^2$ rises from 8.56 to about 20.9 for $m = 100$ (using $z_{1-\alpha/m} \approx 3.29$), a factor of roughly 2.4. Testing all pairs of n components multiplies m by $\binom{n}{2}$ and the required sample by a further logarithmic factor, which is one more reason the interaction structure is rarely screened.

Remark 17.2 (Tolerance is a scientific choice). The tolerance ε must be fixed from the scientific question, as the smallest effect of interest, before data are collected. Choosing ε after the fact to match the achievable precision reproduces the original error in different notation: the claim then reports the resolution of the experiment and not the behavior of the system.

18 Consequences for Interpretation and Reporting

18.1 Domains

Neural network interpretability. Circuit-level analyses of transformers have identified backup components that take over when primary ones are ablated. In the analysis of indirect object identification in a small language model, ablating the main name-moving heads was found to activate other heads with a similar role [13]. The phenomenon of downstream layers adjusting to compensate for an ablated layer has been described as self-repair and studied as a general effect in language models [8]. Earlier work on single-unit ablation asked how much individual units matter for generalization and found that units with strong selectivity were not more important than others [9], which is a form of the null-versus-importance gap in a different vocabulary. In all these cases, a component can fail an ablation test while being causally central, and any inference from ablation to mechanism should be paired with a test of the compensation hypothesis.

Genetics. The discrepancy between genetic mutants and acute knockdowns has been attributed in several cases to transcriptional adaptation, where mutant transcripts trigger up-regulation of related genes [11, 4]. The knockout phenotype is then the phenotype of $\mathcal{R}_\infty$ and the knockdown is closer to $\mathcal{R}_0$. Reporting both is the experimental analogue of the pair $(\Delta^0, \Delta^\infty)$ in Table 4.

Structures and load paths. Structural engineering already contains the concept that nothing happening to the coarse observable, the structure standing, is compatible with substantial reconfiguration of the internal load paths. The reserve capacity of redundant members is a designed compensation, and its consumption is invisible to an inspector who looks only at whether the building stands.

Institutions. Removing a role, rule, or office from an organization and finding that outcomes are unchanged is routinely read as evidence that the item was superfluous. The same logic applies: duties may have been reassigned, informal channels may have absorbed the function, and the reconfiguration may have a cost paid elsewhere in the system.

18.2 A minimal reporting standard

The following items are proposed as a minimum for any study whose conclusion depends on a null ablation. They are listed in Appendix C in checklist form.

- R1. Intervention type and replacement value.** State how the component was removed (zeroed, mean-ablated, resampled, deleted, degraded) and whether other choices were tested.
- R2. Adaptation horizon.** State whether the system had time to reconfigure, and report at least one acute and one chronic condition when both are feasible.
- R3. Observable and resolution.** State the observable, its resolution, and the tolerance ε used to call an effect null, with a justification. Use an equivalence test rather than failure to reject when claiming absence of an effect.
- R4. Operating conditions.** State the regime tested and whether saturation, ceiling, or floor effects are plausible in it. Vary the background where possible.
- R5. Joint ablations.** Report ablations of pairs among the plausible compensators, at least those suggested by the architecture. State the largest joint size tested.
- R6. Effort variables.** Report the state of the non-regulated variables after ablation. A null on the regulated variable accompanied by shifts in effort variables is a compensation, not an absence of function.
- R7. Claim language.** Replace “ C is not necessary” or “ C does nothing” by the exact claim: “no effect exceeding ε on O under conditions X at horizon T for interventions of type I .”

None of these is burdensome for a single study, and they convert the null from a claim of absence into a bounded statement that other work can build on or contradict.

19 Designing for Informative Nulls

The reporting standard of Section 18 governs how existing experiments are described. The results above also bear on how systems and experiments should be built so that a future null carries information. Five principles follow from the theory.

- D1. Expose the effort variables.** Proposition 7.7 places the information about a compensated ablation in the free coordinates, and Theorem 23.2 shows that observing them turns single ablations into a reconstruction. A system that logs only its regulated output discards this information at the source. Recording the state of backup pathways is cheap when planned and often impossible retrospectively.
- D2. Provide ablation as a first-class interface.** An intervention layer that supports several replacement modes (zero, mean, resampled, and restored) and several horizons converts the question “what did C do” into a family of measurements. This includes the ability to restore, since the rebound test of Section 15 depends on it.

- D3. Instrument the reserve.** The tolerance margin of Proposition 11.2 is a property of the system that no single ablation reveals, but it can be monitored. In structural health monitoring, gauges on redundant members detect that one has failed even though the structure stands, because the survivors' loads shift by a predictable factor $n/(n-1)$. The analogous telemetry for other systems is the load on backup components, and a sudden change in it is a signal that reserve has been consumed.
- D4. Test on fresh specimens or restore between tests.** Proposition 11.4 shows that successive ablations on one specimen test progressively weaker systems. Independent replicates, or ablate-restore cycles that return the system to baseline before the next test, keep each ablation a test of the intact system.
- D5. Select ablations adaptively.** With a budget of ablations, choose the next intervention using the residuals of the previous ones. Theorem 13.2 says the worst case is exponential, but under sparsity priors adaptive selection can approach the linear cost seen in Theorem 23.2, where n single ablations suffice.

None of these principles requires knowledge of the architecture in advance. They amount to arranging for the system to keep, and for the experimenter to see, the variables into which a silent ablation transfers its information.

20 Connections to Broader Frameworks

The formal results above are self-contained. Several frameworks developed elsewhere in this research program address the same structure in a different vocabulary, and the correspondences are recorded briefly.

Refusal in SpherePOP. In the SpherePOP formalism the REFUSE primitive is not mere nonparticipation. Refusing one route can change which route becomes admissible, so the effect of a refusal is not the effect of the route being absent from an otherwise unchanged set of routes. This is the operator picture of Theorem 7.8: refusing C is an intervention followed by a change in the admissibility structure of what remains, and the two must be distinguished.

Residuals as topology probes. In the residual-and-admissibility line of work, the remainder after a computation is treated as information and not as error. The residual after ablation plays this role. It is the trace left in the surviving components, and it is a probe of the topology of what remains. The effort variables of Proposition 7.7 and the survivor strains of the load-path example are instances.

Charts in HOLOcoord. An element cannot be assigned significance from its isolated coordinate, because removing it changes the transition relations among the surviving charts. This is the statement that the local role of a component is defined relative to the atlas in which it sits, and that deleting it produces a different atlas. The distinction between $S - C$ and $\mathcal{R}(S, \text{do}(C = 0))$ is a distinction between the atlas with a chart deleted and the atlas re-glued.

History in MEM|8. Persistence of a rendered state after deletion does not establish the absence of historical dependence. The state can retain consequences of a path that is no longer locally visible. In the language of this essay, the observable is unchanged because the effect has moved into an unobserved variable, and the history that produced that variable is what the null conceals.

These correspondences are offered as orientation. The essay does not depend on them, and each reduces, for present purposes, to the theorems above.

21 Objections and Replies

Objection 1: Experimentalists already know this. The caution is widely voiced, and terms such as compensation, redundancy, and backup pathway are standard. The claim of this essay is narrower. The caution is rarely operationalized: reports seldom state the horizon, the tolerance, the replacement value, or whether joint ablations were run, and conclusions are still phrased as absence. Sections 8 through 17 give the quantities that would turn the caution into a reportable result: a rate bound (Corollary 8.2), a margin (Proposition 11.2), a rebound (Proposition 15.3), and a certifiable tolerance (Proposition 17.1).

Objection 2: Compensation is a rare special case. How common compensation is in a given class of systems is an empirical question, and the essay takes no position on the frequency. The point is logical. The inference from null to irrelevance is licensed only if compensation, saturation, degeneracy, coarseness, and level-set coincidence have been excluded, and their frequency in a class of systems is what determines how often the inference fails. Degeneracy and redundancy are widely regarded as pervasive in biological and engineered systems [3]. The experiments proposed here, a second horizon, a joint ablation, a rebound, estimate the frequency for the system at hand rather than assuming it.

Objection 3: A non-null effect still proves necessity. It proves necessity at the tested horizon, for the tested replacement, under the tested conditions, and only if destabilization and distal effects are excluded (Section 14). The symmetry of the argument is the reason the essay treats both directions. The claim is not that ablation is uninformative but that it reports the combined effect of an intervention and a response, and that the two must be separated by design.

Objection 4: This proves too much. If nulls cannot be trusted, nothing can. The argument does not say nulls cannot be trusted. It says which claims a null can support, and it gives those claims a precise form: no effect exceeding ε on O under conditions X at horizon T , certified by an equivalence test with stated power. Such claims are useful, they can be combined across studies, and they can be strengthened by the additional experiments the essay describes. In the linear case, Theorem 23.2 shows that nulls can be turned from an uncertainty into a complete measurement when the observable is rich enough.

Objection 5: The exponential lower bound is a worst-case result. It is. Theorems 13.1 and 13.2 construct adversarial families, and real systems are structured. The remark following them states that priors such as sparsity, modularity, and locality carry the inference in practice. The result is offered as a demand that these priors be stated, since without them the data underdetermine the architecture, and as a warning that the underdetermination is severe when the observable is a single bit.

Objection 6: Adaptation is a nuisance, and the frozen counterfactual $S - C$ is what is wanted. That depends on the question. If the question is mechanistic, what C contributes at the operating point, then Δ^0 is the relevant quantity and the acute regime is appropriate. If the question concerns robustness, what happens to the system when C is lost, then Δ^∞ is relevant. The same intervention answers both questions when the horizon is varied, and the gap between the two answers is the compensation of Lemma 2.6. Treating one question's answer as the other's is the error under discussion.

22 Invariance Is Relative to an Observer

The preceding sections have repeatedly changed the observable in order to change the verdict: a scalar output versus the effort variables, a coarse pass/fail measure versus a fine one, an endpoint versus a curve. These are instances of one relation among observers.

22.1 Refinement

Definition 22.1 (Observers and refinement). An observer is a map $O : \Theta \rightarrow Y$. The observer O_2 *refines* O_1 if $O_1 = f \circ O_2$ for some map f . Then O_1 is coarser.

Proposition 22.2 (Factorization criterion). O_1 *factors through* O_2 *if and only if* $\sim_{O_2} \subseteq \sim_{O_1}$, *that is, $[\theta]_{O_2} \subseteq [\theta]_{O_1}$ for every θ .*

Proof. If $O_1 = f \circ O_2$ and $O_2(\theta) = O_2(\theta')$ then $O_1(\theta) = O_1(\theta')$. Conversely, if the inclusion holds, define $f(O_2(\theta)) = O_1(\theta)$. This is well defined because $O_2(\theta) = O_2(\theta')$ implies $O_1(\theta) = O_1(\theta')$, and it defines f on the image of O_2 . $\square$

Corollary 22.3 (Silence is monotone under coarsening). *If O_2 refines O_1 and an ablation is exactly null for O_2 , it is exactly null for O_1 . The converse fails: an ablation can be null for O_1 and not for O_2 . If f is 1-Lipschitz, the same holds for ε -nulls.*

Proof. $\mathcal{R}_T \sim_{O_2} \theta^0$ implies $\mathcal{R}_T \sim_{O_1} \theta^0$ by the proposition. For the ε version, $d_1(O_1(\mathcal{R}_T), O_1(\theta^0)) = d_1(f(O_2(\mathcal{R}_T)), f(O_2(\theta^0))) \leq d_2(O_2(\mathcal{R}_T), O_2(\theta^0))$. For the failure of the converse, take the homeostat with $O_1 = Y$ and $O_2 = (Y, x_2)$: after chronic ablation $\Delta O_1 = 0$ and $\Delta O_2 = a \neq 0$. $\square$

Every observer is a partition of Θ , and refinement is the partition order. An ablation is null for an observer when the intervention stays within one block, and coarser observers have larger

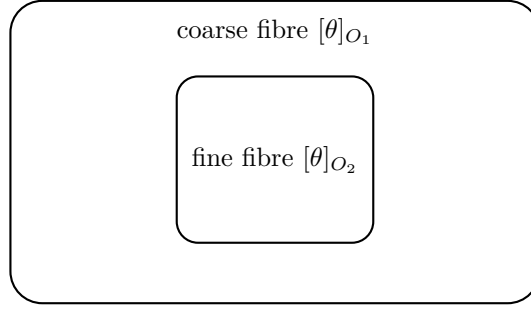


Figure 10: Refining the observer shrinks the equivalence class. Any ablation that stays within the fine fibre is invisible to both observers; one that leaves it but stays in the coarse fibre is visible only to O_2 .

blocks. This is the abstract statement that lies under the earlier results about resolution: a scalar is a very coarse observer of a network, and the survivor vector is a fine one.

22.2 Richer observation shrinks the compatible set

The same order acts on the inference from data to architecture. Let $\mathcal{S}$ be a class of candidate architectures and let $F_1 : \mathcal{S} \rightarrow \mathcal{D}_1$ be an experimental protocol, mapping an architecture to the data it would produce.

Proposition 22.4 (Refined protocols shrink compatible sets). *If $F_2 = (F_1, G)$ appends further measurements G to a protocol F_1 , then for every observed datum $d_2 = (d_1, g)$,*

$$\{S \in \mathcal{S} : F_2(S) = d_2\} \subseteq \{S \in \mathcal{S} : F_1(S) = d_1\}.$$

Proof. Any S with $F_2(S) = (d_1, g)$ has $F_1(S) = d_1$. □

The proposition is elementary and it is the abstract form of the contrast between Theorems 13.1 and 23.2. Under the scalar protocol the compatible set contains at least $2^N - n2^{N/2}$ structures. After the protocol is refined to the survivor vectors, the compatible set in the symmetric linear class is a single architecture. Everything between is a monotone shrinking of the set as observers are added.

22.3 No privileged observer

Two ablation studies of the same system with different observers can both be correct and disagree, exactly as two studies at different horizons can (Corollary 8.3). The disagreement is a comparison of partitions. A report should state its observer, and the reporting standard of Section 18 already does so under R3. The additional recommendation from this section is to report, where possible, a finer observer alongside the coarse one, because Corollary 22.3 guarantees that the finer observer can only reduce silence.

23 From Null Results to Structural Tomography

Section 13 showed that scalar single-ablation data cannot identify architecture. Section 7 showed that the information discarded by a null lives in the residual, that is, in the state of the surviving components. This suggests that the negative result is an artifact of the coarse observable, and that with a richer observable a sequence of ablations and their compensations can reconstruct the hidden network. This section proves a case in which that is exactly true.

23.1 A linear model

Consider a linear equilibrium network. Let K be a real symmetric positive definite $n \times n$ matrix (a conductance, stiffness, or coupling matrix), let $u \in \mathbb{R}^n$ be a known input, and let $x = Gu$ with $G = K^{-1}$ be the equilibrium state. Ablating node i clamps $x_i = 0$ and removes the node's equation, which leaves the reduced system

$$K_{-i,-i} x^{(i)} = u_{-i},$$

where $K_{-i,-i}$ is the principal submatrix omitting row and column i . Let x_{-i} denote the baseline state restricted to the other nodes. The following identity describes the residual.

Lemma 23.1 (Ablation residual). *The residual $r^{(i)} = x^{(i)} - x_{-i}$ satisfies*

$$r^{(i)} = K_{-i,-i}^{-1} K_{-i,i} x_i = -\frac{G_{-i,i}}{G_{ii}} x_i.$$

Proof. The baseline equations restricted to rows $-i$ read $K_{-i,-i} x_{-i} + K_{-i,i} x_i = u_{-i}$. Subtracting $K_{-i,-i} x^{(i)} = u_{-i}$ gives $K_{-i,-i} (x^{(i)} - x_{-i}) = K_{-i,i} x_i$, which is the first equality. For the second, the identity $KG = I$ restricted to the block of rows $-i$ and column i gives $K_{-i,-i} G_{-i,i} + K_{-i,i} G_{ii} = 0$, so $K_{-i,-i}^{-1} K_{-i,i} = -G_{-i,i}/G_{ii}$. $\square$

The residual is a multiple of a column of the Green's function G . It is the response of the surviving network to the removal of node i 's coupling, and it points along the direction in which node i was connected to the rest.

23.2 Identifiability

Theorem 23.2 (Linear structural tomography). *Let K be symmetric positive definite with $G = K^{-1}$ having no zero entries, let u be known, and suppose the baseline state $x = Gu$ has $x_i \neq 0$ for all i . If one observes x and the ablated states $x^{(i)}$ for all $i = 1, \dots, n$, then K is uniquely determined.*

Proof. By Lemma 23.1, $r^{(i)} = -(G_{-i,i}/G_{ii})x_i$, so since $x_i \neq 0$ the normalized column $\tilde{g}_i := G_{-i,i}/G_{ii} = -r^{(i)}/x_i$ is determined by observations. Thus $G_{ji} = G_{ii}\tilde{g}_i[j]$ for all $j \neq i$. Symmetry of G gives $G_{ii}\tilde{g}_i[j] = G_{jj}\tilde{g}_j[i]$, and since G has no zero entries both $\tilde{g}_i[j]$ and $\tilde{g}_j[i]$ are nonzero, so

$$\frac{G_{jj}}{G_{ii}} = \frac{\tilde{g}_i[j]}{\tilde{g}_j[i]}.$$

Taking $i = 1$ determines $d_j := G_{jj}/G_{11}$ for all j (with $d_1 = 1$). Then $G = G_{11}\tilde{D}$ where $\tilde{D}_{ji} = d_i\tilde{g}_i[j]$ for $j \neq i$ and $\tilde{D}_{ii} = d_i$ are known. Finally $x = Gu = G_{11}\tilde{D}u$; since $x_1 \neq 0$, $(\tilde{D}u)_1 \neq 0$ and $G_{11} = x_1/(\tilde{D}u)_1$. Thus G is determined, and $K = G^{-1}$. $\square$

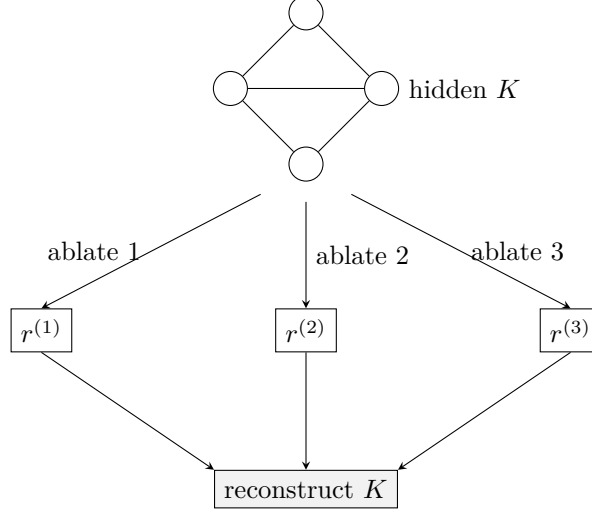


Figure 11: Structural tomography. Each single ablation of the hidden network produces a residual vector $r^{(i)}$ proportional to a column of the Green's function. Under the hypotheses of Theorem 23.2, the collection of residuals determines K .

Theorem 23.2 contrasts with Theorems 13.1 and 13.2. In the latter, the observable is a single bit and n single ablations reveal nothing about structure. In the former, the observable is the full survivor vector, and n single ablations reveal everything. The two theorems bracket the role of resolution: the null on the scalar is uninformative because the information has moved into variables the scalar does not record. Whenever the survivors are recorded, the nulls become a measurement.

Example 23.3 (Parallel supports revisited). Return to n identical parallel supports under load W . Removing one leaves $W/(n-1)$ per survivor against baseline W/n , a ratio $n/(n-1)$. With ratios 2, $3/2$, $4/3$, $6/5$ for $n = 2, 3, 4, 6$ respectively, a single measured strain ratio identifies the number of supports, although the coarse observable (the deck stands) is silent. The single ablation is a null in the binary observable and a full measurement of n in the strain observable.

23.3 Sequences of interventions

The result concerns single ablations in a linear symmetric model. The wider proposal it motivates is that sequences of interventions and the compensations they induce can serve as a form of tomography for invisible networks: a family of ablations $\{a_1, a_2, \dots\}$ paired with the residuals $\{r_1, r_2, \dots\}$ constrains the interaction structure, in the way that a family of projections constrains an image. Several questions arise.

- **Nonlinearity.** Lemma 23.1 has a first-order analogue near an operating point of a nonlinear system, with G replaced by the inverse Jacobian. How large a perturbation is admissible before the local linearization fails, and can residual asymmetry between an ablation and its inverse estimate the nonlinearity?
- **Adaptation.** When the system adapts, the residual depends on the horizon T . The time course of the residual, rather than its endpoint, may encode the adaptation dynamics V , so that the family $\mathcal{R}_T$ is itself an observable of the hidden architecture.
- **Adaptive design.** Given a budget of ablations, which sequence maximizes information? By Theorem 13.2 the worst-case cost is exponential, but under sparsity priors adaptive selection of the next ablation from the previous residuals may achieve far better.
- **Noise.** Identifiability under measurement noise depends on the conditioning of K and of its principal submatrices, and the noise tolerance of the reconstruction is an open quantity.

These questions belong to a sequel whose organizing image is the load path that no one drew: an internal route for force, signal, or influence, invisible in the finished structure and recoverable only from the pattern of how the structure responds when pieces are taken away.

24 Preserved Distinctions

The results of the essay can be summarized as a principle about when distinctions may be discarded.

Preserve distinctions before composing them.

A distinction between two states may be erased by an observer (a coarse output, an average, a threshold) or by a composition of operations (aggregating before comparing, quantizing before storing). Once erased at one level it is recoverable only if some other observer still separates the states. The following instances have appeared above.

Refusal is not absence. Removing a component by subtraction ($\mathcal{R}_0$) and refusing it as an intervention ($\mathcal{R}_T$) are different operations, because the latter is followed by reconfiguration of what remains (Theorem 7.8). The two must not be composed into a single notion of “the system without C ”.

Persistence is not truth. That the observable persists after deletion is a fact about the observer, not about the state. The unchanged reading is compatible with a changed configuration, a running cost, and a reduced margin (Section 6). Persistence of a value certifies nothing about the underlying history.

Recovery is not invention. Under the survivor-vector observer in the symmetric linear class, the compatible architecture is unique and the reconstruction is a function of the data alone (Theorem 23.2): this is recovery. Under the scalar observer the compatible set is exponentially large (Theorem 13.1), and any single architecture selected from it is selected by a prior: this is invention. The same procedure can be either, depending on the observer.

More abstractly, observational equivalence is not identity:

$$O(x) = O(y) \not\Rightarrow x = y.$$

Proposition 24.1 (Recoverability of distinctions). *Let $O_1, \dots, O_J$ be the available observers. A distinction $\theta \neq \theta'$ is recoverable if and only if some O_j has $O_j(\theta) \neq O_j(\theta')$. All distinctions are recoverable if and only if $\bigcap_j \sim_{O_j}$ is the identity relation, that is, the joint observer $(O_1, \dots, O_J)$ is injective. Composing an observer with any map q can only enlarge the erased relation: $\sim_O \subseteq \sim_{q \circ O}$, with strict inclusion whenever q identifies two values in the image of O .*

Proof. The first two claims restate the definitions. If $O(\theta) = O(\theta')$ then $q(O(\theta)) = q(O(\theta'))$, giving the inclusion. If q identifies two distinct values $O(\theta_1) \neq O(\theta_2)$ then $\theta_1 \sim_{q \circ O} \theta_2$ but $\theta_1 \not\sim_O \theta_2$. $\square$

A system is interpretable only insofar as the distinctions erased at one level remain recoverable at another. The practical rule that follows from the proposition is a rule about order: record the un-aggregated state (effort variables, residuals, the state of backups) and the intervention before averaging over seeds, inputs, or specimens, and before thresholding to a pass/fail outcome. Composition can be done afterward from the record. It cannot be undone.

25 Conclusion

The sentence “we removed C , and nothing happened” reports that a finite intervention left a measured observable within tolerance. That is all it reports. It does not report that the sensitivity vanishes, because a finite step is not a derivative and the step may land on the same level set. It does not report that C was redundant, because the system may have re-supplied C ’s contribution rather than possessed a duplicate. It does not report that the system minus C was measured, because in adaptive systems the measured object is $\mathcal{R}(S, \text{do}(C = 0))$, the output of a reconstruction, and $S - C$ is only the frozen counterfactual of it.

The positive content of the null is the compensatory transformation of Corollary 7.5, together with the compensation gap of Lemma 2.6, which bounds from below how much the system did in response. The information lost by a scalar readout is not destroyed. It is relocated into effort variables, into joint ablations, into the time course of the residual, and into the state of the survivors, where Theorem 23.2 shows it can, in a tractable case, be fully recovered.

The geometric reading unifies the results. A null ablation is a return of the trajectory to the observational fibre of the baseline, and an experiment that records only the observable sees the quotient by that fibration. In these terms, *invariance is not absence of change; it is change constrained to remain invisible under a specified observation map*. Output invariance does not

imply configuration invariance, cost invariance (Proposition 9.2), or margin invariance (Proposition 12.3), and a zero local derivative does not imply a null finite ablation (Proposition 16.3), just as a null finite ablation does not imply a zero derivative.

The added structure of timing, margin, and rebound supplies further handles on the same phenomenon. A null at a single horizon bounds the adaptation rate from below (Corollary 8.2), and the disagreement between studies at different horizons is a measurement of the crossover time. The whole ablation profile reveals the tolerance margin that single ablations conceal, and random ablation can miss a small cut with overwhelming probability (Proposition 11.5). The rebound after restoration reads the compensation gap directly (Proposition 15.3). And a null can be certified only at tolerances the noise level and sample size permit (Proposition 17.1).

Two consequences follow for practice. Claims of absence should be written as bounded statements of the form given in the reporting standard, with tolerance, horizon, replacement, and conditions stated. And the null result should be treated as the beginning of an experiment: the natural next measurements are the joint ablation, the second horizon, and the residual, each of which converts a silence into a finding.

The general principle, stated in Section 24, is to preserve distinctions before composing them. Observational equivalence is not identity, refusal is not absence, persistence is not truth, and recovery is not invention. A system remains interpretable only insofar as the distinctions erased at one level remain recoverable at another.

A Proofs of Theorems 13.1 and 13.2

Proof of Theorem 13.1. Let $n \geq 4$, $k = \lfloor n/2 \rfloor$, and let $\binom{[n]}{k}$ denote the k -subsets of $K = [n]$, so $|\binom{[n]}{k}| = N$. For a subfamily $\mathcal{F} \subseteq \binom{[n]}{k}$ define $\varphi_{\mathcal{F}}(z) = 1$ iff $z \supseteq A$ for some $A \in \mathcal{F}$. All members of $\mathcal{F}$ have the same size, hence form an antichain, and $\varphi_{\mathcal{F}}$ is monotone. Distinct subfamilies yield distinct functions: for $A \in \binom{[n]}{k}$ we have $\varphi_{\mathcal{F}}(\mathbf{1}_A) = 1$ iff $A \in \mathcal{F}$, since a k -set contains a member of $\mathcal{F}$ only if it equals that member.

Call $\mathcal{F}$ *good* if for every $i \in [n]$ some $A \in \mathcal{F}$ has $i \notin A$. If $\mathcal{F}$ is good then $\mathcal{F} \neq \emptyset$, so $\varphi_{\mathcal{F}}(\mathbf{1}) = 1$, and for each i the set $[n] \setminus \{i\}$ contains some $A \in \mathcal{F}$, so $\varphi_{\mathcal{F}}(\mathbf{1} - e_i) = 1$. Thus every good subfamily has single-ablation data $(1, \dots, 1)$.

It remains to count good subfamilies. A subfamily is *not* good iff there is i with $i \in A$ for every $A \in \mathcal{F}$, that is, $\mathcal{F} \subseteq \{A \in \binom{[n]}{k} : i \in A\}$. This set has $\binom{n-1}{k-1}$ elements, so at most $2^{\binom{n-1}{k-1}}$ subfamilies are contained in it. By a union bound over i , the number of subfamilies that are not good is at most $n 2^{\binom{n-1}{k-1}}$. Now $\binom{n-1}{k-1} = \frac{k}{n} \binom{n}{k} \leq N/2$ because $k = \lfloor n/2 \rfloor \leq n/2$. Hence at least $2^N - n 2^{N/2}$ subfamilies are good. For $n = 4$ this equals $64 - 32 = 32 > 0$ and the bound is positive for all larger n .

Distinct good subfamilies give distinct structures, and each pair of distinct structures differs at some retained set $\mathbf{1}_A$ with $|A| = k$, that is, under the joint ablation of the $n - k$ components outside A . $\square$

Proof of Theorem 13.2. Let $n \geq 3$, $k = \lfloor n/2 \rfloor$, $N = \binom{n}{k}$. For $B \subseteq \binom{[n]}{k}$ define

$$\varphi_B(z) = \begin{cases} 1 & \text{if } |z| \geq k+1, \\ 1 & \text{if } |z| = k \text{ and } z \in B, \\ 0 & \text{otherwise,} \end{cases}$$

with sets identified with indicator vectors. This is monotone: if $z \subseteq z'$ and $\varphi_B(z) = 1$, either $|z| \geq k+1$ and then $|z'| \geq k+1$, or $z \in B$ and either $z' = z \in B$ or $|z'| \geq k+1$. Also $\varphi_B(\mathbf{1}) = 1$ because $n \geq k+1$. Every single ablation is null, because $|[n] \setminus \{i\}| = n-1 \geq k+1$ when $n \geq 3$. Distinct B give distinct functions, since $\varphi_B(z) = \mathbf{1}[z \in B]$ on k -sets.

Consider any deterministic procedure using $q < N$ queries. Answer every query at a set of size k with 0, and every query at a set of any other size with the value forced by the definition (1 for $|z| \geq k+1$, and 0 for $|z| \leq k-1$). These are consistent with $B = \emptyset$. Since $q < N$, some k -set z^* is never queried, and the same answers are also consistent with $B = \{z^*\}$: every queried k -set differs from z^* and receives 0 under both, and queries at other sizes have B -independent answers. The procedure's output is a function of the answers alone, so it cannot distinguish the two structures, which are distinct. Hence at least N queries are required. $\square$

Remark A.1. The adversary argument in the second proof shows more. The informative queries are exactly the ablations that leave $k = \lfloor n/2 \rfloor$ components, so the hardness lies at the intermediate scale of joint ablation, and it is the scale which small studies do not reach.

B Worked Numerical Examples

B.1 Saturation

With $O = \tanh(c + r)$ and ablation $c : 1 \rightarrow 0$:

r	$O(c = 1)$	$O(c = 0)$	ΔO
0	0.7616	0.0000	0.7616
1	0.9640	0.7616	0.2024
2	0.9951	0.9640	0.0311
3	0.9993	0.9951	0.0042
5	0.99999	0.99991	0.00008

The same component, with the same effect on its input, produces effects on O differing by four orders of magnitude depending on the background. A null at $r = 5$ says nothing about the regime $r = 0$.

B.2 Parallel supports

For n identical parallel supports under load W , per-support load before and after one removal:

n	Before	After	Ratio $n/(n-1)$
2	$W/2$	W	2.000
3	$W/3$	$W/2$	1.500
4	$W/4$	$W/3$	1.333
6	$W/6$	$W/5$	1.200
10	$W/10$	$W/9$	1.111

The coarse observable (the deck stands) is unchanged whenever the survivors' capacity exceeds the post-removal load. The fine observable identifies n from the ratio alone. The information is present in each row and is lost only by recording the coarse observable.

B.3 Two-component OR system

With $v(\emptyset) = 0$ and $v(\{C\}) = v(\{D\}) = v(\{C, D\}) = 1$:

Quantity	C	D
Single-ablation effect $v(K) - v(K \setminus c)$	0	0
Order C, D marginal	1	0
Order D, C marginal	0	1
Shapley value ϕ	1/2	1/2
Joint ablation effect $\Delta_{\{C,D\}}$	1	
Interaction ε_{CD}	1	

B.4 Homeostat

With $y^* = 10$, $x_1 = a = 4$, $k = 1$: baseline $x_2 = 6$ and $Y = 10$. Ablate $x_1 \rightarrow 0$. The static effect is $Y = 6$, so $\Delta^0 = 4$. Under relaxation $x_2(t) = 10 - 4e^{-t}$, which tends to 10 and gives $\Delta^\infty = 0$. At $t = 1$, $x_2 = 6 + 4(1 - e^{-1}) \approx 8.53$ and $Y \approx 8.53$, so the reconfigured effect is $\Delta^1 \approx 1.47$ and $\kappa_1 \approx 2.53$.

Check: $x_2(t)$ solves $\dot{x}_2 = -(x_2 - 10)$ with $x_2(0) = 6$, so $x_2(t) = 10 - 4e^{-t}$; at $t = 1$, $e^{-1} \approx 0.3679$ and $x_2 \approx 8.528$.

C Reporting Checklist

The following items can be appended to the methods section of any study whose conclusions depend on ablation.

-
- ☐ Removal method and replacement value stated; alternatives tested or listed as untested
 - ☐ Acute and chronic horizons reported, or the missing one stated as a limitation
 - ☐ Observable, resolution, and tolerance ε stated with justification
 - ☐ Equivalence test or confidence interval reported for claimed nulls
 - ☐ Operating conditions listed; background varied, or saturation excluded on stated grounds
 - ☐ Pairwise ablations among plausible compensators reported; largest joint size stated
 - ☐ State of unregulated (effort) variables after ablation reported
 - ☐ Claim written in bounded form: effect below ε on O under X at horizon T for intervention I
 - ☐ Priors used to infer architecture from ablation data stated explicitly
-

References

- [1] D. G. Altman and J. M. Bland. Absence of evidence is not evidence of absence. *BMJ*, 311:485, 1995.
- [2] R. E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston, 1975.
- [3] G. M. Edelman and J. A. Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24):13763–13768, 2001.
- [4] M. A. El-Brolosy et al. Genetic compensation triggered by mutant mRNA degradation. *Nature*, 568:193–197, 2019.
- [5] L. H. Hartwell, P. Szankasi, C. J. Roberts, A. W. Murray, and S. H. Friend. Integrating genetic approaches into the discovery of anticancer drugs. *Science*, 278:1064–1068, 1997.
- [6] D. Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.
- [7] K. S. Lashley. In search of the engram. *Symposia of the Society for Experimental Biology*, 4:454–482, 1950.
- [8] T. McGrath, M. Rahtz, J. Kramar, V. Mikulik, and S. Legg. The Hydra effect: Emergent self-repair in language model computations. arXiv:2307.15771, 2023.
- [9] A. S. Morcos, D. G. T. Barrett, N. C. Rabinowitz, and M. Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations*, 2018.
- [10] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

- [11] A. Rossi et al. Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature*, 524:230–233, 2015.
- [12] L. S. Shapley. A value for n -person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, 1953.
- [13] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. arXiv:2211.00593, 2022.