

From Microscopic Capability to Collective Function: An Evidential Ladder and Reducibility Audit for Interacting Units

Kai Wang

Conceptual / Methodological Perspective with Computational Audit Case Studies

Version 0.5 — Preprint-ready candidate — 20 August 2026

Abstract

As active and physical-learning systems acquire sensing, memory, communication, internal regulation, and learning, richer microscopic architectures increasingly produce different macroscopic behavior. The central inferential problem is not whether such mechanisms can change collective dynamics, but what level of claim the evidence supports. We formulate a cumulative evidential ladder for claims of a new useful collective capability: (I) microscopic mechanism availability; (II) a causal collective dynamical effect under matched intervention; (III) response-class expansion relative to an adversarially optimized lower architecture under a hard capability budget; (IV) positive task-specific value after explicit cost accounting; and (V) collective irreducibility relative to a stated family of reducible alternatives. The evidence types are heterogeneous—interventional, set-theoretic, decision-theoretic, and causal—but the composite claim becomes stronger as criteria accumulate. We operationalize the framework through a reducibility audit combining expressive null models, matched resources, information-preserving shuffles, explicit reductions, held-out evaluation, and stopping rules. A Vicsek-type calibration shows that verdicts depend on the reducer family: local alignment sustains $O(1)$ polarization while a matched independent-persistent null decays approximately as $N^{-0.46}$, yet a global-broadcast reducer also sustains $O(1)$ polarization and therefore closes any stronger locality-specific irreducibility claim for that observable. Two worked audit examples then show why stronger claims can be reduced, downgraded, or remain unresolved. A state-matched anticipation assay reduces, under its stated low-dimensional predictive objective, to a single-controller problem. An interaction-only mechanical stiffness-reallocation gate is highly topology-sensitive: across 24 held-out network realizations the pre-specified mean comparison is statistically inconclusive, and a post-hoc ratio sensitivity check can reverse the nominal sign without yielding a stable network-level advantage. We position the framework as a domain-specific synthesis of null-model reasoning, functional reducibility, causal-emergence ideas, and active/physical-learning methodology rather than as a universal theory or certificate of emergence.

Keywords: active matter; physical learning; collective capability; reducibility; null models; causal irreducibility; model comparison; adaptive matter

1. Introduction

Nonequilibrium statistical physics has long shown that simple local interactions can generate macroscopic order, transport, stresses, waves, and phase behavior that are naturally described at a scale above individual units. Active matter extends this program by allowing microscopic constituents to consume energy and generate sustained motion, producing collective states unavailable to passive equilibrium descriptions [1]. Recent work pushes the same logic toward increasingly capable units: active particles may carry internal state spaces [4], exchange or learn policies [5], integrate collision histories [6], perform measurements that alter many-body phases [7], or be embedded in programmable and learning materials [8-13]. Dauchot has summarized this frontier as a trajectory from active to odd to smart matter, emphasizing the shift from spontaneous organization toward task-driven functionality [2].

This expansion creates a methodological difficulty. Suppose a richer microscopic model displays a new pattern, faster recovery, improved robustness, anticipatory preparation, or superior task performance. What has actually been established? The added mechanism may have changed the dynamics. But the same response may still be reachable by a sufficiently expressive lower-class model. A new response may also be costly once sensing, communication, reconfiguration, and actuation are accounted for. Even a useful many-particle response may be reducible to a single predictor, independently adapting units, independently adapting bonds, or a low-dimensional centralized representation.

The problem is not unique to active matter. Causal-emergence work asks whether macro-level descriptions can possess stronger intervention-based causal efficacy than micro-level descriptions [14]. Null-model methodology distinguishes evidence that a process is causally necessary from mere comparison with a baseline [15]. Recent work on higher-order networks similarly asks when a richer interaction representation is dynamically reducible to a simpler one [16]. These traditions share a concern with necessity, explanatory level, and reduction, but they are not normally formulated as an explicit evidential ladder for microscopic-capability claims in active and physical-learning systems.

Here we develop such a domain-specific synthesis. For the specific composite claim that a richer microscopic architecture creates a new, useful, irreducibly collective capability, we organize the evidential burden into five cumulative criteria: a microscopic mechanism must be present; it must causally alter collective dynamics; the resulting response must remain outside the reachable set of an adversarially optimized lower class; at least one such expanded response must have positive task-specific value after explicit cost accounting; and that valuable capability must resist the strongest justified reductions. The underlying evidence is heterogeneous in kind, so “higher” here means that more criteria have been satisfied, not that all questions lie on a single homogeneous scale.

We calibrate the audit against nested reducer families. A standard local-alignment flocking model first shows that collective polar order survives finite-size increase while a matched independent-persistent null decays toward zero. Adding a global-broadcast reducer then shows that polarization itself is not specific to local interaction: the broader reducer reproduces $O(1)$ order. Two worked audit examples demonstrate further claim downgrade or non-establishment. The first begins with a state-matched anticipatory particle assay and shows that its headline signal can be reduced to a one-controller prediction problem under the stated low-dimensional sufficient statistic. The second forces both sensing and actuation through a mechanical network, but after symmetric adversarial optimization and testing across independent network disorder, the data fail to establish a stable expected advantage for the train-selected distributed rule over a local bond-wise reducer.

2. An Architectural Scaffold for Increasingly Capable Units

The following hierarchy is a bookkeeping scaffold, not a novelty claim and not an ordering of collective performance. Related ladders and taxonomies already occur in intelligent-active-particle, smart-matter, and physical-learning literatures [2,3,8].

2.1 Passive interacting units

$$M0: \dot{x}_i = \mu(\sum_j F_{ij} + F_{i\text{ext}}) + \xi_i$$

The passive class contains externally expressed physical degrees of freedom, interaction forces, noise, and external driving. It can already support macroscopic density, stress, pressure, flow, and disturbance propagation. The point is simply that local interaction can produce useful macroscopic variables without internal information processing.

2.2 Active fixed-law units

$$M1: \dot{x}_i = \mu\sum_j F_{ij} + v_0 p_i + \xi_i$$

The active class adds sustained microscopic energy transduction. In an active-Brownian realization, the propulsion direction itself diffuses rotationally. This expands mechanism availability, but it does not imply a universal performance ordering between passive and active models [1].

2.3 Regulated / history-erasing units

$$M2: \dot{s}_i = G(s_i, E_i, N_i; \Theta_0), \quad u_i = A(s_i, E_i, N_i; \Theta_0)$$

Regulated units may contain nonlinear feedback, sensing, transient integration, delay, and finite memory while retaining a fixed response architecture. The term history-erasing is operational: under a specified washout intervention, histories that differ before the washout converge to the same effective state for the subsequent probe.

2.4 History-bearing units

$$M3: \dot{H}_i = K(H_i, s_i, E_i, N_i), \quad u_i = A(s_i, H_i, E_i, N_i)$$

A retained state remains causally available after the washout used to equalize faster states. This distinction is not coordinate-invariant: the variables can always be combined into a larger state vector. It is therefore an experimental architecture constraint, not an ontological partition of dynamical systems. The existence of a history variable is also not itself called adaptation. Functional adaptation is a task- and environment-dependent verdict.

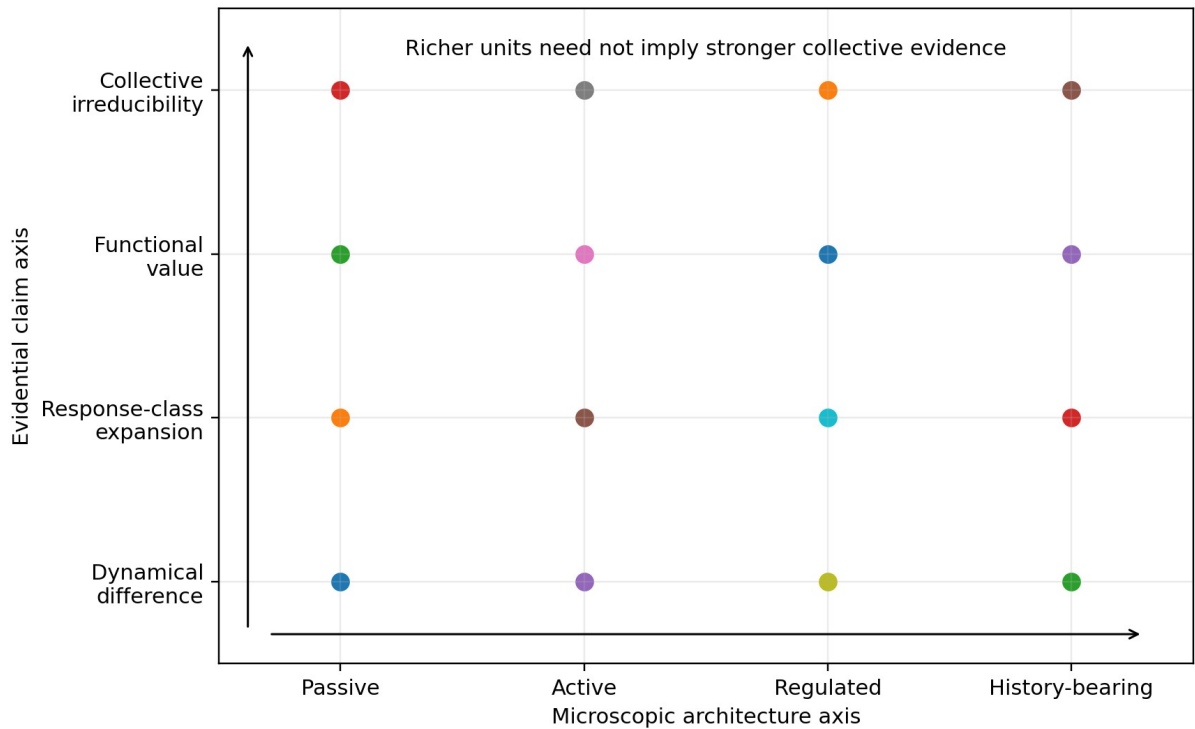


Figure 1. Architecture and evidence are distinct axes. Moving from passive to history-bearing units increases microscopic mechanism availability. It does not determine whether the resulting macroscopic effect is merely different, expands a response class, improves function, or is irreducibly collective.

3. A Cumulative Evidential Ladder for New Collective-Capability Claims

Level I — Microscopic mechanism availability

A new microscopic degree of freedom, energy source, state variable, controller, memory channel, or communication rule has been introduced. This is the architectural prerequisite. It establishes no macroscopic effect by itself.

Level II — Collective dynamical difference

Under a matched intervention in which non-target resources are held fixed, enabling the candidate mechanism causally changes at least one collective observable Y . This criterion is stronger than observing different parameterizations: it asks whether the mechanism itself is responsible for the dynamical difference.

Level III — Collective response-class expansion

Define R_k as the set of collective responses reachable by lower architecture M_k within a hard capability budget: permitted sensors, actuators, interaction range, observation time, and other non-costed physical constraints are fixed, but resource use inside that domain is not yet penalized in the objective. A Level-III claim requires a response Y^* reached by M_{k+1} that has not been reached by the adversarially optimized lower-class family R_k under the tested search. This is a set-theoretic reachability statement, not yet a statement that Y^* is useful. Like Level V, Level III is defeasible: finite search cannot prove that Y^* lies outside every conceivable lower-class realization, so response-class expansion must always be stated relative to an explicit tested lower family, parameter domain, and optimization procedure.

Level IV — Task-specific functional value

For the composite claim considered here, Level IV requires that at least one Level-III expanded response also has positive task-specific value. Resource use that was treated as a hard admissibility constraint at Level III is now explicitly costed or exposed on a Pareto front—for example energy, sensing, communication, reconfiguration work, or time. A useful mechanism that merely makes an already reachable response cheaper may still be scientifically important, but it is an efficiency claim rather than evidence for a new collective capability in this ladder.

Level V — Collective irreducibility

The strongest claim additionally requires failure of the strongest justified reducible alternatives for the same valuable expanded response: single estimators, independent agents, independent interacting elements such as bonds, mean-field models, or low-dimensional centralized summaries. Irreducibility is never asserted simpliciter. It is always relative to an explicit tested reducible family F , resource budget, and intervention set. A finite audit can falsify an irreducibility claim but cannot prove universal inaccessibility to every conceivable reducer.

$$\Delta C > 0 \Rightarrow \Gamma > 0 \Rightarrow \Delta irr > 0$$

Here ΔC denotes a response-class gap relative to the tested lower family, Γ a task-specific functional gap, and Δirr an irreducibility gap relative to the tested reducible family F . For the cumulative composite claim, satisfying Level V presupposes the earlier criteria by definition. The underlying evidence types nevertheless differ: Level II is interventional, Level III set-theoretic and defeasible under finite model search, Level IV decision-theoretic, and Level V causal/reductive and likewise relative to an explicit reducer family. The ladder therefore orders evidential burden for one target claim; it does not assert that all collective phenomena can be ranked on a single scalar axis.

4. Relation to Existing Frameworks

4.1 Active, informational, and smart matter

The architectural progression is intentionally non-novel. Intelligent-active-particle work already discusses increasingly capable active agents [3]; thermodynamically consistent active matter with complex internal states is available [4]; decentralized learning has been given a kinetic theory [5]; collision-history memory can drive population-level phase transitions in physical swarms [6]; and measurement/information-processing rules can themselves generate collective phase behavior [7]. Anticipatory active-matter formalisms additionally treat agents whose present motion depends on prospective states [17]. The present contribution is therefore not another component in this ladder.

4.2 Physical learning and mechanical memory

Material memory is a mature topic spanning hysteresis, return-point memory, training, aging, and cyclic protocols [9]. Physical-learning research goes further by allowing local material parameters to change through experience-dependent rules [8]. Directed aging can train mechanical networks [10], supervised physical changes can implement input-output mappings [11], nonequilibrium protocols can improve adaptability [12], and temporal contrastive learning exploits implicit nonequilibrium memory [13]. These precedents occupy much of the mechanism space that initially motivated our hierarchy.

4.3 Causal emergence, null models, and functional reducibility

The audit is better understood as a synthesis across methodological traditions than as a new theory of emergence. Hoel and colleagues quantify causal emergence by intervention-based effective information and show that a macro model can in some systems possess greater causal efficacy than a micro description [14]. Null-model methodology asks whether a candidate process is actually required to explain an observed pattern [15]. Functional-reducibility work asks when richer interaction representations are dynamically necessary rather than merely descriptively available [16]. Romeo et al. derive information-theoretic bounds on the robustness of self-organized systems, framing organization as noisy encoding and identifying substrate-level information limits [18]. That program is more formal and supplies quantitative bounds; the present audit addresses a different level of inference: how a researcher should downgrade or retain a capability claim after specific alternative architectures are tested. Related relevance criteria for nonequilibrium perturbations similarly ask when an added microscopic feature changes large-scale behavior rather than merely being present [21]. These approaches are complementary, and the present framework claims novelty only for their explicit assembly into an active/physical-learning claim-audit workflow with stopping rules and worked audit examples.

5. A Collective Capability Audit

5.1 Specify the target claim before selecting the baseline

A model comparison that is adequate for detecting a dynamical difference may be useless for testing causal necessity. The investigator should state in advance whether the hypothesis concerns difference, response-class expansion, functional value, or irreducibility.

5.2 Construct the strongest lower-class alternative

All mechanisms not themselves under test should remain available to the lower class. If retained history is the target mechanism, the lower class may still use nonlinear current-state feedback, stochastic policies, static heterogeneity, and state-dependent interactions. The scientific question is whether the architectural addition is necessary, not whether a weak baseline loses.

5.3 Match non-target resources

Energy input, sensing bandwidth, communication range, action bandwidth, reconfiguration work, parameter count, and observation time should be matched when they are not constitutive of the mechanism being tested.

5.4 Use matched causal nulls

A shuffled-history null retains the history variable, parameter count, communication machinery, and action opportunity while destroying only the history-future relation. Analogous nulls should preserve as much of the candidate architecture as possible while deleting the proposed causal link.

5.5 Use state matching cautiously

State matching can isolate retained history from residual physical-state differences. But it can also manufacture an information asymmetry if only one model retains the future-relevant statistic. State-matched

anticipation should therefore be treated as a capability-isolation test, not by itself as evidence of collective irreducibility.

5.6 Attempt explicit reductions

A useful reduction ladder is full collective → independent interacting elements → independent agents → single estimator/controller. The binding reduction depends on the substrate. In a mechanical network, independent bonds may be a stronger null than independent particles because they preserve interaction-generated local state while deleting cross-bond information. Any surviving “irreducibility” statement must name the tested reducible family F ; passing the finite ladder is evidence against those reducers, not a universal certificate.

5.7 Separate information from function

Information acquisition, retention, integration, physical realization, and task outcome are distinct stages. Success at one stage should not be interpreted as evidence for the whole causal chain.

5.8 Test scale dependence and predefine stopping rules

Finite-size analysis should preserve meaningful dimensionless controls. More importantly, a failed decisive null should trigger claim downgrade rather than a bespoke redesign chosen only to restore a positive result.

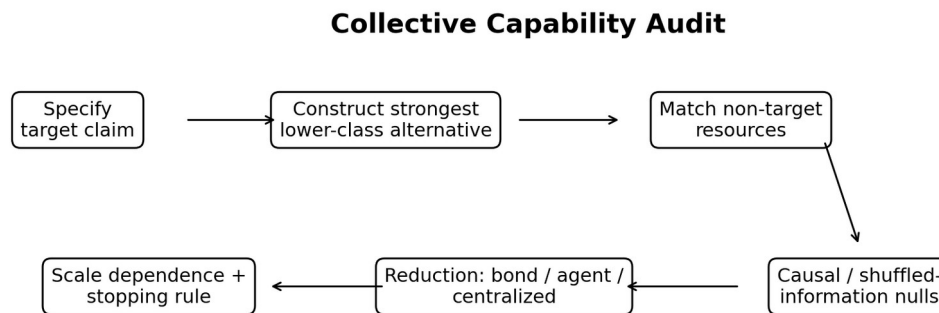


Figure 2. Collective Capability Audit. The sequence is designed to increase the burden of evidence as the claim becomes stronger. Failure at a later stage downgrades the claim rather than invalidating all lower-level observations.

6. Calibration Positive Control: Local Alignment Produces Persistent Collective Polar Order

We used a standard Vicsek-type local-alignment model [19,20] as a calibration. Agents move at fixed speed in a periodic two-dimensional domain and update their heading toward the mean orientation of neighbors within radius R , followed by angular noise. The first reducible null retains identical self-propulsion, angular-noise amplitude, density, observation time, and single-agent persistence but deletes neighbor-orientation information. We then add a stronger centralized reducer in which each agent receives the instantaneous population-mean orientation and aligns to that globally broadcast direction under the same angular-noise budget. Global polarization P is the magnitude of the mean heading vector.

Across $N=128, 256, 512$, and 1024 , the independent-persistent model gives mean polarizations $0.0698, 0.0569, 0.0415$, and 0.0266 , with a log-log slope of approximately -0.463 . The local-alignment model gives $0.972, 0.974, 0.904$, and 0.911 , with slope approximately -0.039 . The global-broadcast reducer gives approximately 0.993 at every tested N , with log-log slope -2.4×10^{-5} . Thus local alignment decisively exceeds the matched independent null, but polarization itself is also realizable by the stronger globally informed reducer.

At $N=1024$, the mean polarization ratio between local alignment and the independent-persistent model is approximately 34.2, while the global-broadcast reducer reaches $P \approx 0.993$ and is about 37.3 times the independent value. This calibration therefore demonstrates both retention and downgrade. The audit retains an interaction-dependent collective-order claim relative to the independent-local family, but it does not retain a stronger locality-specific irreducibility claim once a reducer with global orientation information is admitted. The result is intentionally about reducer-relative inference, not about proving that the Vicsek mechanism is uniquely necessary for polarization.

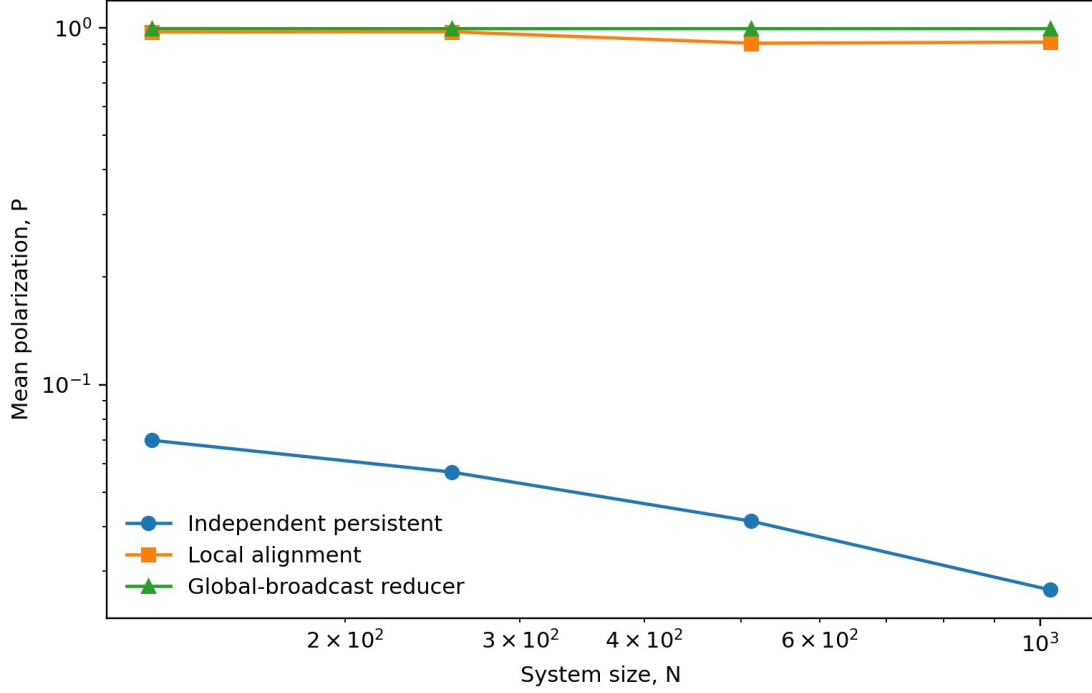


Figure 3. Calibration across nested reducer families. Mean global polarization P versus system size for local-alignment agents, a matched independent-persistent null, and a global-broadcast reducer. The independent null decays close to the $N^{-1/2}$ finite-sampling expectation; local alignment sustains $O(1)$ polar order; and global broadcast also sustains $O(1)$ order. The broader reducer therefore closes any claim that local interaction is uniquely necessary for polarization.

7. Audit Case Study I: Apparent Collective Anticipation

Consider a sequence of binary perturbations $R_n \in \{-1, +1\}$ with temporal persistence $P(R_{n+1}=R_n)=p$. A history-bearing architecture retains a statistic H correlated with the next perturbation after a state-matching intervention, whereas the history-erasing architecture does not. A signed pre-perturbation collective displacement can therefore be nonzero for the history-bearing class and zero for the history-free class in expectation.

The apparent result is weaker than it first appears. Under the stated quadratic objective, and only when the future-relevant information is compressible to the low-dimensional sufficient statistic $b(H)=E[R_{n+1}|H]$, the preparation problem reduces to $J(x|H)=b(H)x+\lambda x^2$ with optimum $x^*=-b(H)/(2\lambda)$. The many-body substrate is therefore unnecessary for that particular capability claim. This reduction should not be generalized to high-dimensional predictive environments or objectives whose optimum is intrinsically many-body.

This does not make the manipulation scientifically useless. It supports the limited statement that retained predictive history can alter future preparation. The collective interpretation is what fails. The correct reduction test preserves the predictive information while removing the collective substrate.

8. Audit Case Study II: Interaction-Only Mechanical Reallocation

8.1 Goal and pre-specified gate

The second test was designed specifically to eliminate the single-agent loophole. Both the local signal and its actuation were forced through a mechanical network. The pre-specified gate asked whether access to one-hop neighboring bond information produced a robust mechanical advantage beyond the strongest justified bond-local reducer. Because the alpha implementation uses a static equilibrium compression proxy rather than temporally integrated memory, the case is presented as a stiffness-reallocation gate, not as a general test of history-bearing learning.

8.2 Mechanical network

The gate used a two-dimensional disordered spring network initialized from a 6×6 node array ($N=36$). Interior and boundary coordinates were weakly randomized, Delaunay triangulation supplied candidate edges, and long edges were removed while preserving connectivity and minimum local degree, targeting mean coordination $z \approx 4.4$. The bottom boundary was fixed and spatially varying downward forces were applied to top-boundary nodes.

Each bond e had rest length $l_{0,e}$ and baseline stiffness $k_0=1$. Compression was softened beyond a threshold $\epsilon_b=0.03$: the incremental stiffness beyond the threshold was βk_e with $\beta=0.05$. Equilibrium states minimized elastic energy minus external work with L-BFGS-B.

8.3 High-dimensional predictive load field

$$r_{n+1} = \rho S r_n + \sqrt{(1-\rho^2)} \epsilon_n, \quad \rho = 0.8$$

S shifts the top-boundary load field by one spatial position. External force component j was $F_{n,j} = -F_0[1+aL \tanh(r_{n,j})]$ with $F_0=0.20$ and $aL=0.50$. For the Gaussian shift process, a component using only its own previous value has prediction MSE 1, whereas access to the appropriate one-hop neighbor has MSE $1-\rho^2=0.36$. A rank- q centralized linear summary has MSE $1-\rho^2q/K$, so matching the one-hop optimum requires retained rank proportional to the number K of spatial load components.

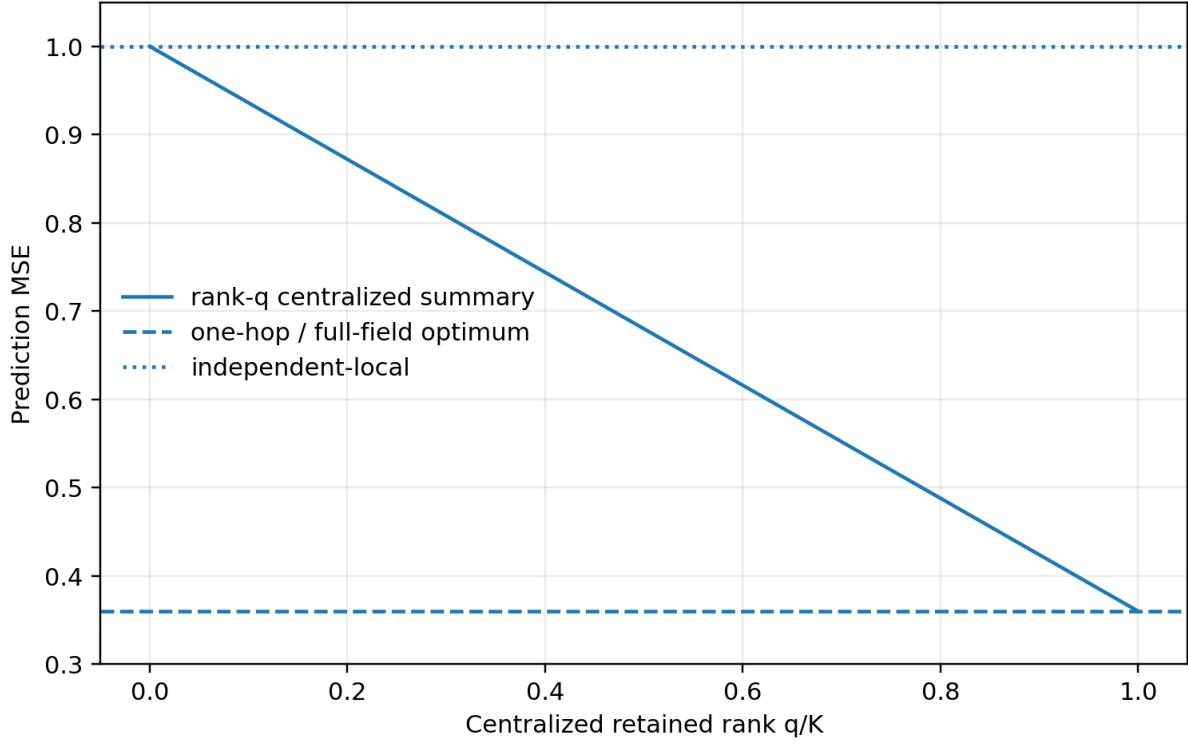


Figure 4. Information-layer benchmark for the spatial Markov field at $\rho=0.8$. An independent-local predictor has MSE 1. One-hop spatial information reaches MSE 0.36. A centralized rank- q summary approaches this optimum only as q approaches the spatial field dimension K .

8.4 Static compression proxy and matched reconfiguration budget

After equilibration under load r_n , each bond received a local compression proxy $H_e = \max(0, -\epsilon_e)$, where $\epsilon_e = (l_e - l_{0,e})/l_{0,e}$. This quantity is interaction-generated but static: it is not a time-integrated history variable and should not be interpreted as demonstrating material memory. A raw reallocation score s was mean-centered and projected to a fixed L2 stiffness budget $W=0.5$, with $0.5 \leq k_e \leq 1.5$. Mean centering prevents a candidate architecture from improving simply by increasing total stiffness.

8.5 Symmetric adversarial families and multi-network evaluation

To remove the asymmetric-optimization concern in v0.2, v0.7 gave both independent and distributed candidates the same nonlinear transformation family. For local proxy H , both could use powers $\gamma \in \{0, 0.5, 1, 2\}$ and threshold quantiles $q \in \{0, 0.5, 0.75\}$. Distributed candidates received only one additional degree of freedom: a one-hop coefficient ratio $r \in \{-1, -0.5, -0.25, 0, 0.25, 0.5, 1\}$; $r=0$ explicitly nests the independent family. Candidate selection used four independent training network realizations (seeds 20–23), one loading episode each. Final inference used 24 previously unseen network realizations (seeds 100–123), each evaluated on four held-out loading episodes. Statistical resampling was performed at the network level, not the load level. Because model selection on a small finite sample can itself be noisy [23], the frozen held-out comparison is accompanied below by a clearly labeled post-hoc ratio-sensitivity analysis that is used diagnostically, not as confirmatory evidence.

The train-selected independent candidate used $\gamma=1$ with the upper-quartile threshold. The train-selected distributed candidate used $\gamma=1$, no threshold, and one-hop ratio -0.25 . These rules were frozen before the 24-network evaluation. An additional episode-shuffle control reassigned the compression-proxy field across loading episodes within each network, preserving the marginal network statistics while disrupting the load-specific proxy–future pairing.

9. Results of the Interaction-Only Gate

9.1 Single-network result and why it was insufficient

The original v0.6a single-network gate is retained as a cautionary intermediate result rather than as the final inference. On one N=36 network, an expressive independent-bond family closed the apparent distributed advantage: distributed-minus-independent CVaR95 was -0.001023 (-0.0377%) over 20 held-out loading seeds, with a loading-seed bootstrap interval spanning zero. Because all 20 loads shared the same network realization, this result could not establish disorder-level generality. It motivated the v0.7 multi-network audit rather than supporting a broad null conclusion.

9.2 Symmetric multi-network held-out comparison

Architecture	Mean CVaR95	Failure fraction	Compliance
Independent-bond, train-selected	2.916413	0.068434	12.816503
Distributed one-hop, train-selected	2.924593	0.067774	12.816770
Episode-shuffled distributed	3.034721	0.069884	12.816503

Across the 24 held-out network realizations, each averaged over four loading episodes, the train-selected independent candidate had mean CVaR95=2.916413 and the train-selected distributed candidate 2.924593. Thus distributed-minus-independent was +0.008180, or +0.2805% relative to the independent value; positive values are worse. A network-level bootstrap 95% interval was [-0.3218, 0.3473], which is far wider than the nominal mean difference and spans both modest advantage and modest disadvantage. The data therefore fail to establish a stable expected advantage or disadvantage for the train-selected distributed rule. Mean compliance differed by only +0.000267 and failure fraction by -0.000659. The episode-shuffled distributed control had mean CVaR95=3.034721, indicating that the load-linked spatial proxy was informative even though the necessity of cross-bond access was not established.

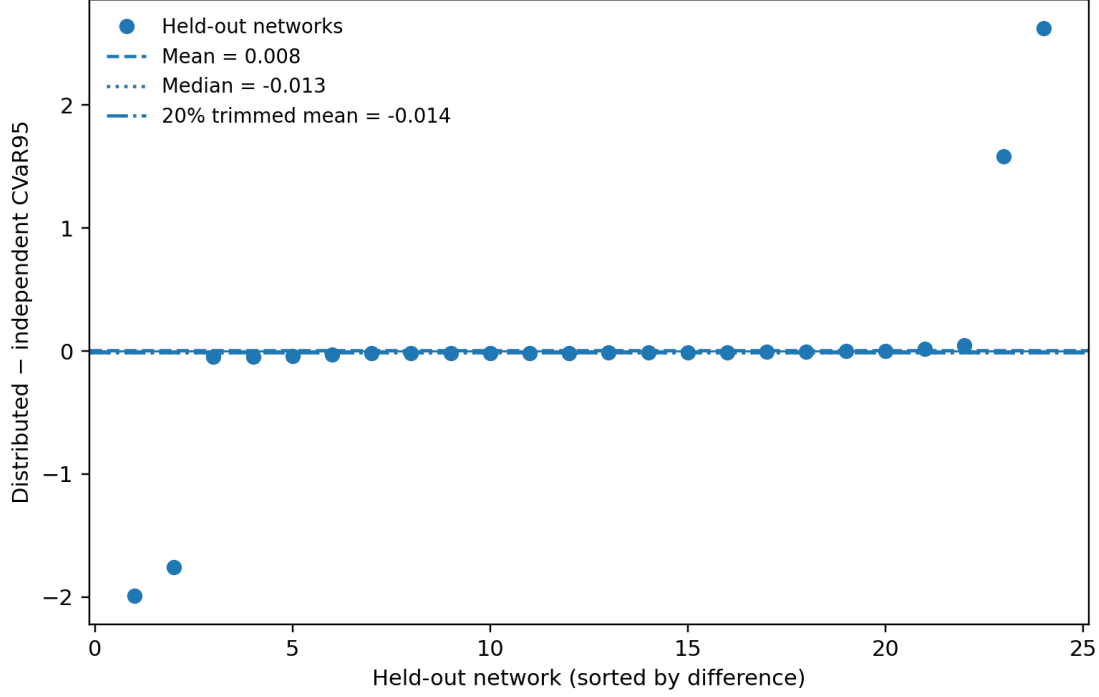


Figure 5. Network-level distributed-minus-independent CVaR95 differences across 24 held-out mechanical-network realizations, each averaged over four loading episodes. Negative values favor the distributed rule. Twenty of 24 differences are negative, but a few large reversals dominate the arithmetic mean. The mean is $+0.0082$, whereas the median is -0.0134 and the 20% trimmed mean is -0.0142 . The wide network-bootstrap interval for the mean spans zero. These summaries are descriptive; the pre-specified decision criterion remains the disorder-averaged mean.

9.3 Disorder heterogeneity and claim downgrade

The disorder-level result is heterogeneous and underpowered rather than evidence of absence. The distributed rule had lower CVaR95 on 20 of 24 networks (83.3%), usually by a small amount, while four network realizations produced large reversals. The arithmetic mean difference was $+0.0082$, but the median was -0.0134 and the 20% trimmed mean -0.0142 . This disagreement between location summaries shows that the conclusion is sensitive to the heavy topology-dependent tails of the fixed-size disorder distribution. Because the pre-specified primary functional criterion was expected CVaR95 across disorder, we do not replace the go/no-go rule post hoc with sign count, median, or trimmed mean. The correct confirmatory conclusion is narrower: the study failed to establish a disorder-generalized expected-performance advantage for the train-selected one-hop rule. The full distribution should nevertheless be reported because disorder-sensitive systems can violate simple self-averaging expectations, particularly near critical or instability-dominated regimes [22].

This conclusion is intentionally narrow. It does not imply that distributed physical learning is generally reducible to independent bonds, nor does it establish that the true expected effect is zero. It shows that the pre-specified gate did not establish the stronger cross-bond necessity claim. Once the independent-bond family closed the confirmatory gap, additional mean-field or centralized mechanical reducers were not required for that negative decision, although they would be relevant to any future attempt at a positive irreducibility claim.

9.4 Post-hoc selection-sensitivity check

The four-network training set is intentionally small and can introduce model-selection noise. To test whether the frozen distributed coefficient was unusually unfavorable, we performed a post-hoc ratio sweep within the same $\gamma=1$, no-threshold distributed subfamily that contained the train-selected rule, evaluating $r \in \{-1, -0.5, -0.25, 0, 0.25, 0.5, 1\}$ on the 24-network test set. This analysis is deliberately non-confirmatory because it uses

the test data for selection. The most favorable observed ratio was $r=-0.5$, with mean $CVaR_{95}=2.889209$. Relative to the frozen independent candidate, the mean difference became -0.027204 (-0.93%), the median difference -0.01311 , and the 20% trimmed mean -0.01199 ; however, the network-bootstrap 95% interval remained wide, $[-0.2568, 0.1824]$, and only 16 of 24 networks favored this post-hoc distributed candidate.

The sensitivity check changes the interpretation but not the pre-specified decision. It shows that the nominal sign of the mean is selection-sensitive and that the original four-network training stage may have chosen a suboptimal distributed coefficient. At the same time, even a test-selected favorable ratio does not yield a precise disorder-level estimate. The defensible conclusion is therefore failure to establish a stable expected advantage, not evidence that no distributed advantage exists. This distinction is exactly why model selection and final evaluation should remain separated [23].

10. Claim Downgrade as a Scientific Outcome

The calibration and two audit cases show that the audit can retain, downgrade, or leave claims unresolved as the reducer family becomes stronger. Local-alignment flocking survives the specified independent-agent reduction, but its polarization is reproduced by a global-broadcast reducer, so the locality-specific irreducibility claim is downgraded. The anticipation claim collapses to a single predictive controller under its low-dimensional sufficient statistic. The mechanical reallocation case is different again: the train-selected comparison is statistically inconclusive across disorder, and post-hoc coefficient selection can reverse the nominal mean without resolving the uncertainty. The appropriate scientific outcome is therefore not a forced binary verdict, but the strongest reducer-relative statement the evidence actually supports.

Possible downgrade of the composite claim: Level V $\rightarrow$ Level IV $\rightarrow$ Level III $\rightarrow$ Level II

The arrows here represent possible claim downgrade, not logical implication. When a stronger null reproduces the target pattern or closes the performance gap, the appropriate action is to retain the weakest statement that still distinguishes the models rather than to discard all observations or redesign the task until the preferred claim returns.

11. Implications for Active Matter and Physical Learning

For active and smart matter, the framework suggests that adding memory, information processing, or learning should not automatically be described as adding a new collective capability. The relevant claim depends on whether the mechanism merely changes dynamics, expands a lower-class response set, provides positive value under explicit costs, or remains necessary after reduction.

For physical learning, the mechanical example highlights a distinction between a many-body substrate and a many-body adaptive computation. A global mechanical response may require a connected network while the reconfiguration rule remains decomposable into local elements. Conversely, the Vicsek calibration shows that an interaction mechanism can remain load-bearing for a macroscopic order parameter relative to a matched independent-agent family.

The same logic can be applied to biological interacting-unit models, but no biological irreducibility claim is tested here; extending the audit to cells or organisms would require domain-specific reducers and interventions.

12. Limitations

The architectural hierarchy is not unique. Different applications may require different axes, and the regulated/history-bearing distinction is operational rather than coordinate invariant.

No finite adversarial search proves that a response is inaccessible to every conceivable lower-class model. Level-III response-class expansion and Level-V irreducibility are both defeasible and always relative to an explicit architecture family, parameter domain, sensing access, resource budget, and optimization procedure.

The computational mechanical gate remains intentionally limited. The v0.7/v0.8 inference spans 24 held-out $N=36$ network realizations but still uses one system size, one compression-softening law, a static equilibrium compression proxy rather than dynamic memory, a compact training set, and a finite hand-specified candidate family. The disorder distribution is strongly topology-sensitive at this size, with a few large reversals dominating the mean; no finite-size self-averaging claim is made. The post-hoc ratio sweep further shows model-selection sensitivity. These limitations preclude a broad statement about distributed physical learning and make the correct result one of non-establishment rather than evidence of absence.

The first anticipation reduction is analytical rather than a faithful one-to-one reproduction of every detail of the earlier soft-particle assay. Its force is limited to the claim it targets: when the task collapses to a low-dimensional predictive sufficient statistic, many-body realization is not necessary for the headline future-conditioned preparation.

The audit is not a universal certificate of emergence or irreducibility. The Vicsek calibration makes this explicit: local alignment decisively beats an independent-persistent reducer for polarization, yet a global-broadcast reducer reaches the same $O(1)$ ordered regime with additional shared information. Every retained Level-III or Level-V claim must therefore name the lower or reducer family it survived.

13. Discussion

As interacting-matter models acquire increasingly sophisticated sensing, memory, learning, and control, it becomes easier to generate impressive macroscopic differences and harder to infer what those differences mean. The central scientific question shifts from whether mechanism X changes the system to what simpler architecture can reproduce the same result.

This perspective separates microscopic architecture from evidential status. Passive, active, regulated, and history-bearing units form one axis describing what mechanisms are available. Dynamical difference, response-class expansion, functional value, and collective irreducibility form another axis describing what has been demonstrated. Movement along the first axis does not guarantee movement along the second.

The methodological ancestry of this proposal is explicit. Null-model reasoning asks whether a proposed process is necessary for a pattern [15]; causal-emergence approaches compare causal efficacy across scales [14]; functional-reducibility work asks when richer interaction descriptions are dynamically necessary [16]; and information-bound approaches derive substrate-level limits on robust self-organization [18]. The present contribution is not a replacement for those theories. It is a claim-audit workflow for active and physical-learning studies: make the target claim cumulative and explicit, give the lower architecture its strongest justified resources, attempt reductions, evaluate on held-out structure as well as stochastic forcing, and stop when the decisive alternative closes the gap.

The computational counterexamples also suggest a practical norm. A weak baseline can make a model look novel; an expressive null can turn the same numerical effect into a mechanism-neutral result; and a small model-selection set can make even the sign of a held-out mean sensitive to the chosen candidate. Stronger collective claims therefore require both adversarial architecture comparison and explicit separation of model selection from final evaluation [23]. In disordered systems, distributional reporting is also essential when a few realizations dominate the mean [22].

The stopping rule is as important as the test. When the decisive lower-class alternative closes the gap, further task redesign may be scientifically justified for a new question, but it should not be presented as if the original hypothesis survived. Claim downgrade is itself a valid research outcome.

14. Conclusion

Increasing microscopic architectural capability does not by itself establish increasing collective capability. A richer model may alter collective dynamics without expanding the reachable response class; it may expand the response class without improving function; and it may improve function while remaining

reducible to a simpler controller or independently adapting elements. The strongest defensible claim is therefore the one that survives the strongest justified reducible alternative under matched resources and held-out evaluation.

Do not infer a new collective capability until the strongest reasonable reducible alternative has failed.

Data and Code Availability

The v0.6a single-network gate and the v0.7/v0.8 robustness and calibration packages contain the code and machine-readable results used in the computational examples. The package includes `multinetwork_gate_v07.py` / `.json` for the symmetric 24-network mechanical evaluation, `ratio_sensitivity_v08.py` / `.json` for the post-hoc coefficient-sensitivity check, `vicsek_positive_control.py` / `.json` for local alignment versus the independent-persistent reducer, and `vicsek_meanfield_rung_v08.py` / `.json` for the global-broadcast reducer. Fixed seeds, candidate families, bootstrap settings, and generated figures are included.

References

- [1] Marchetti, M. C. et al. Hydrodynamics of soft active matter. *Rev. Mod. Phys.* 85, 1143–1189 (2013). <https://doi.org/10.1103/RevModPhys.85.1143>.
- [2] Dauchot, O. From active to odd to smart matter. *Phys. Rev. E* 114, 011001 (2026). <https://doi.org/10.1103/tzyr-t47w>.
- [3] Löwen, H. & Liebchen, B. Towards Intelligent Active Particles. *arXiv:2501.08632* (2025).
- [4] Bebon, R. & Speck, T. Thermodynamics of active matter with internal degrees of freedom. *Commun. Phys.* 9, 162 (2026). <https://doi.org/10.1038/s42005-026-02662-z>.
- [5] Jung, G., Ozawa, M. & Bertin, E. Kinetic Theory of Decentralized Learning for Smart Active Matter. *Phys. Rev. Lett.* 134, 248302 (2025). <https://doi.org/10.1103/5m44-kwhv>.
- [6] Li, S. et al. Informational Memory Shapes Collective Behavior in Intelligent Swarms. *Phys. Rev. Lett.* 136, 138302 (2026). <https://doi.org/10.1103/rt97-ncmf>.
- [7] VanSaders, B., Fruchart, M. & Vitelli, V. Measurement-induced phase transitions in informational active matter. *PNAS Nexus* 5, pgag077 (2026). <https://doi.org/10.1093/pnasnexus/pgag077>.
- [8] Stern, M. & Murugan, A. Learning Without Neurons in Physical Systems. *Annu. Rev. Condens. Matter Phys.* 14, 417–441 (2023). <https://doi.org/10.1146/annurev-conmatphys-040821-113439>.
- [9] Keim, N. C., Paulsen, J. D., Zeravcic, Z., Sastry, S. & Nagel, S. R. Memory formation in matter. *Rev. Mod. Phys.* 91, 035002 (2019). <https://doi.org/10.1103/RevModPhys.91.035002>.
- [10] Pashine, N., Hexner, D., Liu, A. J. & Nagel, S. R. Directed aging, memory, and nature’s greed. *Sci. Adv.* 5, eaax4215 (2019). <https://doi.org/10.1126/sciadv.aax4215>.
- [11] Stern, M., Arinze, C., Perez, L., Palmer, S. E. & Murugan, A. Supervised learning through physical changes in a mechanical system. *Proc. Natl Acad. Sci. USA* 117, 14843–14850 (2020). <https://doi.org/10.1073/pnas.2000807117>.
- [12] Falk, M. J. et al. Learning to learn by using nonequilibrium training protocols for adaptable materials. *Proc. Natl Acad. Sci. USA* 120, e2219558120 (2023). <https://doi.org/10.1073/pnas.2219558120>.
- [13] Falk, M. J., Strupp, A. T., Scellier, B. & Murugan, A. Temporal Contrastive Learning through implicit non-equilibrium memory. *Nat. Commun.* 16, 2163 (2025). <https://doi.org/10.1038/s41467-025-57043-x>.
- [14] Hoel, E. P., Albantakis, L. & Tononi, G. Quantifying causal emergence shows that macro can beat micro. *Proc. Natl Acad. Sci. USA* 110, 19790–19795 (2013). <https://doi.org/10.1073/pnas.1314922110>.
- [15] Bausman, W. C. Modeling: Neutral, Null, and Baseline. *Philos. Sci.* 85, 594–616 (2018). <https://doi.org/10.1086/699021>.
- [16] Lucas, M. et al. Reducibility of higher-order networks from dynamics. *Nat. Commun.* 17, 1551 (2026). <https://doi.org/10.1038/s41467-025-68273-4>.
- [17] Raulin-Foissac, A. & Nicolas, A. Physics of anticipatory active matter, with application to crowd dynamics. *arXiv:2606.14818* (2026).
- [18] Romeo, N. et al. Information bounds the robustness of self-organized systems. *arXiv:2511.01682* (2025).
- [19] Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I. & Shochet, O. Novel Type of Phase Transition in a System of Self-Driven Particles. *Phys. Rev. Lett.* 75, 1226–1229 (1995). <https://doi.org/10.1103/PhysRevLett.75.1226>.

- [20] Ginelli, F. The Physics of the Vicsek model. *Eur. Phys. J. Spec. Top.* 225, 2099–2117 (2016).
<https://doi.org/10.1140/epjst/e2016-60066-8>.
- [21] Garcia Lorenzana, G. et al. When is nonreciprocity relevant? *arXiv:2509.17972* (2025).
- [22] Aharony, A. & Harris, A. B. Absence of Self-Averaging and Universal Fluctuations in Random Systems near Critical Points. *Phys. Rev. Lett.* 77, 3700–3703 (1996). <https://doi.org/10.1103/PhysRevLett.77.3700>.
- [23] Cawley, G. C. & Talbot, N. L. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* 11, 2079–2107 (2010).

End of manuscript v0.5