Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What Organises in Path-Decomposable Neural Network Architectures: On Structural Disposition, Conditional Realisation, and Causal Attribution of Computational Responsibility

Ambarish Moharil

a.moharil@tue.nl

Eindhoven University of Technology

Damian Tamburri

Jheronimus Academy of Data Science

Indika Kumara

Tilburg University

Research Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11721788/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

What Organises in Path-Decomposable Neural Network Architectures:

On Structural Disposition, Conditional Realisation, and Causal Attribution of Computational Responsibility

Ambarish Moharil[1,2,3*], Damian A. Tamburri[1,2] and Indika Kumara[3]

[1*]Safeguard Lab, Eindhoven University of Technology.
[2]Data Engineering Department, Jheronimus Academy of Data Science.
[3]Data Governance Department, Tilburg University.

*Corresponding author(s). E-mail(s): a.moharil@tue.nl;
Contributing authors: d.a.tamburri@tue.nl;
i.p.k.Weerasinghadewage@tilburguniversity.edu;

Abstract

Parametrisations of one neural network architecture can reach nearly the same error yet route their internal computation very differently. Functional error alone is therefore insufficient to show *what learning organises*. For this reason, we describe multilayer perceptrons (MLPs) and convolutional networks (CNNs) at three levels: their *structural disposition*, their *realisation* under a declared condition and their *re-realisation* after *causal* intervention. At each level, a symmetry-invariant path measure yields *intrinsic* observables, together with a *layer-conserved flow* and a corresponding *multiscale flow geometry. Keeping these explanatory layers apart locates each regularity where it holds and provides an intrinsic frame for explaining and interpreting neural networks*. We study 2,847 optimiser trajectories from Bayesian AutoML and 149,040 interventions on 324 networks. Learning, formalised as an *exact exponential tilt*, concentrates the path measure: median effective occupancy falls to 12.9% of its initial value in all 48 MLPs and to 13.0% in 46 of 48 CNNs. The architecture fixes the structural occupancy level, the environment modulates its magnitude and adaptive optimisers drive it down. Within matched MLPs, Adam raises the entropy deficit by 3.323 nats and SGD by only 0.062. Of Adam's rise, 96.7% follows the validation optimum, so *structural and functional time* decouple. The geometry records the same reorganisation, as MLPs contract their scale while CNNs internalise flow

within boxes. All these findings concern the disposition, which states only what is made available. A conditional realisation provably reweights this structural support at a given state but never extends it. On the MPI3D and Fashion-MNIST benchmarks, reference conditions realise a median of 0.68% and 7.82% of the structural occupancy respectively, and the activation class shapes which part is realised. Attenuating the realised carriers moves the response more than activation-matched controls in every network, although the activation law decides whether realised or structural carriers matter more. Learning therefore organises a *symmetry-faithful field of computational possibility*, of which each condition realises a narrower and load-bearing part.

Keywords: Neural network characterisation, eXplainable AI, computational paths, causal intervention, multiscale geometry

1 Introduction

The procedure of *learning* changes the internal organisation of a neural network. Yet the descriptions that are ordinarily used to characterise a trained network do not specify what that organisation is. This highlights the abstractness of current explainability methods. Much of the research on explainability relates the inputs of a network to its outputs (Guidotti et al., 2018). Furthermore, the analysis of the loss landscape shows that different parameter solutions can achieve a comparable loss (Garipov et al., 2018). Reparameterisations can also alter the parameters of a network without altering the function that it represents (Dinh et al., 2017). In our own measurements, two distinct optimisers, Adam and SGD with momentum, reach the same median best test error over 118 matched blocks of five architectures and five statistical learning environments (Section 4). However, the two optimisers differ by an order of magnitude in the internal reorganisation that we measure. Moreover, under Adam, most of this reorganisation accumulates after the validation optimum has been reached. The functional time and the structural time of a network can therefore separate. Consequently, what *learning* organises cannot be read from the function of a network alone. Answering this question therefore first requires an object of organisation that is faithful to the computation of the network and is able to describe the internal organisation.

A network of units and connections naturally suggests a graph-theoretic reading (Boccaletti et al., 2006; Newman, 2003). Within this tradition, complex-network studies have characterised organisational structure through statistics of node degree and through self-similarity under box covering approaches (Song et al., 2007, 2005). Neural networks have also been studied as graphs and as complex networks (La Malfa et al., 2022; You et al., 2020). However, for path-decomposable architectures, an analysis of degree remains trivial. The adjacency of such a network, and hence the degree of each unit, is fixed by its architecture before, during, and after training. Instead, the parameters carried by these fixed connections are the quantities that change. The natural next object is therefore the weighted strength of each node (Barrat et al., 2004). However, such strength statistics are not intrinsic to the computation. In a ReLU network, the incoming weights of a hidden unit can be scaled up and its outgoing weights can be scaled down without any change in the function of the network (Grigsby et al., 2023; Neyshabur et al., 2015). However, the strength of that unit changes. Moreover,

along a complete path from the input to the output, these rescalings cancel. *Complete computational paths therefore remain the more faithful object of organisation,* and this choice of object also exposes a broader gap in the research on explainability (Gonon et al., 2024). In many explanations, the explanandum itself remains implicit. One may ask why a prediction occurred, why a feature mattered, or why an internal unit activated. Attribution methods, for example, assign importance to features or units of a network with respect to an output (Bach et al., 2015; Sundararajan et al., 2017). However, their faithfulness to the underlying computation remains contested (Adebayo et al., 2018; Jacovi & Goldberg, 2020). Mechanism and circuit studies, in turn, intervene in internal components (Conmy et al., 2023; Geiger et al., 2021). These components are usually chosen by hypothesis or by searching within individual models. Moreover, the coordinates in which such explanations are stated are not always invariant under the symmetries of the parameters (Godfrey et al., 2022; Lim et al., 2024). What is missing, therefore, is an account that keeps these three objects apart. The first is the organisation that the trained parameters make available. The second is the part of that organisation that a particular input actually uses. The third is the consequence of changing that part.

In this work, we develop such an account for path-decomposable networks, namely MLPs and CNNs. We call the input-independent organisation of complete paths the *structural disposition* of a network. We call its condition-specific use the *conditional realisation* of that disposition. Our investigation is guided by three research questions.

- **RQ.1** What does *learning* organise among the computational routes that an architecture makes available, and how does that organisation vary with the architecture, the training environment, the optimiser and the activation?
- **RQ.2** How much, and which part, of this learnt organisation does an input use relative to a reference input?
- **RQ.3** Does the network's response depend on the part an input uses, and does the computation reorganise when only that part is weakened?

We pursue these questions on frozen populations of architectures, environments and optimisers since only a population separates what recurs from what depends on task or activation. Each stage then answers the limit of the one before. For RQ.1, we construct the structural measure, establish its invariance and follow it along training (Sections 3 and 4). A disposition, however, says nothing about what an input uses. For RQ.2, we therefore derive the exact finite response of a declared condition and compare its realisation with the disposition (Section 5). Since the decomposition is exact, any gap between them belongs to the computation. A realisation may still be only descriptive. For RQ.3, we attenuate realised and structural carriers with the parameters fixed, to ask which of them carries the response and is load-bearing (Section 6). An explanation can then name its object, a realised computation. It becomes contrastive because the network is held fixed while the condition changes (Miller, 2019). Its faithfulness can thus be tested on the computation itself.

2 Preliminaries

We review four lines of research and state what each leaves open.

Symmetry and path-space representations. Permutations and positive rescalings generate classes of parameters that realise the same function, so quantities defined on individual weights can change while the function is preserved (Dinh et al., 2017; Godfrey et al., 2022; Grigsby et al., 2023; Lim et al., 2024). Along a complete path, compensating rescalings cancel, a property that path-normalised optimisation and path norms use to regularise training and bound capacity (Gonon et al., 2024; Neyshabur et al., 2015). How learning and the optimiser redistribute mass over the paths of a fixed architecture has not been measured on such an invariant object (RQ.1).

Faithfulness and the explanatory object. An explanation is faithful when it reflects what the model computes, and this must be tested for each method (Jacovi & Goldberg, 2020). Tests have found saliency maps that barely change when the parameters are randomised (Adebayo et al., 2018), and reference-based attributions that change when the input is shifted although the output does not (Kindermans et al., 2019). Both tests examine maps from input to output. An input gradient adds up the contributions of all complete paths, and the rescalings above cancel along each path. The gradient is therefore unaffected by them, but for the same reason it cannot show how the computation is distributed inside the network. Likewise, convolution is equivariant to translation by construction (Cohen & Welling, 2016), yet trained CNNs are only approximately invariant to translation, rotation and scale (Lenc & Vedaldi, 2019), so a symmetry of the architecture need not hold for what the network has learnt. An explanation of internal organisation therefore needs an object unchanged by the network's symmetries, which raw activations, weights and gradients do not provide, together with a declared condition (RQ.1 and RQ.2).

Attribution and intervention. Relevance propagation and Integrated Gradients distribute a prediction over the input features, measured against a reference (Bach et al., 2015; Sundararajan et al., 2017). Causal abstraction tests whether internal components are responsible for a behaviour (Geiger et al., 2021), and circuit discovery finds the subgraphs that carry it (Conmy et al., 2023). Pruning instead removes parameters permanently (Frankle & Carbin, 2019). None of these approaches asks which part of the trained network a given input actually uses (RQ.2). Their interventions also fix targets in advance or alter the parameters, so weakening only the part an input uses, with the network otherwise unchanged, has not been tested (RQ.3).

Neural networks as multiscale systems. Trained networks have been studied as weighted graphs whose connections change during learning (La Malfa et al., 2022, 2021; Moharil, 2026; Testolin et al., 2020; You et al., 2020). Box covering measures self-similarity in complex networks by counting how many boxes of a given size cover them (Song et al., 2007, 2005), and Box Covering Analysis Weighted (BCANw) extends it to weighted networks (Wei et al., 2013). Network geometry shows that such scales mean something only relative to a declared distance and grouping rule (Boguñá et al., 2021). For neural networks, distance has come from the topology or the raw weights, and the weights change under the symmetries above. A distance derived from the computation, with scaling claimed only over a validated range of scales, is still missing

(Appendix A), which RQ.1, 2 and 3 investigate, within the three structural, conditional and intervention levels at which they are resolved.

3 Observables and Experimental Programme

This section defines the instruments with which the organisation of a trained network is measured. Each formal statement is followed by the inference it licenses, and the proofs are given in Appendix C.

3.1 Computational carriers in MLPs and CNNs

We write a trained network at training state t as $\mathfrak{N}_t = (\mathcal{N}, \theta_t, \Phi)$, where $\mathcal{N}$ is the architecture with its connectivity rules, θ_t is the learnt parameter state and Φ collects the operation laws that make the architecture executable. Following Section 2, the carrier is placed at the lowest relational level so that it survives the transformations declared to be redescriptions of the architecture. For the architectures studied here, this level is the complete computational path.

Parameter states. For an architecture a, a point $\theta \in \mathcal{W}_a$ is the complete trainable state, and run b has the state $\theta_{b,t}$ at checkpoint t. Training moves one trajectory through the parameter space of its architecture, and we retain t on every quantity derived from the state.

Multilayer perceptrons. The architecture is a finite layered directed graph $G = (V, E)$ with layers $V_0, \dots, V_L$ of widths $n_\ell = |V_\ell|$. A complete path $\pi = (i_0, \dots, i_L)$, with $i_\ell \in V_\ell$, has one directed edge between consecutive units. The family Π of these paths is the structural carrier, and a dense architecture has $N_{\text{topo}} = \prod_{\ell=0}^{L} n_\ell$ of them. At training state t, the state θ_t assigns weight $w_e(t)$ to each edge $e \in E$, fixing how prominently each available relation is represented.

Convolutional networks. Convolution separates a learnt parameter from its executions. Units are indexed by (ℓ, c, s), with channel c and spatial site s. A parameter identity $\vartheta = (\ell, c_{\text{out}}, c_{\text{in}}, \delta)$, with kernel offset δ, carries the value $\theta_\vartheta(t) = \theta^{(\ell)}_{c_{\text{out}}, c_{\text{in}}, \delta}(t)$ at training state t. One use of this coefficient at a destination site s, from $(\ell - 1, c_{\text{in}}, s + \delta)$ to $(\ell, c_{\text{out}}, s)$, is an execution occurrence a with $\rho(a) = \vartheta$. The map ρ is many-to-one, so all occurrences of ϑ carry the same value $\theta_\vartheta(t)$. The spatially expanded execution graph G_{exec} resolves the occurrences individually. A complete CNN carrier is a directed path $\gamma = ((c_0, s_0), \dots, (c_L, s_L))$ through G_{exec} with $s_{\ell-1} = s_\ell + \delta_\ell$, or the corresponding affine relation under stride, dilation or padding. The expansion adds no trainable degrees of freedom. It records repeated uses of shared coefficients, and these uses cease to be equivalent once their realised states differ under a condition (Section 5.3). The compact parameter multigraph is therefore a quotient of G_{exec} under ρ, not the primitive carrier.

What the carrier fixes. In both families, the carrier Π_a is fixed by the architecture and carries no training index, so runs of one architecture are compared as measures on a common carrier, and different architectures only through carrier-normalised summaries such as occupancy. Measures on different carriers differ in support and size, so only quantities normalised by the carrier remain comparable across architectures.

3.2 The structural measure and its invariance

For a complete MLP path π and a complete CNN path γ, the squared structural gain at training state t is

$$Q_t^S(\pi) = \prod_{e \in \pi} w_e(t)^2, \qquad Q_t^S(\gamma) = \prod_{\ell=1}^{L} \theta^{(\ell)}_{c_\ell, c_{\ell-1}, \delta_\ell}(t)^2, \tag{1}$$

so that the CNN score is the product of the squared kernel coefficients that γ traverses. From here on, π denotes a complete carrier of either family, and Π the carrier of the architecture under study. Squaring gives a non-negative score that is smooth in the weights and retains the cancellation produced by positive rescaling. Normalisation over the carrier gives

$$Z_S(t) = \sum_{\pi \in \Pi} Q_t^S(\pi), \qquad \mu_t^S(\pi) = \frac{Q_t^S(\pi)}{Z_S(t)}. \tag{2}$$

Definition 1 (Structural disposition) The structural disposition of run b at training state t is the normalised measure $\mu_{b,t}^S$ induced by $\theta_{b,t}$ on the architectural carrier $\Pi_{a(b)}$. For a single trajectory, we write μ_t^S.

The architecture supplies the possibilities, and θ_t distributes one unit of probability mass across them. The disposition is the probability object $\theta_{b,t} \mapsto \mu_{b,t}^S \in \mathcal{P}(\Pi_{a(b)})$, not the parameter state itself, and no input has entered it. Reading it as a property of the computation requires its invariance under function-preserving transformations.

Theorem 1 (Structural path-gain gauge invariance) *Consider a bias-free feed-forward ReLU MLP at a fixed training state t, and let v be any hidden unit. For some $a > 0$, apply the positive hidden-unit rescaling $w_{u\to v}(t) \mapsto a\, w_{u\to v}(t)$ and $w_{v\to r}(t) \mapsto a^{-1} w_{v\to r}(t)$ to every incoming and outgoing edge of v. Then $Q_t^S(\pi)$ is unchanged for every complete input-to-output path π. Consequently, the structural partition function $Z_S(t)$ (Gonon et al., 2024), the normalised structural measure μ_t^S and every quantity derived solely from that measure are unchanged.*

The invariance extends to any finite set of hidden-unit rescalings (Appendix Corollary 1). A permutation of hidden units also preserves the function, but it moves paths within the carrier instead of fixing them. Both transformations map carriers one to one and scale every score by a common constant, which the normalisation removes.

Theorem 2 (Carrier-preserving symmetry invariance) *Let a transformation map the carriers by a bijection T_g, with $Q'_t(T_g r) = K_g\, Q_t(r)$ for a constant $K_g > 0$ and every carrier r. Then $\mu'_t = (T_g)_{\#}\mu_t$, and every statistic invariant to relabelling the carriers is unchanged.*

Rescaling is the case in which T_g is the identity and $K_g = 1$, and a permutation relabels paths. Invariance is therefore claimed within a declared family of such transformations. A transformation that mixes paths, rather than relabelling them, violates the condition and need not preserve path probabilities, so the invariant must then be sought at a coarser level (Boundary 1).

Occupancy and concentration. The disposition is characterised through its Rényi

entropies (Rényi, 1961), its effective path number and its architecture-normalised occupancy,

$$H_q(t) = \frac{1}{1-q} \log \sum_{\pi} \mu_t^S(\pi)^q, \qquad N_{\text{eff}}^{(q)}(t) = e^{H_q(t)}, \qquad \eta_q(t) = \frac{N_{\text{eff}}^{(q)}(t)}{N_{\text{topo}}}, \tag{3}$$

with Shannon entropy $H_1(t)$ as the limit $q \to 1$. Increasing q weights the dominant paths, and normalising by N_{topo} makes path counts comparable across architectures. $N_{\text{eff}}^{(1)}(t)$ is the number of equiprobable paths with the same Shannon entropy as μ_t^S, and larger orders weight the dominant paths more strongly. Relative to the uniform measure $u_\pi = 1/N_{\text{topo}}$, the entropy deficit is

$$S_{\text{deficit}}(t) = \log N_{\text{topo}} - H_1(t) = D_{\text{KL}}\big(\mu_t^S \,\|\, u\big) = -\log \eta_1(t). \tag{4}$$

Occupancy and deficit are two exact descriptions of one concentration. They separate the multiplicity that the architecture supplies from the multiplicity that the measure occupies. Within the declared symmetry family, Theorem 1 ensures that their change between training states reflects reorganisation rather than gauge. Appendix B.1 gives an exact ensemble reading, with path log-weights $\omega_\pi(t) = \log Q_t^S(\pi)$.

3.3 Exact computation and conserved flow

The number of complete paths grows multiplicatively with depth and width, so the measure cannot be enumerated in networks of practical size. The multiplicative structure that makes the path a faithful carrier also makes sums over the carrier exactly computable.

Theorem 3 (Exact dynamic programming) *If $Q_t(\pi) = \prod_{e\in\pi} q_e(t)$ with $q_e(t) \geq 0$ on a finite layered acyclic graph, layerwise recursion computes $Z(t) = \sum_\pi Q_t(\pi)$ exactly, without enumerating paths.*

For the structural measure, $q_e(t) = w_e(t)^2$, or $q_a(t) = \theta_{\rho(a)}(t)^2$ on the occurrence edges of a CNN. The structural flows of an edge and of a node are the marginals $F_e(t) = \sum_{\pi \ni e} \mu_t^S(\pi)$, $\qquad F_v(t) = \sum_{\pi \ni v} \mu_t^S(\pi)$, so that $F_e(t)$ is the probability that a path sampled from μ_t^S traverses e. With the suffix scores $R_v(t)$, the edge flow factorises exactly as $F_e(t) = L_u(t)\, q_e(t)\, R_v(t)/Z(t)$. This is a combinatorial identity, and it assumes no independence between paths (Proposition 3).

Theorem 4 (Complete-carrier cut conservation) *Let μ_t be a normalised measure on complete carriers at training state t, with marginals $F_e(t) = \sum_{r \ni e} \mu_t(r)$. Let C be a set of elementary execution relations such that every carrier crosses exactly one member of C. Then $\sum_{e\in C} F_e(t) = 1$. Likewise, if every carrier visits exactly one node of a layer V_ℓ, then $\sum_{v\in V_\ell} F_v(t) = 1$.*

Every layer transition E_ℓ is such a cut. The conserved quantity is one unit of normalised carrier mass, not activation magnitude, gradient norm or energy. Training redistributes it within each layer but cannot change its total. The theorem holds for any normalised measure, including the conditional measures μ_t^c of Section 5. The enrichment $\mathcal{F}_e(t) =$

$|E_\ell|\, F_e(t)$ therefore has mean one over each layer transition (Proposition 4), which makes flow comparable across widths and architectures. In a CNN, the occurrence flow of a layer separates exactly by coefficient and by site.

Theorem 5 (Occurrence, identity and reuse decomposition) *Let A be an occurrence of a layer drawn with probability $F_a(t)$, and $\Theta = \rho(A)$. Then $H(A) = H(\Theta) + H(A \mid \Theta)$, and if each of the n_{param} identities has m sites, $S_{\text{occ}} = S_{\text{param}} + S_{\text{reuse}}$ with $S_{\text{param}} = \log n_{\text{param}} - H(\Theta)$ and $S_{\text{reuse}} = \log m - H(A \mid \Theta)$.*

3.4 The induced geometry as a measurement instrument

Geometry asks which form the flow takes across scales of observation. Each edge of the geometry-bearing graph receives the length $\lambda_e^{(p)}(t) = (\mathcal{F}_e(t) + \varepsilon)^{-p}$, with $\varepsilon > 0$ and $p = 1$ in the principal analysis. The length decreases strictly with enrichment (Proposition 8), so strongly used relations become short. The induced distance $d_F^{(p)}(u, v; t)$, the shortest accumulated length on the undirected support, defines a metric. The ruler is part of the observer, and every geometric statement is conditional on (p, ε). Box covering asks how the minimum number $N_B(\ell_B; t)$ of boxes of diameter below ℓ_B in this metric changes with scale (Song et al., 2007, 2005), so a box groups units joined by strong flow, and the weighted box counting analysis approximates it on weighted distances (Wei et al., 2013).

The characteristic scale $\xi_{50}(t)$, defined by $N_B(\xi_{50}(t); t)/N = 0.5$ for N vertices, separates the scale of the covering curve from its shape. On the coordinate $x = \ell_B/\xi_{50}(t)$, the transition-supported maximum $d^*(t)$ of the effective dimension $d_{\text{eff}}^F(x; t) = -d \log N_B / d \log x$, attained at $x_{d^*}(t)$, locates the sharpest compression. An exponent $d_B(t)$ is reported only within a credible scaling window (Appendix A.1). The peak flow lacunarity $\Lambda_F^*(t)$ measures how unevenly mass occupies the boxes (Allain & Cloitre, 1991). The directed inter-box retention is the fraction of directed flow that passes between distinct boxes of the quotient graph. We define further descriptors in Appendix C.10.

Theorem 6 (BCANw equivariance) *If a graph isomorphism φ preserves enrichment, $\mathcal{F}'_{\varphi(e)}(t') = \mathcal{F}_e(t)$, and both graphs use the ruler (p, ε), then φ is an isometry, and the BCANw covering functions and their descriptors coincide up to relabelling.*

A geometric difference measured with one ruler therefore arises from the flow. The ruler is declared rather than derived, so conclusions close to the measure are stronger than claims about any exponent.

3.5 Population construction by AutoML-based Bayesian acquisition

A regularity in one network may originate in its architecture, environment, optimiser or initialisation. We therefore release these sources one at a time, on populations that AutoML-based Bayesian acquisition builds for structural variation rather than performance (Appendix D.1). Each search starts from a stratified Sobol design (Sobol', 1967) and continues with random-forest surrogate acquisition in the manner of SMAC

(Hutter et al., 2011; Moharil et al., 2024).

Architectures. Per family, 600 Sobol architectures and 400 expected-improvement acquisitions ($\xi = 0.15$ for MLPs, 0.03 for CNNs) give 1,008 candidates scored by

$$\begin{aligned} J_{\mathrm{MLP}} &= \beta_1 N_{\mathrm{struct}} + \beta_2 N_{\mathrm{cfg}} + \beta_3 C_{\mathrm{cover}}, & (\beta_1, \beta_2, \beta_3) &= (0.60, 0.25, 0.15), \\ J_{\mathrm{CNN}} &= \beta_1' T[N_{\mathrm{phen}}] + \beta_2' T[N_{\mathrm{cfg}}] + \beta_3' C_{\mathrm{cover}}, & (\beta_1', \beta_2', \beta_3') &= (0.55, 0.25, 0.20), \end{aligned} \tag{5}$$

where N_{struct} is the novelty of a 17-dimensional initial fingerprint of path occupancy, flow heterogeneity and BCANw geometry, N_{phen} that of the structural change after a 60-epoch exposure, N_{cfg} configuration novelty, C_{cover} a reward for under-covered cells and T a robust transformation. Greedy maximin selection freezes 48 per family.

Environments. Each family explores 60 environments, each trained for 80 epochs on four sentinel architectures. Acquisition maximises $A_{\mathrm{env}} = \beta_1^e U + \beta_2^e N_{\mathrm{env}} + \beta_3^e \widehat{N}_{\mathrm{struct}} + \beta_4^e \widehat{I}_{A\times E}$, with $(\beta_1^e, \ldots, \beta_4^e) = (0.30, 0.25, 0.25, 0.20)$, rewarding surrogate uncertainty, environment novelty, predicted structural novelty and predicted sentinel disagreement. All acquisition weights are fixed in advance, and their optimisation remains an open question. In each family, 24 environments are frozen by coverage, structural novelty and disagreement.

Frozen populations. A 12-architecture handoff panel crossed with the 24 environments and two seeds gives 576 trajectories per family, and optimisers are compared in matched blocks. Realisation and intervention use 324 benchmark networks, three MLPs and three CNNs on each of MPI3D and Fashion-MNIST under three optimisers, three activations and three seeds. Every frozen member is retained (Appendix D.2).

4 Structural Disposition and Learning

We now ask what learning does to the structural disposition (RQ.1), using the architecture cohort, the environment matrix and the optimiser panel of Section 3.5 and their CNN counterparts. The changes in the observables on which the section rests are collected in Table E3 (Appendix E.1), and the values cited below are taken from it.

4.1 Concentration across architectures and environments

Learning concentrates the measure in every architecture. In the architecture cohort, the effective occupancy falls in all 48 architectures, with median folds of 0.129 for η_1 and 0.017 for η_2, and the entropy deficit rises in all 48 by a median of 2.051 nats. Edge-flow and node-flow heterogeneity also rise in all 48, with median folds of 1.385 and 1.944 (Table E4). The normalised deficit $S_{\mathrm{deficit}}/\log N_{\mathrm{topo}}$ rises from 0.173 to 0.201, so the effect is not an artefact of path count, and by Theorem 1 these changes reorganise the computation and not its parameter gauge. The concentration preserves the architecture ordering almost exactly, with a Spearman correlation of 0.990 between initial and final η_1 (Figure E2; Table E6). The function settles much earlier than the measure. The median best-test checkpoint is epoch 12, and from that checkpoint to epoch 400 the entropy deficit still rises in all 48 architectures, by a median of 2.037 nats. Within the cohort, depth orders this structural phenotype more strongly than width morphology (Table E7; Figure E4).

The direction persists across environments. In the environment matrix, η_1 falls and the entropy deficit rises in all 576 runs, and η_2 falls in all 528 numerically valid runs (Table E9). The median Spearman correlation of architecture ranks across the 276 environment pairs is 0.993 for $\log \eta_1$ (Table E11). Architecture accounts for a normalised structured share of 0.996 in the final $\log \eta_1$ and S_{deficit}, whereas environment accounts for 0.977 of the post-best deterioration of test error (Table E12). The environment nevertheless modulates the magnitude, since the environment-level correlation between signal-to-noise ratio and entropy-deficit growth is +0.827 (Table E13). Under Adam, the direction of concentration therefore recurs across these architectures and environments. The architecture fixes the level at which it acts, and the environment modulates how far it proceeds.

4.2 Motion of the measure under different optimisers

Having established the direction of change, we ask how the optimiser produces it, the second part of RQ.1. Two training states of one trajectory share the carrier, so the motion between them admits an exact description that holds for any optimiser.

Theorem 7 (Normalised-measure exponential tilt) *Let $Q_{t_0}(\pi) > 0$ and $Q_{t_1}(\pi) > 0$ be the scores of two training states on the same finite carrier Π, with normalised measures μ_{t_0} and μ_{t_1}. Define the finite log-growth field $G(\pi) = \log\big(Q_{t_1}(\pi)/Q_{t_0}(\pi)\big)$. Then*

$$\mu_{t_1}(\pi) = \frac{\mu_{t_0}(\pi)\, e^{G(\pi)}}{\mathbb{E}_{\mu_{t_0}}\left[e^{G}\right]}. \tag{6}$$

Theorem 8 (Finite entropy production under exponential selection) *Under the assumptions of Theorem 7, let $\mu_a(\pi) \propto \mu_{t_0}(\pi)\, e^{aG(\pi)}$ for $0 \le a \le 1$, and let $S_{\text{deficit}}(a) = \log|\Pi| - H(\mu_a)$. Then*

$$\frac{dS_{\text{deficit}}}{da} = \mathrm{Cov}_{\mu_a}\left(G, \log \mu_a\right), \qquad S_{\text{deficit}}(t_1) - S_{\text{deficit}}(t_0) = \int_0^1 \mathrm{Cov}_{\mu_a}\left(G, \log \mu_a\right) da. \tag{7}$$

Concentration therefore increases exactly when the growth field favours paths that are already prominent. The covariance factorises as $C_G = \rho_p(G, \log p)\, \sigma_G\, \sigma_{\log p}$, into an alignment ρ_p between growth and prominence and an intensity $\sigma_G^2 = \mathrm{Var}_p(G)$ of differential growth. The marginal-flow and replicator forms are given in Appendix C.8, and the identity holds numerically with a median absolute residual of 2.16×10^{-8} over 8,619 stored intervals (Appendix D.4).

The optimiser regime scales the concentration. The median entropy-deficit increase is 3.323 nats under Adam, 0.706 under RMSProp, 0.109 under SGD with momentum and 0.062 under SGD (Table E17; Friedman $\chi^2_F(3) = 319.25$, $W = 0.902$), and Adam exceeds every comparator in all 118 blocks (Holm-adjusted $p < 1.3 \times 10^{-20}$; Table E18). Optimiser identity accounts for 93.5% of the variation in entropy-deficit change, whereas the optimiser accounts for only 0.4% of the variation in best-test log error (Table E19). Adam and SGD with momentum reach the same median best test error of 0.212 while ending in markedly different path measures, so functional equivalence does not imply structural equivalence.

Structural time outlasts functional time. The median best-validation epoch is 20 for Adam, 40 for RMSProp, 120 for SGD with momentum and 399 for SGD (Figure E18). After its optimum, Adam adds a median entropy deficit of 3.176 nats, which is 96.7% of its total increase. The median post-optimum deterioration of test error follows the same order (0.1129, 0.0816, 0.0236 and 0.0033), but the co-occurrence is temporal and not causal, since within the Adam arm, the environment still explains about 94% of that deterioration.

Selection intensity separates the optimisers. The alignment $\rho_p(G, \log p)$ follows the optimiser ordering but remains modest in every case. On the interval from epoch 80 to 120, by contrast, the median $\mathrm{Var}_p(G)$ is about 3.81 under Adam and 0.014 under SGD (Figure E22). A small positive alignment acting on a broad growth field therefore produces substantial concentration. Block-fixed-effect trajectory models confirm optimiser-specific evolution of both components (Wald $\chi^2 = 422.1$ for alignment and 8548.3 for $\log \mathrm{Var}_p(G)$; Table E21).

Convolutional networks. In the CNN architecture panel, 46 of 48 architectures concentrate, with a median entropy-deficit increase of 2.039 nats (Holm-adjusted $p = 1.78 \times 10^{-13}$) carried almost entirely by the parameter-identity term, since S_{param} rises by 0.682 of a median rise of 0.685 in S_{occ} while S_{reuse} rises by 0.009, and the deficit moves in the same direction in 258 of 288 architecture-environment cells. The optimiser ordering changes, however. At epoch 360, the median deficit is highest under RMSProp (3.945) and lowest under SGD with momentum (2.746), with $W = 0.373$ reached through crossings (Appendix E.5). Since the comparators are calibrated on training loss alone (Appendix D.2), these results establish optimiser-regime specificity without separating the algorithm from its learning rate.

4.3 Structural geometry without a universal exponent

Concentration is a statement about the measure. We now ask what form the corresponding flow takes across scales, using the instrument of Section 3.4.

The metric contracts around a backbone that deforms late. The characteristic scale contracts in all 48 architectures (median ξ_{50} fold 0.934; Figure E7) and in all 576 runs, preserves the architecture ordering ($\rho_s = 0.976$) and follows the optimiser ordering of concentration ($W = 0.809$). After normalisation by ξ_{50}, the covering relations share a principal fine-to-coarse transition, and the deformation appears beyond the crossover (Figure E7). The crossover height d^* decreases in 36 of 48 architectures and in 454 of 576 runs, with a median within-architecture correlation $\rho_s(S_{\mathrm{deficit}}, d^*) = -0.656$. From the best-test checkpoint to the final one, peak lacunarity increases and inter-box retention decreases in all 48 architectures, so the geometry keeps reorganising after the function has reached its best held-out regime.

Scaling is finite and the exponent is not universal. Every architecture and every run has an accepted finite scaling window at the final snapshot, but the exponent within it ranges from 0.266 to 2.047 in the cohort and learning does not preserve its ordering (Figure E2, $\rho_s = -0.124$). The optimiser hardly orders it ($W = 0.028$; Wald $p = 0.168$), and pooled tails of the microscopic flow prefer a lognormal to a pure power law in 94 of 96 combined MLP and CNN architectures (Figure E11). The exponent

therefore remains conditional on a finite window, while the measure-level laws and the contraction recur across populations.

5 Conditional Realisation

Section 4 establishes how learning organises the structural disposition μ_t^S, into which no input has entered. We now ask what a declared condition realises from it (RQ.2). We keep two kinds of populations apart throughout. The controlled populations are trained on the synthetic teacher tasks of Section 3.5 and establish which structural laws survive the change of measure (Section 5.2). The benchmark case studies train one factorial of 162 networks on each of MPI3D and Fashion-MNIST and test the same relation on real image data (Section 5.3).

5.1 The finite condition and support domination

A condition $c = (x, x^\circ)$ declares an input x and a reference x°, and asks how a network at training state t transmits the difference between the two executions. This transmission is exact at every operation.

Theorem 9 (Exact finite-response identity) *Let $B : \mathbb{R}^m \to \mathbb{R}^n$ be continuously differentiable on the segment joining u° and u, with Jacobian J_B, and let $R_B(u, u^\circ) = \int_0^1 J_B\big(u^\circ + \alpha(u - u^\circ)\big)\, d\alpha$. Then $B(u) - B(u^\circ) = R_B(u, u^\circ)\,(u - u^\circ)$.*

For a coordinatewise activation ϕ, the operator is the secant $s_\phi(z, z^\circ) = \big(\phi(z) - \phi(z^\circ)\big)/(z - z^\circ)$, which tends to the Jacobian as $u^\circ \to u$ (Proposition 5); for ReLU it lies in $(0, 1)$ whenever the contrast crosses the kink. Substituting the identity at every operation and expanding the product of linear maps decomposes the response exactly over paths.

Theorem 10 (Exact conditioned response decomposition) *If every operation of a finite acyclic network is replaced by its finite-response operator under $c = (x, x^\circ)$, then $f_{i_L}(x) - f_{i_L}(x^\circ) = \sum_\pi A^c_{\pi,t}$, where $A^c_{\pi,t}$ is the signed product of the input contrast, the weights and the finite-response coefficients along π.*

MLP paths and spatially unrolled CNN paths are both instances (Corollary 2). The squared amplitude factorises over the structural score,

$$Q_t^c(\pi) = \big(A^c_{\pi,t}\big)^2 = Q_t^S(\pi)\, R_{c,t}(\pi), \qquad \mu_t^c(\pi) = Q_t^c(\pi)/Z_c(t), \tag{8}$$

where $R_{c,t}(\pi) \geq 0$ collects the squared input contrast and secants. A condition with $Z_c(t) = 0$ is flagged and is not assigned a value (Appendix D.4).

Proposition 1 (Support domination) If $Q_t^c(\pi) = Q_t^S(\pi)\, R_{c,t}(\pi)$ with finite $R_{c,t}(\pi) \geq 0$ on a common finite carrier, then $\operatorname{supp} \mu_t^c \subseteq \operatorname{supp} \mu_t^S$, equivalently $\mu_t^c \ll \mu_t^S$.

A condition can select among the paths that the disposition makes available, but it cannot realise a path that the disposition excludes. When $R_{c,t}$ is positive on the whole structural support and varies across it, the mass is redistributed without any change of *carrier* support (Corollary 3), as expected for smooth activations. Since both measures share the carrier and the parameter state, $\Delta_t^{\mathrm{S}\to\mathrm{C}} Y = Y(\mu_t^c) - Y(\mu_t^S)$ is a change of

Table 1: Structural (S) to conditional (C) change on synthetic tasks for MLPs (a) and CNNs (b). Δ, median change; *, same direction in all 24 environments; W_E, environment concordance of the conditional value (regression/classification); ρ_s, architecture rank correlation; W, Kendall's concordance.

(a) MLP environment matrix, architecture cohort and optimiser panel

	Regression			Classification			Env.	Arch.	Opt.
Observable	S	C	Δ	S	C	Δ	W_E	ρ_s	W_C
S_{deficit} ↑	3.030	6.115	+3.140*	2.963	5.668	+2.766*	0.548/0.687	0.965	0.706
Edge-flow CV ↑	1.751	4.227	+2.483*	1.714	3.914	+2.219*	0.346/0.538	0.689	0.707
Node-flow CV ↑	0.405	1.232	+0.828*	0.394	1.173	+0.777*	0.467/0.646	0.605	0.694

(b) CNN architecture, environment and optimiser panels

	Architecture panel				Env.	Arch.	Optimiser W		
Observable	S	C	Δ	r_{rb}	Δ	ρ_s	S	C	Δ
S_{deficit} ↑	3.082	6.003	+2.794	+0.998	+2.903*	0.909	0.037	0.585	0.568
S_{occ} ↑	1.251	2.944	+1.634	+1	+1.660*	0.469	0.019	0.534	0.612
S_{param} ↓	1.161	0.915	−0.136	−0.558	−0.186	0.031	0.016	0.393	0.388
S_{reuse} ↑	0.107	1.915	+1.809	+1	+1.916*	0.406	0.240	0.631	0.648

measure within one trained network, and a fall of η_1 under conditioning is an empirical finding rather than a consequence of the proposition. In CNNs, the chain rule gives $S_{\text{occ}} = S_{\text{param}} + S_{\text{reuse}}$ for the conditional measure as for the structural one (Theorem 5; Appendix C.10), and because the flow-to-length map is nonlinear, aggregating flow before the metric is built generally changes the geometry (Proposition 7). Conditions are nested within a network, so intrinsic observables are summarised per network and population geometry uses the condition-averaged flow.

5.2 Realisation in the controlled populations

The controlled populations reuse the synthetic generators of Section 4, with an MLP cohort of 20 architectures, an activation panel of 960 runs, an environment matrix of 576 runs with paired regression and classification tasks, and an optimiser panel of 288 runs, and with 48 CNNs, 576 CNN environment runs and 30 matched CNN optimiser blocks. Table 1 gives the structural-to-conditional changes, and the full results are in Appendices E.7 and E.8.

Conditioning concentrates the disposition and keeps its architecture order. In the CNN architecture panel, conditioning raises the deficit of the same trained state by a median 2.794 nats ($r_{rb} = 0.998$ over 48 networks), and η_1 falls from 0.046 to 0.002 (Table E30). In the MLP environment matrix, the deficit rises by 3.140 nats under regression and 2.766 under classification, and all 24 environments agree in direction for the deficit, both occupancies and both flow heterogeneities (Table 1a). The architecture order of concentration survives the change of measure, with rank correlation 0.965 in the conditional MLP cohort of 20 architectures under Adam and 0.909 across the CNN environments (Table 1). Over training, that cohort's conditional deficit grows by 8.784 nats against 2.441 for the structural one, in all 20 architectures, so the condition amplifies the concentration that learning produces (Appendix E.7). Environment

keeps its effect (W_E in Table 1a) but loses its order, since no core structural-to-conditional environment-rank correlation survives the Holm family. Realisation also separates the tasks more than the weights do. Under ReLU, the median Hellinger distance between matched regression-trained and classification-trained measures rises from 0.375 to 0.756, and rises in all 24 environments (Appendix E.7).

Activation and optimiser act on different factors of the realised measure. In $Q_t^c = Q_t^S R_{c,t}$, the optimiser tilts the disposition through its growth field (Theorem 7). The activation filters it under a condition through $R_{c,t}$. The activation acts by class. Across 160 matched blocks, the rank of six activations is nearly fixed for hidden input specialisation ($W = 0.996$). It is strong for the conditional deficit ($W = 0.799$) and weak for ξ_{50} ($W = 0.173$; Tables E24 and E26). Piecewise-linear gates realise the largest deficits (5.617 and 5.305 nats), because a zero secant removes paths. Smooth gates follow (4.280 to 4.383), and identity is lowest (4.172). The optimiser acts by mechanism. Adaptive scaling broadens the growth field ($\mathrm{Var}_p(G) \approx 3.81$ under Adam, 0.014 under SGD). Momentum alone barely tilts the measure (0.109 against 0.062 nats; Section 4.2). A broad tilt and a gating filter therefore compound towards concentration. Yet the two factors share an origin. The secants depend on preactivations set by the learnt weights, so the optimiser also enters $R_{c,t}$. Its concordance therefore rises from 0.282 for the structural MLP deficit to 0.706 for the conditional one, and RMSProp overtakes Adam (7.414 against 5.575; Table E27).

In CNNs, realisation selects where shared parameters are used. Structural learning raises S_{occ} almost entirely through S_{param}, by 0.682 of a median rise of 0.685 (Section 4.2), and conditioning reverses this balance. In the architecture panel, S_{occ} rises by 1.634 through S_{reuse} ($+1.809$, $r_{rb} = 1$), while S_{param} falls (-0.136); across environments, S_{reuse} rises in all 24 and S_{param} falls in 23 (Tables E30 and E31). The median rank retention of parameter identities is 0.803, against 0.024 for occurrences, so the learnt coefficients remain recognisable while the spatial sites at which they carry flow are reordered.

Geometry follows the carrier. Under conditioning, the MLP cohort medians of d^* fall from 1.751 to 0.236, peak lacunarity rises from 99.2 to 802.0 and retention falls from 0.541 to 0.066. In the CNN environment panel, d^* also falls in all 24 environments, but lacunarity falls and retention rises (Tables E31 and E22), so the direction of the texture change depends on the carrier. A credible scaling window, present in all 160 structural states of the MLP cohort, survives in only 48 conditional ones (Table E25).

5.3 Benchmark case studies on MPI3D and Fashion-MNIST

Having established the relation on synthetic tasks, we test it on real images. MPI3D is posed as regression of the horizontal position of an object (40 levels), with its other generative factors as nuisance, and Fashion-MNIST as ten-class classification, both at 16×16 resolution. Three MLPs and three CNNs are crossed with SGD with momentum, RMSProp and Adam, with ReLU, tanh and SiLU, and with three seeds, giving 162 networks per benchmark. Each network is observed with fixed parameters in its structural state, under clean conditions and under a deterministic Gaussian sensor shift at 15 dB, with the reference $x^\circ = 0$. Intrinsic observables use 160 shared

Table 2: Structural (S): input-free μ^S; clean input-conditional (C) and 15 dB shifted (Sh): realisations with $x^\circ = 0$. 162 frozen networks per benchmark; W, Kendall's concordance.

		MPI3D (regression)			Fashion-MNIST (classification)		
Networks	Observable	S	C	Sh	S	C	Sh
All	$S_{\text{deficit}}/\log N \uparrow$	0.168	0.384	0.372	0.154	0.255	0.257
	Activation W	0.498	0.849	0.838	0.700	0.931	0.916
MLP	Arch. W	0.115	0.285	0.235	0.531	0.236	0.202
	Local Δ^F	−0.007	−0.248	−0.207	+0.100	+0.043	+0.044
CNN	Arch. W	0.553	0.759	0.759	0.790	0.931	0.931
	$S_{\text{occ}} \uparrow$	1.245	3.134	2.902	1.289	2.376	2.411
	S_{param}	0.937	0.913	0.877	1.047	1.188	1.161
	$S_{\text{reuse}} \uparrow$	0.139	2.210	1.974	0.131	1.049	1.091
	Local Δ^F	−0.049	+0.341	+0.099	−0.108	−0.079	−0.072

endpoint events and geometry 30 preregistered networks per state (Appendix E.9). All results cited below are tabulated in Appendix E.12.

The condition selects a small part of the structural support on both tasks. Under clean conditions, the median η_1 is 0.68% of its structural value on MPI3D and 7.82% on Fashion-MNIST, and every network moves in this direction (Tables E34 and E35). The sensor shift then acts with a task-dependent sign. On MPI3D it raises η_1 ($r_{rb} = +0.910$) and lowers the normalised deficit ($r_{rb} = -0.927$), spreading realised mass over more of the structural support, whereas on Fashion-MNIST it lowers η_1 ($r_{rb} = -0.986$) and raises the deficit ($r_{rb} = +0.975$). What recurs is the sensitivity of the realised measure to the condition at fixed parameters. Conditioning also strengthens the activation ordering on both tasks (from 0.498 to 0.849 and from 0.700 to 0.931; Table E36), as expected from the secants of Theorem 9, whereas the optimiser concordance stays small.

Geometry shares the form of the response but not its scale. From the structural to the clean state, d^*, the transition fraction and retention fall on both benchmarks, while peak lacunarity rises ($r_{rb} = +1.000$ and $+0.535$). The characteristic scale rises on MPI3D ($r_{rb} = +0.677$) and falls on Fashion-MNIST ($r_{rb} = -0.948$; Table E35). Under the shift, MPI3D deforms broadly, whereas on Fashion-MNIST only ξ_{50} changes after correction, and the local mismatch of each family changes sign between the benchmarks (Table E37). In CNNs, the reuse deficit rises from 0.139 to 2.210 on MPI3D and from 0.131 to 1.049 on Fashion-MNIST, while S_{param} changes by less than 0.15 (Table E37), so the reuse mechanism of the controlled populations recurs on real data.

Realised occupancy is not a universal performance surrogate. Beyond the design factors, clean $\log \eta_1$ explains a further 0.106 of log test error on MPI3D (cluster-robust $p = 7.1 \times 10^{-30}$) but only 0.005 on Fashion-MNIST ($p = 0.10$; Table E38). Concentration under a condition and the reuse mechanism of CNNs thus recur across populations, whereas the response to the shift and the informativeness of occupancy depend on the task. Equivariance of the realised geometry is examined with the interventions in the next section.

6 Intervention and Responsibility

Section 5 shows that a condition realises a small, selective part of the structural disposition. We now ask whether that part is load-bearing (RQ.3): does weakening the computation that carries realised flow change the response more than weakening other active computation? We intervene on the 324 frozen benchmark networks of Section 5.3 in 149,040 controlled interventions, followed by 2,592 post-intervention geometries and 32,400 equivariance comparisons (Appendix E.13).

6.1 Execution-only attenuation

The intervention acts on the carrier of realised flow, an execution edge in an MLP and a spatial occurrence a in a CNN. During inference, the contribution u of each selected carrier becomes $g\,u$, with $g = 1 - \alpha$ and $0 \le \alpha < 1$, in both executions of $c = (x, x^\circ)$, and all downstream states are recomputed; parameters and architecture are unchanged. The manifestation displacement $D_{\mathrm{man}} = \left|\widetilde{\Delta f_{i_L}} - \Delta f_{i_L}\right|$ measures how far the declared response moves, and $D_H(\mu_t^c, \widetilde{\mu}_{t,\alpha}^c)$ how far the measure is re-realised.

Proposition 2 (No creation of support; zero-intervention identity) (i) If the pre-intervention and post-intervention scores factor through the same structural score, $Q_t^c = Q_t^S R_{c,t}$ and $\widetilde{Q}_{t,\alpha}^c = Q_t^S \widetilde{R}_{c,t,\alpha}$ with finite non-negative factors, then $\operatorname{supp} \widetilde{\mu}_{t,\alpha}^c \subseteq \operatorname{supp} \mu_t^S$. (ii) For $\alpha = 0$, $\widetilde{\Delta f_{i_L}} = \Delta f_{i_L}$ and $\widetilde{\mu}_{t,0}^c = \mu_t^c$ whenever the measure is defined.

Attenuation thus re-realises the condition within structural possibility, though not necessarily within the previous conditional support, since weakening a negative input can activate a silent ReLU unit (Proposition 11). A vanishing partition function would leave the measure undefined (Proposition 12); it occurred in none of the interventions. The operator suite passed 34 of 34 tests (maximum error 7.08×10^{-9}), and the zero-dose identity and checkpoint immutability were verified on trained networks. At equal budget and dose, a carrier set is more responsible for the response under c when its attenuation displaces that response more.

6.2 Load-bearing realised flow

Realised flow locates load-bearing computation. Under 40 conditions per network and environment, three equal-budget layerwise sets are attenuated with $\alpha = 0.25$: the realised set R, the top tenth of positive realised support; the structural set S, selected from μ_t^S; and the matched set M, drawn from the remaining realised support with matched source activation. Attenuating R displaces the response more than attenuating M in every network of all four populations (rank-biserial 1, Holm-adjusted $p \approx 7.4 \times 10^{-28}$; Table 3a). R also exceeds S in most networks, but this comparison is organised by activation: the fraction of networks with R above S averages 0.940 under ReLU and 0.968 under SiLU but 0.148 under tanh, whose structural set is the more consequential in every dataset, environment and family. Realised flow thus identifies load-bearing computation, while the relative weight of disposition and realisation depends on the activation regime.

Re-realisation is ordered more uniformly than the output. Attenuating R or S broadens the surviving measure, whereas attenuating M sharpens it, and the Hellinger displacement under R exceeds that under S in 88.9% to 99.4% of networks. When

Table 3: Intervention on the frozen benchmark networks. R, S and M are the realised, structural and activation-matched carrier sets; entries are medians over networks, with the fraction of networks positive in brackets; * marks Holm significance in (d).

	Fashion-MNIST		MPI3D	
Observable	Clean	Shifted	Clean	Shifted
(a) Load-bearing realised flow ($\alpha = 0.25$; 162 networks per benchmark)				
D_{man}, R−M	2.730 (1.00)	2.548 (1.00)	0.394 (1.00)	0.267 (1.00)
D_{man}, R−S	1.295 (0.68)	1.142 (0.68)	0.123 (0.74)	0.105 (0.65)
D_{man}, R−S, tanh MLPs	−0.348	−0.388	−0.028	−0.034
D_{man}, R−S, tanh CNNs	−0.324	−0.314	−0.014	−0.078
(b) Dose response of the realised set ($\alpha = 0.10, 0.25, 0.50$)				
D_{man}	1.85, 3.83, 5.90	1.81, 3.61, 5.30	0.20, 0.48, 0.78	0.14, 0.31, 0.49
D_H	0.23, 0.53, 0.84	0.22, 0.53, 0.84	0.28, 0.64, 0.92	0.29, 0.64, 0.91
Strictly increasing D_{man}	0.988	0.969	0.975	0.901
(c) Geometry change after realised attenuation (Holm-significant unless n.s.)				
Δd^*	+0.254	+0.264	+0.060	+0.104
$\Delta \Lambda_F^*$	−125.3	−124.5	−296.4	−289.0
Δ inter-box retention	+0.125	+0.115	+0.042	+0.073
Δd_B^F	+0.219	+0.221	+0.070	+0.089
$\Delta \log \xi_{50}$	+0.148	+0.165	−0.201	−0.040 (n.s.)
Of 10: shifted, R>M, R>S	7, 9, 5	7, 10, 5	7, 10, 4	7, 8, 5
(d) Occurrence minus shared-parameter targeting (matched budget, 81 CNNs)				
ΔD_H	0.037* (0.83)	0.037* (0.86)	0.016 (0.59)	0.024 (0.62)
ΔD_{man}	0.232 (0.61)	0.137 (0.56)	−0.019 (0.43)	0.022* (0.57)

clean-preferential and shifted-preferential sets are crossed with the two conditions, each set reorganises the measure of its own condition more in every network, while the output crossover is weaker and vanishes or reverses under tanh (Appendix E.15). Responsibility also differs from task value: in shifted MPI3D CNNs, attenuating R moves the response most and yet lowers the median task loss.

Responsibility is graded with dose. With each set frozen across dose, raising α from 0.10 to 0.50 increases both outcomes in every population (Table 3b), and every Hellinger increment is positive in all networks. The curve decelerates in a median of 96.3% of ReLU and 88.9% of SiLU networks, whereas 59.3% of tanh networks accelerate. The advantage of R over M widens with dose, while that over S peaks at the middle dose as structural attenuation catches up (Appendix E.16).

6.3 Re-realised geometry and occurrence targeting

Re-realisation has a geometric phenotype. The flow averaged over ten conditions is measured with the fixed instrument of Section 3.4 before and after each attenuation; by Theorem 6, differences under a common ruler arise from the flow. Seven of ten descriptors shift in every population (Table 3c). After attenuating R, the flow covers more steeply (d^* and d_B^F rise), fills boxes more evenly (Λ_F^* falls) and keeps more directed flow between boxes, while ξ_{50} moves in task-dependent directions. The realised change exceeds the matched change on 8 to 10 descriptors but the structural change on only 4 or 5. Tanh networks reorganise as well, with lower lacunarity and higher retention in all eight tanh cells, although their output-level ordering is reversed.

Table 4: Equivariance advantage $A_{\text{eq}} = H_{\text{inv}} - H_{\text{eq}}$ of the realised input-origin field in CNNs: median over networks, with the fraction of networks positive in brackets.

Dataset	Transformation	Clean	Shifted
Fashion-MNIST	Translation, 1 or 2 pixels	0.154 (1.00)	0.191 (1.00)
	Rotation, $\pm 10^\circ$ or $\pm 20^\circ$	0.104 (1.00)	0.133 (1.00)
	Scaling by 0.90	0.038 (1.00)	0.055 (1.00)
	Scaling by 1.10	−0.006 (0.35)	0.009 (0.93)
MPI3D	Vertical translation, 1 or 2 pixels	0.507 (1.00)	0.393 (1.00)

In CNNs, the spatial use of a shared weight matters. Occurrence targeting ranks spatial occurrences by F_a; shared-parameter targeting ranks identities by $F_\vartheta = \sum_{a:\rho(a)=\vartheta} F_a$ and attenuates every reuse of a selected identity, with the budget matched per layer in primitive occurrences. On Fashion-MNIST, occurrence targeting moves the measure further ($p_{\text{Holm}} \leq 4.4 \times 10^{-7}$; Table 3d) while touching 135 to 158 of the 256 first-layer input positions, against 250 to 253; on MPI3D the difference is not significant. The output effect follows activation, +0.589 under ReLU and −0.493 under tanh on clean Fashion-MNIST, so the location of the computation does not fix the size of the output change. Quotienting occurrences to identities discards information relevant to intervention, consistently with the concentration of S_{reuse} under conditioning (Section 5.2).

Equivariance of the realised field is approximate. The input-origin field $P_x(s)$, the marginal of μ_t^c over input positions, is recomputed for a transformed input $T_g x$ and compared with the transported field $(T_g)_\# P_x$ and the untransformed P_x through Hellinger errors H_{eq} and H_{inv}. Translation, rotation and shrinking favour transport in every network, whereas enlargement does not on clean Fashion-MNIST (Table 4). A positive advantage is not exact equivariance, which requires $H_{\text{eq}} = 0$: residuals remain and the worst networks depart visibly from transport (Appendix E.18). When the input moves, the dominant realised mass moves with it. Under shift, transport strengthens on Fashion-MNIST, with a median change of 0.058 for translation, and weakens on MPI3D by 0.127. The realised computation therefore follows the geometry of the input, but how closely it does so is learnt and depends on the data.

7 Discussion

Answers to the research questions. For RQ.1, learning concentrates the measure in all 48 MLPs of the cohort, all 576 environment-matrix runs and 46 of 48 CNNs, with an optimiser- and activation-dependent magnitude. For RQ.2, a condition realises a narrower part of the disposition without extending its support (Proposition 1). For RQ.3, the realised part is load-bearing in every network, while the activation law decides whether realised or structural carriers matter more. These answers imply that an explanation of a trained *network must name the disposition it draws on and the condition under which it is realised*, and must also test that realisation by intervention.

Controlled and real data. The laws found under the designed variation survive on real data. Conditioning concentrates the measure in every controlled environment of both families and in every benchmark network, and the CNN move of specialisation to

spatial reuse recurs on both benchmarks (Section 5.3). The response to a change of condition is task-specific: the sensor shift diffuses the measure on MPI3D and concentrates it on Fashion-MNIST, and conditioning raises peak lacunarity on the benchmarks but lowers it in controlled CNNs. The spatial-reuse result implies that explanations and interventions in CNNs must be resolved at spatial occurrences, since a shared parameter does not locate the computation it serves. Varying kernel size, stride or pooling could then compare CNNs by how a condition selects among their occurrences.

Graded concentration and compression. Because each layer conserves one unit of flow, no enrichment can exceed its layer's edge count, $\mathcal{F}_e \leq |E_\ell|$, and any power law must therefore end in a cutoff. Learning, moreover, reweights paths multiplicatively through an exact tilt (Theorem 7), which favours a lognormal. Correspondingly, the tails of the edge enrichment $\mathcal{F}_e$ are lognormal when pooled and most often truncated within layers (Appendix A.2). Concentration is thus graded, and enrichment falls off gradually without splitting the edges into a few carrying the flow and many carrying almost none. The data therefore give no threshold for compression, so a subnetwork must be defined by the share of flow it keeps. Since intervention shows that the response depends on the realised part, whether such a subnetwork retains the function is the next question. Lottery tickets instead select by weight magnitude (Frankle & Carbin, 2019), which changes under function-preserving rescalings.

The loss & path measure. Adam and SGD with momentum reach almost the same median best test error on the 118 matched MLP blocks of the synthetic environment panel. Yet their entropy-deficit increases are 3.323 and 0.109 nats. The functional objective therefore constrains what the network computes and leaves open how the computation is distributed over its paths, and the optimiser selects within this freedom. A structural criterion, for training or compression, must be faithful to the computation, as the entropy deficit is and weight statistics are not. An optimiser informed by the shape of the measure, and not only by its total Z_S as in path-normalised optimisation, is the natural next investigation. Two CNN architectures do not concentrate, so it should also ask whether *good* function requires concentration. Under Adam, 96.7% of the deficit increase accrues after the validation optimum. After the best-test checkpoint, peak lacunarity rises and retention falls in all 48 MLP architectures. *Function and structure are found to settle on different clocks.* Between the functional optimum and most of the concentration lies a window to stop training with the function retained. Whether stopping at that point helps generalisation remains an open question, since deterioration and concentration co-occur without proven causation (Appendix F.5).

8 Conclusion

We asked what organises in path-decomposable neural networks. Our answer is *computational possibility* ($\mu_t^{S/c}(\pi)$). The architecture supplies a field of complete paths before training, and *learning differentiates this field rather than creating it.* The first gap was the lack of a measure that distributes a network's computation over its internal paths and is unchanged by the symmetries that preserve it. We extend the path norm into such a probability measure, whose exact ensemble form decomposes a trained network into its pathways. Its Rényi entropies and entropy deficit are therefore invariant, and they read as occupancy and concentration. Its marginals give conserved, directed flows, from which a declared ruler derives a geometry that inherits these

properties. *Architectures, environments and training times can thus be compared in one intrinsic frame.* The second gap was conflating *what a network makes available with what an input uses.* Across 2,847 optimiser trajectories and 324 benchmark networks, learning concentrates the disposition in a recurrent direction, with optimiser- and activation-dependent magnitude. Each condition realises a narrower part of it, and attenuating that part changes the response more than activation-matched controls in every network. *Organised possibility is therefore load-bearing.* An explanation of a trained network can now begin from what learning has organised and test what it carries. The construction can extend to Transformers once attention is resolved into a derived response graph, which we take up in subsequent work.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems* (Vol. 31).

Allain, C., & Cloitre, M. (1991). Characterizing the lacunarity of random and deterministic fractal sets. *Physical Review A*, *44*(6), 3552–3558, https://doi.org/10.1103/PhysRevA.44.3552

Bach, S., Binder, A., Montavon, G., et al. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), e0130140, https://doi.org/10.1371/journal.pone.0130140

Barrat, A., Barthélemy, M., Pastor-Satorras, R., et al. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, *101*(11), 3747–3752, https://doi.org/10.1073/pnas.0400087101

Boccaletti, S., Latora, V., Moreno, Y., et al. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*(4–5), 175–308, https://doi.org/10.1016/j.physrep.2005.10.009

Boguñá, M., Bonamassa, I., De Domenico, M., et al. (2021). Network geometry. *Nature Reviews Physics*, *3*, 114–135, https://doi.org/10.1038/s42254-020-00264-4

Cohen, T., & Welling, M. (2016). Group equivariant convolutional networks. *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 2990–2999).

Conmy, A., Mavor-Parker, A., Lynch, A., et al. (2023). Towards automated circuit discovery for mechanistic interpretability. *Advances in neural information processing systems* (Vol. 36, pp. 16318–16352).

Dinh, L., Pascanu, R., Bengio, S., et al. (2017). Sharp minima can generalize for deep nets. *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1019–1028).

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *International conference on learning representations.*

Garipov, T., Izmailov, P., Podoprikhin, D., et al. (2018). Loss surfaces, mode connectivity, and fast ensembling of DNNs. *Advances in neural information processing systems* (Vol. 31).

Geiger, A., Lu, H., Icard, T., et al. (2021). Causal abstractions of neural networks. *Advances in neural information processing systems* (Vol. 34, pp. 9574–9586).

Godfrey, C., Brown, D., Emerson, T., et al. (2022). On the symmetries of deep learning models and their internal representations. *Advances in neural information processing systems* (Vol. 35, pp. 11893–11905).

Gonon, A., Brisebarre, N., Riccietti, E., et al. (2024). A path-norm toolkit for modern networks: Consequences, promises and challenges. *International conference on learning representations.*

Grigsby, E., Lindsey, K., Rolnick, D. (2023). Hidden symmetries of ReLU networks. *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 11734–11760).

Guidotti, R., Monreale, A., Ruggieri, S., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93, https://doi.org/10.1145/3236009

Hutter, F., Hoos, H.H., Leyton-Brown, K. (2011). Sequential model-based optimization for general

algorithm configuration. *Learning and intelligent optimization (LION 5)* (Vol. 6683, pp. 507–523).

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205).

Kindermans, P.-J., Hooker, S., Adebayo, J., et al. (2019). The (un)reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Cham: Springer.

La Malfa, E., La Malfa, G., Caprioli, C., et al. (2022). *Deep neural networks as complex networks.* (Preprint at https://arxiv.org/abs/2209.05488)

La Malfa, E., La Malfa, G., Nicosia, G., et al. (2021). *Characterizing learning dynamics of deep neural networks via complex networks.* (Preprint at https://arxiv.org/abs/2110.02628)

Lenc, K., & Vedaldi, A. (2019). Understanding image representations by measuring their equivariance and equivalence. *International Journal of Computer Vision*, *127*(5), 456–476, https://doi.org/10.1007/s11263-018-1098-y

Lim, D., Putterman, T., Walters, R., et al. (2024). The empirical impact of neural parameter symmetries, or lack thereof. *Advances in neural information processing systems* (Vol. 37, pp. 28322–28358).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38, https://doi.org/10.1016/j.artint.2018.07.007

Moharil, A. (2026). Temporal geometry of deep networks: Hyperbolic representations of training dynamics for intrinsic explainability. *International conference on learning representations.*

Moharil, A., Vanschoren, J., Singh, P., et al. (2024). Towards efficient AutoML: A pipeline synthesis approach leveraging pre-trained transformers for multimodal data. *Machine Learning*, *113*(9), 7011–7053, https://doi.org/10.1007/s10994-024-06568-1

Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256, https://doi.org/10.1137/S003614450342480

Neyshabur, B., Salakhutdinov, R., Srebro, N. (2015). Path-SGD: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems* (Vol. 28, pp. 2422–2430).

Rényi, A. (1961). On measures of entropy and information. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 547–561).

Serrano, M.Á., Krioukov, D., Boguñá, M. (2008). Self-similarity of complex networks and hidden metric spaces. *Physical Review Letters*, *100*(7), 078701, https://doi.org/10.1103/PhysRevLett.100.078701

Sobol', I.M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, *7*(4), 86–112, https://doi.org/10.1016/0041-5553(67)90144-9

Song, C., Gallos, L.K., Havlin, S., et al. (2007). How to calculate the fractal dimension of a complex network: The box covering algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(03), P03006, https://doi.org/10.1088/1742-5468/2007/03/P03006

Song, C., Havlin, S., Makse, H.A. (2005). Self-similarity of complex networks. *Nature*, *433*(7024), 392–395, https://doi.org/10.1038/nature03248

Sundararajan, M., Taly, A., Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3319–3328).

Testolin, A., Piccolini, M., Suweis, S. (2020). Deep learning systems as complex networks. *Journal of Complex Networks*, *8*(1), cnz018, https://doi.org/10.1093/comnet/cnz018

Wei, D.J., Liu, Q., Zhang, H.X., et al. (2013). Box-covering algorithm for fractal dimension of weighted networks. *Scientific Reports*, *3*, 3049, https://doi.org/10.1038/srep03049

You, J., Leskovec, J., He, K., et al. (2020). Graph structure of neural networks. *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 10881–10891).

Statements and Declarations

- **Funding.** This publication is part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth, with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), Deloitte and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the programme LTP KIC 2020-2023. The work was carried out in the SAFEGUARD lab, a collaboration between the Jheronimus Academy of Data Science (JADS) and Deloitte. We thank Eindhoven University of Technology for access to its high-performance computing cluster, and Tilburg University and JADS for research access and support.
- **Competing interests.** The SAFEGUARD lab is co-funded by Deloitte. The funders had no role in the design of the study, the analysis, the interpretation of the results or the decision to publish. The authors declare no other competing interests.
- **Ethics approval and consent.** Not applicable. The study uses synthetic data and the public MPI3D and Fashion-MNIST datasets, and involves no human participants or personal data.
- **Data availability.** MPI3D and Fashion-MNIST are publicly available. The synthetic environments are generated deterministically by the released code. The result tables and the complete results archive are available at https://github.com/a-moharil/pathflow-organisation.
- **Additional Technical Archive.** An additional auxiliary online technical archive consisting of extended experimental results is also provided at https://github.com/a-moharil/pathflow-organisation/blob/main/ECMLPKDD_results_archive.pdf.
- **Code availability.** The code for the path measure, flow, geometry, acquisition and interventions is available at https://github.com/a-moharil/pathflow-organisation.
- **Author contributions.** A.M.: conceptualisation, methodology, software, formal analysis, investigation, visualisation and writing of the original draft. D.A.T.: supervision, funding acquisition and review of the results along with editing the manuscript. I.K.: supervision with review of the results along with editing of the manuscript.

Appendix A From Multiscale Theory to the Experimental Design

This appendix explains how ideas from the theory of dimension, scaling and network renormalisation shape the geometric part of the programme. Each part states the idea briefly and then the design decision that follows from it. The definitions and acceptance rules themselves are given in Appendix C.10.

A.1 Scale regimes over finite windows

The idea. Box counting characterises how an object fills space across scales. At a scale ℓ, it counts the fewest sets of diameter at most ℓ that cover the object. We call this number $N_B(\ell)$. When this count follows a power law, $N_B(\ell) \propto \ell^{-d_B}$, its exponent

defines the box-counting dimension

$$d_B = \lim_{\ell \to 0} \frac{\log N_B(\ell)}{\log(1/\ell)}.$$

A trained network is finite. Its informative scales therefore lie between its smallest distance and its diameter. Within this range we read the local slope $d_B(\ell) = -d \log N_B(\ell)/d \log \ell$. A slope that stays constant over a window marks scale-free behaviour there. A slope that changes (over time) marks scale-dependent behaviour. Both describe the organisation of the network. We ask which regime a state occupies, over which scales, and how learning moves it between regimes. Rescaling by a characteristic scale then asks whether states of different size share one form. Architectures of different complexity, and states at different times, thereby receive a common geometric coordinate.

The design. Every state is covered on a fixed grid of 64 scales, so that states are compared at the same resolution. Repeated plateaus of the box count are compressed before the slope is computed. A slope then reflects a real change in the cover. A window of constant slope is accepted when it passes seven safeguards (Appendix C.10). Two ensure evidence: at least six distinct box counts and a span of at least 0.5 decade. Two exclude flat segments of the staircase: a slope of at least 0.25 and an at least fivefold change of N_B. Three ensure a single exponent: $R^2 \geq 0.97$, a coefficient of variation of the local slope of at most 0.25, and no preference for a quadratic log-log fit ($\Delta\text{BIC} \leq 2$). The existence of a window and its exponent are recorded separately. Each covering curve is also rescaled to $x = \ell/\xi_{50}$, where ξ_{50} is the scale at which the box count falls to half the number of vertices. Collapse onto one master curve is measured by its root-mean-square error in log units.

The findings. Every architecture and every run has an accepted window at the final snapshot. The geometry of each trained state is therefore scale-free over a finite range. The exponent within that range varies from 0.266 to 2.047 across the cohort. Learning reorders it ($\rho_s = -0.124$ between initialisation and the final state), and the optimiser barely orders it ($W = 0.028$). The exponent is thus a property of each state. The form of the cover is shared instead. After rescaling by ξ_{50}, the covering curves follow one fine-to-coarse transition, with a median master-curve error of 0.034 dex over training. The characteristic scale itself contracts, in all 48 architectures (median fold 0.934) and in all 576 runs of the environment matrix. Learning therefore shortens the scale of a common form. Late in training the form deforms beyond the crossover, and the error rises to 0.086 dex. The crossover also softens, in 36 of 48 architectures and 454 of 576 runs. Within an architecture, this softening tracks concentration (median $\rho_s(S_{\text{deficit}}, d^*) = -0.656$), and it does so most consistently in deeper networks. The geometry thus records the reorganisation of the measure as a contraction of scale (Section 4.3).

A.2 A ruler and a cover derived from the computation

The idea. A graph carries no metric of its own. A geometry on it must be built, and it is only as faithful as the quantity it is built from. Hop distance is the usual

choice. In a trained network, however, the edges are fixed by the architecture, so hop distance cannot register learning. A weighted network must turn its weights into lengths, and scale is then defined relative to that choice (Boguñá et al., 2021). We build the lengths from the flow of the path measure. The flow is a marginal of an intrinsic measure. It is unchanged by rescaling, relabelled by permutation, and sums to one in every layer. A geometry built from it inherits these properties. Theorem 6 makes this exact: an isomorphism that preserves the flow is an isometry, so equal flows give equal geometries. The flow is therefore what anchors the faithfulness of the geometry. Box covering on a declared metric tests self-similarity (Song et al., 2007, 2005), and BCANw extends it to weighted shortest paths (Wei et al., 2013). Studies of complex networks usually take the metric from topology or from a hidden space (Serrano et al., 2008; Song et al., 2005). Here it is taken from the computation.

The design. Each edge receives the length $\lambda_e^{(p)}(t) = (\mathcal{F}_e(t) + \varepsilon)^{-p}$, with $p = 1$ in the principal analysis. Strongly used relations become short, and a box groups units joined by strong flow. The same ruler parameters are applied to every state compared, so a difference in geometry comes from the flow and not from the instrument. Finding the minimum cover is equivalent to colouring the graph that joins every pair of vertices too far apart to share a box. Exact colouring is NP-hard (Song et al., 2007). The cover therefore uses greedy colouring, with vertices ordered by descending weighted strength as in BCANw. This ordering is deterministic, so each cover is reproducible. Its count is an upper bound with a staircase shape, and the plateaus are recorded through the transition fraction and the maximum plateau span. Since the ruler is built from flow, the distribution of flow sets the distribution of lengths. We characterise it through the enrichment $\mathcal{F}_e = |E_\ell| F_e$, which compares each edge with the uniform share of its layer. A pure power law, a truncated power law, a lognormal and an exponential are fitted to the same tail and compared by BIC, both pooled over the network and layer by layer.

The findings. The flow is broad, with a characteristic scale. Pooled tails prefer a lognormal in 46 of 48 architectures and a truncated power law in the other two. A pure power law is never preferred, and it trails the best model by a median of about 246 BIC units. The environment matrix agrees, with 495 of 576 pooled tails preferring a lognormal and 81 a truncated power law. Single layers are more varied. Of 918 layers, 47.6% prefer a truncated power law, 26.9% a lognormal, 15.5% an exponential and 10.0% a pure power law. Power-law behaviour is thus a property of individual layers, and pooling across layers yields a lognormal. A lognormal is the typical outcome of a product of many positive factors. An edge flow is such a product, of upstream reach, squared weight and downstream continuation (Section 3), so the tail agrees with the multiplicative structure of the measure. It also agrees with wider network practice, where many apparent power laws are better described by alternatives once compared directly. For the geometry, a lognormal flow gives a lognormal spread of lengths. A few strongly used edges become very short and form a backbone, and the metric contracts around it during training.

A.3 Observables on the coarse network

The idea. Covering decides which vertices are grouped at a scale. Renormalisation decides how relations among the groups are inherited. Each box is contracted to a

supernode, and a quotient rule assigns the relations between boxes. The quotient rule is a separate choice from the covering rule. Repeating the step asks whether the coarse network resembles the fine one. Real networks studied in this way often unfold into self-similar layers (Song et al., 2007), and weighted networks do so too. We ask the same question of a trained computation, and how learning changes the answer.

The design. The covering metric is undirected, while the computational flow is directed. Both are kept on one partition. The cover fixes the boxes, and the directed quotient flow $F'_{AB} = \sum_{u\in A, v\in B} F_{uv}$ between distinct boxes records what the flow does on them. Four observables follow. The flow lacunarity Λ_F and its peak Λ_F^* describe how unevenly mass fills the boxes. The inter-box retention, at four box fractions from $N_B/N \approx 0.5$ to 0.0625, measures how much directed flow passes between boxes rather than within them. The KS statistic compares the distribution of quotient flows across these resolutions. The local mismatch Δ_i^F compares how flow mass and metric support grow around each node, on one common window. Each stage, from measure to ruler, cover and quotient, is declared separately. A change in any observable can therefore be traced to the stage at which it arises.

The findings. The coarse flow resembles the fine flow without repeating it. The median KS distance between resolutions is 0.245 (range 0.177 to 0.320), and it grows with network size ($\rho_s = 0.728$ with parameter count). Retention falls from 0.860 at $N_B/N \approx 0.5$ to 0.189 at 0.0625. At coarse scales, most flow is therefore internal to boxes, so strong flow is grouped rather than spread across the network. The Rényi occupancies also change with scale (median within-architecture coefficient of variation 0.557 for η_1), so concentration depends on the resolution at which it is read. Locally, mass grows more slowly than support. Wherever a common window exists, the median of Δ_i^F is negative in all 48 architectures (cohort median -0.408), and 99.5% of valid nodes are negative. The measure is thus locally concentrated relative to the metric volume around it. These observables keep moving after the function settles. From the best-test checkpoint to the final one, peak lacunarity rises and retention falls in all 48 architectures. Many real networks unfold in a nearly self-similar way under geometric renormalisation, but a trained network is self-similar only in part. Its coarse structure keeps a related shape, while learning draws the flow inward into groups that the cover resolves.

Appendix B Observables and Interpretation

B.1 A statistical-mechanical reading of the path measure

The structural measure of Section 3.2 admits an exact representation in the mathematical form of a statistical ensemble whose states are complete computational routes. Here we explain what those measurements mean, how they connect a static description of the disposition to its motion under training along with its realisation under a condition.

Paths as states with additive log-weights.

For a complete path π at training state t, we define the path log-weight

$$\omega_\pi(t) = \log Q_t^S(\pi) = \sum_{e \in \pi} \omega_e(t), \qquad \omega_e(t) = \log w_e(t)^2. \tag{B1}$$

The log-weight of a path is therefore a sum of local edge log-weights, and a strong edge carries a large log-weight. The structural measure and its partition functions take the normalised exponential form

$$\mu_t^S(\pi) = \frac{e^{\omega_\pi(t)}}{Z_1(t)}, \qquad Z_q(t) = \sum_\pi e^{q\,\omega_\pi(t)} = \sum_\pi Q_t^S(\pi)^q. \tag{B2}$$

The correspondence is exact, and it is ensemble-theoretic rather than physical in its current formulation. Complete paths play the role of states, and their multiplicative gains determine their statistical weights. The additivity of the log-weight is what makes the ensemble tractable. For a dense layered network, the partition function is a product of transfer matrices,

$$Z_q(t) = \mathbf{1}^\top |W_L|^{\circ 2q} \cdots |W_1|^{\circ 2q}\, \mathbf{1}, \tag{B3}$$

where $|W_\ell|^{\circ 2q}$ raises the modulus of every entry of the layer weight matrix to the power $2q$. The dynamic programming of Theorem 3 is the evaluation of this product, and the flow $F_e = L_u\, q_e\, R_v / Z$ is the probability that a sampled state occupies the bond e.

The Rényi order as a sharpening parameter.

In this representation, the Rényi entropies are differences of logarithmic partition functions,

$$H_q(t) = \frac{\log Z_q(t) - q \log Z_1(t)}{1 - q}. \tag{B4}$$

The order q acts as an exponent on the path weights relative to the reference ensemble at $q = 1$. The escort distribution $\mu_t^{(q)}(\pi) = e^{q\,\omega_\pi(t)}/Z_q(t)$ concentrates on the paths of largest log-weight as q increases, and it becomes uniform over the occupied support as $q \to 0$. Increasing q therefore sharpens the ensemble towards its dominant core, while decreasing q flattens it towards the architectural carrier. Since $H_0 = \log |\operatorname{supp} \mu_t^S|$, which equals $\log N_{\text{topo}}$ whenever no path has vanishing weight, the occupancies η_q form an ordered spectrum, $1 = \eta_0 \geq \eta_1 \geq \eta_2$. The Shannon occupancy describes the full ensemble, whereas the order-two occupancy, $N_{\text{eff}}^{(2)} = Z_1^2/Z_2$, describes its dominant core. A training process that lowers η_2 more strongly than η_1 is therefore one that sharpens the dominant core faster than it narrows the ensemble as a whole.

Entropy as expected information.

At the level of a single path, the quantity

$$I_\pi(t) = -\log \mu_t^S(\pi) \tag{B5}$$

is the information associated with encountering that path under the current measure. Its exponential $1/\mu_t^S(\pi)$ is a reciprocal-probability multiplicity. It is the number of carriers that a hypothetical equiprobable ensemble would contain if every carrier had the probability of π. The contribution of the path to the entropy is $h_\pi(t) = \mu_t^S(\pi) I_\pi(t)$, and

$$H_1(t) = \sum_\pi h_\pi(t) = \mathbb{E}_{\pi\sim\mu_t^S}\big[I_\pi(t)\big]. \tag{B6}$$

The Shannon entropy is therefore the expected logarithmic multiplicity encountered when a path is sampled according to the network's own disposition. In the ensemble representation, the same quantity satisfies the exact relation

$$H_1(t) = \log Z_1(t) - \big\langle \omega \big\rangle_{\mu_t^S}, \tag{B7}$$

so that entropy, mean log-weight and the logarithm of the partition function are three views of one ensemble. The Shannon effective number,

$$N_{\text{eff}}^{(1)}(t) = \prod_\pi \left(\frac{1}{\mu_t^S(\pi)}\right)^{\mu_t^S(\pi)}, \tag{B8}$$

is the probability-weighted geometric mean of the reciprocal multiplicities. If the mass were spread uniformly over exactly K paths, it would equal K. It therefore extends the cardinality of an equiprobable carrier to a heterogeneous one.

Typical-set multiplicity.

The effective number also has an asymptotic meaning. If complete paths are drawn independently from μ_t^S, a typical sequence of n draws has probability close to $e^{-nH_1(t)}$, and the typical set contains approximately $e^{nH_1(t)}$ sequences. The quantity $e^{H_1(t)}$ is therefore the exponential multiplicity per draw of the typical ensemble generated by the disposition. This construction concerns repeated sampling from the measure. It does not assume that the paths of the network are statistically independent components of its computation.

Concentration as departure from the uniform ensemble.

The uniform measure $u_\pi = 1/N_{\text{topo}}$ is the reference ensemble of the architecture, in which every available route is equally weighted. The entropy deficit,

$$S_{\text{deficit}}(t) = \log N_{\text{topo}} - H_1(t) = D_{\text{KL}}\big(\mu_t^S \,\|\, u\big), \tag{B9}$$

measures how far the learnt ensemble lies from this reference. It compares the multiplicity that the architecture supplies with the multiplicity that the learnt measure effectively occupies. In this reading, an overparameterised architecture supplies a large repertoire of routes, and learning allocates mass within that repertoire. Concentration is accordingly a selective allocation over a dense carrier. It is neither parameter sparsity, since most weights remain non-zero, nor activation sparsity, since the structural measure is not conditioned on an input, nor architectural pruning.

Training as an exact reweighting of the ensemble.

Between two training states t_0 and t_1 that share path support, the change of path log-weight is $\Delta\omega_\pi = G_\pi$, where $G_\pi = \log\big(Q^S_{t_1}(\pi)/Q^S_{t_0}(\pi)\big)$ is the path-growth field of Section 4.2. The new measure is therefore an exact exponential reweighting of the old one,

$$\mu^S_{t_1}(\pi) = \frac{\mu^S_{t_0}(\pi)\, e^{G_\pi}}{\mathbb{E}_{\mu^S_{t_0}}\big[e^G\big]}. \tag{B10}$$

Along the interpolation $\mu_a \propto \mu^S_{t_0} e^{aG}$, with $0 \le a \le 1$, the entropy deficit moves according to

$$\frac{dS_{\mathrm{deficit}}(a)}{da} = \mathrm{Cov}_{\mu_a}\big(G, \log \mu_a\big) = \mathrm{Cov}_{\mu_a}\big(G,\, \omega_a\big), \tag{B11}$$

where the second form uses $\log \mu_a = \omega_a - \log Z_a$. Concentration therefore occurs precisely when the growth of a path covaries positively with its current log-weight. In that case, paths that are already strong grow further and weak paths fall behind, so that the log-weight profile is stretched rather than merely shifted. A uniform shift of all log-weights, in contrast, leaves the measure unchanged. The optimiser comparisons of Section 4.2 can therefore be read as comparisons of how strongly, and how consistently, each optimiser stretches the profile of route log-weights.

Conditioning as an additive log-factor.

The same reading extends to conditional realisation. Whenever the realised score factorises as $Q^c(\pi) = Q^S(\pi)\, R_c(\pi)$ (Section 5), the realised log-weight is

$$\omega^c_\pi = \omega_\pi + \log R_c(\pi), \tag{B12}$$

so that a condition adds a path-dependent log-factor to the structural profile. For an MLP, $\log R_c(\pi) = 2\log|\Delta x_{i_0}| + 2\sum_\ell \log\big|s^{(\ell),c}_{i_\ell}\big|$ is again a sum of local terms, which is why the conditional measure is computed by the same dynamic programming. A carrier with $R_c(\pi) = 0$ receives realised log-weight $-\infty$. Conversely, a carrier excluded from the structural support already has structural log-weight $-\infty$, and no finite log-factor can raise it. Support domination, $\mathrm{supp}\,\mu_c \subseteq \mathrm{supp}\,\mu^S$, is the ensemble form of this statement. A condition can strengthen, weaken or forbid states of the disposition, but it cannot create states that the disposition does not contain.

Appendix C Formal results and proofs

This appendix collects the formal results of the work in their logical order. The order differs from the order of use in the main text. It begins with the measure on a finite carrier and proceeds through the structural measure, finite response, realisation, symmetry, parameter sharing and geometry. It closes with the motion of the measure under training, intervention and the dependency structure of the whole. Statements made in the main text are repeated with their original numbers, so that the appendix can be read on its own. The remaining statements appear here only and are cited from the main text.

C.1 Carrier, score, measure and support

The proofs use a small common core. Let $\mathcal{R}$ be a finite carrier of computationally complete relational objects. Depending on the architecture, an element $r \in \mathcal{R}$ is a complete MLP path or a spatially resolved CNN execution path. The carrier is fixed by the architecture and carries no training index. At training state t, let $Q_t(r) \geq 0$ be an unnormalised score induced by the parameter state θ_t. Whenever

$$Z(t) := \sum_{r \in \mathcal{R}} Q_t(r) > 0, \tag{C13}$$

the normalised measure and its support are

$$\mu_t(r) = \frac{Q_t(r)}{Z(t)}, \qquad \operatorname{supp} \mu_t = \{r \in \mathcal{R} : Q_t(r) > 0\}. \tag{C14}$$

In the structural MLP construction, $Q_t^S(\pi) = \prod_{e \in \pi} w_e(t)^2$. In the conditional construction, the score is the square of a signed finite-response amplitude evaluated at the same parameter state, $Q_t^c(\pi) = \left(A_{\pi,t}^c\right)^2$. The non-negative measure records prominence. The signed amplitude retains constructive and destructive response.

Lemma 1 (Normalisation identity) Whenever $Z(t) > 0$, the measure μ_t satisfies $\mu_t(r) \geq 0$ for every $r \in \mathcal{R}$ and $\sum_{r \in \mathcal{R}} \mu_t(r) = 1$.

Proof Non-negativity follows from $Q_t(r) \geq 0$ and $Z(t) > 0$. Moreover,

$$\sum_{r \in \mathcal{R}} \mu_t(r) = \frac{1}{Z(t)} \sum_{r \in \mathcal{R}} Q_t(r) = \frac{Z(t)}{Z(t)} = 1. \tag{C15}$$

□

The lemma is elementary, but every conservation statement below derives from it. Conservation is therefore not an additional law imposed on the network. It is the graph projection of a normalised measure on complete carriers.

ReLU as the simplest realisation gate.

Several later statements are most transparent for ReLU, $\sigma(z) = \max\{0, z\}$. The map suppresses negative preactivation exactly and transmits positive preactivation linearly. A route through a unit can therefore be structurally available, with non-zero learnt weights, while a given condition places the unit in its suppressed half-space. Two properties of ReLU are used below. First, it is positively homogeneous, $\sigma(az) = a\,\sigma(z)$ for $a > 0$, which creates the rescaling symmetry of Theorem 1. Second, realisation under a finite contrast is not binary. When the two endpoint preactivations lie on opposite sides of the kink, the finite response lies strictly between zero and one (Appendix C.3). The theory therefore uses finite-response operators in place of an activation mask.

C.2 The structural measure: invariance, computation and conservation

The structural measure must satisfy three requirements before it can serve as an observable. It must not change under redescriptions of the same function. It must be computable although the number of complete paths grows exponentially with depth. Its projections onto the graph must have a fixed total, so that their changes are redistributions. The following results establish each requirement in turn.

Theorem 1 (Structural path-gain gauge invariance, restated). *Consider a bias-free feed-forward ReLU MLP at a fixed training state t, and let v be any hidden unit. For some $a > 0$, apply the positive hidden-unit rescaling $w_{u\to v}(t) \mapsto a\, w_{u\to v}(t)$ and $w_{v\to r}(t) \mapsto a^{-1} w_{v\to r}(t)$ to every incoming and outgoing edge of v. Then $Q_t^S(\pi)$ is unchanged for every complete input-to-output path π. Consequently, the structural partition function $Z_S(t)$, the normalised structural measure μ_t^S and every quantity derived solely from that measure are unchanged.*

Proof A complete feed-forward path either does not pass through v, in which case none of its factors changes, or passes through v exactly once. In the second case, its squared path product contains one incoming and one outgoing factor of v. Under the rescaling,

$$\big(a\, w_{u\to v}(t)\big)^2 \big(a^{-1} w_{v\to r}(t)\big)^2 = w_{u\to v}(t)^2\, w_{v\to r}(t)^2. \tag{C16}$$

All remaining factors are unchanged, so $Q_t^S(\pi)$ is unchanged for every π. Therefore $Z'_S(t) = Z_S(t)$, and hence $\mu_t^{S\prime} = \mu_t^S$. Every functional of the normalised measure alone is therefore invariant at training state t. The reparameterisation is also function-preserving, because ReLU is positively homogeneous. □

Corollary 1 (Invariance extends to any finite set of hidden-unit positive rescalings) If the transformation of Theorem 1 is applied independently to any finite collection of hidden units at training state t, the complete structural path measure μ_t^S remains unchanged.

Proof Each rescaled unit contributes the factor $a_v^2 a_v^{-2} = 1$ to every path that traverses it. A product of finitely many such factors is still one. □

The theorem identifies a precise sense in which the path measure is more intrinsic than isolated edge magnitudes. Its scope is the positive nodewise rescaling group of homogeneous ReLU networks. The general condition under which a transformation leaves the measure unchanged is given in Appendix C.5.

Theorem 3 (Exact dynamic programming, restated). *Let a finite layered acyclic graph at training state t carry non-negative local edge factors $q_e(t) \geq 0$, and let $Q_t(\pi) = \prod_{e\in\pi} q_e(t)$. Then the partition function $Z(t) = \sum_\pi Q_t(\pi)$ can be computed exactly through layerwise dynamic programming, without explicit enumeration of the complete path family.*

Proof Let $L_v(t)$ denote the total score of all directed prefixes from any source to the vertex v. For a source vertex, set $L_v(t) = 1$, or the declared source weight. For every later vertex, set

$$L_v(t) = \sum_{u:\, u\to v} L_u(t)\, q_{u\to v}(t). \tag{C17}$$

By induction on topological depth, $L_v(t)$ equals the sum of products over all complete prefixes ending at v. The base case is immediate. If the claim holds for every predecessor u of v, distributivity gives

$$L_v(t) = \sum_u \Big(\sum_{\rho:\, s \rightsquigarrow u} \prod_{e \in \rho} q_e(t) \Big) q_{u \to v}(t) = \sum_{\rho':\, s \rightsquigarrow v} \prod_{e \in \rho'} q_e(t). \tag{C18}$$

Summing $L_v(t)$ over the target vertices sums each complete path exactly once, so $Z(t) = \sum_{v \in V_{\text{out}}} L_v(t)$. Replacing each factor $q_e(t)$ by $q_e(t)^\alpha$ yields the exact Rényi partition functions by the same argument. □

The recursion needs one pass over the edges per layer. A sum over a number of paths that grows exponentially with depth is therefore evaluated at a cost linear in the number of edges.

Proposition 3 (Forward-backward edge-flow factorisation) Under the assumptions of Theorem 3, let $\mu_t(\pi) = Q_t(\pi)/Z(t)$, and let $R_v(t)$ denote the total score of all suffixes from v to any target. For an edge $e = (u \to v)$, the normalised complete-path marginal is

$$F_e(t) = \sum_{\pi \ni e} \mu_t(\pi) = \frac{L_u(t)\, q_e(t)\, R_v(t)}{Z(t)}. \tag{C19}$$

Proof Every complete path containing e splits uniquely into a prefix ending at u, the edge e and a suffix beginning at v. Because the score factorises multiplicatively, the sum over all such paths is the product of the prefix sum, the factor of e and the suffix sum. Division by $Z(t)$ normalises it. □

An edge therefore carries large flow only when its upstream accessibility, its own factor and its downstream continuation are large together. Flow is a property of an edge in its context, and it is not a function of the edge weight alone.

Theorem 4 (Complete-carrier cut conservation, restated). *Let μ_t be a normalised measure on complete carriers at training state t, with marginals $F_e(t) = \sum_{r \ni e} \mu_t(r)$. Let C be a set of elementary execution relations such that every carrier crosses exactly one member of C. Then $\sum_{e \in C} F_e(t) = 1$. Likewise, if every carrier visits exactly one node of a layer V_ℓ, then $\sum_{v \in V_\ell} F_v(t) = 1$.*

Proof Interchanging the finite sums gives

$$\sum_{e \in C} F_e(t) = \sum_{e \in C} \sum_{r \ni e} \mu_t(r) = \sum_{r \in \mathcal{R}} \mu_t(r) \sum_{e \in C} \mathbf{1}\{e \in r\}. \tag{C20}$$

Every carrier crosses exactly one member of C, so the inner sum is one for every r. The total is therefore $\sum_r \mu_t(r) = 1$ by Lemma 1. The node statement is identical, with node-incidence indicators. □

What is conserved is one unit of normalised carrier mass across a cut. The theorem does not assert conservation of activation magnitude, gradient norm or information in a thermodynamic sense. Training or conditioning can redistribute the unit mass within a cut, but they cannot change its total.

Proposition 4 (Mean-one enrichment) Let C be a finite conserved cut at training state t, and define the architecture-normalised enrichment $\mathcal{F}_e(t) = |C|\, F_e(t)$. Then

$$\frac{1}{|C|} \sum_{e \in C} \mathcal{F}_e(t) = 1. \tag{C21}$$

Proof Substitution and Theorem 4 give $\frac{1}{|C|} \sum_{e \in C} |C|\, F_e(t) = \sum_{e \in C} F_e(t) = 1$. □

The proposition fixes the baseline of every flow comparison. Enrichment above one and depletion below one are measured against a conserved architectural baseline. The number of edges in a layer therefore cannot masquerade as learnt heterogeneity.

C.3 Finite response

A declared condition $c = (x, x^\circ)$ compares two executions of the same network at a fixed training state t. The question is how the difference between the two executions propagates. The results below show that the propagation is exactly linear in the contrast, although the network is nonlinear in its absolute states.

Theorem 9 (Exact finite-response identity, restated). *Let $B : \mathbb{R}^m \to \mathbb{R}^n$ be continuously differentiable on the segment joining u° and u, with Jacobian J_B, and let*

$$R_B(u, u^\circ) = \int_0^1 J_B\big(u^\circ + \alpha(u - u^\circ)\big)\, d\alpha. \tag{C22}$$

Then $B(u) - B(u^\circ) = R_B(u, u^\circ)\,(u - u^\circ)$.

Proof Let $\varphi(\alpha) = B\big(u^\circ + \alpha(u - u^\circ)\big)$ for $0 \le \alpha \le 1$. By the chain rule, $\varphi'(\alpha) = J_B\big(u^\circ + \alpha(u - u^\circ)\big)(u - u^\circ)$. The fundamental theorem of calculus gives $B(u) - B(u^\circ) = \int_0^1 \varphi'(\alpha)\, d\alpha$. The contrast $u - u^\circ$ does not depend on α, so it can be taken outside the integral. □

The operator R_B is the average of the Jacobian along the straight segment between the two endpoint states. For a scalar coordinatewise nonlinearity ϕ, it reduces to the secant

$$s_\phi(z, z^\circ) = \frac{\phi(z) - \phi(z^\circ)}{z - z^\circ}, \qquad z \neq z^\circ.$$

For ReLU with $z^\circ < 0 < z$, the secant equals $z/(z - z^\circ) \in (0, 1)$. It records how much of the finite contrast lies on the transmitting side of the kink.

Proposition 5 (Local Jacobian limit) If B is continuously differentiable at u, then $\lim_{u^\circ \to u} R_B(u, u^\circ) = J_B(u)$.

Proof As $u^\circ \to u$, every point of the joining segment converges to u uniformly in $\alpha \in [0, 1]$. By continuity of J_B, the integrand converges uniformly to $J_B(u)$. The limit can therefore be taken through the integral, which gives $J_B(u)$. □

The finite and the local responses therefore answer related but distinct questions. The finite operator describes an actual contrast across a non-zero region of state space. The Jacobian describes the sensitivity around one state.

Theorem 10 (Exact conditioned response decomposition, restated). *Consider a finite acyclic neural computation evaluated at the two executions of a condition $c = (x, x^\circ)$. Replace every differentiable operation by its finite-response operator from Theorem 9, retaining every dependency of the chosen execution structure. Then the target contrast is an exact linear function of the source contrasts. If the conditioned system is represented by a directed acyclic response graph, the target contrast admits the exact sum over complete response paths*

$$\Delta y = \sum_{\pi \in \Pi} A^c_{\pi,t}, \tag{C23}$$

where $A^c_{\pi,t}$ is the signed product of the source contrast and of the finite-response coefficients encountered along π.

Proof Order the operations topologically. At each operation, Theorem 9 gives an exact linear relation between its output contrast and the contrasts of its parents, evaluated at the two actual endpoint states. Substituting these relations recursively expresses the target contrast as a linear combination of the source contrasts. On a directed acyclic graph, distributivity expands the composition of these linear maps into a sum over complete source-to-target paths. Each monomial of the expansion is the signed amplitude of one path. Every local relation is exact at the two endpoint states, so the expansion introduces no approximation. □

Corollary 2 (Architecture-specific reductions) The finite-path decomposition of an MLP and the decomposition over the spatially unrolled execution graph of a CNN are instances of Theorem 10, provided the carrier retains every dependency required at the chosen level of description.

For an MLP with coordinatewise activations, the amplitude of a path $\pi = (i_0, \dots, i_L)$ is

$$A^c_{\pi,t} = \Delta x_{i_0} \prod_{\ell=1}^{L} w^{(\ell)}_{i_\ell i_{\ell-1}}(t) \prod_{\ell=1}^{L-1} s^{(\ell),c}_{i_\ell}(t)$$

. Hence $Q^c_t(\pi) = (A^c_{\pi,t})^2 = Q^S_t(\pi) R_{c,t}(\pi)$, with $R_{c,t}(\pi) = \Delta x^2_{i_0} \prod_{\ell=1}^{L-1} \left(s^{(\ell),c}_{i_\ell}(t)\right)^2$. Over 6,000 validation cases, the path expansion reconstructs the output contrast to machine precision (Appendix D.4). The theorem does not state that the network is globally linear. Once a condition is declared, the difference between its two executions propagates exactly through a condition-specific linear response system.

C.4 Disposition and realisation

The factorisation $Q^c_t = Q^S_t R_{c,t}$ relates the realised score of a path to its structural score. The next results state what a condition can and cannot do to the structural support.

Proposition 1 (Support domination, restated). *If* $Q_t^c(\pi) = Q_t^S(\pi)\, R_{c,t}(\pi)$ *with finite* $R_{c,t}(\pi) \geq 0$ *on a common finite carrier, then* $\operatorname{supp} \mu_t^c \subseteq \operatorname{supp} \mu_t^S$*, or equivalently* $\mu_t^c \ll \mu_t^S$.

Proof If $\pi \notin \operatorname{supp} \mu_t^S$, then $Q_t^S(\pi) = 0$. Hence $Q_t^c(\pi) = 0$ for every finite $R_{c,t}(\pi) \geq 0$, and $\pi \notin \operatorname{supp} \mu_t^c$. On a finite carrier, inclusion of supports is equivalent to absolute continuity. Every set of zero structural measure contains only structurally null paths, and it therefore also has zero conditional measure. □

Remark 1 The statement concerns supports. Probability measures are not subsets of one another, and the measure-theoretic form of the statement is absolute continuity on a common carrier. For architectures with coupled operations, the factorisation must be stated on the response carrier that the conditional theory actually uses.

Corollary 3 (Redistribution without change of support) It is possible that $\operatorname{supp} \mu_t^c = \operatorname{supp} \mu_t^S$ while $\mu_t^c \neq \mu_t^S$.

Proof Choose $R_{c,t}(\pi) > 0$ for every structurally supported π, but not constant on the support. Every structural path then remains supported. Normalisation of $Q_t^S(\pi) R_{c,t}(\pi)$ nevertheless changes the relative probabilities, unless $R_{c,t}$ is constant μ_t^S-almost everywhere. □

A condition can therefore act purely by redistributing positive mass over an unchanged support. This is the typical regime of smooth activations, whose secants do not vanish. Under ReLU, a condition can in addition remove paths from the support.

C.5 Symmetry

Theorem 1 concerns one family of transformations. The general question is when a changed representation of the network is the same measured organisation. The answer depends on whether the transformation relabels carriers or mixes them.

Theorem 2 (Carrier-preserving symmetry invariance, restated). *At a fixed training state* t*, let* g *be a transformation of the network representation that induces a bijection*

$$T_g : \mathcal{R} \to \mathcal{R}' \tag{C24}$$

between the carriers before and after the transformation. Let Q_t *and* Q'_t *be the non-negative carrier scores, with positive normalising constants* $Z(t)$ *and* $Z'(t)$*, and suppose that*

$$Q'_t(T_g r) = K_g\, Q_t(r) \qquad \text{for all } r \in \mathcal{R}, \tag{C25}$$

where $K_g > 0$ *is independent of* r*. Then the normalised measure transforms only by pushforward,*

$$\mu'_t = (T_g)_{\#}\mu_t. \tag{C26}$$

Consequently, every statistic invariant to carrier relabelling is unchanged, including the Shannon entropy, the Rényi entropies, the effective occupancy and any symmetric functional of the carrier probabilities.

Proof The transformed partition function is

$$Z'(t) = \sum_{r' \in \mathcal{R}'} Q'_t(r') = \sum_{r \in \mathcal{R}} Q'_t(T_g r) = K_g Z(t). \tag{C27}$$

Therefore $\mu'_t(T_g r) = K_g Q_t(r) / \big(K_g Z(t)\big) = \mu_t(r)$, which is the pushforward relation of Equation (C26). Statistics invariant to relabelling depend only on the multiset of carrier probabilities, and they are therefore preserved. □

The common factor K_g is essential. A transformation may rescale every score, since normalisation removes a common factor. It may not rescale scores differently on different carriers.

Corollary 4 (Permutation invariance) Any hidden-unit permutation that induces a bijection of complete carriers and preserves corresponding carrier scores satisfies Theorem 2 with $K_g = 1$.

Corollary 5 (Positive ReLU rescaling as a carrier-preserving symmetry) The transformation of Theorem 1 is a special case of Theorem 2, with T_g equal to the identity on path labels and $K_g = 1$.

Boundary 1 (Path-mixing transformations) If a function-preserving transformation mixes several old carriers into several new carriers, rather than inducing a carrierwise bijection that satisfies Equation (C25), carrier-level probabilities need not be invariant. The intrinsic object must then be sought at a coarser invariant level, such as an appropriate bundle, quotient, operator or response subspace. For this reason, ordinary scalar path identity is not assumed to be universal across all neural architectures.

C.6 Parameter sharing

In a CNN, one learnt coefficient is used at many spatial sites. The realised computation lives on these execution occurrences. The compact parameter description is obtained by mapping each occurrence to its identity through the sharing map ρ. Let A be an execution occurrence drawn according to the normalised occurrence flow of a CNN transition, structural or conditional, and let $\Theta = \rho(A)$ be its parameter identity. This probabilistic reading adds no assumption. Every complete path crosses exactly one occurrence of the transition, so the occurrence flows of the transition form a distribution (Theorem 4).

Theorem 5 (Occurrence, identity and reuse decomposition, restated). *For the deterministic quotient $\Theta = \rho(A)$,*

$$H(A) = H(\Theta) + H(A \mid \Theta). \tag{C28}$$

With $H_{\mathrm{occ}} = H(A)$, $H_{\mathrm{param}} = H(\Theta)$ and $H_{\mathrm{reuse}} = H(A \mid \Theta)$, the specialisations S_{occ}, S_{param} and S_{reuse} of Appendix C.10 satisfy $S_{\mathrm{occ}} = S_{\mathrm{param}} + S_{\mathrm{reuse}}$ whenever every identity has the same number of execution sites.

Proof Because Θ is a function of A, the pair (A, Θ) carries no more uncertainty than A, so $H(A, \Theta) = H(A)$. The chain rule gives $H(A, \Theta) = H(\Theta) + H(A \mid \Theta)$. Equating the two expressions proves the entropy identity. With $n_{\text{occ}} = n_{\text{param}}\, m$ occurrences for n_{param} identities of common multiplicity m, the uniform values satisfy $\log n_{\text{occ}} = \log n_{\text{param}} + \log m$. Subtracting the entropy identity gives the decomposition of the specialisations. □

S_{param} measures the concentration of flow over which learnt coefficient is used. S_{reuse} measures the remaining concentration over where it is used, once the coefficient is known. The decomposition is an exact information-theoretic separation induced by the quotient ρ.

Proposition 6 (Occurrence-to-identity pushforward flow) Let $F_a(t)$ be a normalised flow on execution occurrences and ρ the map to parameter identities. The identity flow is the pushforward $F_\vartheta(t) = \sum_{a:\rho(a)=\vartheta} F_a(t)$. If a complete carrier can use one identity more than once, the identity-flow count is $F_\vartheta(t) = \sum_\pi \mu_t(\pi)\, N_\vartheta(\pi)$, where $N_\vartheta(\pi)$ is the number of occurrences of ϑ along π.

Proof The first statement is the definition of the pushforward of the occurrence measure under ρ. For the second, sum the occurrence indicators of the same identity along each carrier, and interchange the finite sums over carriers and occurrences. □

Proposition 7 (Aggregation and metric construction do not commute) Let two parallel occurrences a_1, a_2 share one parameter identity and carry flows $F_1 \geq F_2 > 0$, and let $\lambda(F) = (F + \varepsilon)^{-p}$ with $p > 0$. Then $\lambda(F_1 + F_2) \neq \min\{\lambda(F_1), \lambda(F_2)\}$.

Proof Since λ is strictly decreasing, $\min\{\lambda(F_1), \lambda(F_2)\} = \lambda(F_1)$. Because $F_1 + F_2 > F_1$, strict monotonicity gives $\lambda(F_1 + F_2) < \lambda(F_1)$. □

Quotienting occurrence flow before the geometry is built is therefore not equivalent, in general, to building the occurrence geometry and quotienting afterwards. The correct order is to resolve spatial execution, then to condition and measure, and only then to aggregate by parameter identity.

C.7 From conserved flow to geometry

The geometry is downstream of the measure. A declared ruler converts flow into edge length. Once the ruler is fixed, several geometric consequences are exact. The choice of the ruler itself is not forced by probability theory. The results below are therefore conditional on the declared ruler.

Proposition 8 (Flow-length antitonicity) For $\lambda_p(\mathcal{F}) = (\mathcal{F} + \varepsilon)^{-p}$ with $\varepsilon > 0$ and $p > 0$,

$$\frac{d\lambda_p}{d\mathcal{F}} = -p\,(\mathcal{F} + \varepsilon)^{-p-1} < 0. \tag{C29}$$

Hence greater normalised flow enrichment necessarily produces a shorter elementary edge under the declared ruler.

Proof Differentiate directly. Since $p > 0$ and $\mathcal{F} + \varepsilon > 0$, every factor multiplying the leading minus sign is positive. □

Theorem 6 (BCANw equivariance, restated). *Let $\varphi : G \to G'$ be a graph isomorphism between two geometry-bearing conditioned graphs, whose flow fields are evaluated at training states t and t'. Suppose that corresponding edges have equal normalised enrichment,*

$$\mathcal{F}'_{\varphi(e)}(t') = \mathcal{F}_e(t), \tag{C30}$$

and that both graphs use the same ruler parameters (p, ε). Then φ is an isometry of the induced shortest-path metric spaces. Consequently, the exact BCANw covering functions are identical at every scale, up to relabelling, and every descriptor determined solely by those metric-measure data is preserved.

Proof Equal enrichments imply equal corresponding edge lengths under the common ruler, $\lambda'_{\varphi(e)}(t') = \lambda_e(t)$. Every path P from u to v maps bijectively to a path $\varphi(P)$ from $\varphi(u)$ to $\varphi(v)$ with the same total length. Taking minima over the two path sets gives $d_{G'}\big(\varphi(u), \varphi(v); t'\big) = d_G(u, v; t)$, so φ is an isometry. A subset is an admissible box at scale ℓ_B in G if and only if its image is an admissible box at the same scale in G'. Minimum covers therefore have equal cardinality, and the exact covering functions are identical. Any descriptor constructed only from the preserved metric and the transported normalised measure is consequently invariant. □

The theorem concerns exact flow-preserving isomorphisms. Transformations of the input, such as translation or rotation of an image, are not of this kind for a trained CNN. Their effect on the realised field is therefore measured, and it is not implied by the theorem (Section 6.3).

Proposition 9 (Population averaging and nonlinear geometry do not commute) Let $F^{(1)}, \ldots, F^{(n)}$ be condition-specific flow fields on a common graph, and let $\mathcal{G}$ denote the map from a flow field to a BCANw descriptor. In general,

$$\mathcal{G}\Big(\frac{1}{n}\sum_{i=1}^{n} F^{(i)}\Big) \neq \frac{1}{n}\sum_{i=1}^{n} \mathcal{G}\big(F^{(i)}\big). \tag{C31}$$

Proof It suffices to exhibit one non-commuting component of $\mathcal{G}$. The ruler is nonlinear for $p \neq 0$, so

$$\Big(\frac{F_1 + F_2}{2} + \varepsilon\Big)^{-p} \neq \frac{1}{2}\Big[(F_1 + \varepsilon)^{-p} + (F_2 + \varepsilon)^{-p}\Big] \tag{C32}$$

for generic $F_1 \neq F_2$. Averaging flow before the edge lengths are built therefore already differs from averaging lengths built per condition. Shortest-path minimisation and covering add further nonlinearities, so equality cannot hold in general. □

For this reason, the population geometry and the median input-resolved geometry are reported as separate quantities.

C.8 Motion of the measure under training

The results so far concern one state. The next results describe how the measure moves between two training states that share a common positive carrier support. They hold for any pair of such states, and they do not depend on which optimiser produced them.

Theorem 7 (Normalised-measure exponential tilt, restated). *Let $Q_{t_0}(\pi) > 0$ and $Q_{t_1}(\pi) > 0$ be the scores of two training states on the same finite carrier Π, with normalised measures μ_{t_0} and μ_{t_1}. Define the finite log-growth field*

$$G(\pi) = \log \frac{Q_{t_1}(\pi)}{Q_{t_0}(\pi)}.$$

Then

$$\mu_{t_1}(\pi) = \frac{\mu_{t_0}(\pi)\, e^{G(\pi)}}{\mathbb{E}_{\mu_{t_0}}\left[e^G\right]}. \tag{C33}$$

Proof By the definition of G, $Q_{t_1}(\pi) = Q_{t_0}(\pi)\, e^{G(\pi)}$. Therefore

$$Z(t_1) = \sum_{\pi} Q_{t_0}(\pi)\, e^{G(\pi)} = Z(t_0)\, \mathbb{E}_{\mu_{t_0}}\left[e^G\right]. \tag{C34}$$

Substituting into $\mu_{t_1}(\pi) = Q_{t_1}(\pi)/Z(t_1)$ gives the stated form. □

A path gains probability exactly when its growth exceeds the ensemble average of e^G. A growth common to all paths cancels in normalisation and leaves the measure unchanged.

Proposition 10 (Exact marginal-flow update under tilting) Let $I_e(\pi)$ indicate that the path π uses the elementary relation e. Under the assumptions of Theorem 7,

$$F_e(t_1) = \frac{\mathbb{E}_{\mu_{t_0}}\left[I_e\, e^G\right]}{\mathbb{E}_{\mu_{t_0}}\left[e^G\right]}. \tag{C35}$$

If $F_e(t_0) > 0$, then

$$\frac{F_e(t_1)}{F_e(t_0)} = \frac{\mathbb{E}_{\mu_{t_0}}\left[e^G \mid e \in \pi\right]}{\mathbb{E}_{\mu_{t_0}}\left[e^G\right]}. \tag{C36}$$

Thus a relation gains normalised flow if and only if the paths traversing it receive greater-than-average multiplicative amplification.

Proof By definition, $F_e(t_1) = \sum_\pi I_e(\pi)\, \mu_{t_1}(\pi)$, and inserting Theorem 7 gives the first expression. Since $\mathbb{E}_{\mu_{t_0}}[I_e\, e^G] = F_e(t_0)\, \mathbb{E}_{\mu_{t_0}}[e^G \mid e \in \pi]$, dividing by $F_e(t_0)$ gives the second. □

Theorem 8 (Finite entropy production under exponential selection, restated). *Under the assumptions of Theorem 7, let $\mu_a(\pi) \propto \mu_{t_0}(\pi)\, e^{aG(\pi)}$ for $0 \le a \le 1$, and let*

$S_{\text{deficit}}(a) = \log|\Pi| - H(\mu_a)$. *Then*

$$\frac{dS_{\text{deficit}}}{da} = \mathrm{Cov}_{\mu_a}\big(G, \log \mu_a\big), \qquad S_{\text{deficit}}(t_1) - S_{\text{deficit}}(t_0) = \int_0^1 \mathrm{Cov}_{\mu_a}\big(G, \log \mu_a\big)\, da. \tag{C37}$$

Proof Write $\psi(a) = \log \mathbb{E}_{\mu_{t_0}}[e^{aG}]$, so that $\log \mu_a(\pi) = \log \mu_{t_0}(\pi) + aG(\pi) - \psi(a)$. Differentiating the normalised exponential family gives

$$\frac{d\mu_a(\pi)}{da} = \mu_a(\pi)\big(G(\pi) - \mathbb{E}_{\mu_a}[G]\big). \tag{C38}$$

Since $S_{\text{deficit}} = \log|\Pi| + \sum_\pi \mu_a(\pi) \log \mu_a(\pi)$ and $\sum_\pi d\mu_a(\pi)/da = 0$,

$$\frac{dS_{\text{deficit}}}{da} = \sum_\pi \frac{d\mu_a(\pi)}{da} \log \mu_a(\pi) = \sum_\pi \mu_a(\pi)\big(G(\pi) - \mathbb{E}_{\mu_a}[G]\big) \log \mu_a(\pi), \tag{C39}$$

which is the covariance. Integrating from $a = 0$ to $a = 1$, where $\mu_0 = \mu_{t_0}$ and $\mu_1 = \mu_{t_1}$ by Theorem 7, gives the integral form. □

The deficit therefore increases along the interpolation exactly when growth covaries positively with the current log-probability, that is, when already prominent paths grow faster. A field that favours weak paths decreases the deficit.

Corollary 6 (Infinitesimal replicator form) For a differentiable positive score $Q_t(\pi)$, let $g_t(\pi) = \frac{d}{dt} \log Q_t(\pi)$. Then

$$\dot\mu_t(\pi) = \mu_t(\pi)\big(g_t(\pi) - \mathbb{E}_{\mu_t}[g_t]\big), \qquad \dot S_{\text{deficit}}(t) = \mathrm{Cov}_{\mu_t}\big(g_t, \log \mu_t\big). \tag{C40}$$

Proof Differentiating $\log \mu_t(\pi) = \log Q_t(\pi) - \log Z(t)$ gives $\dot\mu_t(\pi)/\mu_t(\pi) = g_t(\pi) - \dot Z(t)/Z(t)$, and $\dot Z(t)/Z(t) = \mathbb{E}_{\mu_t}[g_t]$. The entropy statement follows as in the proof of Theorem 8, with g_t in place of G. □

The empirical question about an optimiser is therefore not whether these laws hold, since they hold whenever their assumptions are satisfied. It is which growth field the optimiser produces, and whether that field is aligned with the existing path hierarchy strongly and persistently enough to increase concentration.

C.9 Execution-only intervention

An attenuation I_α with dose $0 \le \alpha < 1$ multiplies each selected contribution by $g = 1 - \alpha$ during inference. It rewrites neither the parameter state nor the architectural adjacency, and it recomputes the downstream execution of both members of the condition. The question is which supports the intervened realisation can occupy. Proposition 2 of the main text is restated in its two parts.

Proposition 2(i) (No creation of structural support, restated). *Suppose that a fixed architecture and parameter state admit a common structural carrier with score Q_t^S, and that the pre-intervention and post-intervention conditional scores factor through it,*

$$Q_t^c(\pi) = Q_t^S(\pi)\, R_{c,t}(\pi), \qquad \widetilde Q_{t,\alpha}^c(\pi) = Q_t^S(\pi)\, \widetilde R_{c,t,\alpha}(\pi), \tag{C41}$$

with finite non-negative realisation factors. Then $\operatorname{supp} \widetilde{\mu}^c_{t,\alpha} \subseteq \operatorname{supp} \mu^S_t$.

Proof If $\pi \notin \operatorname{supp} \mu^S_t$, then $Q^S_t(\pi) = 0$. Hence $\widetilde{Q}^c_{t,\alpha}(\pi) = 0$, and π cannot belong to the support of the post-intervention measure. □

Execution-only attenuation can therefore reorganise realisation within structural possibility. It cannot place normalised flow on a structurally unsupported carrier.

Proposition 11 (Post-intervention support need not be nested in conditional support) Under the assumptions of Proposition 2(i), it is not true in general that $\operatorname{supp} \widetilde{\mu}^c_{t,\alpha} \subseteq \operatorname{supp} \mu^c_t$. A carrier that was structurally available but conditionally unrealised before intervention can be realised after it.

Proof Consider a downstream ReLU unit with preactivation $z = b + a_1 + a_2$, where a_1 is the contribution selected for attenuation and a_2 another incoming contribution. Choose values with $z < 0$ and $a_1 < 0$, so that the unit is inactive before intervention. After attenuation, $\widetilde{z} = b + (1 - \alpha)a_1 + a_2$, which becomes positive for sufficiently large α. Every structurally available carrier through this unit that had zero response before intervention can then acquire a positive response. No parameter edge has been created. The intervention has changed which part of the existing structural support is realised. □

Both realisations are therefore constrained to structural possibility, $\operatorname{supp} \mu^c_t \subseteq \operatorname{supp} \mu^S_t$ and $\operatorname{supp} \widetilde{\mu}^c_{t,\alpha} \subseteq \operatorname{supp} \mu^S_t$. The second support need not lie inside the first. Intervention can expose a latent structural route, but it cannot manufacture a new one.

Proposition 2(ii) (Zero-intervention identity, restated). *For* $\alpha = 0$*, the gate is* $g = 1$ *and, in exact arithmetic, the intervened execution is identical to ordinary inference. Whenever the conditional measure is defined,* $\widetilde{\Delta f_{i_L}} = \Delta f_{i_L}$ *and* $\widetilde{\mu}^c_{t,0} = \mu^c_t$*, and every deterministic descriptor computed from the execution agrees exactly.*

Proof With $\alpha = 0$, every selected contribution is multiplied by one, so no elementary operation differs from baseline inference. By forward determinism, every downstream state is identical, and so are all finite-response operators, carrier amplitudes, normalised measures and deterministic summaries. □

The identity is the reference against which every non-zero dose is compared. It also serves as a numerical test of the intervention operator.

Proposition 12 (Degeneracy criterion) Let $Z_c(t) = \sum_\pi \left(A^c_{\pi,t}\right)^2$. Then $Z_c(t) = 0$ if and only if $A^c_{\pi,t} = 0$ for every carrier π. In that case the conditional measure is undefined, and it is not the zero measure.

Proof Each summand is non-negative, and a finite sum of non-negative numbers vanishes if and only if every summand vanishes. If $Z_c(t) = 0$, the normalisation $\mu^c_t = Q^c_t / Z_c(t)$ is undefined. Assigning zero mass to every carrier would violate Lemma 1. □

The propositions apply only while the intervention changes execution. Retraining, weight editing, dynamic parameter generation and condition-dependent routing alter the structural score or the carrier itself, and the structural object must then be redefined.

C.10 Geometric descriptors

Section 3.4 introduces the flow-induced ruler, the BCANw cover, the characteristic scale $\xi_{50}(t)$, the crossover descriptors $d^*(t)$ and $x_{d^*}(t)$, and the peak flow lacunarity. This appendix gives the remaining definitions together with the rule under which each descriptor is reported. All geometric descriptors of one training state are computed from a single BCANw cover of the flow-induced metric with $p = 1$, evaluated on a fixed grid of 64 scales. A descriptor that fails its acceptance rule is recorded as undefined and is never set to zero.

Microscopic heterogeneity.

Within a conserved cut C, the enrichment $\mathcal{F}_e(t) = |C|\, F_e(t)$ has mean one (Proposition 4), so its dispersion measures learnt inequality independently of layer width. The edge-flow heterogeneity $\mathrm{CV}_{\mathrm{edge}}(t)$ is the coefficient of variation of $\mathcal{F}_e(t)$ over the edges, and the node-flow heterogeneity $\mathrm{CV}_{\mathrm{node}}(t)$ is the same statistic for the node enrichment $\mathcal{F}_i(t) = |V_\ell|\, F_i(t)$. Their folds are final over initial values.

The tail of the distribution of $\mathcal{F}_e(t)$ is examined by model comparison rather than by assuming a power law. Four models are fitted to the same tail, namely a pure power law, a truncated power law, a lognormal and an exponential, and their Bayesian information criteria are compared descriptively. A preferred model therefore describes the tail and does not identify a generative mechanism. Pooled fits use all edges of a network, and layer-resolved fits are reported separately.

Normalised covering and its collapse.

On the intrinsic coordinate, the normalised covering function of a training state is

$$g_t(x) = \frac{N_B\big(x\,\xi_{50}(t); t\big)}{N}. \tag{C42}$$

The covering family of one architecture collapses when $g_t(x) \approx G(x)$ across checkpoints for a single function G. A pure power law $G(x) \propto x^{-d}$ is only a special case of such a collapse, so collapse is examined before any exponent is fitted. The master-curve error is the root-mean-square deviation of $\log_{10} g_t(x)$ from $\log_{10} G(x)$ over the common grid of x, and it is reported in dex. Because a BCANw count is a staircase in ℓ_B, repeated plateaus are compressed before $d^F_{\mathrm{eff}}(x; t)$ is differentiated. The maximum $d^*(t)$ is therefore supported by a genuine transition of the box count and is not produced by the boundary of a plateau.

Finite scaling window.

An interval $[x_1, x_2]$ is treated as a credible single-exponent regime, and the slope $d_B^F(t)$ of the fit $\log N_B \approx a - d_B^F(t) \log x$ is reported, only if the fit passes every one of the following safeguards:

1. the interval contains at least six distinct box-count values;
2. it spans at least 0.5 decade of scale;
3. the fitted slope satisfies $d_B^F(t) \geq 0.25$;
4. N_B changes at least fivefold across the interval;
5. the log-log fit has $R^2 \geq 0.97$;
6. the coefficient of variation of the local slope inside the interval is at most 0.25;
7. a quadratic log-log alternative is not materially preferred, that is, $\Delta\mathrm{BIC} = \mathrm{BIC}_{\mathrm{linear}} - \mathrm{BIC}_{\mathrm{quadratic}} \leq 2$.

The thresholds define the estimator and carry no theoretical status. They exclude flat staircase segments and windows that straddle a curved crossover, which is why the existence of a window and the value of its exponent are reported as separate results.

Mass on the boxes.

The node flow is normalised to a measure on the geometry-bearing graph, $\nu_v(t) = F_v(t)/\sum_u F_u(t)$. A box B of the cover at scale x carries the mass $M_B(x;t) = \sum_{v\in B} \nu_v(t)$, and the flow lacunarity is

$$\Lambda_F(x;t) = \frac{\langle M_B(x;t)^2 \rangle}{\langle M_B(x;t) \rangle^2}, \tag{C43}$$

where the averages run over the boxes of that cover. Since $\Lambda_F - 1$ is the squared coefficient of variation of the box mass, larger values indicate a more unequal allocation of mass across boxes. The peak flow lacunarity $\Lambda_F^*(t)$ is the maximum of $\Lambda_F(x;t)$ over intermediate scales.

Local support and mass growth.

Around a node i, the metric ball $B_i(r) = \{j : d_F(i,j) \leq r\}$ contains $n_i(r) = |B_i(r)|$ nodes and the mass $m_i(r) = \sum_{j\in B_i(r)} \nu_j(t)$. The local support exponent $d_i^F = d\log n_i / d\log r$ and the local mass exponent $\alpha_i^F = d\log m_i / d\log r$ are fitted on the same radial window, and their difference

$$\Delta_i^F = \alpha_i^F - d_i^F \tag{C44}$$

compares the growth of flow mass with the growth of available support. A negative value means that mass accumulates more slowly than support as the ball expands, so the measure is locally concentrated relative to the available metric volume. A common window is accepted only if it contains at least seven points, spans at least 0.5 decade, gives $R^2 \geq 0.95$, and shows at least a fourfold growth of n_i and a threefold growth of m_i. Fitting both exponents on one window prevents Δ_i^F from subtracting slopes

measured over incomparable intervals. Two summaries are reported, the median of Δ_i^F over valid nodes and the valid local paired fraction $V_\Delta(t)$, which is the fraction of evaluated nodes that admit an accepted common window.

Directed flow after coarse graining.

The boxes of a cover are contracted to the nodes of a quotient graph. The geometry that defines the boxes is undirected, whereas the flow carried between them remains directed. For distinct boxes $A \neq B$, the quotient flow is

$$F'_{AB}(t) = \sum_{u \in A} \sum_{v \in B} F_{uv}(t). \tag{C45}$$

The inter-box retention at a box scale is the fraction of the total directed flow that passes between distinct boxes, $\sum_{A \neq B} F'_{AB}(t) \big/ \sum_{(u,v) \in E} F_{uv}(t)$. The remainder is internal to the boxes. Retention is evaluated at four target box fractions from $N_B/N \approx 0.5$ to $N_B/N \approx 0.0625$, and the mean inter-box retention is its average over these resolutions. To compare the shape of the coarse flow across resolutions, each quotient flow is normalised by its mean, $X_{AB}(\ell) = F'_{AB}/\langle F' \rangle_\ell$. The quotient-flow KS statistic is the mean pairwise Kolmogorov-Smirnov distance between the samples $X(\ell)$ of the four resolutions. A small value indicates approximate distributional collapse under quotienting, and this statement does not require a power law.

The covering staircase.

The transition fraction is the number of distinct compressed box-count plateaus divided by the 64 raw scales. The maximum plateau span is the longest interval, in decades of scale, over which the compressed box count does not change. Both describe the resolution structure of the cover rather than the allocation of mass or flow.

Sharing decomposition in CNNs.

For a hidden transition ℓ of a CNN, let A_ℓ be the execution occurrence traversed by a carrier path drawn from the structural measure, and let $\Theta_\ell = \rho(A_\ell)$ be its parameter identity. Write $F_a(t)$ for the flow through occurrence a and $F_\vartheta(t) = \sum_{a:\rho(a)=\vartheta} F_a(t)$ for the flow through identity ϑ. By Theorem 5, these entropies satisfy the exact identity

$$H^{(\ell)}_{\text{occ}} = H^{(\ell)}_{\text{param}} + H^{(\ell)}_{\text{reuse}}, \qquad H^{(\ell)}_{\text{reuse}} = H(A_\ell \mid \Theta_\ell) = -\sum_a F_a(t) \log \frac{F_a(t)}{F_{\rho(a)}(t)}, \tag{C46}$$

where $H^{(\ell)}_{\text{occ}} = H(A_\ell)$ and $H^{(\ell)}_{\text{param}} = H(\Theta_\ell)$. The reuse entropy measures how uncertain the spatial site of execution remains once the parameter identity is known. Each entropy is turned into a specialisation by subtracting it from its uniform value,

$$S^{(\ell)}_{\text{occ}} = \log n^{(\ell)}_{\text{occ}} - H^{(\ell)}_{\text{occ}}, \qquad S^{(\ell)}_{\text{param}} = \log n^{(\ell)}_{\text{param}} - H^{(\ell)}_{\text{param}}, \qquad S^{(\ell)}_{\text{reuse}} = \log m_\ell - H^{(\ell)}_{\text{reuse}}, \tag{C47}$$

where $n^{(\ell)}_{\text{occ}}$ and $n^{(\ell)}_{\text{param}}$ count the occurrences and identities, and m_ℓ is the reuse multiplicity of each identity. For a stride-one transition in which every identity has

Table D1: The four acquisition procedures of the structural programme.

Procedure	Discovery population	Frozen	Information sought
MLP architecture	1,008 realised architectures (600 Sobol, 400 SMAC, 8 replays)	48	Novelty of the initial complete-path and BCANw phenotype
CNN architecture	1,008 architectures (600 Sobol, 400 adaptive, 8 references)	48	Novelty of learning-induced structural change and of configuration
MLP environment	36 Sobol and 24 adaptive environments	24	Uncertainty, structural-response novelty and sentinel disagreement
CNN environment	34 Sobol, 2 anchors and 24 adaptive environments	24	The same criteria, evaluated through CNN sentinels

the same number of execution sites, $n_{\mathrm{occ}}^{(\ell)} = n_{\mathrm{param}}^{(\ell)} m_\ell$, so that $S_{\mathrm{occ}}^{(\ell)} = S_{\mathrm{param}}^{(\ell)} + S_{\mathrm{reuse}}^{(\ell)}$. An increase of S_{param} therefore records specialisation over which shared coefficients carry flow. An increase of S_{reuse} records specialisation over the spatial sites at which a coefficient is used.

Appendix D Experimental Protocols and Ledgers

D.1 AutoML-based Bayesian acquisition

D.1.1 Design principle

We investigate how learning organises computation through the evolution of the parameter state. A population restricted to a few conventional networks could make the observed relations reflect particular depths, widths, connectivity patterns or sharing arrangements. The environment population faces the same problem, and it must remain separate from the finite input conditions under which a learnt disposition is later realised. We therefore construct both populations by AutoML-based Bayesian acquisition. Each search begins with a space-filling design and continues with surrogate-guided acquisition. The acquired quantity depends on the scientific question of the experiment, so the four procedures share a design principle without sharing one objective (Table D1).

Conventional hyperparameter optimisation seeks $x^\star = \arg\min_{x \in \mathcal{X}} L_{\mathrm{val}}(x)$ and so concentrates on configurations expected to perform well. A structural law can be tested only on a population that keeps the variation such a search removes. We therefore distinguish three populations. The discovery population contains the configurations evaluated during acquisition. The frozen population contains the configurations selected for controlled longitudinal evaluation. The statistical population identifies the units over which comparisons and tests are performed. Discovery observations are adaptively selected, and they are never treated as inferential replicates,

$$\text{adaptive search} \neq \text{statistical replication} \neq \text{population sampling}. \tag{D48}$$

D.1.2 MLP architecture acquisition

Architecture space.

We consider bias-free, fully connected ReLU MLPs with input dimension 32 and a scalar output. A configuration is

$$\mathsf{A} = (L, P_{\text{target}}, m, \omega, \tau), \tag{D49}$$

where L is the hidden depth, P_{target} the target parameter budget, m the width morphology, and ω and τ the strength and curvature of the width profile. The search covers $9 \le L \le 32$ and $2 \times 10^4 \le P_{\text{target}} \le 1.8 \times 10^5$, with hidden widths restricted to $[8, 512]$. For the normalised layer position $z_j = j/(L-1)$, $j = 0, \dots, L-1$, the five morphologies have the relative profiles

$$\begin{aligned}
&\text{constant: } p_j = 1, \qquad \text{taper: } p_j = \omega^{z_j^\tau}, \qquad \text{expansion: } p_j = \omega^{(1-z_j)^\tau}, \\
&\text{bottleneck: } p_j = 1 - (1-\omega)\exp\Big[-\tfrac{(j-(L-1)/2)^2}{2(L/6)^2}\Big], \\
&\text{hourglass: } p_j = \omega + (1-\omega)\,|2z_j - 1|^\tau,
\end{aligned} \tag{D50}$$

with a lower bound on the Gaussian width for shallow bottlenecks. For the non-constant families, $\omega \in [0.45, 0.95]$ and $\tau \in [0.60, 2.00]$. A multiplicative scale is searched so that, after rounding to integer widths $\mathbf{h} = (h_1, \dots, h_L)$, the realised count

$$P(\mathsf{A}) = 32h_1 + \sum_{j=1}^{L-1} h_j h_{j+1} + h_L \tag{D51}$$

minimises its logarithmic discrepancy from P_{target}.

Admissibility.

Admissibility is tested on the realised graph rather than the nominal configuration. The median hidden width must be at least 18 and the tenth-percentile width at least 12. No more than 20% of hidden layers may sit at either width limit. The realised count must lie between 0.80 and 1.25 times the target, and the graph must contain at most 6,800 nodes. Duplicate realised width sequences are rejected.

Initial structural fingerprint.

Each candidate is characterised before learning. Under a common Kaiming initialisation with one model seed across configurations, we compute the structural measure $\mu_0^S(\pi) \propto \prod_{e\in\pi} w_e(0)^2$, its effective occupancies, its conserved flows and the BCANw organisation of the induced metric. These define a fingerprint $\mathbf{s}_0(\mathsf{A}) \in \mathbb{R}^{17}$. It contains the logarithmic Shannon and order-two occupancies, the edge-flow coefficient of variation, the scale ratios ξ_{25}/ξ_{50} and ξ_{10}/ξ_{50}, the finite-window exponent with its span and credibility indicator, the number of covering crossovers, the locations, slopes and prominences of the first two crossover peaks, and the location and magnitude of the peak flow lacunarity. Architectures with similar size or occupancy can therefore be

distinguished by their covering transitions, characteristic scales and coarse mass distributions. Missing or inadmissible coordinates are imputed by the medians of the Sobol population, and all coordinates are standardised by its means and standard deviations,

$$\mathbf{z}_0(\mathsf{A}) = \text{Standardise}\big[\,\text{Impute}(\mathbf{s}_0(\mathsf{A}))\big]. \tag{D52}$$

This transformation is fixed once the Sobol population is complete.

Sobol design.

Six depth strata, four logarithmic budget strata and the five morphologies define 120 cells. Within each cell, a scrambled four-dimensional Sobol sequence assigns the depth, logarithmic budget, strength and curvature of five configurations, which gives $120 \times 5 = 600$ architectures. An inadmissible or duplicate point is replaced by a new draw within the same cell, so all 120 cells are covered. The 600 fingerprints form the fixed reference cloud $\mathcal{Z}_0 = \{\mathbf{z}_0(\mathsf{A}_i)\}_{i=1}^{600}$.

Discovery score.

Structural novelty is the mean distance to the three nearest members of $\mathcal{Z}_0$, excluding the candidate itself when it belongs to the cloud, mapped to a bounded score,

$$N^{\text{raw}}_{\text{struct}}(\mathsf{A}) = \frac{1}{3} \sum_{\mathsf{B} \in \mathcal{N}_3(\mathsf{A};\mathcal{Z}_0)} \big\|\mathbf{z}_0(\mathsf{A}) - \mathbf{z}_0(\mathsf{B})\big\|_2, \qquad N_{\text{struct}}(\mathsf{A}) = 1 - \exp\big[-N^{\text{raw}}_{\text{struct}}(\mathsf{A})\big]. \tag{D53}$$

The configuration novelty $N_{\text{cfg}}(\mathsf{A})$ is the distance to the nearest evaluated configuration in the normalised coordinate space, with the morphology as an ordinal index and the budget on a logarithmic scale. The coverage reward is $C_{\text{cover}}(\mathsf{A}) = 1/(1+n_{\text{cell}}(\mathsf{A}))$, where n_{cell} counts the configurations already evaluated in the cell. These define

$$J_{\text{MLP}}(\mathsf{A}) = 0.60\, N_{\text{struct}}(\mathsf{A}) + 0.25\, N_{\text{cfg}}(\mathsf{A}) + 0.15\, C_{\text{cover}}(\mathsf{A}), \qquad \mathcal{L}_{\text{MLP}}(\mathsf{A}) = 1 - J_{\text{MLP}}(\mathsf{A}). \tag{D54}$$

The largest weight falls on the novelty of the initial structural phenotype. The cost $\mathcal{L}_{\text{MLP}}$ is the scalar objective supplied to the surrogate.

SMAC refinement.

The Bayesian stage uses SMAC with a 96-tree random-forest surrogate. The 600 Sobol evaluations are replayed into its history, so acquisition begins from an evaluated phenotype population. With surrogate mean $\mu(\mathsf{A})$, uncertainty $\sigma(\mathsf{A})$ and best observed cost $\mathcal{L}^*$, expected improvement in minimisation form is

$$\text{EI}(\mathsf{A}) = I(\mathsf{A})\,\Phi\big(z(\mathsf{A})\big) + \sigma(\mathsf{A})\,\phi\big(z(\mathsf{A})\big), \qquad I(\mathsf{A}) = \mathcal{L}^* - \mu(\mathsf{A}) - \xi, \quad z(\mathsf{A}) = \frac{I(\mathsf{A})}{\sigma(\mathsf{A})}, \tag{D55}$$

with $\xi = 0.15$. Its precise evaluation is delegated to the SMAC engine. Four controls prevent local collapse. A quarter of proposals come from a random design. Every fourth accepted proposal must come from one of the least-covered cells. A hard guard

excludes cells whose occupancy exceeds that of the least-covered cell by more than three. Inadmissible or duplicate proposals receive a penalty cost of 1.25, which informs the surrogate without a BCANw evaluation. The search stops after 400 accepted acquisitions. The execution record shows the SMAC engine, 400 accepted acquisitions, 1,911 rejected or penalised proposals and coverage of all 120 cells. Eight further configurations arising during the cold-start replay contribute distinct architectures, so the deduplicated map contains 1,008 realised architectures. Occupancy per cell has median 8 and range 6 to 14, and the map spans 19,786 to 179,586 parameters.

Selection of the 48 architectures.

Selection preserves variation across the initial phenotype space by greedy maximin on the fixed fingerprints. With $\bar{\mathbf{z}}_0$ the mean fingerprint of the map, the first architecture is the one closest to the centre and the second the one farthest from it,

$$\mathsf{A}_1 = \arg\min_{\mathsf{A}} \|\mathbf{z}_0(\mathsf{A}) - \bar{\mathbf{z}}_0\|_2, \qquad \mathsf{A}_2 = \arg\max_{\mathsf{A}\neq\mathsf{A}_1} \|\mathbf{z}_0(\mathsf{A}) - \bar{\mathbf{z}}_0\|_2, \tag{D56}$$

and every further architecture maximises its distance to the selected set S_j,

$$\mathsf{A}_{j+1} = \arg\max_{\mathsf{A}\notin S_j} \min_{\mathsf{B}\in S_j} \left\|\mathbf{z}_0(\mathsf{A}) - \mathbf{z}_0(\mathsf{B})\right\|_2, \tag{D57}$$

until 48 are selected. The cohort is frozen before training. All 48 architectures follow the same 400-epoch protocol, and none is removed for its later concentration, performance or geometry.

D.1.3 CNN architecture acquisition

Carrier and architecture space.

In a CNN, one trainable coefficient takes part in many spatial occurrences, and the structural measure is defined on the paths of the expanded execution graph (Section 3.1). Two CNNs with similar parameter counts can therefore differ in their execution graphs, receptive fields and reuse multiplicities. Architectural diversity must account for spatial execution as well as depth and channel width. A configuration is

$$\mathsf{A} = (m, L, \beta, k, \omega, \tau, z_c), \tag{D58}$$

namely the channel morphology, the number of trainable convolutional layers $3 \leq L \leq 7$, a compute factor $\beta \in \{0.5, 1, 2, 4\}$ relative to the multiply-accumulate cost of a reference CNN, the kernel family k, the profile strength and curvature, and the profile centre. The kernel families use 3×3 hidden convolutions throughout, alternate 3×3 and 1×1 convolutions, or place the spatial convolutions before the pointwise ones. A

final 1×1 convolution produces the scalar map. The morphology profiles are

$$\begin{aligned}
&\text{constant: } p_j = 1, \qquad \text{expansion: } p_j = e^{\omega\tau(z_j - 1/2)}, \qquad \text{taper: } p_j = e^{-\omega\tau(z_j - 1/2)},\\
&\text{bottleneck: } p_j = 1 - \omega \exp\Big[- \tfrac{(z_j - z_c)^2}{2\sigma^2} \Big], \quad \sigma = 0.10 + \tfrac{0.16}{\tau},\\
&\text{hourglass: } p_j = 1 - \omega\,[4 z_j (1 - z_j)]^{\tau}.
\end{aligned} \tag{D59}$$

Each profile is clipped below at 0.18 and normalised by its geometric mean. A search over integer base widths from 2 to 40 fixes the realised channel sequence. A candidate is admissible when its output resolution is at least 4×4 and its expanded graph has at most 6,500 nodes. The realised channel and kernel sequences define its identity, so nominally distinct configurations that realise the same CNN are rejected.

Sobol design and references.

Five morphologies, five depths, four compute factors and three kernel families define 300 strata, each receiving two configurations, which gives 600. A scrambled three-dimensional Sobol sequence sets $\omega = 0.12 + 0.76u_1$, $\tau = 0.50 + 1.80u_2$ and $z_c = 0.25 + 0.50u_3$, and the compute factor is perturbed within its stratum as $\beta_{\text{eff}} = \beta \exp[0.12(u_3 - 0.5)]$. Where rounding or the execution limits prevent enough unique configurations, admissible random configurations supplement the design and are marked as replacements. Eight named reference and control architectures are evaluated before refinement, so that the surrogate also sees known constructions.

Learning-induced structural response.

Unlike the MLP search, the CNN search measures novelty after learning. Every candidate is trained for 60 epochs on one canonical synthetic spatial-regression environment, with 16×16 inputs, ReLU, Adam at learning rate 3×10^{-3}, no weight decay, 4,000 training and 1,000 validation observations, and one reproducible seed per architecture. The structural measure is computed at initialisation and after this exposure, and the response vector is

$$\mathbf{r}_{\text{CNN}}(\mathsf{A}) = \big(\Delta S_{\text{def}},\ \Delta \log \eta_1,\ \Delta \log \eta_2,\ \Delta \overline{\text{CV}}_{\text{occ}},\ \Delta \overline{\text{CV}}_{\text{param}},\ \Delta \overline{S}_{\text{param}},\ \Delta \overline{S}_{\text{reuse}}\big)^{\top}. \tag{D60}$$

The first three coordinates describe the redistribution of the path measure. The next two describe flow heterogeneity across occurrences and across shared identities. The last two describe the parameter-identity and spatial-reuse contributions to the entropy deficit (Section 5.3). Training and validation losses are retained as diagnostics and are not acquisition objectives.

Discovery score.

Response coordinates are robustly scaled on the current archive $\mathcal{D}_n^{\text{CNN}}$. The phenotype novelty is the median distance to the eight nearest archived responses, $N_{\text{phen}}(\mathsf{A}) = \text{median}_{\mathsf{B} \in \mathcal{N}_8(\mathsf{A})} \|\tilde{\mathbf{r}}(\mathsf{A}) - \tilde{\mathbf{r}}(\mathsf{B})\|_2$. The configuration novelty N_{cfg} is the minimum distance to archived configurations in an eight-dimensional encoding of normalised depth, logarithmic count of multiply-accumulate operations, logarithmic parameter count,

ordinal morphology and kernel family, strength, scaled curvature and centre. The coverage reward is $C_{\mathrm{cover}} = 1/\sqrt{1 + n_{\mathrm{cell}}}$. Both novelties are transformed by the medians and interquartile ranges of their earlier values V_n,

$$T_n(v) = \max\left\{0,\ 1.5 + \frac{v - \mathrm{median}(V_n)}{\max[\mathrm{IQR}(V_n), 10^{-8}]}\right\}, \tag{D61}$$

and the discovery score is

$$J_{\mathrm{CNN}}(\mathsf{A}) = 0.55\, T_n[N_{\mathrm{phen}}(\mathsf{A})] + 0.25\, T_n[N_{\mathrm{cfg}}(\mathsf{A})] + 0.20\, C_{\mathrm{cover}}(\mathsf{A}). \tag{D62}$$

Like J_{MLP}, this score depends on the archive, and its normalisation changes as the archive grows.

Random-forest expected improvement.

The search follows SMAC in a direct implementation. A 160-tree random forest predicts J_{CNN} from the encoded configuration and is refitted after each accepted evaluation. Its tree predictions give the mean $\hat{\mu}_n(\mathsf{A})$ and the uncertainty $\hat{\sigma}_n(\mathsf{A})$. With J_n^* the largest observed score and $\xi = 0.03$, expected improvement in maximisation form is

$$\mathrm{EI}_n(\mathsf{A}) = I_n\, \Phi\Big(\frac{I_n}{\hat{\sigma}_n}\Big) + \hat{\sigma}_n\, \phi\Big(\frac{I_n}{\hat{\sigma}_n}\Big), \qquad I_n(\mathsf{A}) = \hat{\mu}_n(\mathsf{A}) - J_n^* - \xi, \tag{D63}$$

with a numerical floor on $\hat{\sigma}_n$. Candidates are ranked by

$$\mathcal{A}_{\mathrm{CNN,arch}}(\mathsf{A}) = \mathrm{EI}_n(\mathsf{A}) + 0.12\, N_{\mathrm{cfg}}(\mathsf{A}) + 0.08\, C_{\mathrm{cover}}(\mathsf{A}) \tag{D64}$$

over a pool of up to 768 admissible configurations. At 75% of steps the top candidate is selected, and at the remaining 25% a candidate is drawn with probability proportional to its coverage reward. Each selected architecture receives the 60-epoch exposure, its response and score are archived, and the surrogate is refitted. After 400 accepted acquisitions, the discovery population contains $600 + 400 + 8 = 1{,}008$ architectures, distinct from the MLP map despite the equal count.

Discovery geometry and selection of the 48 architectures.

A subset of 240 discovered architectures, chosen by maximin coverage of configuration and response and including the references, receives BCANw evaluation at initialisation and after the exposure. These geometric measurements characterise the discovered transformations but do not enter the surrogate. For the longitudinal cohort, each viable candidate is represented by its configuration vector and its robustly scaled response, $\mathbf{z}_{\mathrm{CNN}}(\mathsf{A}) = [\mathbf{v}(\mathsf{A}), \tilde{\mathbf{r}}_{\mathrm{CNN}}(\mathsf{A})] \in \mathbb{R}^{15}$. Six named references are retained when viable, and the remaining members follow the maximin rule of Equation (D57) with $\mathbf{z}_{\mathrm{CNN}}$ in place of $\mathbf{z}_0$. The cohort thus preserves separation in both configuration and learning-induced response, whereas the MLP cohort is selected on initial fingerprints. The frozen design comprises $48 \times 2 = 96$ trajectories over 360 epochs, with structural states at epochs $\{0, 1, 2, 5, 10, 20, 40, 80, 120, 160, 240, 360\}$ and geometry at $\{0, 20, 80, 160, 240, 360\}$. A nested 12-architecture panel and a six-architecture sentinel set are selected for the later

experiments. Selection is a coverage procedure and does not retain the highest-scoring architectures.

D.1.4 MLP environment acquisition

Environment space.

For a fixed architecture, the environment acts through the parameter trajectory, $e \mapsto \theta_{a,e,s}(t) \mapsto \mu^S_{a,e,s}(t)$, and does not enter the structural measure as an input condition. An MLP environment is

$$e = (\mathcal{T}, d_{\text{inf}}, r, \text{SNR}, \kappa), \tag{D65}$$

with a linear, shallow nonlinear or deep compositional teacher $\mathcal{T}$, $4 \le d_{\text{inf}} \le 32$ informative inputs, feature correlation $0 \le r \le 0.90$, signal-to-noise ratio $2 \le \text{SNR} \le 50$ sampled logarithmically, and interaction strength $0 \le \kappa \le 1$. The ambient dimension is 32, with 60,000 training, 10,000 validation and 10,000 test observations. Three teacher families, four informative-dimension strata and three SNR strata give 36 cells, each receiving one scrambled Sobol configuration.

Sentinels and structural response.

The 12-architecture handoff panel is selected on standardised architectural and structural descriptors, seeded with extreme representatives of the morphology families and completed by farthest-point coverage. A further farthest-point selection within this panel gives four sentinels. Each candidate environment trains the sentinels for 80 epochs, and each sentinel contributes the response

$$\begin{aligned}\mathbf{r}_{\text{MLP}}(\mathsf{A}, e) = \big(&\Delta \log \eta_1,\ \Delta \log \eta_2,\ \Delta S_{\text{def}},\ \Delta \text{CV}_E,\ \Delta \text{CV}_V,\\ &\Delta \log \xi_{50},\ \Delta d_B,\ \Delta d^*,\ \Delta x_{d^*},\ \Delta \log \Lambda^*_F\big)^\top.\end{aligned} \tag{D66}$$

The first five coordinates describe the path measure and its flows, and the remaining five the induced geometry. An environment can produce familiar concentration while changing the characteristic scale, the crossover strength or the coarse distribution of mass. Concatenating the four sentinels gives $\mathbf{R}(e) \in \mathbb{R}^{4\times 10}$, a 40-dimensional response.

Acquisition.

A multi-output random forest with 160 trees, minimum leaf size 2, bootstrap sampling and feature fraction 0.8 predicts $\widehat{\mathbf{R}}(e) = \hat{f}_n(\mathbf{x}(e))$ from the encoded environment $\mathbf{x}(e)$, which combines the normalised continuous coordinates with a one-hot teacher family. Four quantities are computed for each candidate. The uncertainty is the mean tree dispersion across response coordinates, $U(e) = \text{mean}_j \, \text{SD}_b[\widehat{R}^{(b)}_j(e)]$. The environment novelty is $N_{\text{env}}(e) = d_e^{-1/2} \min_{e' \in \mathcal{E}_n} \|\mathbf{x}(e) - \mathbf{x}(e')\|_2$, with d_e the encoding dimension. The predicted structural novelty $\widehat{N}_{\text{struct}}(e)$ is the mean distance from the standardised predicted response to its three nearest observed responses. The predicted disagreement is the mean pairwise distance between the standardised sentinel predictions,

$$\widehat{I}_{A\times E}(e) = \frac{2}{K(K-1)} \sum_{a<a'} \big\|\tilde{\mathbf{r}}_a(e) - \tilde{\mathbf{r}}_{a'}(e)\big\|_2, \qquad K = 4. \tag{D67}$$

This is a discovery quantity, distinct from the interaction effects later estimated on the frozen factorial. Each component is mapped to $[0, 1]$ by a robust transformation $\mathcal{R}$ between the fifth and ninety-fifth percentiles of the candidate pool, and

$$\mathcal{A}_{\mathrm{MLP,env}}(e) = 0.30\,\mathcal{R}[U(e)] + 0.25\,\mathcal{R}[N_{\mathrm{env}}(e)] + 0.25\,\mathcal{R}[\widehat{N}_{\mathrm{struct}}(e)] + 0.20\,\mathcal{R}[\widehat{I}_{A\times E}(e)]. \tag{D68}$$

No scalar expected improvement is used here. The surrogate predicts a multivariate response, and the acquisition combines uncertainty, novelty and disagreement explicitly. Pools contain up to 4,096 candidates. A quarter of proposals explore under-covered cells at random, every fourth adaptive proposal is directed to the least-covered cells, and a cell-density guard applies. The remaining proposals maximise $\mathcal{A}_{\mathrm{MLP,env}}$. Validation and test losses are measured on the 240 sentinel-environment evaluations but never enter the acquisition.

Freezing the 24 environments.

After 36 Sobol and 24 adaptive evaluations, the panel is built in three stages. Twelve environments are chosen by farthest-point coverage of the environment space, seeded with representatives of the teacher families. Six are added with the largest observed structural novelty, and six more with the largest observed disagreement among the remainder. The frozen panel contains 8 Sobol-origin and 16 adaptive environments. It retains all teacher families (7 linear, 10 shallow nonlinear, 7 deep compositional) and spans $d_{\mathrm{inf}} = 4$ to 31, $r = 0.003$ to 0.898, SNR = 2.04 to 48.73 and $\kappa = 0.004$ to 0.961. Selection raises the median structural novelty from 5.879 to 6.593 and the median disagreement from 4.263 to 4.859. The full experiment comprises $24 \times 12 \times 2 = 576$ trajectories over 400 epochs.

D.1.5 CNN environment acquisition

Environment space and sentinels.

A CNN environment is

$$e = (\mathcal{T}, q_{\mathrm{info}}, r, \mathrm{SNR}, \lambda_{\mathrm{int}}, \lambda_{\mathrm{loc}}), \tag{D69}$$

with a linear-spatial, shallow nonlinear or deep compositional teacher, informative spatial fraction $0.10 \le q_{\mathrm{info}} \le 1$, spatial correlation $0 \le r \le 0.90$, $2 \le \mathrm{SNR} \le 50$, and interaction strength and localisation in $[0, 1]$. The initial population contains 34 scrambled Sobol configurations and two canonical anchors, one predominantly distributed and one strongly localised, and the adaptive stage adds 24, so that the map contains 60 environments. Four sentinels are chosen from the frozen 12-architecture panel by maximin separation on architectural and response descriptors. Each candidate trains them for 80 epochs on 6,000 training and 2,000 validation observations, with architecture, activation, optimiser and exposure fixed, and structural states recorded at epochs $\{0, 5, 20, 80\}$. Each sentinel yields the seven-coordinate response $\mathbf{r}_{\mathrm{CNN}}(\mathsf{A}, e)$ defined above. The responses are summarised by their coordinatewise medians, and the observed disagreement is the median pairwise distance between robustly scaled sentinel responses.

Acquisition.

Two random forests, each with 240 trees, minimum leaf size 2 and feature fraction 0.8, predict the observed structural novelty $\widehat{N}_{\mathrm{struct}}(e)$ and the observed disagreement $\widehat{I}_{A\times E}(e)$. The uncertainty combines their tree dispersions, $U(e) = (\sigma_N(e) + \sigma_I(e))/2$, and the environment novelty is $N_{\mathrm{env}}(e) = \min_{e'\in\mathcal{E}_n} \|\mathbf{x}(e) - \mathbf{x}(e')\|_2$. With the same robust transformation,

$$\mathcal{A}_{\mathrm{CNN,env}}(e) = 0.30\,\mathcal{R}[U(e)] + 0.25\,\mathcal{R}[N_{\mathrm{env}}(e)] + 0.25\,\mathcal{R}[\widehat{N}_{\mathrm{struct}}(e)] + 0.20\,\mathcal{R}[\widehat{I}_{A\times E}(e)]. \tag{D70}$$

The weights match the MLP procedure, but the surrogate targets and estimators differ. The MLP search predicts the full response matrix, whereas the CNN search predicts novelty and disagreement separately, so equal weights do not imply equal predictions or uncertainties. Each step draws a pool of up to 1,024 admissible candidates. At 75% of steps the top candidate is selected, and at the remaining 25% a candidate is drawn at random from the most novel fifth of the pool. The archive is updated after each evaluation, until 24 adaptive environments have been evaluated.

Freezing the 24 environments.

Eight initial environments, including the two anchors, are chosen by maximin coverage of the environment space. The remaining sixteen are chosen greedily from the adaptive population by

$$J_{\mathrm{panel}}(e) = 0.40\,\widetilde{N}_{\mathrm{env}}(e) + 0.35\,\widetilde{N}_{\mathrm{struct}}(e) + 0.25\,\widetilde{I}_{A\times E}(e), \tag{D71}$$

where the environment novelty is taken relative to the current panel and all three terms are robustly scaled over the remaining candidates at each step. The frozen panel therefore contains 2 anchors, 6 Sobol and 16 adaptive environments. The full experiment comprises $24 \times 12 \times 2 = 576$ trajectories over 360 epochs, with geometry at epochs $\{0, 80, 160, 360\}$ and additional intermediate geometry for the sentinels.

D.1.6 Conditional cohorts, optimiser panels and benchmarks

MLP conditional cohort.

A network can be structurally informative yet fail to give a usable finite response after training. The MLP conditional cohort is therefore acquired with a functional-viability gate. Its discovery population contains 32 Sobol and 48 adaptive architectures, with depths 3 to 20, budgets of 1.5×10^4 to 1.2×10^6 and six morphologies including an irregular-width family. Each candidate is trained for 120 epochs on a fixed synthetic regression with ReLU, Adam and one common seed. The objective is the best validation error, $\mathcal{L}_{\mathrm{C1}}(\mathsf{A}) = \min_{0\le t\le 120} \mathrm{MSE}_{\mathrm{val}}(\mathsf{A}, t)$, minimised by a 320-tree random forest with expected improvement ($\xi = 0.01$) and a quarter of proposals random. Candidates are then restricted by a viability threshold at the 75th percentile of the finite objective values. Twenty architectures are selected by maximin separation of standardised descriptors, beginning with a maximally separated pair and adding representatives of missing morphologies. Performance therefore decides eligibility, while coverage decides

membership. The cohort comprises $20 \times 8 = 160$ ReLU trajectories of 280 epochs, together with six morphology-spanning architectures under tanh with the same eight seeds, which gives 208 trajectories.

CNN conditional cohort.

The CNN conditional experiment performs no new search. It inherits the 12 frozen architectures, which are trained on the canonical spatial-regression environment with the same optimiser, activation, horizon and input resolution, giving $12 \times 4 = 48$ trajectories. Conditional panels and geometry are applied only after this population is fixed.

Optimiser panels.

The optimiser experiments inherit their panels from the earlier phases. For the MLPs, six architectures and six environments are selected by maximin separation of their descriptors and structural responses, and five of each form the optimiser subset. The Adam arm contains $6 \times 6 \times 5 = 180$ trajectories, and the nested optimiser arm contains $5 \times 5 \times 4 \times 5 = 500$, sharing 125 Adam cells, so that 555 unique trajectories were intended. Adam is locked at learning rate 10^{-3}. The comparators are calibrated on training loss alone, on a maximin 2×2 subset with seed 0 for 12 epochs, giving 3×10^{-4} for RMSProp, 3×10^{-3} for SGD with momentum 0.9 and 10^{-2} for SGD. Validation loss, test loss and every path-flow descriptor are excluded from calibration. Within a block, all optimisers share the initialisation and the minibatch order. Of the 555 trajectories, 543 completed and 12 diverged, and the 118 blocks completed by all four optimisers enter the matched comparison. For the CNNs, the 576 environment blocks are each crossed with four optimisers, giving 2,304 trajectories with no divergences. Adam is locked at 0.003, RMSProp uses 0.003 and both SGD variants 0.03, from a global training-loss calibration blind to structure. The optimiser experiments are therefore inherited-panel comparisons with a limited operating-point calibration, not further Bayesian searches.

Benchmarks and interventions.

The 324 benchmark networks form declared factorials on MPI3D and Fashion-MNIST (Appendix E.9). The intervention experiments act on these frozen networks. Their targets are set by the structural or realised flows, never by acquisition scores, so the mechanisms that construct populations stay separate from the rules that probe computational responsibility.

D.2 Frozen populations and the scope of inference

The acquisition fixes the coordinates within which the investigations take place. It does not establish the reported laws of concentration, geometric deformation or conditional realisation. These arise only from the frozen evaluations, in the order

$$\begin{aligned}&\text{admissible space} \to \text{Sobol population} \to \text{adaptive acquisition} \to \text{frozen cohort}\\ &\quad \to \text{longitudinal evaluation} \to \text{environment factorial} \to \text{realisation} \to \text{intervention}.\end{aligned} \tag{D72}$$

The two discovery maps of 1,008 architectures are adaptively constructed and are not samples from a prespecified distribution, so they are not 1,008 independent replicates. Architecture-level inference uses the frozen 48-architecture cohorts. For the structural CNN cohort, the two seeds are aggregated within each architecture, and the architecture is the unit. The environment experiments retain seed-level observations within their balanced architecture-environment cells and examine directional recurrence, rank preservation and the decomposition of variation. A maximin-selected cohort supports statements about recurrence across the diverse architectures investigated. Without a further sampling argument, it does not establish how often a property holds over all MLPs or CNNs.

The frozen panels are recorded experimental artefacts. Their exact compositions, including the provenance of every member, are taken from the recorded panel files rather than re-derived from the selection routines, and they are listed in the panel catalogues of this appendix. For the CNN architecture cohort, the recorded provenance is 7 anchors or controls, 21 Sobol architectures and 20 adaptive architectures.

D.3 Design ledger

Table D2 connects each claim of Sections 4 to 6 to the observable that measures it, the exact statement that licenses its interpretation, the population on which it is evaluated and the inferential unit of its test.

D.4 Numerical fidelity and validity

The constructions of Section 3 are exact. Their implementation is therefore audited by residuals that must vanish in exact arithmetic, and these residuals are treated as consistency checks rather than as empirical findings.

Residuals.

For a quantity B obtained both by dynamic programming and by enumeration on the same carrier, we record

$$\varepsilon_B^{\text{DP-enum}} = \left|B_{\text{DP}} - B_{\text{enum}}\right|, \tag{D73}$$

together with a relative residual where the reference magnitude permits one. Under a condition $c = (x, x^\circ, a)$, let $\mathcal{Y}_c^{\text{dir}} = g_a(x) - g_a(x^\circ)$ be the output contrast obtained from two forward evaluations, and let $\mathcal{Y}_c^{\text{path}} = \sum_\pi \widehat{A}_\pi^c$ be its reconstruction from the numerically evaluated path amplitudes. The reconstruction and normalisation residuals are

$$\varepsilon_{\text{rec}}^c = \left|\mathcal{Y}_c^{\text{dir}} - \mathcal{Y}_c^{\text{path}}\right|, \qquad \varepsilon_{\text{norm}}^c = \Big|\sum_\pi \widehat{\mu}^c(\pi) - 1\Big|. \tag{D74}$$

Observed magnitudes.

In the activation-wide MLP validation, the signed path expansion reconstructs the declared output contrast with median relative error 3.96×10^{-16} and maximum 2.41×10^{-12}, and the partition, entropy, edge-flow and node-flow computations agree with brute-force enumeration at the same scale. In the MLP optimiser experiment, the finite entropy-production identity has median absolute residual 2.16×10^{-8} over 8,619

Table D2: Design ledger. Each claim of Sections 4 to 6 is listed with the observable that measures it, the exact statement that licenses its interpretation, the population on which it is evaluated and the inferential unit of its test.

Claim	Observable	Licensed by	Population	Unit	Sec.
Learning concentrates the disposition	$\eta_q(t)$, $S_{\text{deficit}}(t)$	Thm. 1	48 MLP; 48 CNN architectures	Architecture	4.1
The direction persists across environments	Signs of $\Delta^{\text{time}}_{b;0\to T}$ of η_q, S_{deficit}	Thm. 1	$12 \times 24 \times 2$ per family	Environment	4.1
The optimiser sets the magnitude	$\Delta^{\text{time}}_{b;0\to T} S_{\text{deficit}}$, selection covariance	Tilt and entropy-production theorems (Sec. 4.2)	118 MLP blocks; 576 CNN blocks	Matched block	4.2
The geometry contracts without a universal exponent	$\xi_{50}(t)$, retention, $d_B(t)$	Thm. 6	48 MLP architectures; $12 \times 24 \times 2$ per family	Architecture; environment	4.3
Realisation narrows the disposition	$\eta_q^c(t)/\eta_q(t)$, d^*, Λ^*_F at fixed t	Support domination (Sec. 5.1)	324 benchmark networks	Trained network	5.2
CNN specialisation lies in spatial reuse	Reuse deficit, A_{eq}	Entropy decomposition (Sec. 5.3)	162 benchmark CNNs	Trained network	5.3
Realised carriers are load-bearing	Output displacement, D_H	No creation of support; zero-dose identity (Sec. 6.1)	324 frozen networks	Trained network	6.2
Intervention reorganises the geometry; occurrence targeting dominates	BCANw observables, D_H	Thm. 6; entropy decomposition	324 frozen networks; their CNNs	Trained network	6.3

stored intervals, with 95th percentile 6.15×10^{-6} and maximum 4.29×10^{-4}. The exponential-tilt reconstruction of the final edge and node flows has maximum absolute residuals 1.11×10^{-16} and 1.67×10^{-16}.

Validity of conditional measures.

The realised partition function Z^c vanishes exactly when every carrier amplitude vanishes, and the conditional measure is then undefined rather than zero (Appendix C.9). A condition is recorded as valid when $V(c) = \mathbf{1}\{Z^c > 0\} = 1$. Invalid contrasts are excluded only where a quantity is undefined, and the fraction of valid conditions is reported as an observable of the trained network and of the declared conditional population. Degenerate cases are retained as observed outcomes. In the CNN programme, these include a run with conditional valid fraction 0.617, a network that is fully degenerate at epochs 240 and 360, an environment-phase endpoint panel with 88.0% of jobs fully valid, and an optimiser block with 33 of 128 valid endpoint contrasts.

Validity of geometric descriptors.

Geometry passes through admissibility gates before any descriptor is computed. Finite-window scaling credibility is a separate and stricter gate, so that a computed geometry is never promoted to a scaling law by default. The exponent $d_B(t)$ is reported only

Table E3: Structural reorganisation under learning in MLPs and CNNs. Architecture columns report median changes from initialisation to the final checkpoint, followed by the number of architectures changing in the indicated direction. Environment columns report directional counts, and optimiser columns report Kendall's concordance W. The MLP populations comprise 48 architectures, 576 environment runs and 118 matched optimiser blocks; the CNN populations comprise 48 architectures, 288 architecture-environment cells and 576 matched optimiser blocks. MLP changes of η_1, η_2 and flow CVs are final-to-initial folds, whereas CNN changes are endpoint differences, with η_1 and η_2 expressed as logarithmic differences. Per-optimiser medians are reported in Appendix E.4 and Appendix E.5.

	MLP			CNN		
Observable	Arch.	Env.	Opt. W	Arch.	Env.	Opt. W
$\Delta S_{\text{deficit}} \uparrow$	2.051 (48/48)	576/576	0.902	2.039 (46/48)	258/288	0.373
$\eta_1 \downarrow$	$\times$0.129 (48/48)	576/576	–	$-$2.039 (46/48)	258/288	0.373
$\eta_2 \downarrow$	$\times$0.017 (48/48)	528/528	0.909	$-$2.542 (46/48)	254/288	0.364
Edge/occ.-CV $\uparrow$	$\times$1.385 (48/48)	576/576	0.895	1.932 (46/48)	257/288	0.369
Node-flow CV $\uparrow$	$\times$1.944 (48/48)	576/576	0.918	–	–	–
Parameter-CV $\uparrow$	–	–	–	1.687 (46/48)	259/288	0.368
$S_{\text{occ}} \uparrow$	*No weight sharing*			0.685 (46/48)	259/288	0.360
$S_{\text{param}} \uparrow$				0.682 (46/48)	260/288	0.356
$S_{\text{reuse}} \uparrow$				0.009 (38/48)	22/24[†]	0.276

[†] For S_{reuse}, 22 of 24 environments have a positive mean change across the 12 architectures; other CNN counts refer to individual architecture-environment combinations. CNN η_1 and η_2 entries are logarithmic changes. CNN occurrence-flow and parameter-flow heterogeneity are distinct observables. A dash denotes an unestablished quantity, not a zero effect.

when a credible finite scaling window satisfies the declared criteria (Appendix C.10). Undefined exponents and other undefined quantities are never imputed or set to zero. The benchmark ledgers report the gate pass fractions, the geometry-carrier node counts relative to the declared cap of 5,200 nodes and the conditional valid fraction per design cell (Appendix E.9).

Appendix E Extended Results

E.1 Overview of the structural results

This subsection collects in one table the changes under learning on which Section 4 rests, for the MLP and CNN structural populations. Each entry compares the final checkpoint with initialisation within one population, so the table reports changes over training and not changes of measure; the structural-to-conditional changes of the principal observables are given in Table 1. The detailed tables and figures follow in Appendices E.2 to E.5.

E.2 MLP structural disposition: architecture

This subsection reports the architecture cohort of Section 4 in full. Tables E4 and E5 give the medians and sign counts summarised in Table E3, Table E6 gives the preservation of architecture ordering, and Table E7 and Figures E4 and E5 separate the roles

Table E4: Cohort-wide summary across the 48 longitudinal architectures. IQR is the 25th–75th percentile interval.

Metric	Median	IQR	Range
Best test MSE	0.564	0.559–0.573	0.542–0.610
Final test MSE	0.844	0.795–0.881	0.596–0.984
Best-test epoch	12	8–16	5–60
Final − best test MSE	0.276	0.232–0.308	0.048–0.399
$\eta_1(T)/\eta_1(0)$	0.129	0.094–0.168	0.051–0.331
$\eta_2(T)/\eta_2(0)$	0.017	0.008–0.030	0.002–0.126
$\Delta S_{\text{deficit}}$	2.051	1.787–2.367	1.105–2.980
Edge-CV fold	1.385	1.272–1.469	1.144–1.777
Node-CV fold	1.944	1.663–2.106	1.244–2.537
$\xi_{50}(T)/\xi_{50}(0)$	0.934	0.903–0.953	0.845–0.998
Final d_B^F	1.417	0.477–1.702	0.266–2.047
Final d^*	2.150	1.974–2.247	1.768–3.679
Local median Δ_i^F	−0.408	−0.498 to −0.308	−0.811 to −0.090
Valid local paired fraction	0.232	0.138–0.314	0.045–0.640
Final quotient-flow KS	0.245	0.212–0.270	0.177–0.320
Mean inter-box retention	0.220	0.174–0.271	0.117–0.424

of depth and width morphology. Within the cohort, peak lacunarity falls with depth ($\rho_s = -0.859$), the fraction of nodes with a valid local exponent rises ($\rho_s = 0.869$), and relative edge heterogenisation decreases ($\rho_s = -0.858$). The morphology groups overlap substantially. Among the final descriptors, only the quotient-flow KS statistic separates them (Kruskal-Wallis $H = 10.14$, $p = 0.038$), and that statistic correlates more strongly with network size than with morphology (Table E8). Figure E7 shows the longitudinal geometry discussed in Section 4.3. The temporal master-curve error has a cohort median of 0.0337 dex and rises to 0.0864 dex at the final snapshot, and the coupling $\rho_s(S_{\text{deficit}}, d^*)$ is negative in 38 of 48 architectures and strongest in the deepest band.

The remaining figures of the cohort document the same results at finer resolution. Figure E1 gives the path entropy of every architecture through training, and Figure E3 gives representative path-flow pages from the five morphology families. Figure E6 compares the final geometry across architectures, Figures E8 and E9 decompose it over scale and training time, and Figure E10 reports the validated local geometry at the final snapshot. Figure E12 shows a representative directed renormalisation, and Figure E13 relates late functional deterioration to the structural changes.

E.3 MLP structural disposition: environment

This subsection reports the environment matrix in full. Tables E9 and E10 give the medians and sign counts, Table E11 the preservation of architecture ranks across the 276 environment pairs, Table E12 the normalised structured shares of architecture and environment, and Table E13 the environment-level associations with signal-to-noise

Table E5: Sign consistency of the dominant architecture-wide changes.

Change	Count	Fraction
$\eta_1(T) < \eta_1(0)$	48/48	1.000
$\eta_2(T) < \eta_2(0)$	48/48	1.000
$S_{\text{deficit}}(T) > S_{\text{deficit}}(0)$	48/48	1.000
Edge-CV(T) >Edge-CV(0)	48/48	1.000
Node-CV(T) >Node-CV(0)	48/48	1.000
$\xi_{50}(T) < \xi_{50}(0)$	48/48	1.000
$d^*(T) < d^*(0)$	36/48	0.750
$x_d^*(T) > x_d^*(0)$	40/48	0.833
Architecture-median local $\Delta_i^F < 0$	48/48	1.000

Table E6: Spearman preservation of architecture ordering from initialisation to the final snapshot.

Descriptor	ρ_s
η_1	0.990
η_2	0.992
ξ_{50}	0.976
d_B^F	−0.124
d^*	0.397
x_d^*	0.282

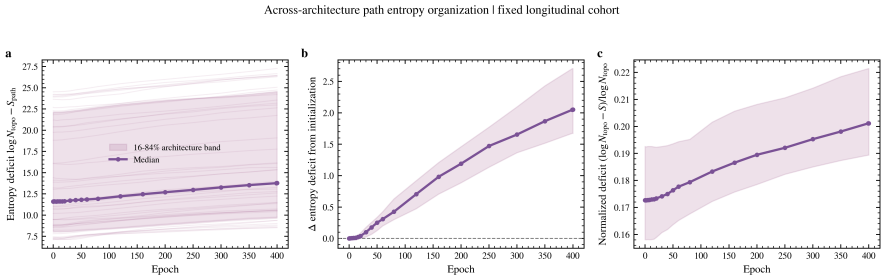


Fig. E1: Across-architecture path entropy through training. Faint curves are individual architectures; the darker curve and band summarise the unchanged 48-network cohort. The entropy deficit and its depth/path-count normalised form rise throughout training rather than saturating at the epoch of best test performance.

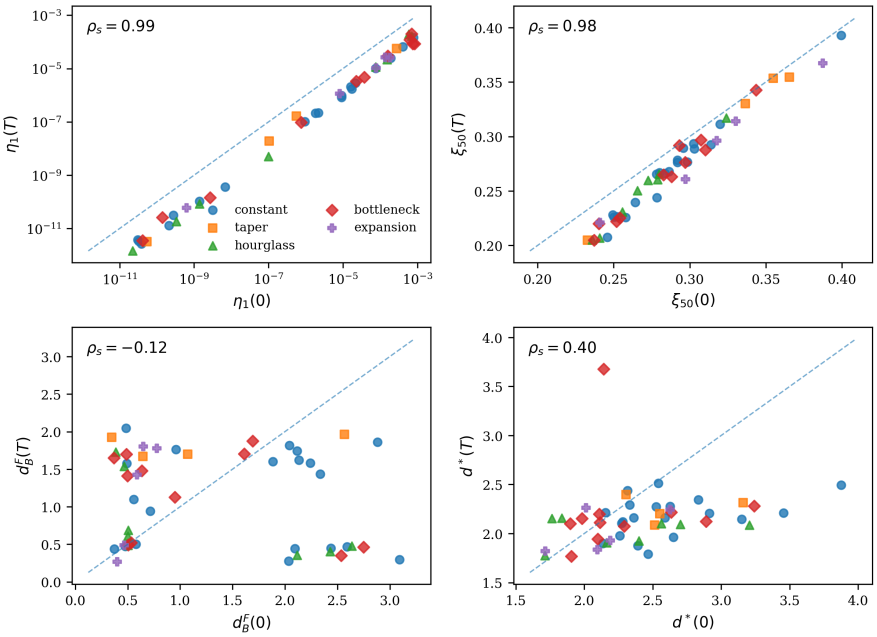


Fig. E2: Initialisation-to-final preservation across the 48 architectures. Path-space occupancy and the characteristic intrinsic scale preserve architecture ordering strongly, whereas the finite-window exponent is substantially reorganised by learning. The dashed diagonal is equality, not a fitted model.

Table E7: Median structural behaviour of the architecture cohort in three hidden-depth bands. Folds are final over initial values, Λ_F^* is the final peak flow lacunarity, V_Δ the fraction of nodes with a valid local paired fit, and Δ^F the median local mass-support mismatch. Deformation is the total structural deformation. The last column is the median within-architecture temporal Spearman correlation between entropy deficit and crossover strength d^*.

Depth	n	Edge-CV fold	Node-CV fold	Λ_F^*	V_Δ	Δ^F	Deformation	$\rho_s(S_{\text{deficit}}, d^*)$
9–12	18	1.479	2.099	95.0	0.135	−0.503	20.64	−0.126
13–20	13	1.383	1.956	57.3	0.242	−0.469	16.78	−0.656
21–32	17	1.229	1.611	31.0	0.351	−0.262	10.66	−0.819

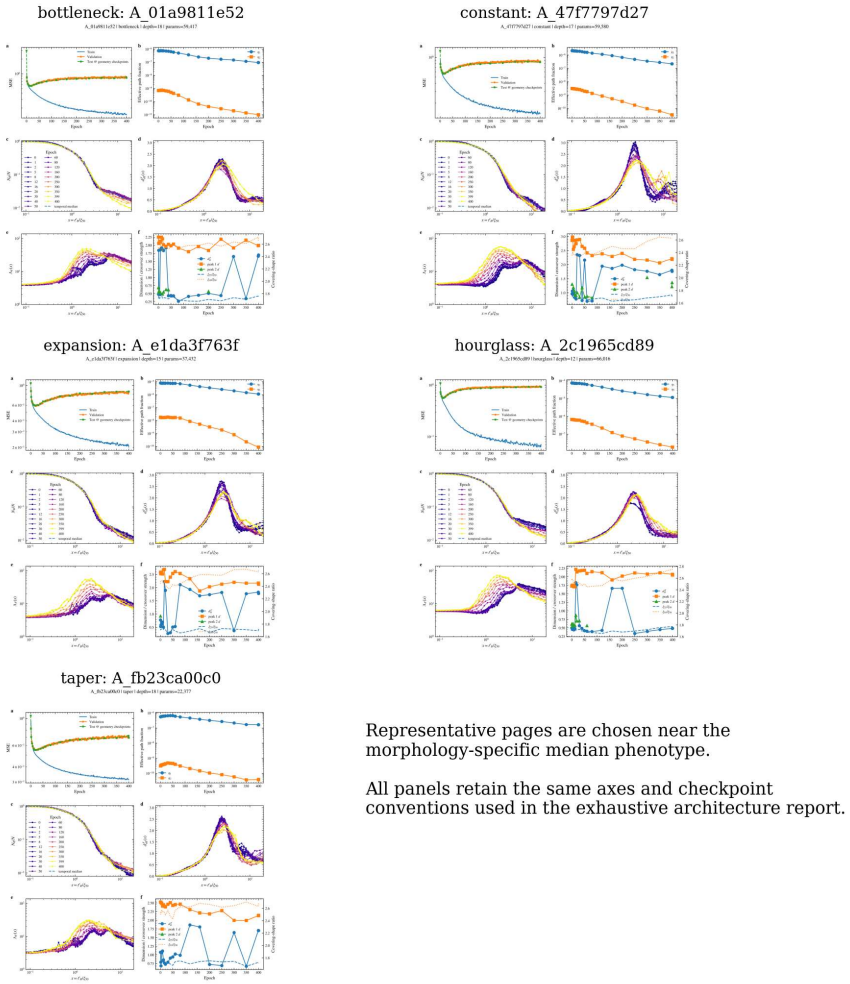


Fig. E3: Representative path-flow pages from the five morphology families. Each page uses the same axes, checkpoint colours and definitions. The normalised covering shape and crossover remain recognizable, while the finite-window and lacunarity responses vary substantially across architectures.

Table E8: Median results by architecture morphology. Small taper and expansion subgroups make morphology-specific inference descriptive rather than definitive.

Morphology	n	Best test	Final test	ΔS	η_1 fold	d_B^F	d^*	Local Δ^F	KS
Constant	20	0.5642	0.8513	2.194	0.111	1.267	2.184	−0.432	0.238
Taper	4	0.5561	0.7401	1.639	0.194	1.813	2.259	−0.443	0.218
Hourglass	8	0.5627	0.8332	2.341	0.104	0.549	2.089	−0.340	0.274
Bottleneck	11	0.5669	0.8405	2.047	0.129	1.409	2.122	−0.348	0.246
Expansion	5	0.5909	0.8576	1.971	0.139	1.425	1.930	−0.478	0.201

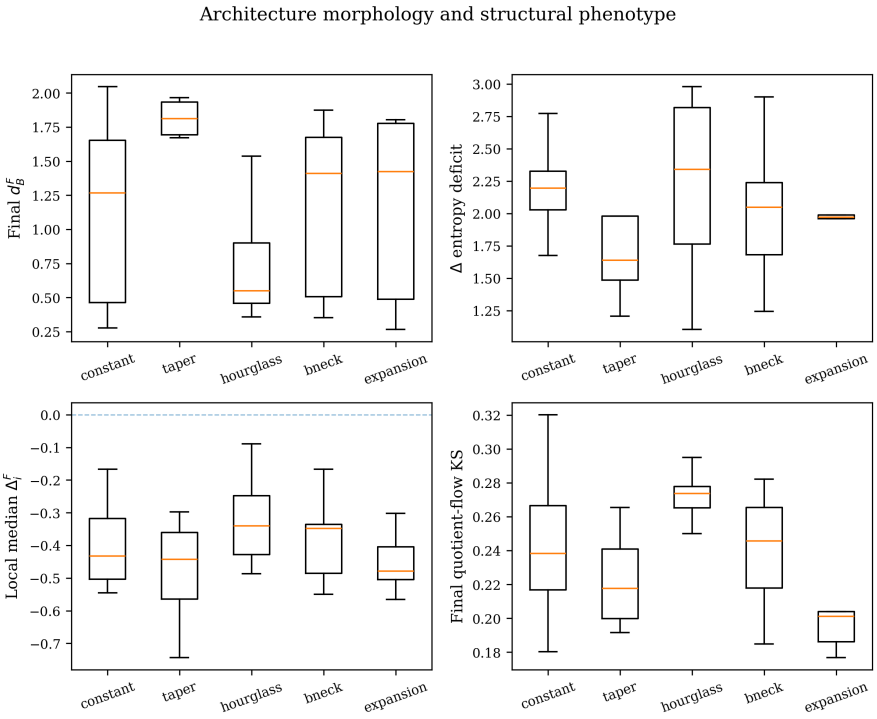


Fig. E4: Architecture morphology and structural phenotype. The expanded cohort shows substantial overlap in finite-window exponent and entropy concentration. The most visible morphology-level separation is instead in quotient-flow KS, and even there capacity remains an important covariate.

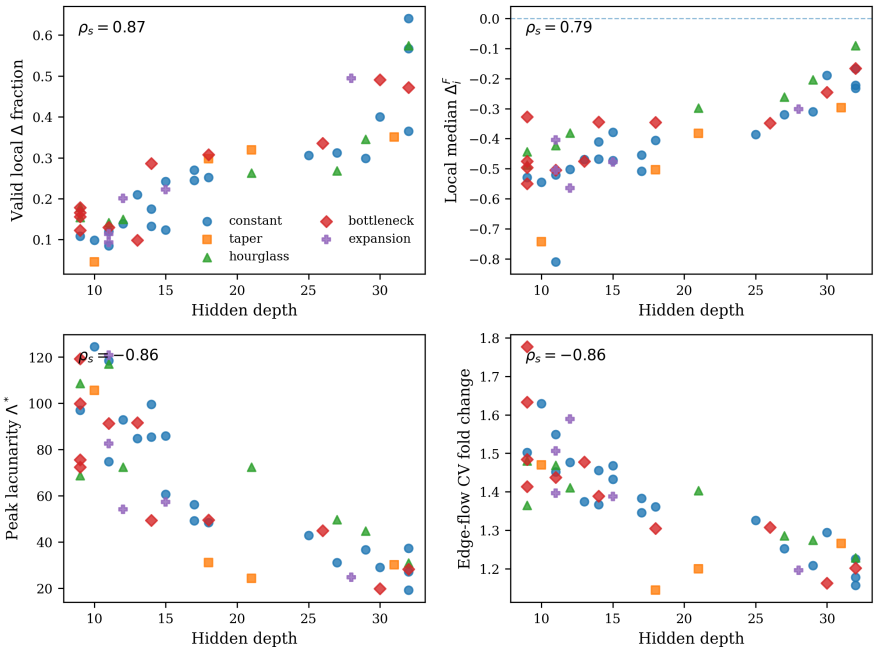


Fig. E5: Depth dependence of the structural phenotype. Deeper architectures support valid common-window local exponents over a larger fraction of nodes, have lower peak lacunarity, show less negative local flow-support mismatch, and undergo weaker relative edge-flow heterogenisation.

ratio. In this population the median η_1 fold is 0.040 and the median entropy-deficit increase is 3.218 nats, both larger than in the architecture cohort.

Figure E14 shows the replication of the cohort's sign laws, Figure E16 the architecture-environment response matrices, and Figure E17 the population trajectories over the 576 runs. Table E14 gives descriptive medians by teacher family, and Table E15 follows the structured variance shares through training.

E.4 MLP structural disposition: optimiser

This subsection reports the optimiser panel in full. Table E18 gives the omnibus and pairwise tests of the optimiser ordering, and Table E19 the descriptive variance shares. Figure E18 compares structural and functional time. RMSProp adds 0.610 nats of entropy deficit after its optimum, or 92.1% of its total increase, and the median best test errors are 0.212, 0.241, 0.212 and 0.217 for Adam, RMSProp, SGD with momentum and SGD. Table E21 reports the cluster-robust trajectory models for the state, the geometry and the selection field. Figure E22 shows the alignment and intensity of the selection field on the six intervals stored identically for every optimiser. On the

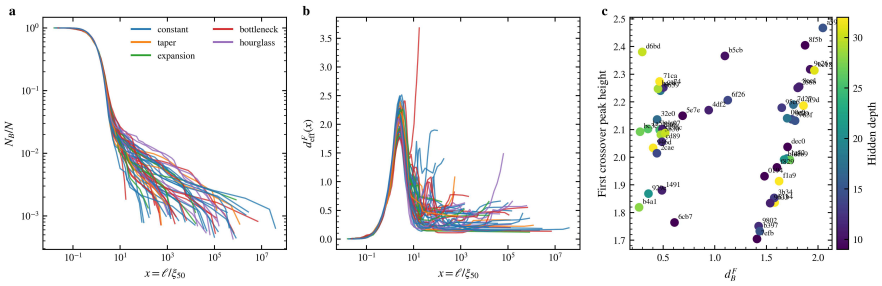


Fig. E6: **Final architecture-wide comparison of the path-flow geometry in the architecture cohort. (a)** Final normalised BCANw covering curves, N_B/N versus the intrinsic scale $x = \ell/\xi_{50}$, for the 48 longitudinal architectures. Curves are coloured by architecture morphology. Although the architectures share a broadly recognizable fine-to-coarse covering transition, the coarse-scale branches remain substantially architecture dependent. **(b)** Corresponding effective-dimension responses $d^F_{\mathrm{eff}}(x)$. The architectures exhibit a common crossover-like form but differ strongly in the height, width and coarse-scale continuation of the response. **(c)** Final finite-window exponent d^F_B versus crossover peak height d^*, with colour indicating hidden depth. The broad occupation of this plane illustrates that the existence of a credible finite scaling window does not imply convergence to a universal scaling exponent or crossover strength.

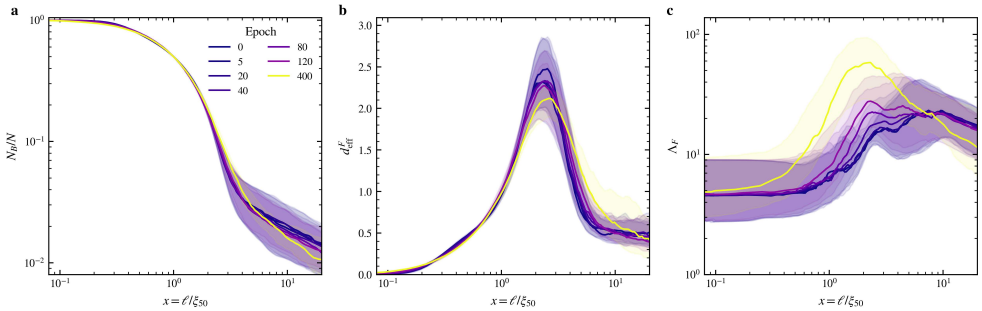


Fig. E7: Learning deforms the crossover and coarse-scale structure of the geometry around a persistent backbone. Cohort-level geometry of the architecture cohort at representative checkpoints; shaded regions show the dispersion across the 48 architectures. (a) The normalised covering relation N_B/N against $x = \ell/\xi_{50}$ stays aligned through the principal transition, while architecture-dependent dispersion grows beyond the crossover at later epochs. (b) The effective dimension $d^F_{\mathrm{eff}}(x)$ develops an intermediate-scale maximum whose height and coarse-scale continuation change through training, so that the late trajectory is marked by crossover softening rather than convergence to one dimension. (c) The flow lacunarity $\Lambda_F(x)$ becomes stronger and shifted relative to initialisation.

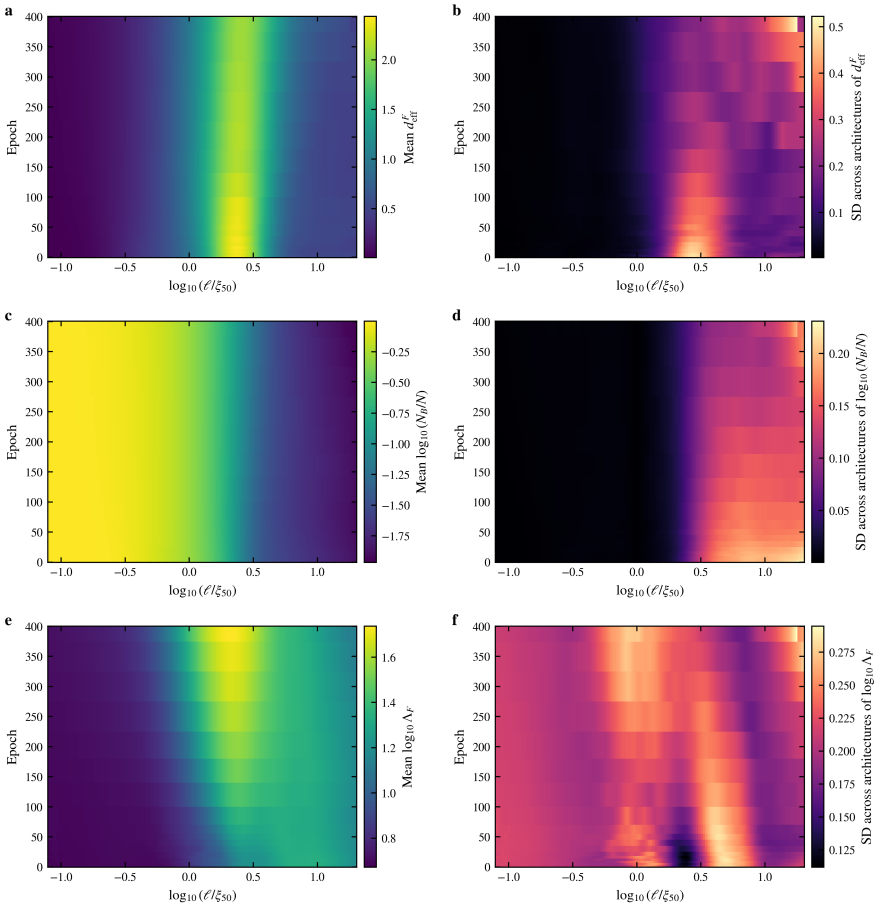


Fig. E8: **Scale-time decomposition of the population geometry in the architecture cohort.** The horizontal coordinate is intrinsic logarithmic scale, $\log_{10}(\ell/\xi_{50})$, and the vertical coordinate is training epoch. Left-column panels show the architecture-wide mean; right-column panels show the corresponding standard deviation across the fixed 48-architecture cohort. **(a,b)** Mean and between-architecture standard deviation of the effective dimension d_{eff}^F. A persistent intermediate-scale response is visible throughout training, while architecture-to-architecture disagreement becomes increasingly localised toward the crossover and coarse-scale regime. **(c,d)** Mean and standard deviation of $\log_{10}(N_B/N)$. The normalised covering backbone remains comparatively stable on the fine-scale side, whereas between-architecture deformation develops primarily after the main covering transition. **(e,f)** Mean and standard deviation of $\log_{10}\Lambda_F$. The lacunarity field undergoes a pronounced temporal redistribution and exhibits substantial architecture-dependent variation near the evolving intermediate and coarse scales. The figure therefore separates the common population trajectory from the scale-localised architectural dispersion that is hidden by any single finite-window exponent.

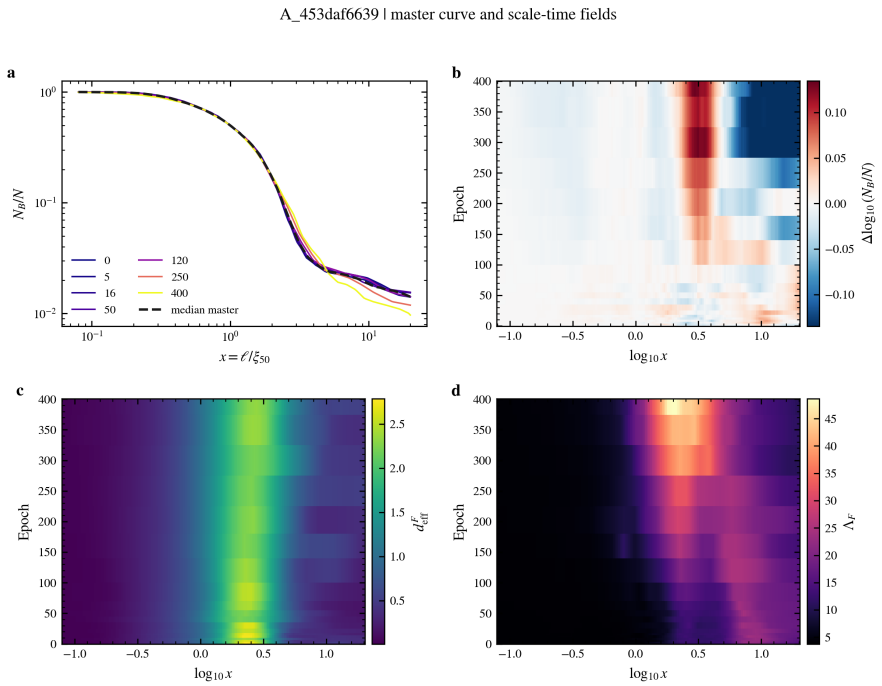


Fig. E9: Representative scale-time organisation from one member of the exhaustive cohort. The normalised covering relation remains close to a common intrinsic shape through much of training, while the effective-dimension and lacunarity fields reveal structured late-time deformation.

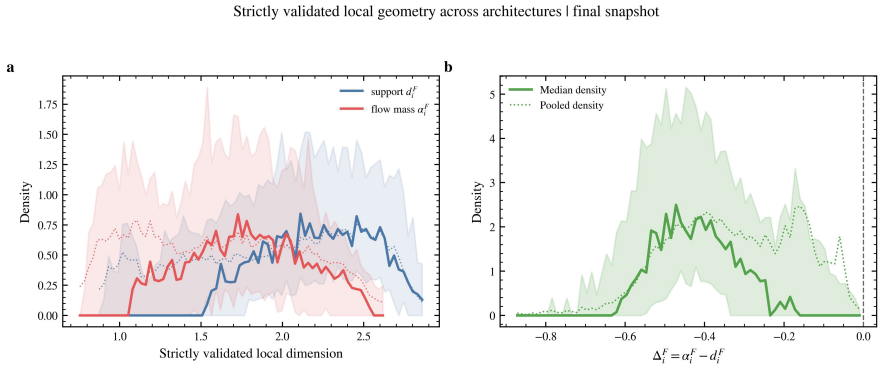


Fig. E10: Strictly validated local geometry at the final snapshot. Panel (a) summarises the architecture distribution of support and flow-mass dimension densities. Panel (b) shows the density of $\Delta_i^F = \alpha_i^F - d_i^F$. The architecture-level median density lies entirely on the negative side of zero.

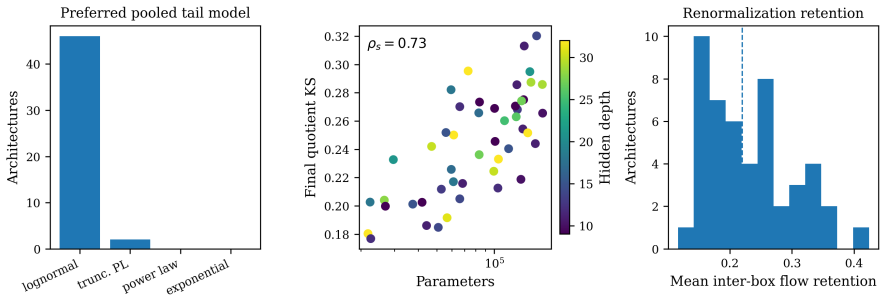


Fig. E11: Microscopic-tail and renormalisation summary. Left: preferred pooled tail models across architectures. Centre: final quotient-flow KS versus parameter count, coloured by depth. Right: distribution of mean inter-box flow retention.

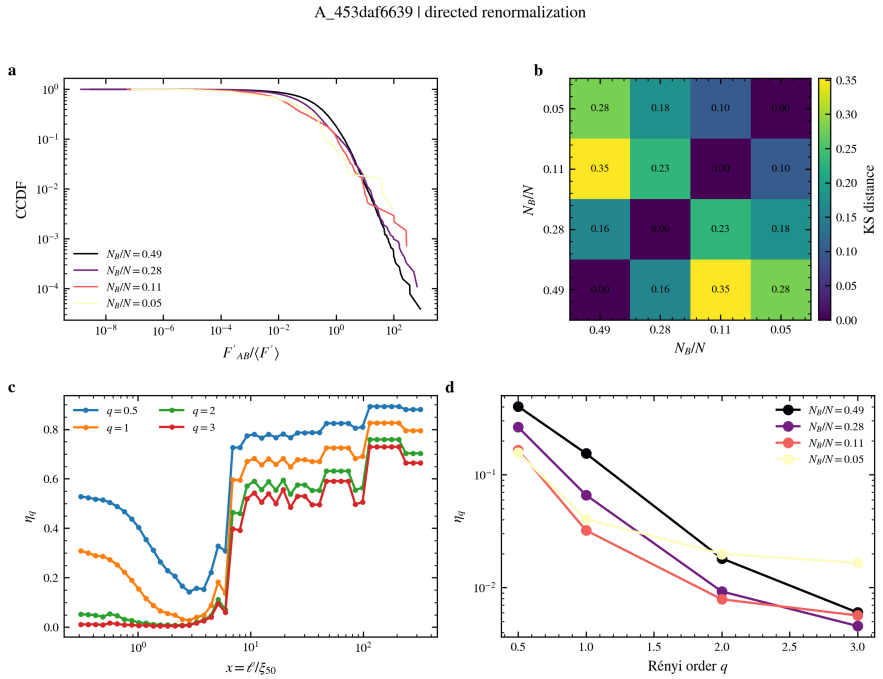


Fig. E12: Representative directed renormalisation page from the exhaustive analysis. The quotient-flow CCDFs remain broadly related across scales, while KS distances, Rényi effective fractions and inter-box retention show that the measure is materially reorganised rather than sitting at a fixed point.

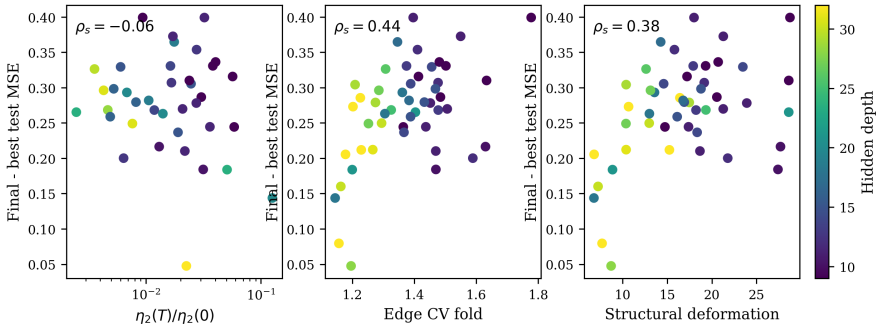


Fig. E13: Cross-architecture relationship between late functional deterioration and three structural changes. These associations remain useful for context but are weaker than the matched within-architecture evidence and are partly organised by depth.

Table E9: Summary of the environment matrix across the 576 longitudinal architecture-environment-seed runs. IQR is the 25th–75th percentile interval. The order-two path fraction has 528 numerically valid fold ratios; all other rows shown here have 576 valid observations.

Metric	Median	IQR	Range
Best test MSE	0.195	0.096–0.264	0.039–0.480
Final test MSE	0.282	0.129–0.397	0.049–0.806
Best-test epoch	20	16–40	5–80
Final − best test MSE	0.086	0.037–0.142	0.009–0.345
$\eta_1(T)/\eta_1(0)$	0.040	0.027–0.058	0.004–0.132
$\eta_2(T)/\eta_2(0)$	0.0023	0.0011–0.0046	0.0000–0.0237
$\Delta S_{\text{deficit}}$	3.218	2.851–3.622	2.028–5.515
Δ edge-flow CV	0.756	0.567–0.947	0.361–1.646
Δ node-flow CV	0.339	0.279–0.388	0.189–0.595
$\xi_{50}(T)/\xi_{50}(0)$	0.922	0.900–0.940	0.836–0.988
Final d_B^F	1.271	0.426–1.569	0.251–2.139
Δd_B^F	0.103	−0.272 to 0.904	−1.972 to 1.579
Final d^*	1.915	1.779–2.090	1.433–4.002
Δd^*	−0.356	−0.555 to −0.049	−1.429 to 2.103
Final x_d^*	2.719	2.494–3.060	2.040–386.926
Local median Δ_i^F	−0.402	−0.487 to −0.241	−0.656 to 0.014
Valid local paired fraction	0.252	0.202–0.338	0.083–0.487
Final quotient-flow KS	0.233	0.212–0.256	0.148–0.325
Mean inter-box retention	0.216	0.166–0.290	0.097–0.467

Table E10: Replication in the environment matrix of the directional laws found in the architecture cohort. The denominator is the number of numerically valid runs for the corresponding quantity.

Direction from the architecture cohort	Consistent runs	Fraction
$\eta_1(T) < \eta_1(0)$	576/576	1.000
$\eta_2(T) < \eta_2(0)$	528/528	1.000
$S_{\mathrm{deficit}}(T) > S_{\mathrm{deficit}}(0)$	576/576	1.000
edge-flow CV increases	576/576	1.000
node-flow CV increases	576/576	1.000
$\xi_{50}(T) < \xi_{50}(0)$	576/576	1.000
architecture-run median $\Delta_i^F < 0$	571/576	0.991
$d^*(T) < d^*(0)$	454/576	0.788

Table E11: Architecture-rank universality across the 24 selected environments. For each descriptor the table reports the median Spearman correlation across all $\binom{24}{2} = 276$ environment pairs, together with the 16th–84th percentile interval.

Descriptor	Median ρ_s	16–84% interval
$\log \eta_1$	0.993	0.986–1.000
$\log \eta_2$	0.993	0.986–1.000
$\log \xi_{50}$	0.521	0.294–0.706
d_B^F	0.357	0.105–0.601
d^*	0.538	0.252–0.713
median Δ_i^F	0.930	0.867–0.965
quotient-flow KS	0.804	0.622–0.902

interval from epoch 80 to 120, the median $\mathrm{Var}_p(G)$ is about 3.81 under Adam, 0.88 under RMSProp, 0.035 under SGD with momentum and 0.014 under SGD. Within the Adam arm, 99.92% of 3,780 measured intervals have positive starting selection covariance.

Table E16 reports run completion, which determines the 118 matched blocks, and Table E17 and Figure E19 give the matched endpoints. Figures E20 and E21 follow the state and the geometry on the checkpoint grids shared by all optimisers. Figure E23 and Table E20 give the descriptive factor decompositions for the full panel and for the balanced Adam arm.

Table E12: Final-snapshot variance decomposition in the environment matrix. U_A, U_E and $U_{A\times E}$ are the normalised structured shares shown in Figure E15, i.e. each is normalised by $\sigma_A^2 + \sigma_E^2 + \sigma_{A\times E}^2$. The last column reports the seed/residual share of the total estimated variance and must be read separately; large values indicate that the structured-bar normalisation can otherwise overstate certainty.

Metric	U_A	U_E	$U_{A\times E}$	residual/total
$\log \eta_1$	0.996	0.004	0.000	0.003
$\log \eta_2$	0.994	0.006	0.000	0.007
S_{deficit}	0.996	0.004	0.000	0.003
edge CV	0.913	0.084	0.002	0.113
node CV	0.900	0.100	0.000	0.095
$\log \xi_{50}$	0.590	0.324	0.086	0.533
d_B^F	0.956	0.044	0.000	0.802
d^*	0.635	0.272	0.093	0.598
x_d^*	0.684	0.236	0.079	0.968
Λ^*	0.905	0.095	0.000	0.174
median Δ_i^F	0.999	0.001	0.000	0.082
valid local Δ fraction	0.957	0.026	0.018	0.143
quotient-flow KS	0.983	0.017	0.000	0.348
inter-box retention	0.983	0.017	0.000	0.347
post-best test deterioration	0.014	0.977	0.009	0.011

Table E13: Environment-level Spearman associations. Each structural response is first averaged across the 12 architectures and two seeds within each of the 24 environments; the reported coefficient then relates that environment mean to the corresponding continuous environment control. These are descriptive associations within the selected environment panel, not independent causal effects.

Response	d_{inf}	Feature corr. r	SNR	Interaction
$\Delta S_{\text{deficit}}$	−0.231	+0.129	+0.827	+0.110
η_1 fold	+0.254	−0.120	−0.812	−0.126
η_2 fold	+0.281	−0.063	−0.794	−0.168
ξ_{50} fold	−0.185	−0.540	−0.841	−0.337
Δd_B^F	−0.265	−0.355	−0.588	−0.123
Δd^*	−0.266	−0.646	−0.671	−0.278
median Δ_i^F	+0.443	+0.449	−0.197	−0.145
quotient-flow KS	+0.164	+0.199	−0.609	−0.097
post-best test deterioration	−0.153	−0.309	−0.938	−0.059

Table E14: Descriptive medians of the environment matrix by teacher family. The selected environments are not a balanced factorial sample over the continuous controls, so these rows summarise the realised panel and are not interpreted as isolated teacher-family effects.

Family	n	Best test	Post det.	η_1 fold	ΔS	ξ_{50} fold	d_B^F	Δd^*	med. Δ_i^F	KS
Deep compositional	168	0.222	0.111	0.044	3.120	0.922	1.309	−0.312	−0.414	0.234
Linear	168	0.211	0.111	0.040	3.216	0.926	1.251	−0.376	−0.398	0.232
Shallow nonlinear	240	0.112	0.049	0.036	3.315	0.919	1.271	−0.383	−0.399	0.233

Table E15: Temporal evolution of the structured variance shares in the environment matrix. Each cell reports $U_A/U_E/U_{A\times E}$ at the indicated checkpoint. The path-entropy scaffold remains architecture-dominated throughout, whereas the metric and crossover geometry acquire appreciable environmental dependence during training. Seed and residual variance are not included in these normalised structured shares.

Metric	Epoch 0	Epoch 80	Epoch 200	Epoch 399
$\log \eta_1$	1.00/0.00/0.00	1.00/0.00/0.00	1.00/0.00/0.00	1.00/0.00/0.00
$\log \xi_{50}$	1.00/0.00/0.00	0.52/0.46/0.02	0.41/0.56/0.03	0.58/0.32/0.10
edge CV	1.00/0.00/0.00	0.89/0.09/0.02	0.88/0.12/0.00	0.91/0.08/0.00
node CV	1.00/0.00/0.00	0.92/0.07/0.01	0.91/0.09/0.00	0.90/0.10/0.00
d^*	1.00/0.00/0.00	1.00/0.00/0.00	0.98/0.02/0.00	0.71/0.29/0.01
d_B^F	1.00/0.00/0.00	0.24/0.00/0.76	0.89/0.11/0.00	0.92/0.05/0.03

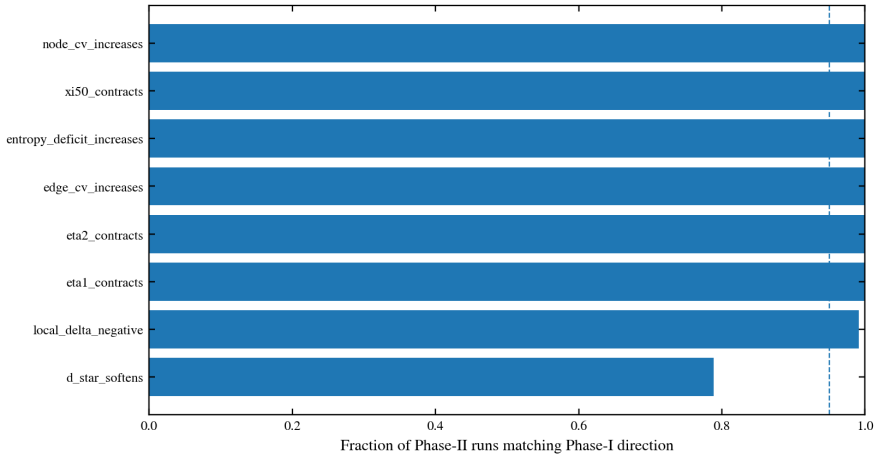


Fig. E14: Replication in the environment matrix of the sign laws of the architecture cohort. The dashed vertical line marks a 95% consistency reference. Path concentration, entropy-deficit growth, edge/node heterogenisation and ξ_{50} contraction are exact in the valid environment-matrix sample; the negative local mismatch is near-universal, while crossover softening is a strong tendency rather than an exact law.

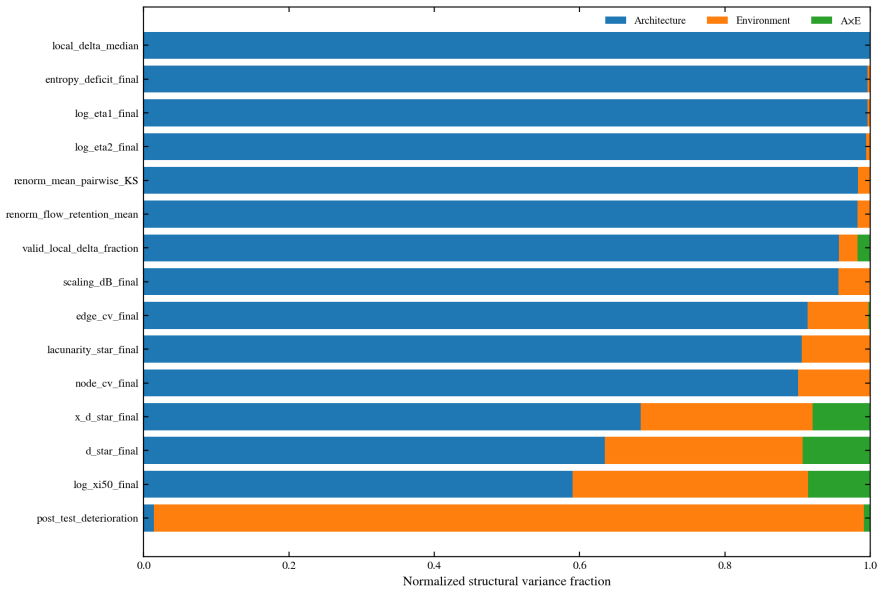


Fig. E15: Normalised structured variance decomposition of the environment matrix at the final snapshot. Blue, orange and green segments represent architecture, environment and architecture-environment interaction after normalizing by the sum of those three structured components. Seed/residual variance is intentionally omitted from the bars and is reported separately in Table E12. Path-space and local-flow descriptors are overwhelmingly architecture-organised within the structured component, whereas post-best test deterioration is overwhelmingly environment-organised.

Table E16: Run completion in the optimiser panel. Cochran's $Q = 17.54$, $p = 0.000548$ across the 125 intended repeated blocks. Seven SGD and five momentum-SGD runs diverged; Adam and RMSProp had none. Divergences are retained as a secondary optimiser-stability outcome and successful-run comparisons use the 118 blocks completed by all four optimisers.

Optimiser	Intended	Completed	Diverged
Adam	125	125	0
RMSProp	125	125	0
SGD+momentum	125	120	5
SGD	125	118	7

Table E17: Matched five-seed endpoint summary of the optimiser panel. Entries are medians, with interquartile ranges in brackets on the line below, over the 118 architecture-environment-seed blocks in which all four optimisers completed. Best-test values are selected on the exact common validation-checkpoint grid, never on test loss.

Quantity	Adam	RMSProp	SGD+momentum	SGD
Runs	118	118	118	118
$\Delta S_{\text{deficit}}$	3.323	0.706	0.109	0.062
	[2.894, 3.881]	[0.515, 0.929]	[0.015, 0.197]	[0.001, 0.122]
$\eta_1(T)/\eta_1(0)$	0.036	0.494	0.897	0.940
	[0.021, 0.055]	[0.395, 0.598]	[0.822, 0.985]	[0.885, 0.999]
Edge-CV fold	1.452	1.081	1.008	1.004
	[1.357, 1.677]	[1.054, 1.165]	[1.000, 1.027]	[0.999, 1.014]
Node-CV fold	2.075	1.281	1.044	1.022
	[1.941, 2.276]	[1.212, 1.461]	[1.010, 1.093]	[1.006, 1.059]
$\xi_{50}(T)/\xi_{50}(0)$	0.932	0.972	1.004	1.001
	[0.913, 0.947]	[0.956, 0.988]	[0.994, 1.010]	[0.994, 1.009]
Δd^*	−0.857	−0.161	−0.049	−0.076
	[−1.026, −0.581]	[−0.301, 0.062]	[−0.182, 0.075]	[−0.228, 0.052]
$\Delta\Lambda^*$	43.740	6.860	−0.545	−0.576
	[33.229, 78.152]	[1.872, 16.469]	[−3.274, 2.227]	[−3.911, 1.304]
Final local Δ_i^F	−0.177	−0.095	−0.051	−0.049
	[−0.361, −0.139]	[−0.212, −0.083]	[−0.064, −0.042]	[−0.061, −0.041]
Coarse-flow retention	0.510	0.699	0.741	0.741
	[0.319, 0.560]	[0.593, 0.726]	[0.696, 0.763]	[0.713, 0.771]
Best test MSE	0.212	0.241	0.212	0.217
	[0.085, 0.325]	[0.092, 0.335]	[0.082, 0.324]	[0.087, 0.329]
Final − best test MSE	0.113	0.082	0.024	0.003
	[0.032, 0.242]	[0.023, 0.193]	[0.000, 0.107]	[0.000, 0.019]

Table E18: Primary paired omnibus tests on the 118 complete optimiser blocks. The Friedman test treats optimiser as a repeated condition within each architecture-environment-seed block; Kendall $W = \chi_F^2/[n(k-1)]$ summarises concordance of optimiser ordering. The very small W for d_B^F is as important as the large effects for measure-level variables.

Outcome	$\chi_F^2(3)$	p	Kendall W
$\Delta S_{\text{deficit}}$	319.25	$< 10^{-60}$	0.902
$\Delta \log \eta_2$	321.96	$< 10^{-60}$	0.909
Edge-CV log-fold	316.75	$< 10^{-60}$	0.895
Node-CV log-fold	324.96	$< 10^{-60}$	0.918
ξ_{50} log-fold	286.31	$< 10^{-60}$	0.809
Δd^*	188.92	1.05×10^{-40}	0.534
$\Delta\Lambda_F^*$	256.62	2.42×10^{-55}	0.725
Δd_B^F	9.97	0.0189	0.028
Best test MSE	247.04	2.86×10^{-53}	0.698
Post-best test deterioration	311.27	$< 10^{-60}$	0.879

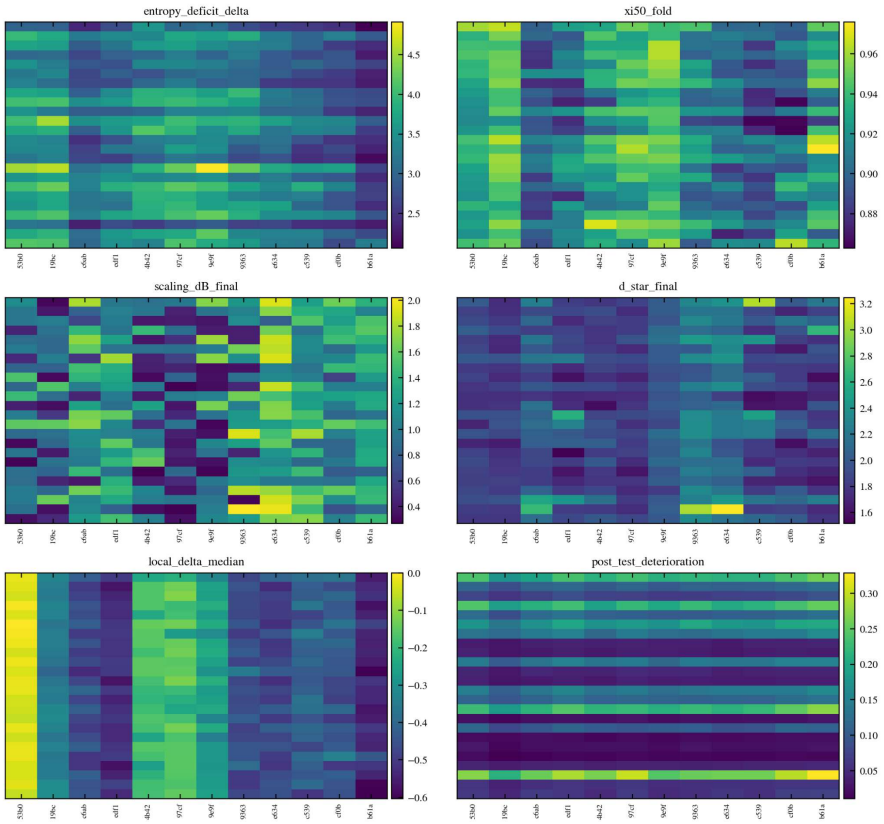


Fig. E16: Architecture-environment response matrices for six quantities of the environment matrix. Columns are the 12 fixed architectures and rows are the 24 selected environments. The figure makes the variance decomposition concrete: some descriptors retain strong architecture-specific vertical structure, post-best functional deterioration is dominated by environment-level horizontal bands, and the geometric exponents show a more mixed response.

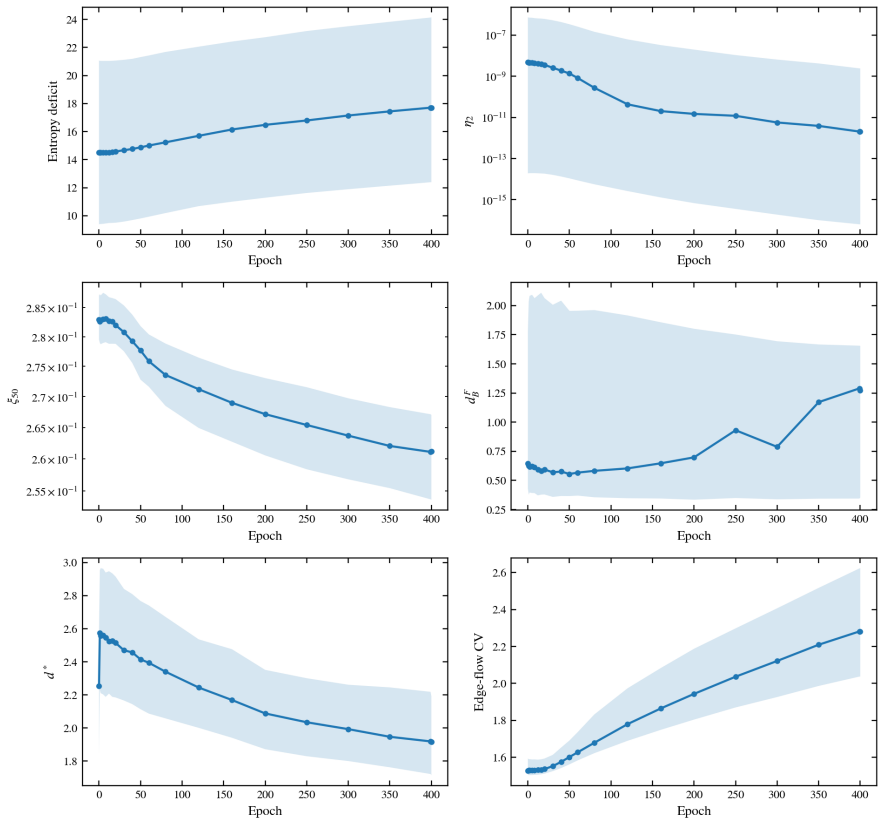


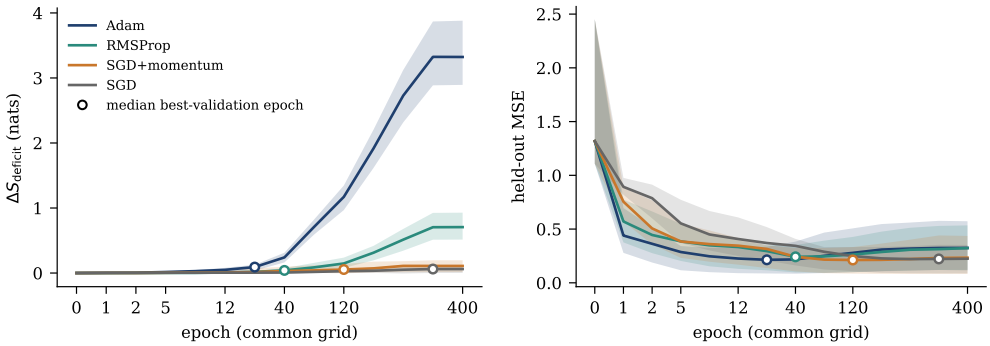
Fig. E17: Population-level trajectories of the environment matrix across 576 runs. Curves show the cohort median and the shaded bands summarise architecture-environment-seed variation. Entropy deficit, order-two path concentration, ξ_{50} contraction and edge-flow heterogeneity exhibit persistent common directions. The geometric slopes are more structured: d^* shows an early sharpening followed by long softening, while d_B^F has a pronounced late response and wide between-run dispersion.

Table E19: Descriptive fixed-factor Type-II sum-of-squares decomposition (per cent of total model-plus-residual sum of squares) on the 118 completely matched blocks. Architecture, environment, seed and the displayed interactions are design factors, not draws from a super-population; these percentages are therefore descriptive of the mapped panel rather than causal population variance components.

Outcome	Opt.	Arch.	Env.	Seed	A×E	O×A	O×E	Resid.
$\Delta S_{\text{deficit}}$	93.5	0.1	1.2	0.1	0.1	1.7	1.8	1.5
Edge-CV log-fold	82.4	6.4	0.7	0.3	0.1	6.3	1.0	2.8
Node-CV log-fold	89.7	3.3	0.9	0.5	0.2	2.1	0.6	2.8
Metric-scale log-fold	64.7	5.9	2.6	0.3	0.7	5.8	1.7	18.3
Crossover change	53.3	3.1	1.1	1.0	0.6	7.2	2.9	30.8
Lacunarity-peak change	71.9	6.5	0.5	0.8	0.1	14.2	1.5	4.4
Finite-window exponent change	0.7	7.9	0.1	15.3	1.7	2.3	1.7	70.4
Best-test log MSE	0.4	0.0	99.2	0.0	0.1	0.0	0.2	0.1

Table E20: Descriptive factor decomposition within the fully balanced 180-run Adam mechanistic arm. Because every Adam run uses the same dynamic partition, Adam-only cumulative J is comparable here even though it is not used for raw cross-optimiser comparison on the original checkpoint grid.

Adam outcome	Arch.	Env.	Seed	A×E	Resid.
$\Delta S_{\text{deficit}}$	19.5	53.9	3.8	2.4	20.3
Median selection covariance	31.6	50.3	1.9	4.8	11.5
Median selection intensity	44.0	33.9	1.6	13.0	7.4
Cumulative J (Adam only)	96.6	0.4	0.2	1.0	1.9
ξ_{50} log-fold	37.8	14.1	0.5	4.0	43.7
Δd^*	35.6	11.0	4.6	4.4	44.4
Post-best test deterioration	3.3	94.1	0.1	1.8	0.7



(a) Entropy-deficit growth

(b) Held-out error

Fig. E18: The function settles early, while the measure keeps concentrating at a rate set by the optimiser. Median and interquartile range over the 118 matched blocks of the optimiser panel, on the common checkpoint grid. Open circles mark each optimiser's median best-validation epoch (20, 40, 120 and 399). (a) $\Delta S_{\text{deficit}}$ from initialisation. (b) Held-out mean squared error.

Table E21: Optimiser-specific motion of the measure, its geometry and its selection field in the optimiser panel. Each model has a fixed intercept for the matched block (architecture, environment, seed) and an optimiser-by-time interaction, and the table reports the cluster-robust joint Wald test of that interaction. State models use the 14 common checkpoints, geometry models the seven common geometry checkpoints, and dynamic models the six epoch intervals stored identically for every optimiser.

Level	Outcome	Obs.	Blocks	Wald χ^2	p
State	S_{deficit}	6608	118	10222.5	$< 10^{-99}$
	log edge CV	6608	118	3227.7	$< 10^{-99}$
	log node CV	6608	118	18471.6	$< 10^{-99}$
Geometry	$\log \xi_{50}$	3304	118	932.6	$< 10^{-99}$
	d^*	3304	118	934.7	$< 10^{-99}$
	$\log \Lambda_F^*$	3304	118	3159.6	$< 10^{-99}$
	d_B^F	3304	118	23.6	0.168
Selection	Alignment $\rho_p(G, \log p)$ (Fisher z)	2832	118	422.1	1.52×10^{-80}
	Selection covariance	2832	118	4221.7	$< 10^{-99}$
	$\log \mathrm{Var}_p(G)$	2832	118	8548.3	$< 10^{-99}$
	$\log J$	2832	118	9379.7	$< 10^{-99}$
	Interval $\Delta S_{\text{deficit}}$	2832	118	1647.9	$< 10^{-99}$

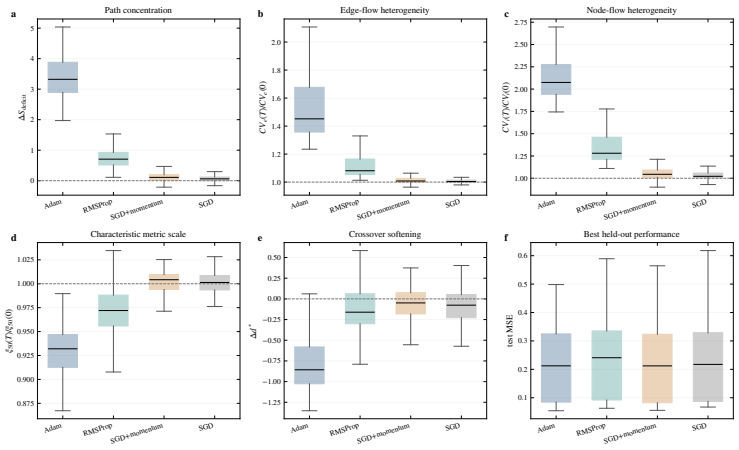


Fig. E19: **Matched five-seed optimiser endpoints.** All panels use the 118 architecture-environment-seed blocks completed by all four optimisers. **(a)** Entropy-deficit growth. **(b,c)** Edge- and node-flow coefficient-of-variation folds. **(d)** Characteristic metric-scale fold. **(e)** Change in crossover peak height d^*. **(f)** Test MSE evaluated at the validation-selected checkpoint on the exact common state grid. Horizontal dashed lines mark no structural change where applicable. Structural reorganisation follows the strong ordering Adam $\gg$ RMSProp $\gg$ momentum-SGD $\simeq$ SGD, whereas best held-out performance occupies a much narrower optimiser range.

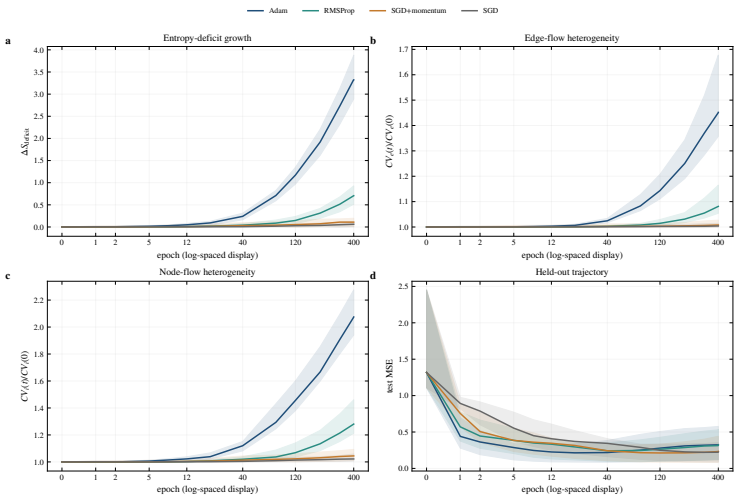


Fig. E20: **Exact common-state trajectories.** Median and interquartile range over the 118 complete optimiser blocks, evaluated only on the common state grid. **(a)** Entropy-deficit growth from initialisation. **(b,c)** Edge- and node-flow CV folds. **(d)** Held-out MSE. The horizontal coordinate is displayed as $\log_{10}(1+t)$ only to retain early checkpoints; all tick labels are actual epochs.

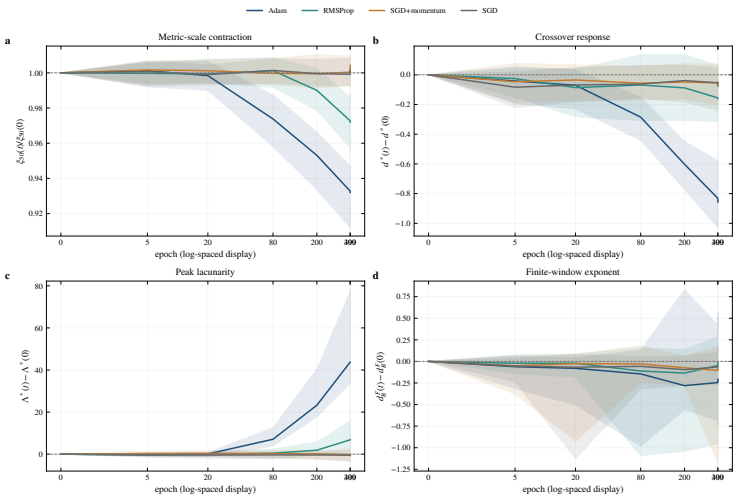


Fig. E21: **Exact common-geometry trajectories.** Median and interquartile range over the 118 complete blocks at the geometry checkpoints shared by all optimisers. **(a)** ξ_{50} fold relative to initialisation. **(b)** Change in crossover peak height d^*. **(c)** Change in peak flow lacunarity Λ^*. **(d)** Change in strict finite-window d_B^F. Optimiser-dependent contraction, softening and lacunarity emerge clearly; the finite-window exponent remains broad and does not acquire a common optimiser trajectory.

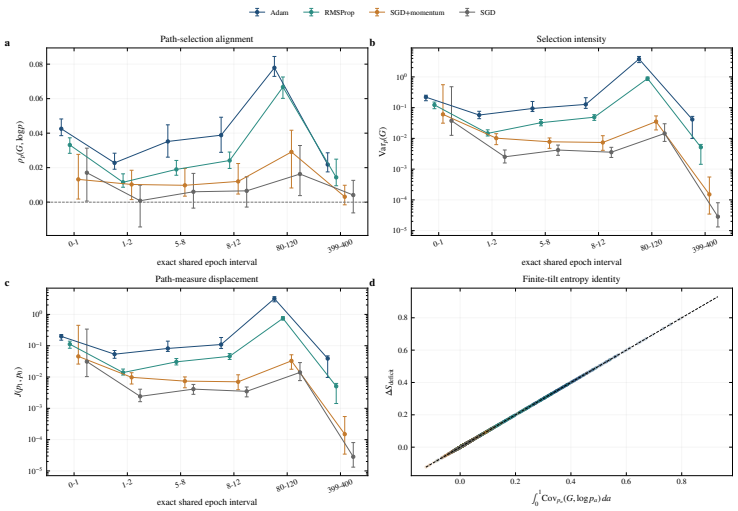


Fig. E22: **Optimiser dynamics on exact shared macro intervals.** **(a)** Path-selection alignment $\rho_p(G, \log p)$. **(b)** Selection intensity $\mathrm{Var}_p(G)$ on a logarithmic axis. **(c)** Jeffreys path-measure displacement J on a logarithmic axis. Points and bars are median and interquartile range over the 118 complete paired blocks. **(d)** Numerical validation of the exact finite-tilt entropy identity; the dashed line is equality.

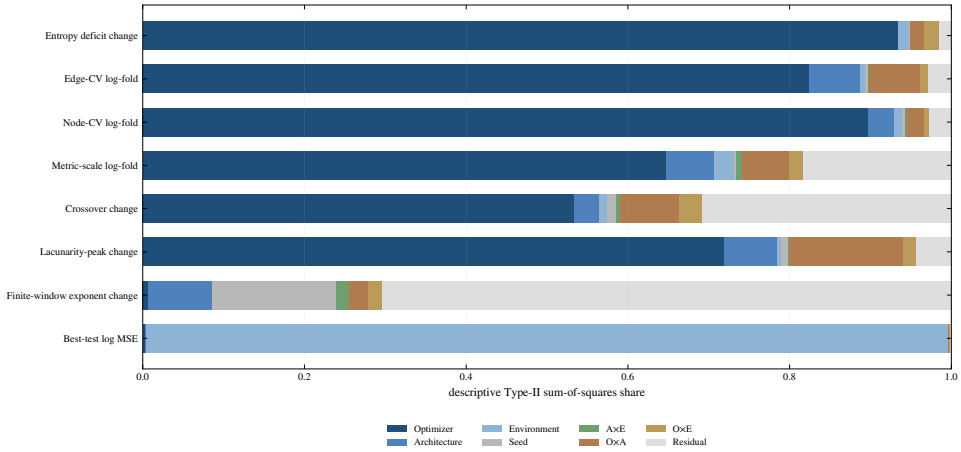


Fig. E23: **Descriptive factor decomposition in the matched optimiser panel.** Bars show Type-II sum-of-squares shares from the fixed-factor Type-II decomposition, normalised by total model-plus-residual sum of squares. The figure is descriptive of the selected panel, not a causal population variance decomposition. Optimiser dominates path concentration and marginal-flow reorganisation; environmental variation dominates best held-out performance; d_B^F remains largely residual/seed/architecture contingent.

E.5 CNN structural disposition

In the CNN architecture panel, 46 of 48 architectures move in the concentrating direction for the entropy deficit, $\log \eta_1$ and $\log \eta_2$, with a median entropy-deficit increase of 2.039 nats (Holm-adjusted $p = 1.78 \times 10^{-13}$). The two counter-directional architectures remain in the analysis, so the law is recurrent and not logically universal. The median deficit first relaxes from 4.094 to 3.959 by epoch 5 and then rises to 6.108 by epoch 360. The increase in occurrence specialisation is carried almost entirely by the parameter-identity term S_{param}, while the spatial-reuse term S_{reuse} changes little.

Across the 24 environments of the CNN environment panel, the entropy deficit moves in the same direction in 258 of 288 architecture-environment cells (0.896), and architecture carries most of the structured variance ($U_A = 0.976$). The finite-window exponent shows no systematic endpoint shift ($p \approx 0.679$), although inter-box retention decreases in all 24 environment means.

In the CNN optimiser panel, the median entropy deficit at epoch 360 is 3.945 under RMSProp, 3.713 under Adam, 2.841 under SGD and 2.746 under SGD with momentum ($W = 0.373$). The ordering emerges through crossings. Optimiser concordance is zero at initialisation, rises above 0.18 by epoch 10, falls to about 0.01 near epoch 80 and grows to 0.373 by epoch 360. Adam is held at its earlier operating point, and the comparators are calibrated on training loss alone (Appendix D.2).

E.6 Overview of the realisation results

Table 1 in Section 5.2 gives the structural-to-conditional changes of concentration, flow heterogeneity and the CNN sharing decomposition in the controlled (synthetic) populations. Table E22 completes it for the BCANw descriptors. Each entry compares the structural and the conditional measure of the same trained state. The direction of geometric change differs between the two families, so the rows carry no arrow and the CNN entries state the direction counted. The detailed tables and figures follow in Appendix E.7 for MLPs and in Appendix E.8 for CNNs, and the benchmark case studies are reported in Appendices E.9 to E.12.

E.7 MLP realisation in the controlled populations

This subsection reports the conditional analyses of the synthetic MLP populations summarised in Section 5.2. The architecture cohort contains 20 architectures with eight seeds under one environment and Adam (160 ReLU runs), with a six-architecture tanh sentinel of 48 runs. The activation panel crosses 20 configurations, two environments, four seeds and six activation laws. The environment matrix crosses 24 frozen environments, six architectures, two seeds and two paired tasks, and the optimiser panel crosses six architectures, two environments, three seeds, two activation classes and four optimisers. Table E23 gives the sampling scale of the last three; inputs are nested within trained runs throughout, so the inferential unit is the matched model block.

Table E22: Structural-to-conditional change of the BCANw descriptors in the controlled populations (synthetic tasks). Arch. ρ_s is the Spearman correlation between the structural and conditional architecture orderings (MLP: median over seeds, 20 architectures). Env. gives the CNN median same-state change $\Delta^{\mathrm{S}\to\mathrm{C}}$ over the 24 environments and the number of environments moving in the stated direction. Opt. W is Kendall's W of the optimiser ordering. The MLP optimiser entries are conditional W at epoch 280. The CNN optimiser entries are conditional W on a compact subset of 4 to 12 blocks, of which only retention survives the Holm correction. A dash means that the source does not establish the quantity.

	MLP			CNN		
Observable	Arch. ρ_s	Env. Δ $(n/24)$	Opt. W	Arch. ρ_s	Env. Δ $(n/24)$	Opt. W
ξ_{50}	-0.041	–	0.296	–	-0.054 (23 $\downarrow$)	0.047
d^*	-0.398	–	0.396	–	-0.270 (24 $\downarrow$)	0.175
Λ_F^*	0.498	–	0.506	–	-317.2 (24 $\downarrow$)	0.119
Inter-box retention	-0.447	–	0.506	–	$+0.108$ (24 $\uparrow$)	0.478

Table E23: Designs of the conditional MLP populations. **(a)** Activation panel. **(b)** Environment matrix. **(c)** Optimiser panel. Input-level observations are repeated conditional realisations nested within trained runs.

(a)

Quantity	Count
Candidate configurations	96
Retained configurations	20
Environments	2
Seeds per cell	4
Activation laws	6
Completed runs	960
Matched blocks	160
Conditional-input rows	491,520
Population geometries	2,880
Input-resolved geometries	2,304
Finite-response checks	6,000

(b)

Quantity	Count
Frozen environments	24
Fixed architectures	6
Tasks	2
Seeds	2
ReLU trajectories	576
Tanh trajectories	96
Total trajectories	672
Finite-response checks	2,000
Paired task states	336
Input-resolved geometries	768

(c)

Quantity	Count
Architectures	6
Environments	2
Seeds	3
Activation classes	2
Optimisers	4

Quantity	Count
Core trajectories	288/288
Extension runs	48/48
Training horizon	280 epochs
State checkpoints	11
Geometry checkpoints	6

Table E24: Architecture and activation effects on the conditional MLP measure. **(a)** Architecture effects in the architecture cohort at the best-validation state (eight matched seed blocks, 20 architectures), with Holm adjustment. **(b)** Friedman effects across the six activation laws of the activation panel (160 matched blocks).

(a)

Observable	χ^2_F	W_A
Structural S_{deficit}	149.193	0.982^{***}
Conditional S^c_{deficit}	148.636	0.978^{***}
Conditional η^c_2	148.671	0.978^{***}
$D_{\text{KL}}(\mu^c \,\|\, \mu^S)$	147.100	0.968^{***}
$H^2(\mu^c, \mu^S)$	147.100	0.968^{***}
Struct./cond. edge rank	137.314	0.903^{***}
Conditional Λ^*_F	145.971	0.960^{***}
Conditional d^*	127.407	0.838^{***}
Conditional retained flow	120.058	0.790^{***}
Conditional CV_{edge}	116.707	0.768^{***}
Conditional CV_{node}	117.921	0.776^{***}
Conditional ξ_{50}	107.614	0.708^{***}
Hidden input specialisation	95.021	0.625^{***}
Conditional local Δ^F	95.357	0.627^{***}
Best test MSE	122.457	0.806^{***}

(b)

Observable	χ^2_F	W
Hidden input specialisation	796.62	0.996
Test MSE	764.54	0.956
Cond./struct. edge rank	725.68	0.907
$D_{\text{KL}}(\mu^c \,\|\, \mu^S)$	696.21	0.870
Conditional edge-flow CV	659.66	0.825
Conditional node-flow CV	646.31	0.808
Conditional entropy deficit	639.05	0.799
Local support-mass Δ^F	369.72	0.462
Crossover height d^*	367.52	0.459
Peak lacunarity Λ^*_F	330.73	0.413
Finite-window d^F_B	173.65	0.220
Characteristic scale ξ_{50}	138.23	0.173

Holm-adjusted significance in (a): $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; unmarked, $p \geq 0.05$. Every architecture effect has $p \leq 1.47{\times}10^{-11}$.

Table E25: Survival of the structural architecture laws under conditioning in the architecture cohort. The structural column refers to the 48-architecture structural cohort of Section 4; the matched structural companion is computed on the same networks as the conditional result and supplies the direct structural-to-conditional bridge.

Observable	Structural cohort	Matched structural	Conditional	Reading
Path concentration	$S_{\text{deficit}} \uparrow$, $\eta_{1,2} \downarrow$ in 48/48	Same signs in 20/20	Same signs in 20/20; larger amplitude	Preserved and amplified
Edge and node heterogeneity	Both increase in 48/48	Both increase in 20/20	Both increase in 20/20	Sign preserved, rank attenuated
ξ_{50}	Contracts in 48/48	Contracts in 20/20	9 contract, 11 expand	Structural law not preserved after realisation
d^*	Decreases in 36/48	Decreases in 17/20	Decreases in 15/20	Direction survives more than ordering
Λ^*_F	Late increase	Increases in 20/20	Increases in 15/20	Partially preserved
Retained quotient flow	Late decrease	Decreases in 20/20	Decreases in 15/20	Sign tendency survives; ordering can reverse
Finite scaling	Widespread; exponent non-universal	Credible in 160/160	Credible in 48/160	Existence-conditioned



Fig. E24: Activation panel. **(a)** Matched activation effects are strongest in realisation-specific probability and flow coordinates. **(b)** Activation-conditioned medians show the distinct realised regimes behind those effects. **(c)** The fixed-design decomposition separates scaffold-associated absolute organisation from activation-associated departure, heterogeneity and specialisation. **(d)** Activation ordering is most stable across environments for specialisation and flow-level variables and progressively less stable for downstream BCANw descriptors. **(e)** Characteristic scale is the strongest geometric bridge from structural disposition to conditional realisation, while crossover, lacunarity, local mismatch and validity-gated finite-window scaling are increasingly condition sensitive.

Table E26: Activation dependence of the conditional MLP measure in the activation panel. **(a)** Median best-checkpoint values by activation; these are distinct coordinates of realised computation and not an overall ranking of activations. **(b)** Normalised fixed-design (Type-II) shares.

(a)

Activation	Test MSE	S^c_{deficit}	D_{KL}	Edge CV	Specialisation	Edge-rank ρ_s
ReLU	0.1495	5.6167	3.1510	3.7352	0.5189	0.8070
Leaky-ReLU	0.1379	5.3049	2.7230	3.6083	0.4648	0.8846
tanh	0.1671	4.3405	1.7548	2.5715	0.1577	0.9113
sigmoid	0.0944	4.2801	1.6644	2.5836	0.0590	0.7513
SiLU	0.1056	4.3828	1.6040	2.8761	0.2516	0.9747
identity	0.7685	4.1723	0.6348	2.1209	0.0131	0.9999

(b)

Observable	Activation	Configuration	Act. × config.	Residual	Model R^2
Conditional entropy deficit	0.229	0.705	0.052	0.007	0.993
$D_{\text{KL}}(\mu^c \,\|\, \mu^S)$	0.734	0.123	0.098	0.013	0.987
Conditional edge CV	0.804	0.032	0.070	0.058	0.942
Hidden input specialisation	0.946	0.023	0.018	0.006	0.994
ξ_{50}	0.044	0.782	0.126	0.036	0.964
d^*	0.249	0.244	0.218	0.216	0.784
Λ^*_F	0.086	0.741	0.106	0.034	0.966
d^F_B	0.117	0.193	0.141	0.397	0.603

E.8 CNN realisation in the controlled populations

This subsection reports the conditional analyses of the synthetic CNN populations. The architecture panel contains 12 frozen architectures with four seeds (48 networks; 67,584 conditional input observations). The environment panel crosses the 24 frozen environments with the 12 architectures and two seeds (576 trajectories; 663,552 input observations), and the environment is the inferential unit for environment claims. The optimiser panel is a compact, exactly matched campaign of five architectures, three environments, two seeds and four optimisers (30 blocks, 120 trajectories), whose design is given in Table E29, together with its results. Because it is smaller than the structural optimiser panel, p-values are not compared across the two.

Table E27: Environment and optimiser panels. **(a)** Same-state structural-to-conditional deformation in the environment matrix; values are medians over the 24 frozen environments. **(b)** Four-optimiser omnibus at the best checkpoint of the optimiser panel; conditional variables use the complete valid four-optimiser blocks.

(a)

Observable	Struct.	Cond.	$\Delta^{\mathrm{S}\to\mathrm{C}}$
Regression			
S_{deficit}	3.030	6.115	+3.140
η_1	0.0483	0.00232	−0.0462
Edge CV	1.751	4.227	+2.483
Node CV	0.405	1.232	+0.828
Classification			
S_{deficit}	2.963	5.668	+2.766
η_1	0.0517	0.00349	−0.0486
Edge CV	1.714	3.914	+2.219
Node CV	0.394	1.173	+0.777

(b)

Response	Blocks	W_O	Holm p
Test MSE	72	0.303	1.53×10^{-13}
Cond. entropy deficit	66	0.706	3.80×10^{-29}
Cond. edge CV	66	0.707	3.80×10^{-29}
Cond. node CV	66	0.694	7.82×10^{-29}
KL to structural	66	0.787	1.71×10^{-32}
Hidden specialisation	66	0.661	1.73×10^{-27}
Cond./struct. edge rank	66	0.697	6.89×10^{-29}
Struct. entropy deficit	72	0.282	1.07×10^{-12}

Table E28: Integrated MLP survival from structural disposition (architecture cohort, environment matrix and optimiser panel of Section 4) to conditional realisation (the corresponding conditional populations).

Resolution	Structural populations	Conditional populations	Reading
Complete-path concentration	Common direction over architectures and environments; optimiser-dependent rate	Architecture law amplified; recurrent across environments; strong optimiser effect	Strongest surviving law is a redistribution direction, not one endpoint
Architecture	Near-perfect occupancy-rank scaffold	Near-unit concentration effects; absolute scale configuration dominated in the activation panel	Architecture fixes the carrier and much of the probability and metric scaffold
Activation and input	Enters μ_t^S only through training	Strong specialisation, KL and flow effects; input-resolved non-commutation	Realisation adds route-use variability absent structurally
Environment	Modulates magnitude and geometry; occupancy ranks highly portable	Strong effects but weak structural-to-conditional environment-rank bridge	Environment remains a factor while realisation reorganises its ordering
Task	Encoded only through the learnt structural state	Conditional task distance exceeds structural task distance	Finite realisation amplifies task separation
Optimiser	Structural tilts and optimiser-specific motion	Strong conditional state and motion effects with changed ordering	Optimiser phenotype survives; rank is object and activation dependent
ξ_{50}	Strong scaffold and recurrent contraction	Strong bridge in the activation panel but condition sensitive elsewhere	A bridge variable, not a universal conditional law
d_B^F	Finite windows common; exponent non-universal	Window existence depends on architecture, task, activation and optimiser	Strict scaling remains validity gated and secondary
Time after functional saturation	Continued structural motion	Continued realised motion	Internal organisation has a later clock than predictive improvement
Population versus input geometry	Not applicable	Non-commutation repeatedly observed	The conditional field cannot be replaced by average-flow geometry

Table E29: CNN optimiser panel. **(a)** Architectures, which vary in depth, sharing pattern, width, output resolution and receptive field (RF). **(b)** Environments. **(c)** Optimiser arms and learning rates. **(d)** Kendall's W of the four-optimiser ordering at the structural endpoint, at the conditional endpoint and for the structural-to-conditional deformation, with Holm adjustment (30 matched blocks). **(e)** Input specialisation at the endpoint. **(f)** BCANw geometry of the population-conditional flow on the compact geometry subset, with Holm adjustment.

(a)

Architecture	Morphology	Depth	Kernels	Channels	Params	Occurrences	Out. res.	RF
`C1_shallow3`	constant	3	[3,3,1]	[4,4]	184	28,368	12	5
`C2_medium3`	anchor	4	[3,3,3,1]	[4,8,8]	908	106,928	10	7
`C3_bottleneck`	bottleneck	4	[3,3,3,1]	[6,3,6]	384	50,712	10	7
`C4_pointwise`	anchor	4	[1,1,1,1]	[4,8,8]	108	27,648	16	1
`C5_deep3`	anchor	5	[3,3,3,3,1]	[4,8,8,12]	1,776	162,192	8	9

(b)

Environment	Teacher	Info. frac.	Spatial corr.	SNR	Interaction	Localisation
`E_CAN_DISTRIBUTED`	deep compositional	0.850	0.200	25	0.350	0.050
`E_B0_3da229f0a0`	shallow nonlinear	0.667	0.552	39.963	0.946	0.038
`E_B0_103c076226`	shallow nonlinear	0.892	0.730	5.216	0.916	0.958

(c)

Optimiser	Rate
Adam	0.003
RMSProp	0.003
SGD with momentum	0.030
SGD	0.010

(d)

	Kendall's W		
Observable	Structural	Conditional	Deformation
Path entropy	0.037	0.585^{***}	0.568^{***}
Entropy deficit	0.037	0.585^{***}	0.568^{***}
S_{occ}	0.019	0.534^{***}	0.612^{***}
S_{param}	0.016	0.393^{***}	0.388^{***}
S_{reuse}	0.240^{**}	0.631^{***}	0.648^{***}

(e)

Observable	N	χ^2_F	W
Node MI	30	58.680	0.652^{***}
Channel MI	30	26.760	0.297^{***}
Spatial MI	30	38.200	0.424^{***}

(f)

Observable	N	χ^2_F	W
ξ_{50} and $\log \xi_{50}$	12	1.700	0.047
d^*	4	2.100	0.175
x_{d^*}	4	1.200	0.100
Crossover lacunarity	12	4.300	0.119
Quotient-flow KS	8	18.750	0.781^{**}
Inter-box retention	12	17.200	0.478^{**}

Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; unmarked, $p \geq 0.05$. The p values are Holm-adjusted in (d) and (f) and unadjusted in (e). In (d), the structural orderings have $p = 1$ except for S_{reuse} ($p = 0.001$), and every conditional and deformation ordering has $p \leq 5.11 \times 10^{-7}$. In (e), every $p \leq 6.61 \times 10^{-6}$. In (f), the unmarked rows have $p = 1$, and the marked rows have $p = 0.002$ and 0.004.

Table E30: CNN architecture panel: same-state structural-to-conditional endpoint changes at epoch 360. **(a)** Intrinsic and sharing observables over the 48 networks. **(b)** BCANw geometry of the population-conditional flow, where N counts the networks with a defined value. r_{rb} is the paired rank-biserial effect.

(a)

Observable	Struct.	Cond.	$\Delta^{\mathrm{S}\to\mathrm{C}}$	r_{rb}
Path entropy	11.874	9.354	−2.794	−0.998
Entropy deficit	3.082	6.003	+2.794	+0.998
η_1	0.046	0.002	−0.043	−0.998
η_2	0.013	5.59×10^{-4}	−0.013	−0.997
S_{occ}	1.251	2.944	+1.634	+1
S_{param}	1.161	0.915	−0.136	−0.558
S_{reuse}	0.107	1.915	+1.809	+1
$\eta_{1,\mathrm{occ}}$	0.313	0.059	−0.251	−1
$\eta_{1,\mathrm{param}}$	0.346	0.412	+0.036	+0.612
$\eta_{1,\mathrm{reuse}}$	0.901	0.152	−0.746	−1

(b)

Observable	N	$\Delta^{\mathrm{S}\to\mathrm{C}}$	r_{rb}	Holm p
$\log \xi_{50}$	48	−0.581	−0.187	0.317
d^*	44	1.650	−0.749	3.23×10^{-5}
x_{d^*}	44	11.517	−0.329	0.230
Transition fraction	48	0.681	−0.283	0.317
Cross. lacunarity	48	427.802	−0.660	2.13×10^{-4}
Quotient-flow KS	44	0.389	−0.418	0.075
Inter-box retention	48	0.295	+0.679	1.33×10^{-4}
Local Δ^F	46	−0.066	+0.600	0.001
Credible d_B^F	16	1.174	+0.456	0.317

Table E31: CNN environment panel over the 24 environments. **(a)** Intrinsic and sharing observables: recurrence of the structural-to-conditional deformation, and structural-to-conditional architecture-rank persistence. **(b)** Recurrence of the deformation for the BCANw geometry. In (a) and (b), Δ is the median environment-level change, + and − count the environments with a positive or negative median change, and the mark gives the Holm-adjusted sign test; the rank correlation ρ is summarised by its quantiles over environments and by the share of environments with $\rho > 0.7$. **(c)** Phenotype-environment associations surviving the main Holm family; they are descriptive relations over the designed panel and not causal effects of the coordinates.

(a)

	Deformation			Architecture-rank persistence ρ					
Observable	Δ	+	−	Median	Q25	Q75	Min	Max	Share ρ>0.7
Path entropy	-2.903^{***}	0	24	0.923	0.862	0.939	0.769	0.972	1
Entropy deficit	$+2.903^{***}$	24	0	0.909	0.853	0.937	0.706	0.986	1
η_1	-0.009^{***}	0	24	0.937	0.909	0.960	0.825	0.993	1
η_2	-0.001^{***}	0	24	0.958	0.937	0.972	0.853	0.993	1
S_{occ}	$+1.660^{***}$	24	0	0.469	0.383	0.603	0.175	0.839	0.125
S_{param}	-0.186^{***}	1	23	0.031	−0.089	0.260	−0.427	0.483	0
S_{reuse}	$+1.916^{***}$	24	0	0.406	0.278	0.463	−0.252	0.594	0

(b)

Observable	Δ	+	−
$\log \xi_{50}$	-0.054^{***}	1	23
d^*	-0.270^{***}	0	24
x_{d^*}	-18.685^{***}	0	24
Transition fraction	-0.063^{***}	0	24
Cross. lacunarity	-317.232^{***}	0	24
Quotient-flow KS	-0.052^{***}	2	22
Inter-box retention	$+0.108^{***}$	24	0
Local Δ^F	$+0.043^{***}$	23	1
Credible d_B^F	−0.041	9	15

(c)

Phenotype	Descriptor	ρ_s
Cond. entropy deficit	Interaction	0.782^{***}
Cond. η_1	Interaction	-0.762^{**}
Cond. η_2	Interaction	-0.676^{*}
$\eta_{1,\text{param}}$	Spatial corr.	-0.672^{*}
ρ_{param} retention	Interaction	-0.667^{*}

Holm-adjusted significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; unmarked, $p \geq 0.05$. In (a) and (b), every marked row has $p \leq 7.18\times10^{-5}$, and the unmarked row has $p = 0.307$. In (c), the adjusted p values are 5.34×10^{-4}, 0.001, 0.024, 0.027 and 0.030 in row order.

Table E32: CNN validity boundaries and integrated result. **(a)** Validity and degeneracy cases retained in the conditional CNN populations; each is kept and excluded only where the corresponding quantity is mathematically undefined. **(b)** Integrated result by resolution; "preserved" refers to a recurrent direction or rank organisation, not to exact numerical invariance.

(a)

Population	Case	Boundary	Treatment
Architecture panel	One network (seed 3) at $t = 360$	Conditional valid fraction 0.617	Retained
Environment panel	`C4_pointwise`, seed 1, environment `E_SOB_0fb26ea74e`	Fully degenerate at $t = 240$ and $t = 360$	Retained; one geometry skipped
Environment panel	Endpoint jobs	88.0% fully valid	Partial validity shown explicitly
Optimiser panel	SGD with momentum, `C3_bottleneck`, `E_BO_3da229f0a0`, seed 1	33/128 valid endpoint contrasts	Retained
All populations	Strict d_B^F	Credible window required	Never imputed

(b)

Resolution	Structurally established	Under realisation	Reading
Global complete paths	Recurrent concentration; strong architecture memory; environment robust	Strong conditional concentration; global architecture order largely retained	Strongest quasi-invariant phenotype
Shared parameter identity	Dominant source of structural occurrence specialisation	Parameter rank partly retained; S_{param} deformation weak or negative	Identity remains available but is not the dominant selector
Spatial reuse	Small structural motion; very stable architecture rank	Dominant positive specialisation; occurrence ordering largely lost	Principal CNN mechanism of conditional selection
Environment	Small main effect; larger architecture-by-environment modulation	Global law survives; interaction strength modulates realised concentration	Context changes expression more than the global law
Optimiser	Changes route, endpoint and BCANw organisation	Conditional phenotype and deformation strongly optimiser dependent; order reorganised	Realisation is optimiser dependent
BCANw geometry	Coherent flow reorganisation; d_B^F non-universal	Recurrent conditional deformation; population non-commutation	Multiscale organisation is real; exponent claims remain gated

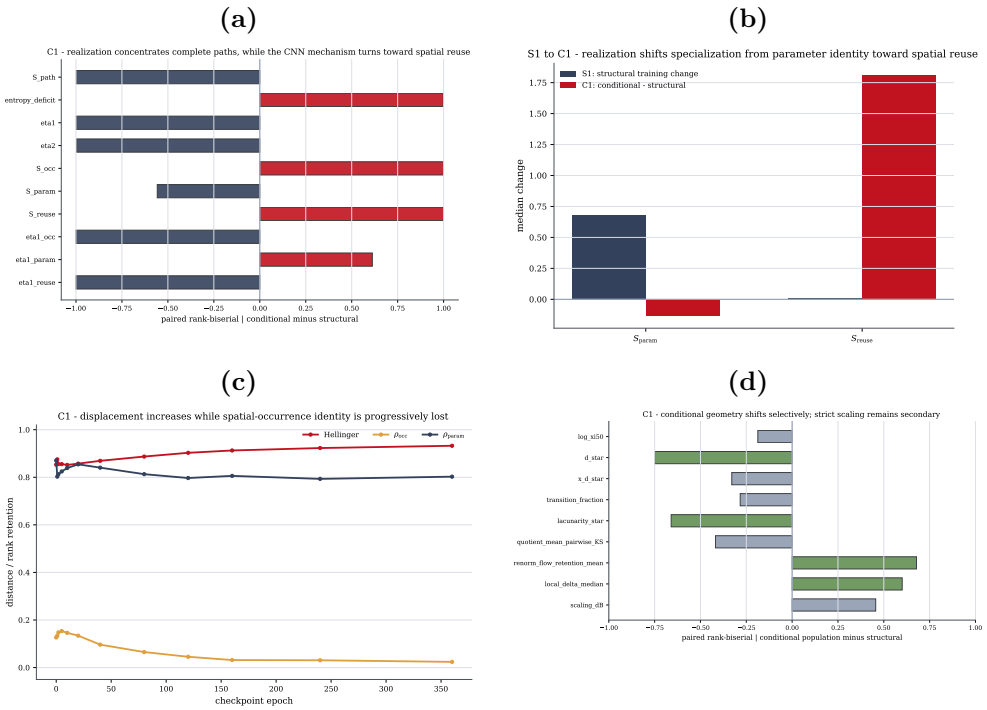


Fig. E25: CNN architecture panel under conditional realisation. **(a)** Same-state structural-to-conditional endpoint effects: complete paths become strongly concentrated, while the sharing decomposition shows a negative parameter-specialisation change and a very large positive reuse-specialisation change. **(b)** Structural training raises occurrence specialisation almost entirely through S_{param}, whereas conditioning produces the opposite pattern, with a dominant positive S_{reuse} deformation; the bars are different contrasts and are compared mechanistically, not as identical estimands. **(c)** Structural-to-realised distance increases through training while occurrence-rank retention approaches zero; shared-parameter identity remains much more recognisable. **(d)** Structural-to-population-conditional BCANw effects at the endpoint are selective and validity gated, and strict d_B^F does not carry the central result.

E.9 Benchmark case studies: design and validity

The two benchmark case studies of Section 5.3 share one factorial and one analysis pipeline (Table E33). MPI3D-Toy enumerates object colour, shape, size, camera height, background colour, horizontal position and vertical position with cardinalities $(6, 6, 2, 3, 3, 40, 40)$; the target is the horizontal position, standardised on the training set, and the other factors act as structured nuisance variation. Fashion-MNIST images are mapped to $[0, 1]$. In both, images are area-resized to 16×16 and standardised by the training mean and standard deviation, and the training, validation and test subsets contain 20,000, 5,000 and 5,000 examples, stratified over the 40 position levels on MPI3D. The MLPs have hidden widths M1 $[512, 384, 256, 128, 64]$, M2 $[384, 384, 384, 256, 128]$ and M3 $[512, 256, 64, 256, 128]$. The CNNs use valid convolutions and global spatial averaging with channels and kernels C1 $[4, 8, 8, 12]$, $[3, 3, 3, 3, 1]$; C2 $[4, 8, 3, 8, 10]$, $[3, 3, 1, 3, 3, 1]$; and C3 $[3, 6, 8, 8, 10, 8]$, $[3, 3, 3, 1, 3, 1, 1]$. The families are not parameter matched (median 4.68×10^5 and 1.57×10^3 parameters on Fashion-MNIST), so family performance is used only as a competence check. Training runs for at most 160 epochs with batch size 256, zero weight decay and learning rates 10^{-2} for SGD with momentum and 10^{-3} for RMSProp and Adam.

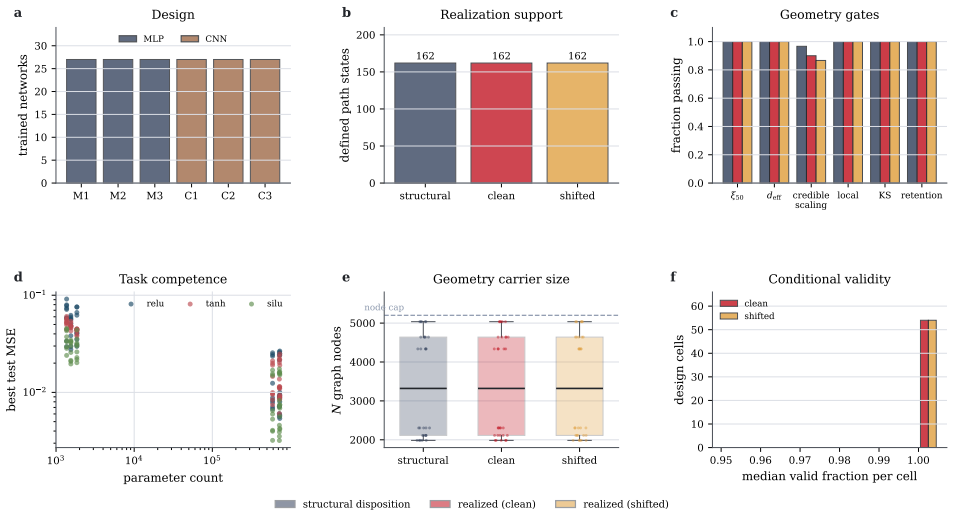
The shifted state applies deterministic additive Gaussian noise at 15 dB in raw image space and reapplies the clean training normalisation, with the reference $x^\circ = 0$ held fixed, so the clean and shifted states differ only through x. The geometry cohort is preregistered as Adam across all activations together with ReLU across all optimisers at seed 0, giving 30 networks per state. BCANw uses 64 scales, a 5,200-node cap, transition-supported d_{eff}^F, common-window local fits, quotient-flow statistics and an input-resolved sentinel on M2 and C2 under Adam. Before the full runs, a preflight of 216 initialisation and stress states on MPI3D and of 108 initialisation and 108 five-epoch stress states on Fashion-MNIST passed every support gate. All geometries satisfy the construction gates; the stricter finite-window credibility criterion holds for 96.7%, 90.0% and 86.7% of the structural, clean and shifted MPI3D geometries and for all Fashion-MNIST geometries (Figure E26). Across the 162 Fashion-MNIST networks, the median best test cross-entropy is 0.451 and the median accuracy 0.849.

E.10 Case study: MPI3D

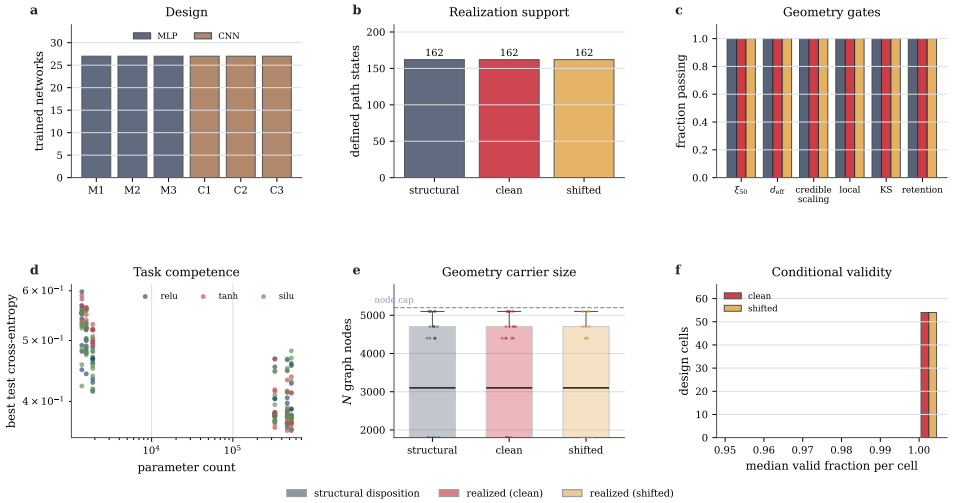
The MPI3D cohort medians and matched tests are reported with those of Fashion-MNIST in Appendix E.12. The figure below resolves them over networks and activations. Figures over scales and carriers are given in the results archive.

E.11 Case study: Fashion-MNIST

The Fashion-MNIST cohort medians and matched tests are reported with those of MPI3D in Appendix E.12. The figure below follows the MPI3D figure. Further figures are given in the results archive.



(i) MPI3D



(ii) Fashion-MNIST

Fig. E26: Design and numerical validity ledgers of the two benchmarks. In each ledger: **(a)** the 162-network factorial, with 27 networks per architecture; **(b)** all networks yield structural, clean and shifted path states; **(c)** geometry-gate pass fractions, with finite-window scaling credibility separated from the basic admissibility gates; **(d)** best test error (MSE on MPI3D, cross-entropy on Fashion-MNIST) against parameter count, shown only as a competence check because the MLP and CNN budgets are not matched; **(e)** geometry-carrier node counts relative to the 5,200-node cap; **(f)** conditional valid fraction per design cell. Boxes are medians and interquartile ranges with 1.5 IQR whiskers. Counts are exact, and undefined quantities are never set to zero.

Table E33: Controlled alignment and intentional differences between the two benchmark case studies. The factorial and the analysis are held fixed wherever possible; task semantics and the data-generating structure differ by design.

Design element	MPI3D	Fashion-MNIST
Task	Regression of normalised horizontal position	Ten-class image classification
Data manifold	Controlled generative image factors	Natural benchmark image distribution
Architectures	M1, M2, M3, C1, C2, C3	
Optimisers	SGD with momentum, RMSProp, Adam	
Activations	ReLU, tanh, SiLU	
Seeds	3	
Trained networks	162	162
Path states	486	486
Endpoint population	160, balanced over position levels	160, 16 per class
Population geometries	90	90
Input-resolved geometries	36	36
Local node rows	11,520	11,520
Degenerate realisation states	0	0
Shift	Deterministic additive Gaussian sensor perturbation at 15 dB in raw image space	

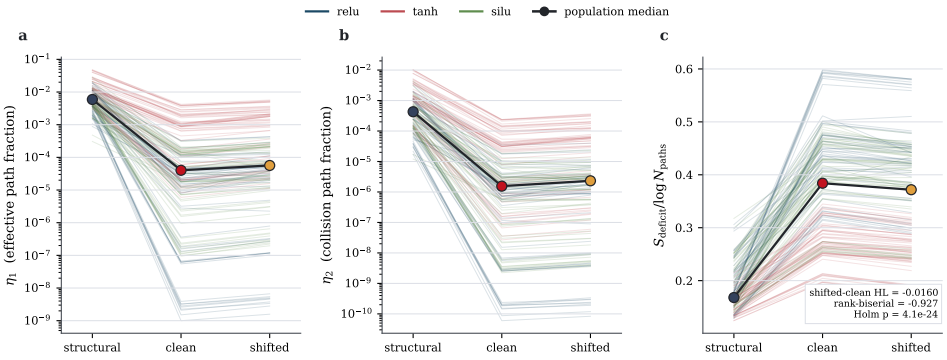


Fig. E27: MPI3D: matched structural, clean and shifted states. Thin lines are networks coloured by activation; heavy lines and markers are cohort medians. **(a)** η_1. **(b)** η_2. **(c)** Normalised entropy deficit, with the shifted-minus-clean Hodges-Lehmann estimate, rank-biserial effect and Holm-adjusted Wilcoxon p. Each network is its own block. The structural-to-clean contraction is followed by a smaller but highly concordant de-concentration under the sensor shift, which is a redistribution of realised mass.

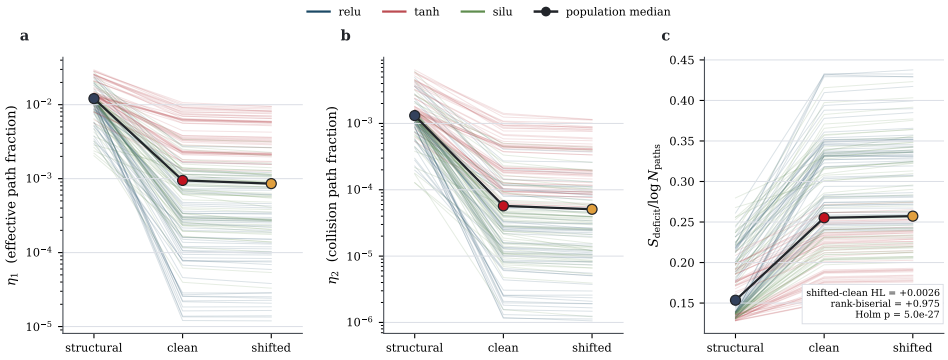


Fig. E28: Fashion-MNIST: matched structural, clean and shifted states. Each faint line is one network coloured by activation, and the heavy line is the median. **(a)** η_1. **(b)** η_2. **(c)** Normalised entropy deficit, with the shifted-minus-clean Hodges-Lehmann estimate, rank-biserial effect and Holm-adjusted paired p. The sensor shift produces a further, highly concordant concentration of the conditional measure.

Table E34: State medians on both benchmarks. Intrinsic rows are medians over the 162 networks of the per-network median over 160 endpoint events; geometric rows are medians over the 30 preregistered geometries of each state. The η_1 and η_2 rows are given in units of 10^{-5} and 10^{-6} respectively.

	MPI3D			Fashion-MNIST		
Observable	Structural	Clean	Shifted	Structural	Clean	Shifted
η_1 ($\times 10^{-5}$)	591.7	4.020	5.652	1210	94.64	85.65
η_2 ($\times 10^{-6}$)	429.0	1.554	2.326	1315	57.23	50.84
$S_{\mathrm{def}}/\log N$	0.1680	0.3841	0.3718	0.1535	0.2552	0.2572
Path entropy (nats)	24.015	16.901	17.262	23.643	19.660	19.612
ξ_{50}	0.313	0.403	0.345	0.306	0.290	0.295
d^*	1.635	0.501	0.795	2.919	1.045	1.046
Transition fraction	0.844	0.563	0.453	0.922	0.414	0.398
Λ_F^*	445.2	1105.8	1011.2	178.0	319.2	312.0
Inter-box retention	0.325	0.065	0.106	0.682	0.358	0.390
Quotient-flow KS	0.205	0.178	0.135	0.155	0.192	0.199
Pooled local Δ^F	−0.047	+0.004	−0.028	−0.017	−0.004	−0.018

E.12 Cross-benchmark comparison and architecture profiles

The two case studies share the architectures, optimisers, activations, seeds, endpoint populations, shift, geometry cohort and BCANw implementation (Table E33), so their results can be aligned directly. Tables E34 to E38 report the state medians, the matched deformations, the factor concordances, the carrier-resolved local and sharing observables and the post-hoc performance models. Table E39 classifies each principal claim as recurrent or contingent.

Table E35: Matched deformations on both benchmarks. **(a)** Intrinsic observables over the 162 networks: Hodges-Lehmann paired shift (HL) and paired rank-biserial effect r_{rb}, with Holm adjustment within benchmark and contrast. **(b)** Geometric observables over the 30 geometries: paired rank-biserial effect r_{rb}, with Holm adjustment. Positive effects mean that the second state of the contrast is larger.

(a)

		MPI3D		Fashion-MNIST	
Contrast	Observable	HL	r_{rb}	HL	r_{rb}
Clean − structural	η_1	-0.007	-1.000^{***}	-0.011	-1.000^{***}
	η_2	-6.55×10^{-4}	-1.000^{***}	-0.001	-1.000^{***}
	$S_{\mathrm{def}}/\log N$	$+0.184$	$+1.000^{***}$	$+0.095$	$+1.000^{***}$
	Path entropy	-5.073	-1.000^{***}	-2.567	-1.000^{***}
Shifted − clean	η_1	$+5.27\times10^{-5}$	$+0.910^{***}$	-1.04×10^{-4}	-0.986^{***}
	η_2	$+1.06\times10^{-6}$	$+0.833^{***}$	-1.06×10^{-5}	-1.000^{***}
	$S_{\mathrm{def}}/\log N$	-0.016	-0.927^{***}	$+0.003$	$+0.975^{***}$
	Path entropy	$+0.475$	$+0.935^{***}$	-0.070	-0.980^{***}

(b)

	Clean − structural		Shifted − clean	
Observable	MPI3D	Fashion-MNIST	MPI3D	Fashion-MNIST
ξ_{50}	$+0.677^{**}$	-0.948^{***}	-0.755^{***}	$+0.665^{**}$
d^*	-0.970^{***}	-0.888^{***}	$+0.669^{**}$	$+0.234$
Transition fraction	-0.731^{**}	-0.961^{***}	-0.684^{**}	-0.209
Λ_F^*	$+1.000^{***}$	$+0.535^{*}$	-0.987^{***}	-0.187
Inter-box retention	-1.000^{***}	-0.656^{**}	$+1.000^{***}$	-0.239
Quotient-flow KS	-0.024	$+0.729^{***}$	-0.359	$+0.385$
Local Δ^F	$+0.213$	-0.157	-0.239	-0.049

Holm-adjusted significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; unmarked, $p \geq 0.05$. In (a), every adjusted p is at most 3.80×10^{-20}.

Architecture profiles.

To retain the identity of each carrier, every benchmark-architecture pair is summarised by a profile of twelve descriptors (Table E40). Intrinsic descriptors use architecture-level medians over the 27 optimiser-activation-seed cells, and geometric descriptors use the five geometry-cohort cells of each architecture. Each raw descriptor z_{jk} of profile j is converted to a percentile over the twelve benchmark-architecture profiles,

$$R_{jk} = 100\,\frac{\mathrm{rank}(z_{jk}) - 1}{12 - 1}, \tag{E75}$$

so that $R_{jk} = 100$ means only that the profile has the largest value of the named property among the twelve, and not that the architecture is better. The profiles are a coordinate transform of the estimators above and add no inferential evidence; a radial difference is not a confidence interval, and two similar shapes can differ in absolute magnitude. Figures E29 and E30 show that one architecture can occupy

Table E36: Factor concordance on both benchmarks: Kendall's W from matched-block Friedman tests. **(a)** Activation and optimiser, with 54 blocks per comparison. **(b)** Architecture within each family, over 27 matched optimiser-activation-seed blocks for the three architectures of the family.

		MPI3D			Fashion-MNIST		
Factor	State	η_1	η_2	$S_{\text{def}}/\log N$	η_1	η_2	$S_{\text{def}}/\log N$
(a)							
Activation	Structural	0.498	0.521	0.498	0.700	0.684	0.730
	Clean	0.849	0.799	0.849	0.931	0.874	0.931
	Shifted	0.838	0.778	0.838	0.916	0.849	0.916
Optimiser	Structural	0.138	0.206	0.138	0.285	0.305	0.276
	Clean	0.212	0.218	0.212	0.108	0.133	0.108
	Shifted	0.241	0.182	0.241	0.133	0.157	0.133
(b)							
MLP architecture	Structural	0.115	0.066	0.635	0.531	0.523	0.931
	Clean	0.285	0.273	0.203	0.236	0.339	0.759
	Shifted	0.235	0.289	0.174	0.202	0.325	0.759
CNN architecture	Structural	0.553	0.605	0.388	0.790	0.759	0.368
	Clean	0.759	0.759	0.203	0.931	0.931	0.680
	Shifted	0.759	0.824	0.339	0.931	0.931	0.680

Table E37: Carrier-resolved observables on both benchmarks. **(a)** Local support-mass mismatch Δ^F by family, with the within-family clean-minus-structural median displacement, rank-biserial effect and Holm-adjusted paired p. **(b)** CNN sharing decomposition: S_{occ}, S_{param} and S_{reuse} are entropy deficits and the η_1 columns the corresponding effective fractions; medians pool the CNN layer-state rows of each benchmark and state.

(a)

	Family	S	C	Sh	$\Delta_{\text{C}-\text{S}}$	r_{rb}	p
MPI3D	MLP	−0.007	−0.248	−0.207	−0.258	−0.917	0.0012
	CNN	−0.049	+0.341	+0.099	+0.385	+1.000	0.0001
FMNIST	MLP	+0.100	+0.043	+0.044	−0.067	−0.900	0.0017
	CNN	−0.108	−0.079	−0.072	+0.036	+0.633	0.0603

(b)

	State	S_{occ}	S_{param}	S_{reuse}	η_1^{occ}	η_1^{param}	η_1^{reuse}
MPI3D	S	1.245	0.937	0.139	0.288	0.392	0.870
	C	3.134	0.913	2.210	0.045	0.403	0.112
	Sh	2.902	0.877	1.974	0.057	0.420	0.144
FMNIST	S	1.289	1.047	0.131	0.276	0.351	0.877
	C	2.376	1.188	1.049	0.096	0.311	0.357
	Sh	2.411	1.161	1.091	0.092	0.318	0.343

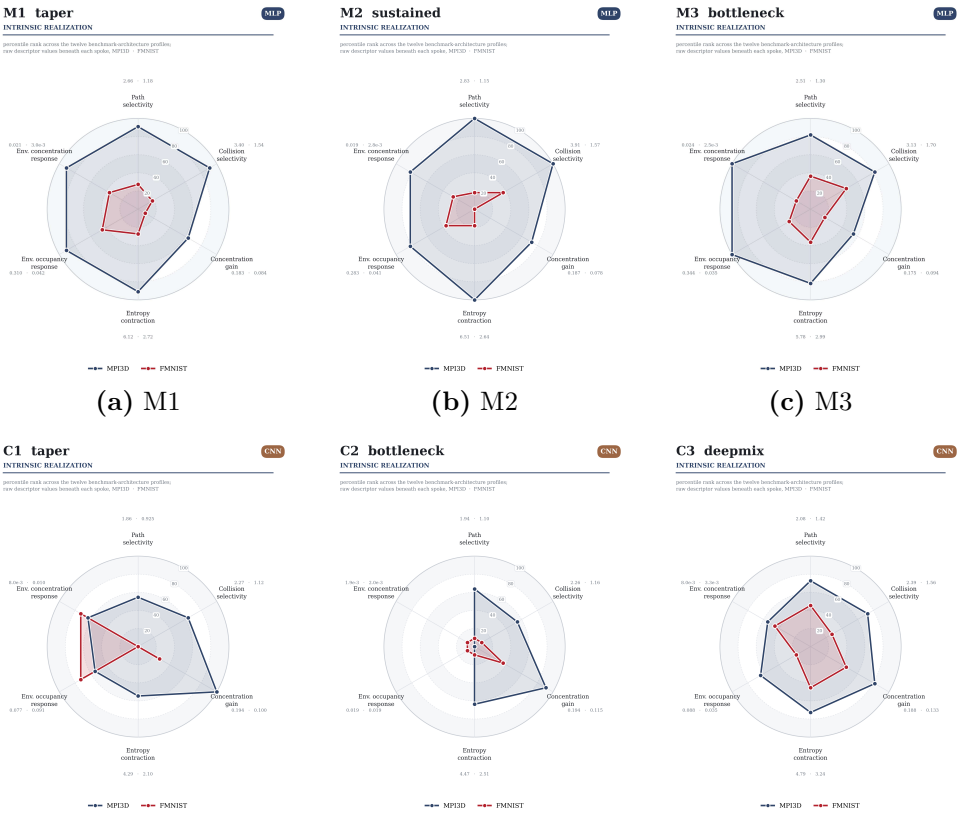
Table E38: Design-adjusted post-hoc performance models. MPI3D models log test MSE and Fashion-MNIST log test cross-entropy; covariance is cluster robust over the architecture-optimiser-activation cells. ΔR^2 is relative to the factors-only model, whose R^2 is 0.8145 on MPI3D and 0.8515 on Fashion-MNIST.

Benchmark	State	R^2	ΔR^2	$\beta_{\log \eta_1}$	95% interval	p_{cluster}
MPI3D	Structural	0.8201	0.0056	$+0.1075$	$[-0.0017, +0.2168]$	0.0537
	Clean	0.9208	0.1063	$+0.1507$	$[+0.1247, +0.1768]$	7.09×10^{-30}
	Shifted	0.9213	0.1068	$+0.1534$	$[+0.1254, +0.1815]$	8.01×10^{-27}
Fashion-MNIST	Structural	0.8531	0.0016	$+0.0192$	$[-0.0172, +0.0557]$	0.3013
	Clean	0.8564	0.0049	-0.0146	$[-0.0320, +0.0029]$	0.1024
	Shifted	0.8552	0.0037	-0.0126	$[-0.0303, +0.0051]$	0.1642

Table E39: Empirical status of the principal cross-benchmark claims. A claim is recurrent when the relation or direction agrees on both benchmarks, and contingent when its sign, strength or functional consequence changes with the task or the carrier.

Phenomenon	MPI3D	Fashion-MNIST	Status
Structural-to-clean support	Severe contraction	Severe contraction	Recurrent
Normalised concentration after conditioning	Strong increase	Strong increase	Recurrent
Activation ordering after conditioning	Strengthens markedly	Strengthens markedly	Recurrent
CNN spatial-reuse specialisation	Strong occurrence and reuse concentration	Strong occurrence and reuse concentration	Recurrent
Structural-to-clean d^* and transition support	Both decrease strongly	Both decrease strongly	Recurrent
Clean-to-shifted intrinsic response	De-concentrates	Concentrates further	Contingent
Clean-to-shifted geometry	Broad deformation	Mostly stable except ξ_{50}	Contingent
Local Δ^F carrier signs (clean)	MLP < 0, CNN > 0	MLP > 0, CNN < 0	Contingent
Realised η_1 beyond design factors	Large incremental R^2	Negligible incremental R^2	Functionally contingent

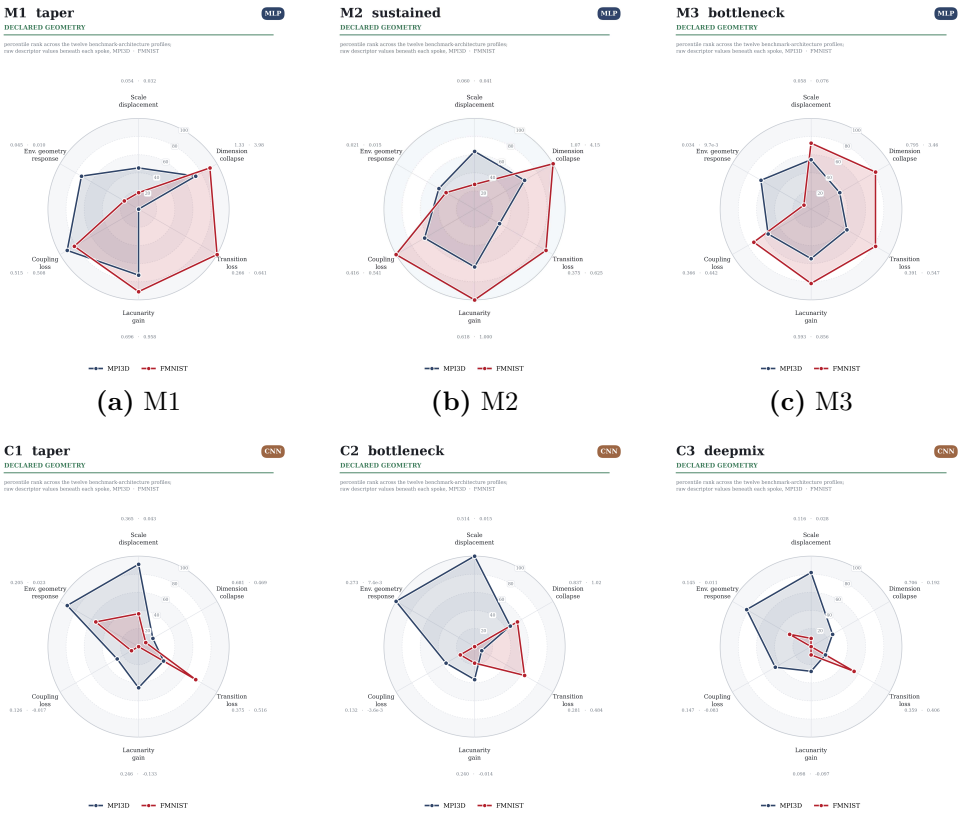
very different regions of the profile space on the two tasks, while family shapes recur. MPI3D MLPs occupy high intrinsic-selectivity and environmental-response percentiles, Fashion-MNIST MLPs show large dimension and transition collapse, MPI3D CNNs show comparatively large scale and environmental geometric reconfiguration, and Fashion-MNIST CNNs are geometrically more restrained.



(a) M1 **(b)** M2 **(c)** M3

(d) C1 **(e)** C2 **(f)** C3

Fig. E29: Intrinsic architecture profiles on MPI3D and Fashion-MNIST, on the common percentile scale of Table E40; the outer radius denotes stronger selectivity or environmental response, not better performance. **(a)** M1 lies near the top of the twelve profiles on MPI3D for path and collision selectivity, entropy contraction and both environmental responses, and is much less extreme on Fashion-MNIST. **(b)** M2 is exceptionally selective on MPI3D and occupies a substantially lower selectivity percentile on Fashion-MNIST. **(c)** On MPI3D, M3 combines strong selectivity with the largest environmental intrinsic response of the twelve profiles. **(d)** On MPI3D, C1 is distinguished more by concentration gain than by maximal selectivity, while on Fashion-MNIST it is weakly selective but environmentally responsive. **(e)** On MPI3D, C2 has minimal environmental intrinsic displacement despite substantial clean selectivity. **(f)** C3 is the most selective Fashion-MNIST CNN, while its MPI3D profile ranks substantially higher in absolute selectivity.



(a) M1 **(b)** M2 **(c)** M3

(d) C1 **(e)** C2 **(f)** C3

Fig. E30: Geometric architecture profiles on MPI3D and Fashion-MNIST, on the common percentile scale of Table E40. **(a)** On Fashion-MNIST, M1 shows strong dimension collapse, transition loss, lacunarity gain and coupling loss. **(b)** Fashion-MNIST M2 is dominated by the same four deformations, while MPI3D M2 deforms in a more balanced way. **(c)** M3 redistributes between intrinsic susceptibility and multiscale deformation according to the task. **(d)** MPI3D C1 shows a large scale displacement and environmental geometric response, which path selectivity alone does not summarise. **(e)** MPI3D C2 has the strongest scale displacement and environmental geometric reconfiguration of the twelve profiles, despite its small intrinsic environmental response. **(f)** MPI3D C3 retains substantial environmental and scale deformation, whereas Fashion-MNIST C3 is comparatively stable.

Table E40: Definition of the profile axes. The plotted radius is the cross-benchmark percentile R_{jk} of the raw descriptor; a larger radius means a stronger value of the named property.

Axis	Raw architecture-level descriptor
Intrinsic profile	
Path selectivity	$-\log_{10}\left(\eta_1^{\text{clean}}/\eta_1^{\text{struct}}\right)$
Collision selectivity	$-\log_{10}\left(\eta_2^{\text{clean}}/\eta_2^{\text{struct}}\right)$
Concentration gain	$(S_{\text{def}}/\log N)^{\text{clean}} - (S_{\text{def}}/\log N)^{\text{struct}}$
Entropy contraction	$S_{\text{path}}^{\text{struct}} - S_{\text{path}}^{\text{clean}}$
Environmental occupancy response	$\left\lvert \log_{10}\left(\eta_1^{\text{shift}}/\eta_1^{\text{clean}}\right)\right\rvert$
Environmental concentration response	$\left\lvert (S_{\text{def}}/\log N)^{\text{shift}} - (S_{\text{def}}/\log N)^{\text{clean}}\right\rvert$
Geometric profile	
Scale displacement	$\left\lvert \log_{10}\left(\xi_{50}^{\text{clean}}/\xi_{50}^{\text{struct}}\right)\right\rvert$
Dimension collapse	$d^{*,\text{struct}} - d^{*,\text{clean}}$
Transition loss	$\tau^{\text{struct}} - \tau^{\text{clean}}$, with τ the transition fraction
Lacunarity gain	$\log_{10}\left(\Lambda_F^{*,\text{clean}}/\Lambda_F^{*,\text{struct}}\right)$
Coupling loss	$\text{Retention}^{\text{struct}} - \text{retention}^{\text{clean}}$
Environmental geometry response	Mean absolute clean-to-shifted displacement over ξ_{50}, d^*, Λ_F^*, retention and quotient-flow KS, with logarithmic relative changes for positive-scale quantities

Table E41: The intervention programme. Each experiment inherits the objects of the previous ones and changes one coordinate of the question.

Experiment	Manipulation	Question addressed
Realisation-guided intervention	Realised, structural and matched sets at $\alpha = 0.25$	Whether conditional realisation identifies load-bearing computation beyond structural prominence and generic active support
Condition crossover	Target identified under the clean or the shifted condition, evaluated under both	Whether computational responsibility is condition specific rather than merely network specific
Dose response	$\alpha \in \{0.10, 0.25, 0.50\}$ on frozen sets	Whether the effect is graded rather than peculiar to one magnitude
Multiscale follow-up	BCANw before and after intervention at $\alpha = 0.25$	Whether re-realisation changes the multiscale representation
Occurrence targeting and equivariance	Spatial occurrence against shared parameter identity; transformed inputs	Whether causal specificity survives parameter sharing, and whether the realised field moves with the input

E.13 Intervention programme and fidelity

The intervention programme of Section 6 reuses the 324 frozen benchmark networks of Section 5.3 without retraining [1]. Every intervention multiplies the contribution of each selected carrier by $g = 1 - \alpha$ during inference, applies the same gate to both members of the condition $c = (x, x^\circ)$ with $x^\circ = 0$, and recomputes every downstream state. The carrier is an execution edge in an MLP and a spatial occurrence $a = (o, i, \delta_y, \delta_x, y, x)$ in a CNN, whose shared identity is $(o, i, \delta_y, \delta_x)$. Three equal-budget layerwise sets are compared. The realised set contains the top tenth of positive realised support, the structural set an equal budget selected from the structural measure towards the same output, and the matched set the same number of carriers from the remaining positive realised support, matched on source-activation magnitude. Table E41 lists the five experiments and Table E42 their audits.

Fidelity is established before any population is analysed. Operator-level tests compare quantities that can be computed in more than one way: path measures, flow marginals, MLP edge semantics, CNN occurrence semantics, normalisation and the intervention operator itself. A preflight on trained checkpoints then exercises the complete pipeline, including the zero-intervention identity and the immutability of the stored parameters. An undefined realisation is recorded as undefined and never set to zero. Post-intervention measures, marginals and displacements are recomputed from the attenuated execution and are not obtained by adjusting their baseline values.

E.14 Realisation-guided intervention

Forty balanced conditions per network are evaluated in the clean and shifted environments, and the three sets are attenuated with $\alpha = 0.25$, giving 77,760 interventions.

[1] Figure titles keep the experiment and question codes of the results archive; RQ3 and RQ7 there denote the crossover and occurrence-targeting questions, not the research questions of this paper.

Table E42: Design and numerical audits. **(a)** Realisation-guided intervention. **(b)** Condition crossover. **(c)** Dose response; the selected set is frozen across dose. **(d)** Multiscale follow-up.

(a)

Quantity	Value
Frozen trained networks	324
Training horizon	160 epochs
Diverged trainings	0
Controlled interventions	77,760
Defined baseline realisations	77,760
Defined intervened realisations	77,760
Intervention-induced degeneracies	0
Target-audit rows	155,520
Operator self-tests	34/34
Maximum operator-test error	7.08×10^{-9}
Maximum trained-sentinel error	1.91×10^{-6}

(b)

Quantity	Value
Frozen checkpoints reused	324
Paired conditions per checkpoint	10
Interventions	12,960
Applicable interventions	12,960
Defined post-intervention realisations	12,960
Intervention-induced degeneracies	0
Pair-level contrast rows	16,200
Model-level rows	1,620
Primary statistical tests	72
Operator self-tests	34/34
Real-checkpoint crossover preflight	48/48

(c)

Quantity	Value
Frozen trained networks	324
Conditions per network	10
Environments	2
Intervention sets	3
Dose strengths	3
Controlled interventions	58,320
Defined baseline realisations	58,320
Defined post-intervention realisations	58,320
Intervention-induced degeneracies	0
Frozen-set groups checked	19,440
Mask hashes changed across dose	0
Target counts changed across dose	0
Operator self-tests	34/34
Dose-schedule tests	10/10
Trained dose preflight	156/156
Missing Hellinger entries	12 (0.0206%)

(d)

Quantity	Value
Frozen trained networks	324
Conditions averaged per state	10
Environments	2
Geometry states per network and environment	4
Declared BCANw states	2,592
BCANw states computed	2,592
BCANw scales	64
Characteristic-scale bracketed fraction	1.000
Effective-dimension supported fraction	1.000
Strict scaling-credible states	2,187
Strict scaling-credible fraction	0.844
Attenuation strength	$\alpha = 0.25$
Selected contribution retained	0.75

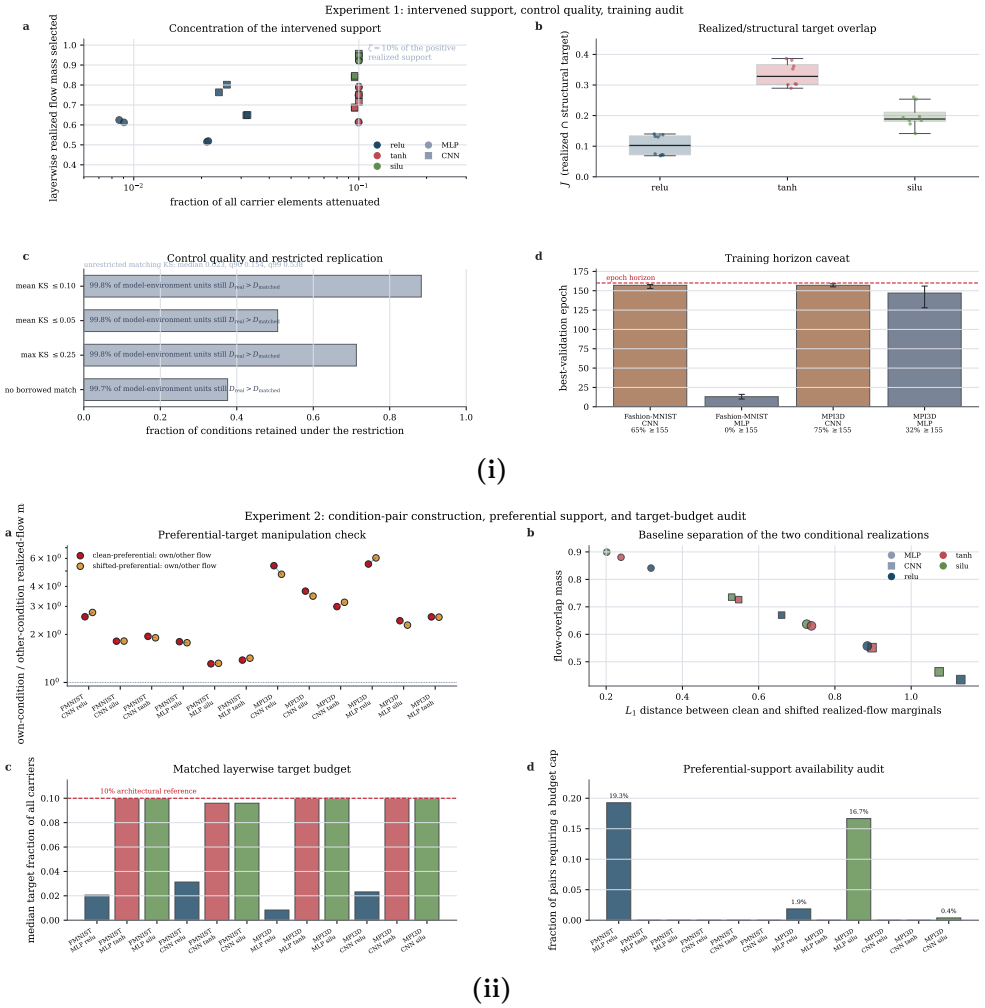


Fig. E31: Targets and controls. **(i)** Realisation-guided intervention: (a) fraction of all carriers attenuated against the layerwise realised-flow mass captured by the realised set, with colour for activation and marker for family; (b) realised-structural target overlap by activation, largest under tanh; (c) restricted replications under progressively stricter activation matching, with the fraction of conditions retained and the fraction of model-environment units in which the realised set still exceeds the matched set; (d) best-validation epoch by dataset and family, where CNN cohorts often reach the training horizon. **(ii)** Condition crossover: (a) own-condition to other-condition realised-flow mass ratio of the clean-preferential and shifted-preferential sets; (b) baseline separation of the clean and shifted flow marginals; (c) median fraction of carriers selected by the matched layerwise budget; (d) fraction of paired items for which the available preferential support caps that budget.

Table E43 reports the primary contrasts, their activation dependence and the within-network coupling of internal and output displacement. The realised minus matched contrast has rank-biserial effect 1 in all four populations (Holm-adjusted $p \approx 7.38 \times 10^{-28}$), and the median shift-minus-clean change of this contrast is -0.206 on Fashion-MNIST and -0.068 on MPI3D. With mean matching distance at most 0.10, 88.3% of conditions remain and 99.8% of model-environment units keep the realised set above the matched set; with borrowing disallowed, 37.6% remain and 99.7% keep the ordering. Across the 24 activation-resolved strata, greater realised-structural overlap accompanies a smaller realised advantage ($\rho_s \approx -0.61$, $p \approx 1.5 \times 10^{-3}$). The realised Hellinger displacement is about 0.54 on Fashion-MNIST and 0.64 to 0.65 on MPI3D, against 0.42 to 0.50 for the structural set and 0.13 to 0.20 for the matched set, and the realised increase of η_1 exceeds the structural increase by factors of 1.67, 1.76, 1.72 and 1.46. The dimensionless manifestation forms bands near 0.55, 0.33 and 0.08 for the three sets. On Fashion-MNIST, the realised set flips 13% of clean MLP and 33% of clean CNN predictions, against 2% and 6% for the matched set. In shifted MPI3D CNNs, the median task-loss change is -0.0588 for the realised set, -0.0279 for the structural set and -0.0121 for the matched set.

E.15 Condition crossover

Each network is evaluated on ten paired items, each a clean input x_{clean} and its 15 dB shifted version x_{shifted}, with conditions $c_A = (x_{\text{clean}}, 0)$ and $c_B = (x_{\text{shifted}}, 0)$. For a carrier u, the signed difference $\delta_u = F_u^A - F_u^B$ ranks carriers by the condition that realises them preferentially, and two equal-budget layerwise sets are formed, $\mathcal{T}_A$ with the largest positive and $\mathcal{T}_B$ with the most negative δ_u. Each set is attenuated with $\alpha = 0.25$ under each condition, giving the cells D_{AA}, D_{BA}, D_{AB} and D_{BB} (set first, evaluation condition second) and 12,960 interventions. The clean-side advantage is $A_{\text{adv}} = D_{AA} - D_{BA}$, the shifted-side advantage $B_{\text{adv}} = D_{BB} - D_{AB}$, and the mean crossover $C_{\text{mean}} = (A_{\text{adv}} + B_{\text{adv}})/2$; each side compares target origin within one evaluation condition. The median own-condition to other-condition flow-mass ratio of the preferential sets is about 1.81 on Fashion-MNIST and 3.3 on MPI3D, and the median L_1 distance between the clean and shifted marginals is about 0.423 and 0.891, with overlap masses 0.788 and 0.554. On Fashion-MNIST the clean-side advantage is supported after correction ($p_{\text{Holm}} = 6.70 \times 10^{-6}$) and the mean crossover is small and positive ($p_{\text{Holm}} = 1.98 \times 10^{-3}$); on MPI3D both sides are positive ($p_{\text{Holm}} = 3.50 \times 10^{-6}$ and 1.23×10^{-3}) and so is their mean ($p_{\text{Holm}} = 1.86 \times 10^{-10}$). The median fraction of pairs with both sides positive is about 9.4% on Fashion-MNIST and 27.4% on MPI3D, reaching 46.3% for CNN ReLU and 45.6% for CNN SiLU. Across MPI3D networks the Hellinger and manifestation mean crossovers are positively associated ($\rho_s \approx 0.46$), whereas the pooled Fashion-MNIST association is negative and the family-separated ones are near zero. In MPI3D CNNs, the mean crossover correlates with the baseline L_1 separation at $\rho_s \approx 0.47$, while in Fashion-MNIST MLPs the association is $\rho_s \approx -0.40$. MPI3D CNN ReLU and SiLU combine positive manifestation crossovers of 0.166 and 0.181 with task-loss crossovers of about -0.065 and -0.067.

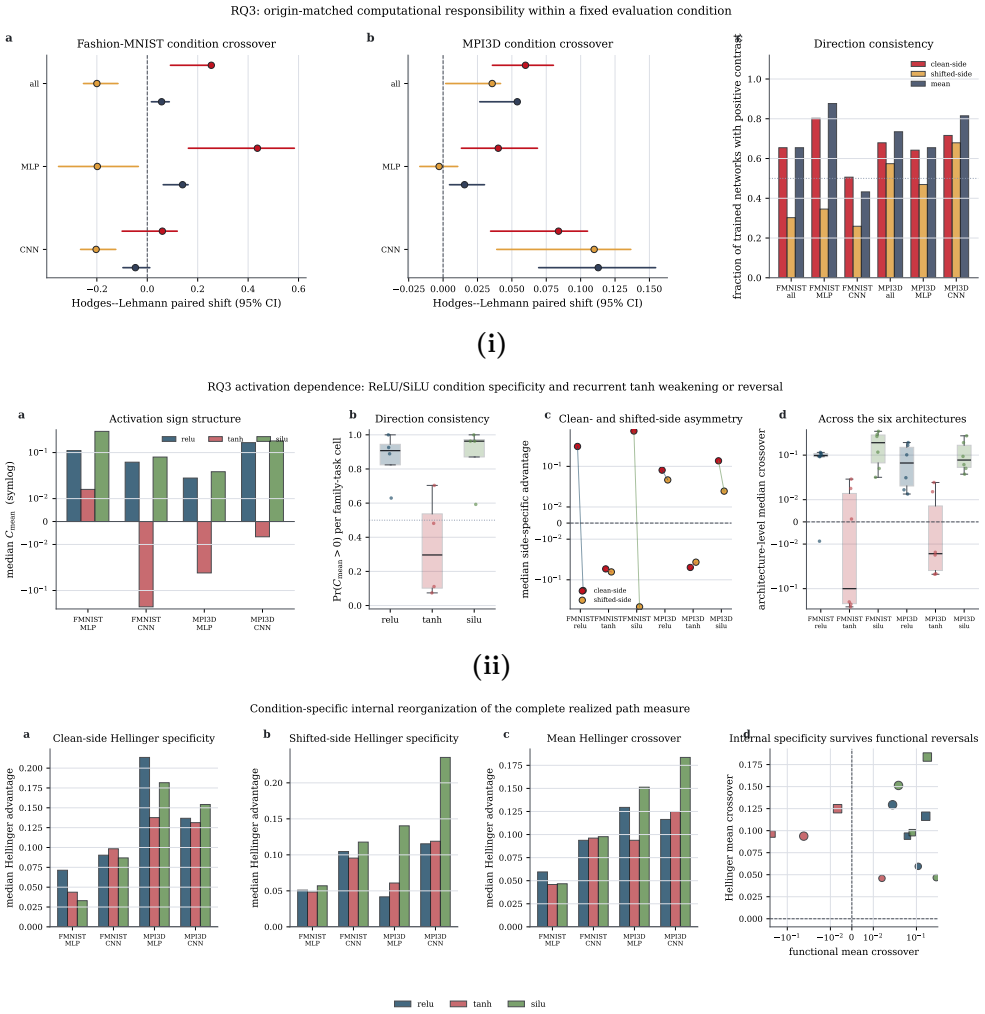


Fig. E32: Condition crossover: primary contrasts, activation and internal specificity. **(i)** (a,b) Hodges-Lehmann paired effects with bootstrap 95% intervals for Fashion-MNIST and MPI3D: clean-side advantage (red), shifted-side advantage (ochre) and their mean (slate); (c) fraction of networks with a positive clean-side, shifted-side or mean crossover at pooled and family-resolved levels. **(ii)** (a) Median mean crossover by dataset, family and activation; (b) direction consistency of the mean crossover across the four family-task cells for each activation; (c) clean-side and shifted-side advantages by activation; (d) architecture-resolved crossover medians across the six architectures. **(iii)** (a) Clean-side, (b) shifted-side and (c) mean Hellinger crossover by dataset, family and activation; (d) family-activation manifestation crossover against Hellinger crossover. The path-measure response is origin specific even where the output coordinate expresses that specificity weakly or in another direction.

Table E43: Realisation-guided intervention. **(a)** Primary contrasts over networks: R − M is realised minus matched and R − S realised minus structural manifestation displacement. **(b)** Median R − S by activation. **(c)** Within-network conditionwise correlation $\rho_s(D_H, D_{\text{man}})$ under the realised set.

(a)

Dataset	Env.	R − M	P(R>M)	R − S	P(R>S)
FMNIST	clean	2.730	1.000	1.295	0.679
FMNIST	shifted	2.548	1.000	1.142	0.679
MPI3D	clean	0.394	1.000	0.123	0.735
MPI3D	shifted	0.267	1.000	0.105	0.648

(b)

Dataset	Env.	Family	ReLU	tanh	SiLU
FMNIST	clean	MLP	2.042	−0.348	1.604
FMNIST	clean	CNN	1.208	−0.324	1.626
FMNIST	shifted	MLP	1.939	−0.388	1.529
FMNIST	shifted	CNN	1.054	−0.314	1.470
MPI3D	clean	MLP	0.354	−0.028	0.157
MPI3D	clean	CNN	0.221	−0.014	0.150
MPI3D	shifted	MLP	0.296	−0.034	0.136
MPI3D	shifted	CNN	0.136	−0.078	0.227

(c)

Dataset	Env.	Family	Median ρ_s	$P(\rho_s>0)$
FMNIST	clean	MLP	−0.150	0.321
FMNIST	clean	CNN	0.207	0.765
FMNIST	shifted	MLP	−0.148	0.284
FMNIST	shifted	CNN	0.108	0.667
MPI3D	clean	MLP	−0.026	0.481
MPI3D	clean	CNN	0.071	0.531
MPI3D	shifted	MLP	0.068	0.593
MPI3D	shifted	CNN	0.057	0.667

E.16 Dose response

Each realised, structural and matched set is selected once from the baseline state and attenuated with $\alpha \in \{0.10, 0.25, 0.50\}$, so that it retains 90%, 75% and 50% of its contribution, under ten conditions in each environment (58,320 interventions). Table E45 reports the curves, their monotonicity, their shape and the dose evolution of the set contrasts. For the realised set, the median middle-minus-low differences are 2.032, 1.887, 0.272 and 0.167, and the high-minus-middle differences 1.666, 1.549, 0.263 and 0.164, positive in at least 98.8% and 90.1% of networks respectively. Because the dose intervals differ, curve shape is compared through the local slopes $s_{\text{low}} = (D(.25) - D(.10))/0.15$ and $s_{\text{high}} = (D(.50) - D(.25))/0.25$; a curve decelerates when $s_{\text{high}} < s_{\text{low}}$. On clean Fashion-MNIST the structural increments are 1.136 and 1.941, so the structural curve catches up at strong attenuation. The realised median $\Delta\eta_1$ on clean Fashion-MNIST moves from about 0.00081 to 0.00398 and back to 0.00319,

Table E44: Condition crossover. **(a)** Primary network-level summaries, with each side-specific advantage computed within a fixed evaluation condition. **(b)** Mean crossover by family and activation, with Holm adjustment. **(c)** Hellinger condition specificity. **(d)** Fashion-MNIST mean crossover on the declared output coordinate and on the full output vector.

(a)

		A_{adv}		B_{adv}		C_{mean}	
Dataset	Level	Median	$P(>0)$	Median	$P(>0)$	Median	$P(>0)$
Fashion-MNIST	All	0.163	0.654	−0.172	0.302	0.051	0.654
Fashion-MNIST	MLP	0.330	0.802	−0.136	0.346	0.111	0.877
Fashion-MNIST	CNN	0.001	0.506	−0.174	0.259	−0.038	0.432
MPI3D	All	0.062	0.679	0.018	0.574	0.037	0.735
MPI3D	MLP	0.045	0.642	−0.002	0.469	0.015	0.654
MPI3D	CNN	0.081	0.716	0.076	0.679	0.094	0.815

(b)

Dataset	Family	Act.	C_{mean}	$P(>0)$
FMNIST	MLP	ReLU	0.111^{***}	0.926
FMNIST	MLP	tanh	0.015	0.704
FMNIST	MLP	SiLU	0.296^{***}	1.000
FMNIST	CNN	ReLU	0.061	0.630
FMNIST	CNN	tanh	−0.231	0.074
FMNIST	CNN	SiLU	0.079	0.593
MPI3D	MLP	ReLU	0.027^{***}	0.889
MPI3D	MLP	tanh	−0.041	0.111
MPI3D	MLP	SiLU	0.038^{***}	0.963
MPI3D	CNN	ReLU	0.166^{***}	1.000
MPI3D	CNN	tanh	−0.007	0.481
MPI3D	CNN	SiLU	0.181^{***}	0.963

(c)

	FMNIST		MPI3D	
Side	Median	$P(>0)$	Median	$P(>0)$
Clean	0.072	0.994	0.151	1.000
Shifted	0.073	0.988	0.105	0.926
Mean	0.072	1.000	0.128	1.000

(d)

Family	Act.	Declared	Full vector	$P(>0)$
MLP	ReLU	0.111	0.275	0.963
MLP	tanh	0.015	−0.102	0.074
MLP	SiLU	0.296	0.731	1.000
CNN	ReLU	0.061	0.707	0.963
CNN	tanh	−0.231	−0.306	0.074
CNN	SiLU	0.079	0.886	1.000

Holm-adjusted significance in (b): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; unmarked, $p \geq 0.05$. Every marked cell has $p \leq 2.61 \times 10^{-6}$, and every unmarked cell has $p \geq 0.166$.

while its Hellinger displacement keeps rising. On clean Fashion-MNIST, the realised minus structural advantage is 1.166, 2.287 and 2.565 under ReLU, 1.364, 1.885 and 1.353 under SiLU, and −0.106, −0.439 and −1.105 under tanh. Optimiser-level strictly increasing fractions range from 88.9% to 100%, and architecture-level ones from 70.4% for shifted MPI3D C2 to 100%. Fashion-MNIST prediction flips under clean realised attenuation rise from 7.5% to 21.8% and 46.8%, against 6.4%, 17.3% and 37.3% for the structural set and 2.0%, 4.9% and 11.4% for the matched set. In shifted MPI3D CNNs the realised task-loss change is −0.089, −0.187 and −0.102 under ReLU, −0.095, −0.182 and −0.033 under SiLU, and −0.012, −0.015 and +0.050 under tanh. The middle dose reaches about 61% to 68% of the high-dose manifestation and 63% to 70% of the high-dose Hellinger displacement.

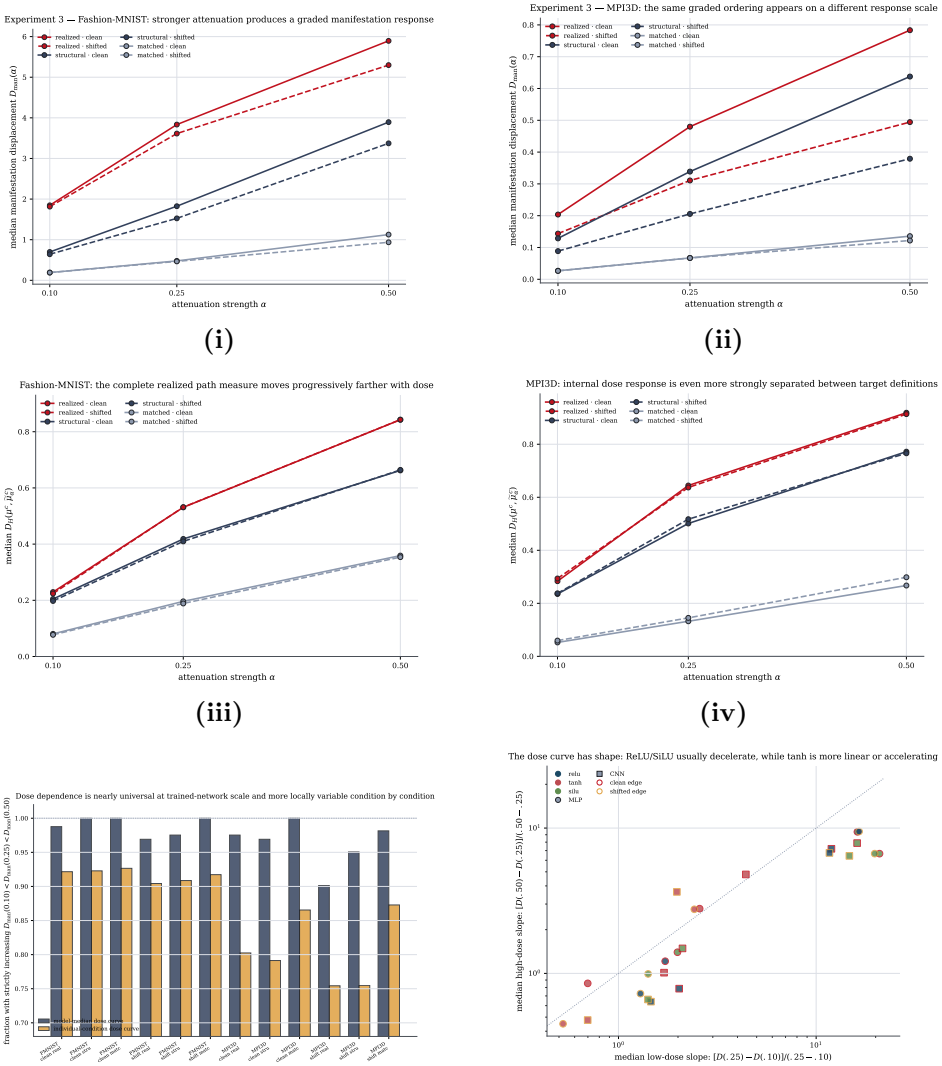


Fig. E33: Dose response: curves, monotonicity and shape. **(i,ii)** Manifestation dose curves on Fashion-MNIST and MPI3D; solid lines are clean and dashed lines shifted evaluation, with realised (red), structural (slate) and matched (grey) sets, and every population-median curve rises with dose. **(iii,iv)** Hellinger displacement of the conditional measure, which increases at every dose for all three sets, with the realised set farthest and the matched set nearest. **(v)** Strict monotonicity at network scale (slate) and condition scale (ochre); individual conditions vary more, particularly on shifted MPI3D. **(vi)** Low-dose against high-dose slope of the realised set in each dataset-environment-family-activation cell; ReLU and SiLU lie mostly below the diagonal, whereas tanh is more often near or above it.

Table E45: Dose response. **(a)** Median manifestation and Hellinger displacement at $\alpha = 0.10$, 0.25 and 0.50, and the fraction of strictly increasing curves at network and condition scale. **(b)** Dose evolution of the realised minus matched (R − M) and realised minus structural (R − S) contrasts. **(c)** Curve shape for the realised set over the eight dataset-environment-family cells of each activation.

(a)

			D_{man}			D_H			Strictly increasing	
Dataset	Env.	Set	0.10	0.25	0.50	0.10	0.25	0.50	Network	Condition
FMNIST	clean	realised	1.847	3.834	5.895	0.229	0.531	0.843	0.988	0.922
		structural	0.698	1.824	3.894	0.205	0.418	0.662	1.000	0.923
		matched	0.190	0.478	1.126	0.081	0.196	0.359	1.000	0.927
	shifted	realised	1.814	3.613	5.298	0.224	0.531	0.843	0.969	0.904
		structural	0.642	1.525	3.372	0.198	0.410	0.664	0.975	0.909
		matched	0.191	0.465	0.936	0.077	0.189	0.354	1.000	0.917
MPI3D	clean	realised	0.204	0.480	0.783	0.284	0.644	0.919	0.975	0.802
		structural	0.129	0.339	0.638	0.235	0.501	0.772	0.969	0.791
		matched	0.027	0.067	0.136	0.053	0.132	0.267	1.000	0.865
	shifted	realised	0.144	0.311	0.494	0.294	0.637	0.914	0.901	0.754
		structural	0.089	0.206	0.379	0.238	0.517	0.766	0.951	0.755
		matched	0.026	0.067	0.122	0.059	0.145	0.298	0.981	0.873

(b)

FMNIST		R − M		R − S	
Env.	α	Median	$P(>0)$	Median	$P(>0)$
clean	0.10	1.555	0.994	0.950	0.735
	0.25	3.159	0.994	1.468	0.673
	0.50	4.383	0.988	1.012	0.648
shifted	0.10	1.513	0.994	0.825	0.728
	0.25	3.009	0.994	1.363	0.660
	0.50	4.294	0.994	0.980	0.654

MPI3D		R − M		R − S	
Env.	α	Median	$P(>0)$	Median	$P(>0)$
clean	0.10	0.173	1.000	0.062	0.765
	0.25	0.400	1.000	0.116	0.735
	0.50	0.605	1.000	0.136	0.698
shifted	0.10	0.113	0.994	0.044	0.710
	0.25	0.230	0.994	0.106	0.679
	0.50	0.339	0.969	0.072	0.599

(c)

Curve shape	ReLU	tanh	SiLU
Decelerating	0.963	0.407	0.889
Accelerating	0.037	0.593	0.111

E.17 Multiscale geometry after intervention

For each network and environment, the normalised flows of ten balanced conditions are averaged carrier by carrier, and the BCANw instrument of Section 3.4 is applied to four states: baseline, and after realised, structural and matched attenuation at $\alpha = 0.25$. The instrument is identical across states (64 scales, enrichment ruler with $p = 1$, undirected shortest paths, directed quotient flow), so state differences arise from the realised flow. Ten descriptors are analysed: $\log \xi_{50}$, d^*, x_{d^*}, the credible d_B^F, Λ_F^*, the median local Δ_i^F, inter-box retention, quotient-flow KS, the transition fraction and the maximum plateau span (Appendix C.10). Table E46 reports the state medians, the realised shifts with their recurrence over the 24 family-activation

strata, the breadth of the effect and the coupling to Hellinger displacement. In the stricter slice with matching distance at most 0.10, the realised geometric change still exceeds the matched change in about 81% to 100% of networks for the five recurrent descriptors. Within tanh, lacunarity decreases and retention increases significantly in all eight family-task-environment cells, d_B^F increases in six and d^* in four. On MPI3D, the shift makes the realised increases of d^*, retention and d_B^F larger and the $\log \xi_{50}$ and x_{d^*} responses less negative, whereas on Fashion-MNIST no environment contrast survives correction.

E.18 Occurrence targeting and equivariance in CNNs

Occurrence targeting. A convolutional preactivation $z_{o,s}^{(\ell)} = \sum_{i,\delta} w_{oi\delta}^{(\ell)} h_{i,s+\delta}^{(\ell-1)}$ uses each stored coefficient at every compatible position, so a shared identity $\vartheta = (\ell, o, i, \delta)$ has many occurrences $a = (\ell, o, i, \delta, s)$ with $\rho(a) = \vartheta$. The intervention multiplies selected primitive terms by 0.75. Occurrence-resolved targeting ranks occurrences by their realised flow F_a^c, while shared-parameter targeting ranks identities by $F_\vartheta^c = \sum_{a:\rho(a)=\vartheta} F_a^c$ and attenuates every reuse of a selected identity. The budget is matched per layer: if the selected identities in layer ℓ expand to B_ℓ occurrences, the occurrence arm also attenuates B_ℓ occurrences. The contrasts $\Delta D_H = D_{H,\mathrm{occ}} - D_{H,\mathrm{param}}$ and $\Delta D = D_{\mathrm{occ}} - D_{\mathrm{param}}$ are summarised per CNN by their median over ten conditions and tested across the 81 CNNs of each dataset-environment cell with a two-sided Wilcoxon signed-rank test, with Holm correction over the eight primary tests. Under shifted evaluation on Fashion-MNIST, the activation-resolved manifestation advantages are $+0.451$ under ReLU, $+0.319$ under SiLU and -0.376 under tanh. In the first convolution, occurrence targeting touches roughly 135 to 158 of the 256 input positions, against 250 to 253 for shared-parameter targeting. Field A, the input-origin marginal $P_{\mathrm{in}}^c(s) = \sum_k F_{(0,k,s)}^c$, shows from which positions realised mass originates, and Field B projects the selected first-layer occurrence flow to source-image coordinates; both are diagnostics and not the intervention.

Equivariance. For a transformation T_g of the input, the recomputed field $P_{T_g x}$ is compared with the transported field $(T_g)_\# P_x(s) = P_x(T_g^{-1} s)$ and with the untransformed P_x, through $H_{\mathrm{eq}} = D_H(P_{T_g x}, (T_g)_\# P_x)$ and $H_{\mathrm{inv}} = D_H(P_{T_g x}, P_x)$, and $A_{\mathrm{eq}} = H_{\mathrm{inv}} - H_{\mathrm{eq}}$. Fashion-MNIST is evaluated under one-pixel and two-pixel translations in four directions, rotations of $\pm 10^\circ$ and $\pm 20^\circ$ and scaling by 0.90 and 1.10, and MPI3D under one-pixel and two-pixel vertical translations, giving 32,400 comparisons. The identity transformation returns numerically zero error. The shifted-minus-clean change of A_{eq} is significant in every tested cell, with median changes of about 0.058, 0.023 and 0.011 for Fashion-MNIST translation, rotation and scale, and -0.127 for MPI3D translation.

Appendix F Extended Discussion

F.1 The findings as reweightings of one ensemble

Appendix B.1 states the formal correspondence between the structural measure and an ensemble whose states are complete paths. Here we use that correspondence to read

Table E46: Multiscale geometry after intervention. **(a)** Population medians before and after the three interventions. **(b)** Median realised shifts from baseline; * marks global Holm significance, and the last three columns give recurrence over the 24 family-activation-dataset-environment strata, with the number of recurrent strata moving up (+) and down (−). **(c)** Breadth of the effect over the ten descriptors. **(d)** Across-network Spearman correlation between Hellinger displacement and the absolute change of a descriptor; * marks post-hoc Holm significance.

(a)

Dataset	Env.	State	d^*	Λ_F^*	Retention	d_B^F	$\log \xi_{50}$
FMNIST	clean	baseline	1.323	425.3	0.353	1.018	−1.336
		realised	1.648	259.8	0.478	1.208	−1.186
		structural	1.543	247.2	0.462	1.163	−1.146
		matched	1.259	447.9	0.331	0.987	−1.347
	shifted	baseline	1.283	418.6	0.362	0.997	−1.328
		realised	1.708	250.9	0.489	1.203	−1.186
		structural	1.540	241.9	0.466	1.136	−1.149
		matched	1.331	442.0	0.346	1.002	−1.340
MPI3D	clean	baseline	0.616	1259.0	0.047	0.394	−0.383
		realised	0.781	832.9	0.133	0.463	−0.595
		structural	0.734	1040.6	0.097	0.472	−0.434
		matched	0.636	1301.2	0.039	0.409	−0.348
	shifted	baseline	0.746	976.1	0.108	0.424	−0.736
		realised	0.913	674.9	0.197	0.567	−0.935
		structural	0.838	760.8	0.156	0.508	−0.812
		matched	0.723	1023.4	0.098	0.432	−0.712

(b)

	FMNIST		MPI3D		Recurrence		
Descriptor	clean	shifted	clean	shifted	Strata	+	−
d^*	0.254^*	0.264^*	0.060^*	0.104^*	16/24	16	0
Λ_F^*	-125.3^*	-124.5^*	-296.4^*	-289.0^*	21/24	0	21
$\log \xi_{50}$	0.148^*	0.165^*	-0.201^*	−0.040	16/24	10	6
Inter-box retention	0.125^*	0.115^*	0.042^*	0.073^*	20/24	20	0
Credible d_B^F	0.219^*	0.221^*	0.070^*	0.089^*	14/23	14	0
Quotient-flow KS	0.019^*	0.014^*	0.016	0.023^*	6/24	4	2
Transition fraction	−0.016	−0.016	-0.016^*	-0.016^*	4/24	0	4
x_{d^*}	-0.497^*	-0.591^*	-1.867^*	-0.797^*	5/24	0	5
Maximum plateau span	0.024	−0.0010	0.0000	0.037	2/24	2	0
Median local Δ_i^F	0.0049	0.0044	−0.017	−0.0085	2/24	2	0

(c)

	FMNIST		MPI3D	
	clean	shifted	clean	shifted
Shifted	7/10	7/10	7/10	7/10
R > M	9/10	10/10	10/10	8/10
R > S	5/10	5/10	4/10	5/10

(d)

	FMNIST		MPI3D	
Descriptor	clean	shifted	clean	shifted
d^*	-0.42^*	-0.35^*	-0.31^*	−0.12
Λ_F^*	0.41^*	0.37^*	0.26^*	0.43^*
$\log \xi_{50}$	-0.41^*	-0.35^*	0.42^*	0.33^*
Max. plateau span	0.63^*	0.53^*	0.44^*	0.49^*
Credible d_B^F	-0.46^*	-0.38^*	0.25	0.18

Table E47: Occurrence targeting and equivariance in CNNs. **(a)** Primary contrasts over 81 CNNs per cell; positive values favour occurrence-resolved targeting. **(b)** Baseline flow-capture advantage of occurrence targeting and post-intervention Field-A redistribution. **(c)** Equivariance advantage A_{eq} by transform family. **(d)** Selected transform-level advantages.

(a)

Dataset	Env.	Outcome	Median	HL	95% CI	$P(>0)$	r_{rb}
FMNIST	clean	ΔD	0.232	0.217	$[-0.006, 0.478]$	0.605	0.319
		ΔD_H	0.037	0.040	$[0.029, 0.050]$	0.827	0.720^{***}
	shifted	ΔD	0.137	0.114	$[-0.128, 0.252]$	0.556	0.182
		ΔD_H	0.037	0.038	$[0.026, 0.048]$	0.864	0.692^{***}
MPI3D	clean	ΔD	−0.019	−0.019	$[-0.038, 0.035]$	0.432	−0.159
		ΔD_H	0.016	−0.000	$[-0.012, 0.029]$	0.593	−0.003
	shifted	ΔD	0.022	0.066	$[-0.013, 0.109]$	0.568	0.357^{*}
		ΔD_H	0.024	0.009	$[0.003, 0.030]$	0.617	0.097

(b)

		Flow advantage		Field-A L_1 advantage	
Dataset	Env.	All layers	First layer	Median	$P(>0)$
FMNIST	clean	1.477	0.312	0.125	0.938
	shifted	1.495	0.325	0.113	0.938
MPI3D	clean	2.101	0.445	0.122	0.728
	shifted	2.026	0.441	0.107	0.815

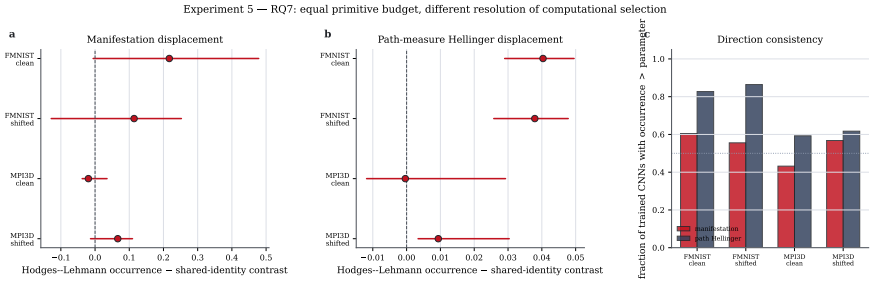
(c)

		Clean				Shifted			
Dataset	Family	A_{eq}	CI	$P(>0)$	r_{rb}	A_{eq}	CI	$P(>0)$	r_{rb}
FMNIST	translation	0.154	$[0.151, 0.158]$	1.000	1.000	0.191	$[0.186, 0.195]$	1.000	1.000
	rotation	0.104	$[0.103, 0.106]$	1.000	1.000	0.133	$[0.132, 0.135]$	1.000	1.000
	scale	0.014	$[0.011, 0.016]$	0.988	0.999	0.032	$[0.029, 0.037]$	1.000	1.000
MPI3D	translation	0.507	$[0.483, 0.533]$	1.000	1.000	0.393	$[0.368, 0.412]$	1.000	1.000

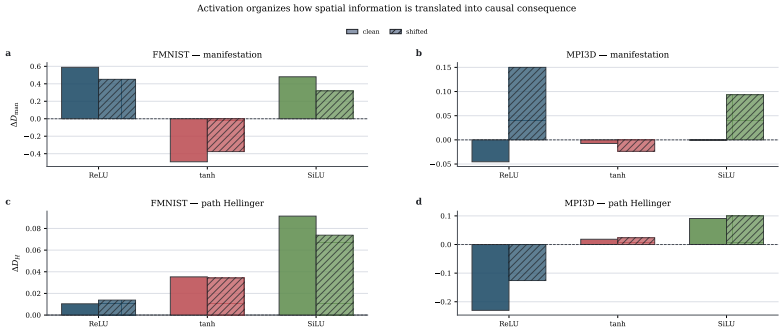
(d)

		Clean		Shifted	
Dataset	Transform	A_{eq}	$P(>0)$	A_{eq}	$P(>0)$
FMNIST	right, 2 px	0.161^{***}	1.000	0.187^{***}	1.000
	rotation 20° anticlockwise	0.147^{***}	1.000	0.177^{***}	1.000
	scale 0.90	0.038^{***}	1.000	0.055^{***}	1.000
	scale 1.10	-0.006^{**}	0.346	0.009^{***}	0.926
MPI3D	down, 2 px	0.497^{***}	1.000	0.364^{***}	1.000

Holm-adjusted significance in (a) and (d): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; unmarked, $p \geq 0.05$. In (a), the marked contrasts have $p = 1.45\times10^{-7}$, 4.36×10^{-7} and 3.12×10^{-2}, and the unmarked ones have $p \geq 6.25\times10^{-2}$. In (d), every marked advantage has $p \leq 2.00\times10^{-12}$ except scale 1.10 on clean Fashion-MNIST, with $p = 2.99\times10^{-3}$.



(i)



(ii)



(iii)

Fig. E34: Occurrence targeting. **(i)** Occurrence-resolved against shared-parameter targeting under the same layerwise primitive budget and 25% attenuation: the Hellinger contrast asks which intervention reorganises the realised execution more, and the manifestation contrast which changes the final response more. **(ii)** Activation-resolved effects, circles for clean and squares for shifted evaluation: Fashion-MNIST manifestation separates into ReLU and SiLU regimes that favour occurrence targeting and a tanh regime that favours shared-parameter targeting. **(iii)** Where the equal budget is placed: occurrence targeting captures more baseline realised flow with a narrower first-layer spatial support, whereas the shared-parameter arm spreads its budget across the reuses of the selected coefficients.

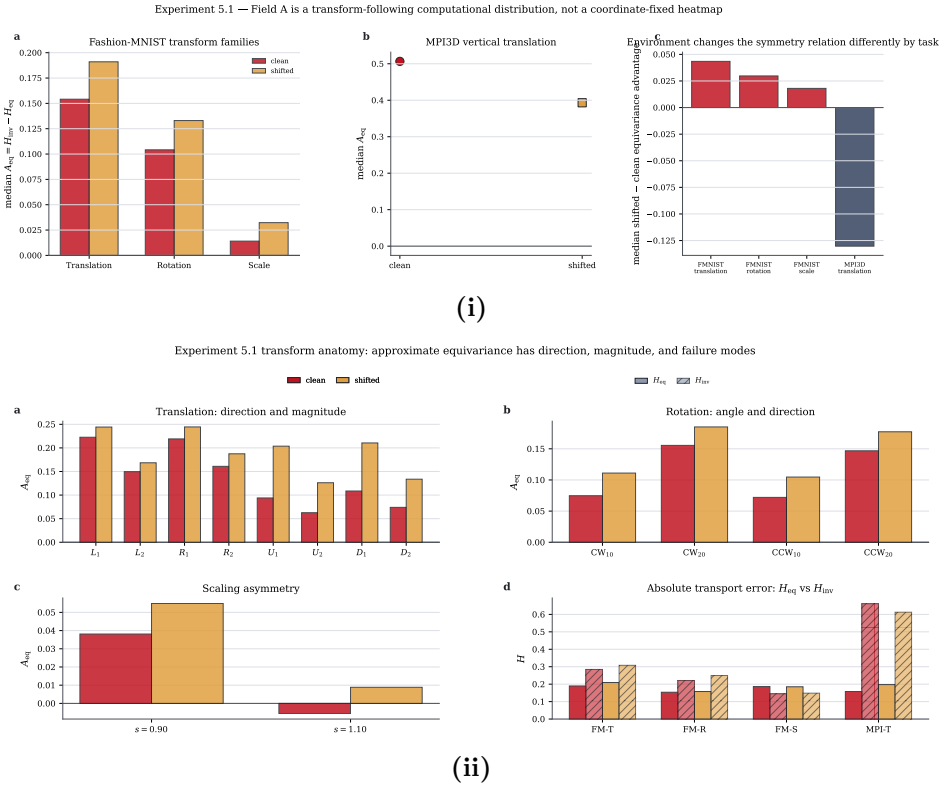


Fig. E35: Equivariance of the realised input-origin field. **(i)** Family-level comparison; a positive advantage means that the recomputed field of the transformed input is closer to the transported baseline than to the baseline left at its original coordinates. **(ii)** Transformation-resolved anatomy, with translation and rotation by direction and magnitude and scale-down and scale-up separately; comparing the equivariant error with the invariant error separates a relative preference for transport from exact geometric fidelity.

the findings of Sections 4 to 6 together. The reading rests on exact relations only. The architecture fixes the states. The multiplicative gain of each path fixes its statistical weight. The partition function normalises those weights.

The architecture as reference.

The uniform measure over the architectural carrier weights every available path equally. The entropy deficit is the relative entropy of the learnt measure with respect to this reference. The occupancy $\eta_1 = e^{-S_{\text{deficit}}}$ expresses the same quantity as the fraction of the carrier that the measure effectively occupies. Concentration is therefore a departure from the reference of the architecture itself. Across environments, the architecture ranks of $\log \eta_1$ are preserved ($\rho_s = 0.993$). Each architecture thus keeps its position relative to its reference while learning displaces it. The Rényi order

q weights the dominant core of the ensemble more strongly as it increases. In the architecture cohort, the median η_2 falls to 0.017 of its initial value, whereas η_1 falls to 0.129. Learning therefore sharpens the dominant core faster than it narrows the ensemble as a whole.

Learning as selection.

Between two training states with common support, the new measure is an exact exponential tilt of the old one by the log-growth field G. The change of deficit equals the integrated covariance between G and the current log-probability (Theorem 7). In the infinitesimal limit, this relation is a replicator equation. A path then gains mass when its growth exceeds the average growth of the ensemble. The identity describes how concentration occurs. It does not determine whether concentration occurs. Its sign is set by the alignment of growth with prominence. This alignment is positive at the start of 99.92% of the 3,780 Adam intervals. Its magnitude is set by the intensity $\mathrm{Var}_p(G)$. On the interval from epoch 80 to 120, this intensity is 3.81 under Adam and 0.014 under SGD. Post-optimal concentration has a direct reading in this language. After the function has reached its best held-out regime, the optimiser continues to select among paths. Networks with markedly different measures attain nearly the same error. This continued selection is therefore not visible in the loss.

Realisation and intervention as further reweightings.

A condition multiplies the structural weight of every path by its realisation factor $R_{c,t}$. The realised ensemble is therefore a reweighting of the structural ensemble on the same states. A factor of zero removes a state. No factor introduces a state outside the structural support (Proposition 1). Execution-only attenuation multiplies the realised contributions of the targeted carriers by the gate g. It thus produces a third ensemble within the same structural support (Proposition 2). The load-bearing result can be restated in this language. Reweighting the realised core moves the ensemble further than reweighting equally active states outside that core.

Universality of direction.

The ensemble has no thermodynamic limit, and the networks studied here are finite. Universality is therefore not sought in a common terminal state or in a common exponent. Within the tested regimes, the recurrent regularity is the direction in which learning reweights the ensemble. Its terminal state and its finite-window exponent vary across populations. Its magnitude depends on the optimiser and is modulated by the environment. Robustness also weakens with distance from the measure. In the environment matrix, the residual share of variance is approximately 0.3% for $\log \eta_1$. It rises to 53% for $\log \xi_{50}$, 60% for d^* and 80% for d_B^F. Claims on the measure therefore carry the strongest language of this work. Claims on finite-window exponents are reported as conditional responses of the geometry.

F.2 The probabilistic status of the path measure

The structural measure normalises squared path gains. Its partition function is therefore the squared ℓ_2 path norm. The squaring is deliberate. It yields a non-negative

weight that is invariant under positive hidden-unit rescaling. It also measures structural magnitude rather than signed cancellation. The probabilistic reading states that a path drawn from the measure crosses an edge e with probability F_e. Every path crosses each cut exactly once. The flows of a layer therefore form a distribution. This reading adds no model assumption. In particular, it does not assert that the network samples paths or that paths execute independently.

The conditional measure is obtained by normalising the squared finite-response amplitudes $Q_t^c(\pi) = |A_t^c(\pi)|^2$ of the complete paths. The amplitudes sum exactly to the output contrast. In general, however,

$$\Big|\sum_{\pi} A_t^c(\pi)\Big|^2 \neq \sum_{\pi} \big|A_t^c(\pi)\big|^2.$$

The measure is therefore not an additive partition of the output change. It measures the magnitude with which each path participates in that change. The signed amplitudes retain constructive and destructive interaction between paths. They are available for every condition, and they are not analysed in this work. A condition whose realised partition function vanishes has no measure. Such a condition is flagged rather than assigned zero mass.

The carrier itself is fixed by the architecture and by the declared symmetry family. A function-preserving transformation may map carriers bijectively and scale their weights by a common factor. The normalised measure is then preserved up to relabelling. A transformation may instead mix several carriers into several others. Scalar path identity is then no longer intrinsic. The measure must consequently be stated on a coarser invariant object.

F.3 The role of controlled synthetic environments

Each synthetic environment is a separately generated learning problem from one controlled family. For MLPs, it is specified by the teacher family, the number of informative coordinates, the feature correlation, the signal-to-noise ratio and the strength of an explicit interaction term. For CNNs, spatial coordinates are added. These coordinates can be varied while the architecture panel, the optimiser and the training protocol stay fixed. Named benchmarks do not permit this control. The 60 discovery environments were obtained by a space-filling initial design followed by Bayesian acquisition. The acquisition rewarded surrogate uncertainty, coverage of the task space, novelty of the structural response and disagreement between architectures. Validation and test losses were excluded from it. The frozen panel of 24 is therefore enriched for environments that challenge a structural law. It is not enriched for environments on which a network performs well. This design allowed the structural findings to be stated as recurrent laws. The direction of concentration survives substantial changes in the statistical problem. Architecture accounts for a normalised structured share of 0.996 in $\log \eta_1$ and in S_{deficit}. Environment, by contrast, accounts for 0.977 of the post-best deterioration of test error. The signal-to-noise ratio further modulates the magnitude of concentration ($\rho_s = +0.827$). The same design supports the realisation findings. Conditioning concentrates the MLP measure in all 24 environments, and it

preserves the architecture ordering of concentration (rank correlation 0.965 for MLPs and 0.909 for CNNs). These separations rest on the crossed design. A small number of fixed datasets could not have provided them.

The benchmarks play a complementary role. They test whether the realisation laws hold outside the synthetic family. They also expose behaviour that the family does not contain. The sensor shift diffuses the realised measure on MPI3D and concentrates it on Fashion-MNIST. On the benchmarks, conditioning raises peak lacunarity and lowers retention. In the controlled CNN panel, it lowers lacunarity and raises retention. The synthetic panels therefore establish what recurs under controlled variation. The benchmarks establish which of those recurrences survive a change of data. The panels are designed rather than randomly sampled. Every claim is accordingly made about the designed population, and its extension to other regimes remains an empirical question.

F.4 Beyond path-decomposable architectures

Two ingredients of this work do not depend on path decomposability. The first is the exact finite response of an operation between two declared executions. It turns a nonlinear computation into a linear system in the contrast variables. The second is the carrier rule. It places the carrier at the lowest relational level that survives the transformations declared to be redescriptions of the architecture. In a Transformer, the primitive execution is not a path system. Query and key coordinates meet multiplicatively. Softmax couples an entire row of scores. A primitive relation may therefore join several realised operands. The realised execution is first described as a typed hypergraph of operation instances. Each operation instance is then replaced by its exact finite-response operator. The conditioned computation thereby becomes a finite, directed and acyclic linear system. Its complete response paths again carry a normalised measure. This carrier is derived from the conditioned response graph. It is distinct from the parameter graph, which records learned identities but not the relations that attention generates from the input.

Symmetry plays a larger role in this setting. Some function-preserving transformations of attention change the basis of internal coordinates. They can mix scalar paths, so that a single response path need not have an invariant identity. The carrier rule then moves the measure to a coarser object on which the symmetry acts without mixing. The formal development of this extension and its empirical tests are the subject of subsequent work. No claim about Transformers is made here.

F.5 Open questions

The findings open several precise lines of investigation. Post-optimal concentration co-occurs with functional deterioration over time. The exact tilt identity names the quantity that a causal test would act on, namely the covariance between growth and current prominence (Theorem 8). An intervention that lowers this covariance while keeping the optimiser fixed would therefore test whether the relation is causal. A layerwise weight-shuffling null would preserve the marginal statistics of the weights and remove their alignment across layers. It would thus measure how much of the concentration is carried by learnt coordination between layers. The optimiser regimes

compared here include their learning rates, and a design that matches the learning rate across algorithms would separate the two contributions.

The activation dependence of responsibility is established with precision. Under tanh, the realised carriers remain load-bearing, since they displace the response more than activation-matched carriers in every network. The structural set, however, is the more consequential target in every dataset, environment and family. The fraction of networks with the realised set above the structural set is 0.148 under tanh, against 0.940 under ReLU and 0.968 under SiLU. The dose response differs in the same way. It decelerates in 96.3% of ReLU and 88.9% of SiLU networks, whereas 59.3% of tanh networks accelerate. The mechanism behind this activation law is the natural next question. The signed amplitudes are available for every condition, and they allow a direct test of whether cancellation between realised paths contributes to it. Finally, the realised field of CNNs favours transport under translation, rotation and shrinking of the input in every network. Under enlargement, it does so in 93% of networks with shifted Fashion-MNIST input and in 35% with clean input. The same populations therefore allow the conditions under which training brings this equivariance closer to exactness to be characterised directly.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ECMLPKDDsubmissionfinal22119.pdf