ILLUSTRATED RESEARCH OVERVIEW OCTOBER 2026

Learning, Structure, and Neural Computation

Research of Liu Ziyin and collaborators

A neural network's task rarely specifies a unique internal solution. Hidden units can be exchanged, factors rescaled, and a computation distributed across layers in many ways. Training selects among these possibilities. In biological systems, local plasticity must also assemble the machinery for learning, while the positions and biophysics of synapses help determine a circuit's computation.

Liu Ziyin's research examines these questions through mathematics, statistical physics, machine learning, and neuroscience. Phase-transition theory explains qualitative changes in learned states; stochastic thermodynamics constrains fluctuations and irreversible dynamics. A recurring method is to isolate a mechanism in a solvable model, identify its collective variables, and derive testable predictions. The aim is to find scientific principles of artificial and biological intelligence.

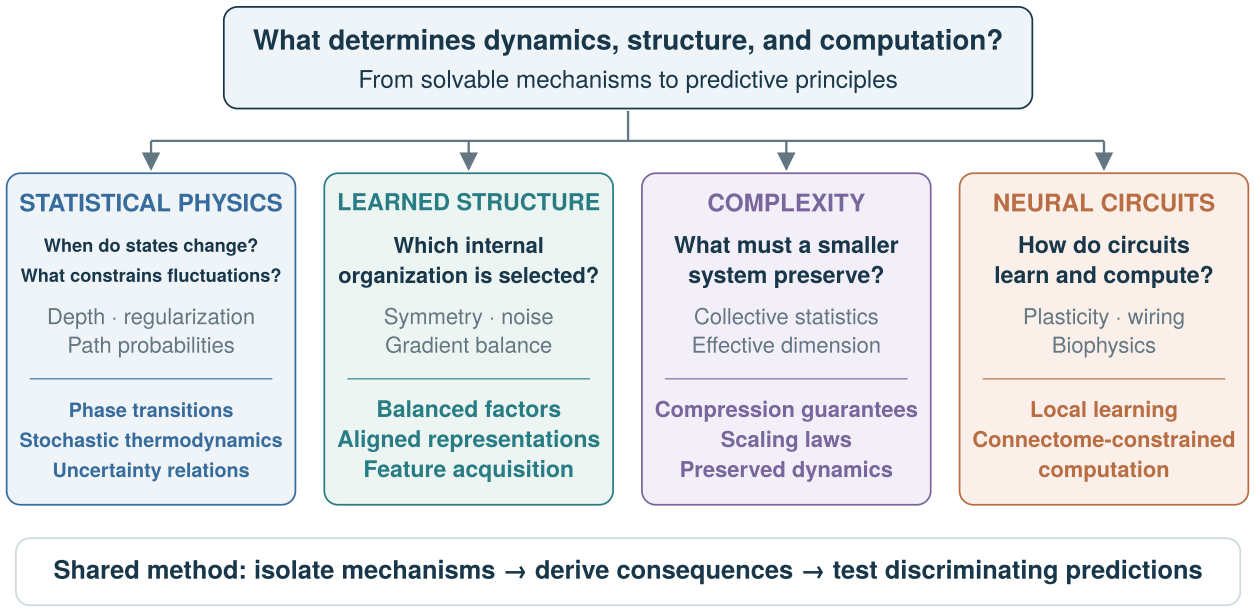


Figure 1. Organization of the research program. Statistical physics, learning structure, effective complexity, and neural circuits form four connected lines. Phase transitions and stochastic thermodynamics supply explicit results as well as tools for studying learning.

A guide to the argument

The linked contents provide entry points into each topic. One route begins with finite-step training, phase transitions, symmetry, and thermodynamics. Another follows representations, topology, scaling, and compression. The biological chapters lead from local learning rules to connectome-based computation. The final sections discuss limitations and propose five AI and five neuroscience questions.

The figures explain concepts and organization: what is related, what is selected, and how a mechanism connects levels of description. They are original schematics based on the cited work. English and Chinese editions share equations, figure numbering, and references; hypotheses, conditional theorems, and biological models retain their distinct evidential status.

Contents

Research overview 1

1 Core concepts and a reading framework 4
1.1 Five distinctions that organize the subject . . . 4
1.2 How to read the mathematical notation . . . 4
1.3 Reading routes and standards of explanation . . . 4

2 Why finite training steps matter 5
2.1 An exact calculation . . . 5
2.2 Structured noise selects solutions and controls escape . . . 5

3 Phase transitions: depth changes how learning appears 6
3.1 A depth-dependent classification . . . 6
3.2 Why a deeper model can keep a collapsed basin . . . 6
3.3 Exact solutions connect the classification to optimization . . . 6

4 Symmetry turns noise into structured selection 7
4.1 A scalar calculation exposes the mechanism . . . 7
4.2 From the example to general results . . . 7

5 Entropic forces predict internal decompositions 8
5.1 The identity task makes selection unusually transparent . . . 8
5.2 A scalar calculation gives the intuition . . . 8

6 Stochastic thermodynamics and the arrow of learning 9
6.1 Probability ratios make time direction measurable . . . 9
6.2 A universal uncertainty relation beyond steady states . . . 9
6.3 Four views of algorithmic irreversibility . . . 9

7 Representation learning as matrix alignment 10
7.1 What is supposed to align? . . . 10
7.2 A mechanism and a structured way for it to fail . . . 10
7.3 Alignment is informative, but task dependent . . . 10

8 Topology, feature learning, and collapse 11
8.1 Small updates can preserve more than the loss . . . 11
8.2 Feature learning has multiple dynamical mechanisms . . . 11
8.3 Different collapses require different explanations . . . 11

9 Neural quadratic forms: from steps to scaling 12
9.1 A common local description of different architectures . . . 12
9.2 How a spectral tail gives a scaling exponent . . . 12

10 Why symmetric systems can be compressed 13
10.1 The exact finite idea . . . 13
10.2 From crowded objects to a polylogarithmic bound . . . 13
10.3 Preserving learning requires preserving the update . . . 13

11 Compressing a fixed deep network 14
11.1 Match input derivatives rather than every hidden coordinate . . . 14
11.2 Two guarantees that should remain distinct . . . 14
11.3 Scaling laws also have invariants . . . 14

12 How a circuit can assemble its learning rule 15
12.1 What self-assembles, and what the theorem requires . . . 15

13 Local plasticity: mechanism, signature, and time **16**
13.1 Heterosynaptic computation is a design principle . . . 16
13.2 Why a Hebbian-looking update need not be a Hebbian rule . . . 16
13.3 A learning rule must also work at the right time . . . 16

14 From a connectome to relative-motion computation **17**
14.1 Different compartments can perform different computations . . . 17
14.2 Why shunting can normalize a signal . . . 17
14.3 What a connectome contributes to explanation . . . 17

15 Current limitations and open boundaries **18**
15.1 From solvable regimes to realistic training . . . 18
15.2 From learned organization to generalization . . . 18
15.3 From compact representations to practical savings . . . 18
15.4 From plausible biological rules to identifiable mechanisms . . . 18
15.5 From connected mechanisms to an integrated theory . . . 18

16 Future research: five questions for AI **19**
Q1. Which phase boundaries predict collapse and recovery of learning in modern networks? . . . 19
Q2. Which irreversible quantities constrain nonstationary training with momentum or Adam? . . . 19
Q3. Which representation-alignment relations causally improve transfer? . . . 19
Q4. Can a scaling exponent be derived without assuming a power-law spectrum? . . . 19
Q5. When does dynamical compression produce a genuine computational saving? . . . 19

17 Future research: five questions for neuroscience **20**
Q6. Which local plasticity mechanisms produce coordinated gradient-like learning? . . . 20
Q7. Which spatial inhibitory mechanism explains relative-motion computation in the fly circuit? . . . 20
Q8. Which fish connectome constraints make sensory integration robust across conditions? . . . 20
Q9. How much mammalian cortical function is predicted by anatomically constrained dynamics? . . . 20
Q10. Which effective circuit variables transfer across flies, fish, and mammals? . . . 20

References **21**

1 Core concepts and a reading framework

These studies connect scientific questions to models, derivations, and testable predictions. Learning theory, physics, and neuroscience offer complementary starting points. Five distinctions help organize the ideas and the relationships among the papers.

1.1 Five distinctions that organize the subject

1. **A function and its parameters.** The prediction $f_\theta(x)$ is what the network computes. The weights θ describe one implementation. Several weight settings can implement the same function, while responding differently to a training update.

2. **An objective and a learning rule.** A loss $L(\theta)$ scores predictions. SGD, gradient flow, and an adaptive optimizer are different procedures for changing weights. Specifying the score does not specify the path or the internal solution.

3. **A representation and an output.** A hidden activity $h(x)$ is an intermediate description of the input. It can retain, discard, or reorganize information that matters for another task, even when the original output is already accurate.

4. **A symmetry and a selected organization.** A symmetry is a transformation that leaves a specified function or loss unchanged. Training can still favor balanced, sparse, or aligned states within those freedoms. Always ask which quantity the symmetry preserves.

5. **A wiring diagram and a computation.** A connectome records connections. A circuit's operation also depends on synaptic location, activity, conductances, and timing. A biological explanation must connect those ingredients to an observable response.

1.2 How to read the mathematical notation

Notation	Interpretation
$\mathbb{E}[\cdot]$	Average over the specified data or training randomness.
∇L, $\nabla^2 L$	Gradient: local direction of increase. Hessian: how that direction changes, or local curvature.
$\mathrm{Cov}(g)$	Fluctuations of the gradient about its mean. A second moment $\mathbb{E}[hh^\mathsf{T}]$ also contains the mean contribution.
Eigenvectors, eigenvalues	Directions or modes, and the strengths associated with them. Their meaning depends on which operator is being studied.
$=$, $\simeq$, scaling	An exact identity, an approximation in a stated regime, or a relation describing growth or decay.

1.3 Reading routes and standards of explanation

For any route, identify the phenomenon, minimal model, assumptions, and a prediction that separates mechanisms. Distinguish an exact theorem, an effective approximation, a computational observation, and a biological hypothesis.

Entry point	Suggested route
Physics	Phase transitions → uncertainty relations → irreversible training.[27, 33, 10]
Learning	Finite steps → noise equilibrium → representations → NQF.[1, 5, 11, 16]
Complexity	Symmetric compression → analytic fixed-network compression.[17, 18]
Neuroscience	Local learning → fly circuitry → fish and mouse resources.[20, 24, 34, 35]

Linear algebra, calculus, probability, dynamical systems, and synaptic physiology provide complementary foundations. The scalar example on page 8 connects function equivalence to noise selection.

2 Why finite training steps matter

SGD updates the weights θ using g_t, the average gradient of a randomly sampled small batch of examples:

$$\theta_{t+1} = \theta_t - \eta g_t, \quad \mathbb{E}[g_t \mid \theta_t] = \nabla L(\theta_t), \quad C(\theta) = \mathrm{Cov}(g_t \mid \theta). \tag{1}$$

Here η is the learning rate and C describes the geometry of gradient noise. Gradient flow replaces these finite jumps with continuous motion. The finite-step and minibatch-noise papers study the discrete update directly and establish when continuous-time approximations are controlled.[1, 2]

2.1 An exact calculation

Take $L(\theta) = \frac{1}{2}\theta^{\mathsf{T}} H \theta$ with $H \succ 0$ and $g_t = H\theta_t + \xi_t$, where the independent additive noise has zero mean and covariance C. Writing $A = I - \eta H$, one obtains $\Sigma_{t+1} = A\Sigma_t A^{\mathsf{T}} + \eta^2 C$. Thus, when $\rho(A) < 1$,

$$H\Sigma + \Sigma H - \eta H \Sigma H = \eta C. \tag{2}$$

Here Σ is the stationary parameter covariance and $\rho(A)$ the largest absolute eigenvalue of A; stability requires it to be below one. The continuous approximation drops the third term. In one dimension,

$$\Sigma_{\text{discrete}} = \frac{\eta c}{h(2 - \eta h)}, \qquad \frac{\Sigma_{\text{discrete}}}{\Sigma_{\text{small step}}} = \frac{2}{2 - \eta h}, \quad 0 < \eta h < 2. \tag{3}$$

The discrepancy grows as the update approaches its stability boundary. Different curvature directions are amplified by different factors.

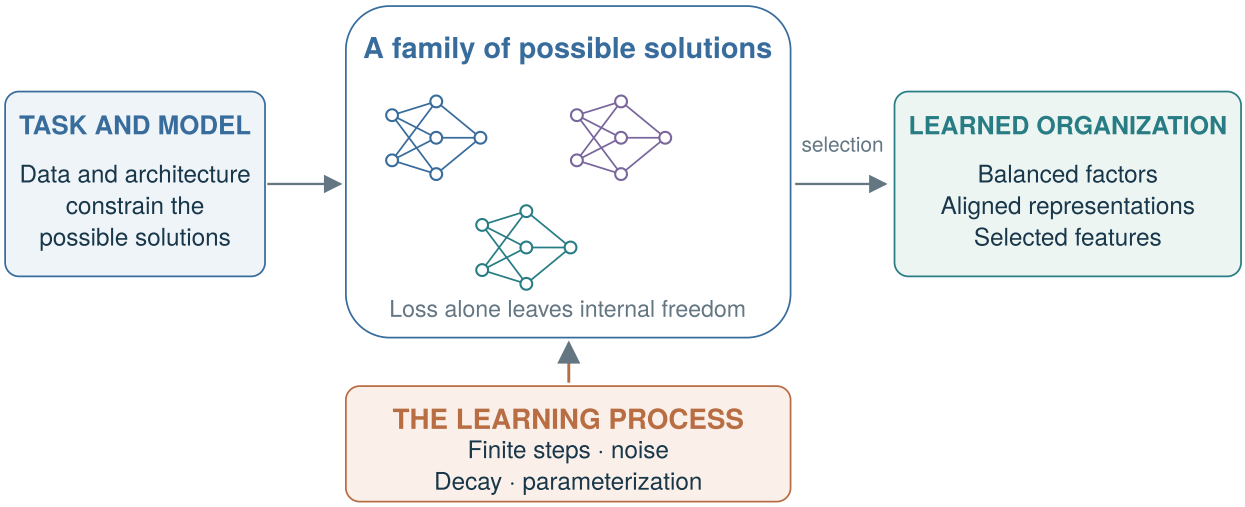


Figure 2. The objective constrains; the learning process selects. A task and architecture allow a family of internal solutions. Finite steps, noise, regularization, and parameterization determine which organizations training favors. The network glyphs are conceptual alternatives.

2.2 Structured noise selects solutions and controls escape

Minibatch noise depends on parameters, data, and residuals. In the 2026 exactly solvable model, isotropic versus anisotropic label noise can select flatter or sharper solutions; the operative quantity is gradient fluctuation.[3]

Work with Takashi Mori and colleagues analyzes multiplicative SGD noise for mean-square loss. A stochastic time change exposes a logarithmic loss barrier, giving power-law escape in its regime.[36] Noise structure thus changes both solution selection and the escape from metastable states.

3 Phase transitions: depth changes how learning appears

The phase-transition work with Masahito Ueda asks when prediction fitting overcomes the pressure to keep a model simple. Its central object is the globally minimized regularized objective,

$$\mathcal{V}(\gamma) = \min_{\theta}\{L_{\text{data}}(\theta) + \gamma R(\theta)\}. \tag{4}$$

An order parameter distinguishes a trivial solution from a learned one. A nonanalytic change in $\mathcal{V}$ or its derivatives defines the transition order. This is a zero-temperature landscape theory: minimization can produce a singularity even at finite network size.[27]

3.1 A depth-dependent classification

In the paper's regularized deep linear model with stochastic neurons and an informative input–target signal, ridge regression has no transition for positive regularization. One hidden layer gives a second-order onset; more than one hidden layer permits a first-order switch between trivial and learning phases. A zeroth-order discontinuity arises at zero regularization in the singular infinite-depth limit. At finite depth, the globally optimized value remains continuous; a training-path jump can instead reflect local trapping.[27]

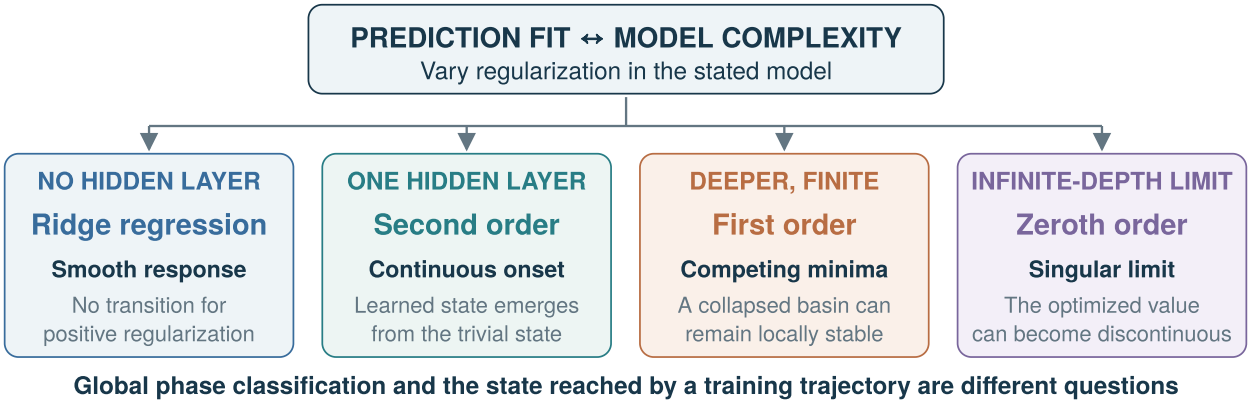


Figure 3. Depth changes the organization of the transition.[27] The four branches classify the stated model as regularization varies. Continuous onset, competing minima, and a singular limiting value are distinct mechanisms. The boxes organize results; no plotted transition curves or measured thresholds are implied.

3.2 Why a deeper model can keep a collapsed basin

A balanced scalar chain makes the mechanism visible. For m weight layers, target $a > 0$, equal factors $w \geq 0$, and quadratic decay,

$$\begin{aligned}\mathcal{J}_m(w;\gamma) &= \tfrac{1}{2}(w^m - a)^2 + \tfrac{\gamma m}{2}w^2 \\ &= \tfrac{a^2}{2} + \tfrac{\gamma m}{2}w^2 - aw^m + \tfrac{1}{2}w^{2m}.\end{aligned} \tag{5}$$

For $m = 2$, $w_*^2 = \max(a - \gamma, 0)$: the learned factors turn on continuously. For $m > 2$ and $\gamma > 0$, the positive quadratic term stabilizes the origin even when a nonzero solution has lower loss. Learning then requires leaving a locally stable collapsed state. This calculation illustrates the landscape mechanism rather than a claim that every abrupt learning curve is a phase transition.

3.3 Exact solutions connect the classification to optimization

With Botao Li and Xiangming Meng, the exact-solution work characterizes deep-linear minima with decay and stochastic neurons, exposing the special role of the origin.[32] The phase paper's linear-origin theorem extends this local perspective to nonlinear networks satisfying its activation and origin assumptions.[27] A complementary PRE result derives stationary distributions in diagonal linear stochastic gradient flow, with noise-induced transitions, broken ergodicity, and fluctuation inversion.[25] These are distinct from the sequential feature-activation dynamics studied by NQF.[16]

4 Symmetry turns noise into structured selection

The mirror-symmetry theorem starts from $\ell_0(\theta) = \ell_0((I - 2OO^\mathsf{T})\theta)$, with $O^\mathsf{T}O = I$. The columns of O specify orthonormal directions whose signs are reflected. The unchanged, or fixed, subspace is

$$O^\mathsf{T}\theta = 0.$$

Gradients preserve this subspace. With a lower Hessian bound, sufficiently strong weight decay places every minimum in it. The framework gives sparsity from rescaling symmetry, low rank from rotation symmetry, and homogeneous ensembling from permutation symmetry.[4]

4.1 A scalar calculation exposes the mechanism

Let $\ell_t = \frac{1}{2}(uv - y_t)^2$ and $e_t = uv - y_t$. A simultaneous SGD step is $u' = u - \eta e_t v$, $v' = v - \eta e_t u$. Expanding the squared norms gives the exact identity

$$D' = \left(1 - \eta^2 e_t^2\right) D, \qquad D = u^2 - v^2. \tag{6}$$

The cross terms cancel. Population gradient flow conserves D; finite steps change it. Where the conditional second moment is finite and $0 < \eta^2 \mathbb{E}[e_t^2 \mid u, v] < 1$, the conditional mean imbalance contracts. Thus equal product functions can acquire balanced internal factors.

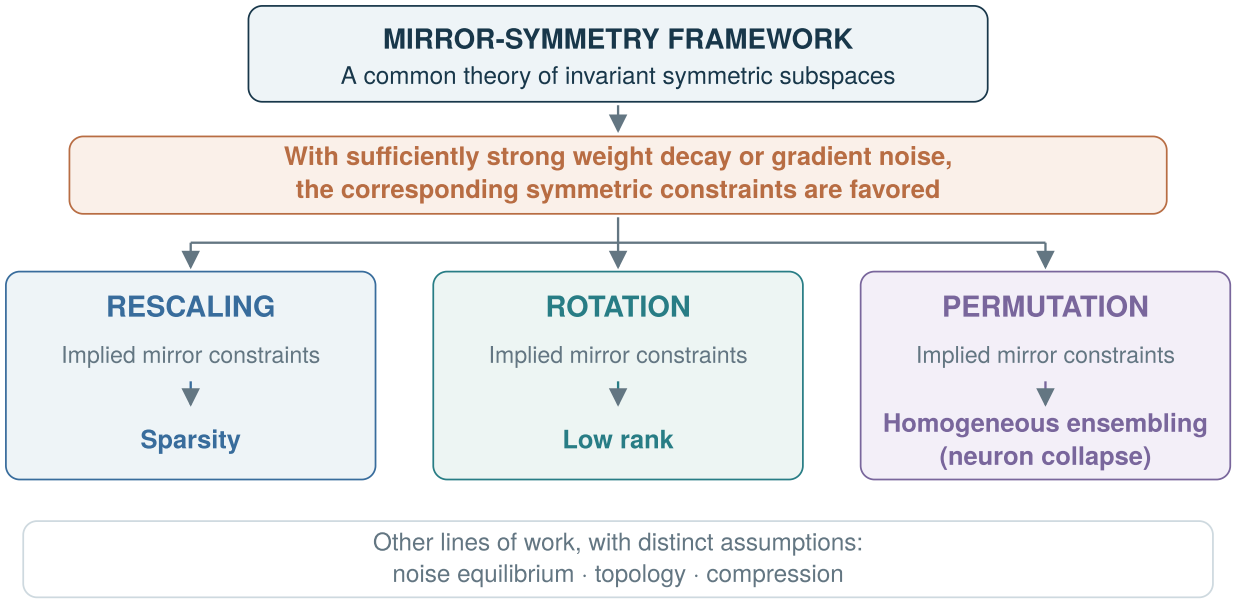


Figure 4. Mirror symmetry organizes the constraint results.[4, 5, 15, 17] Rescaling, rotation, and permutation imply mirror constraints; the common noise-or-decay conditions apply to all three corollaries. Noise equilibrium, topology, and compression are separate uses of symmetry with their own assumptions.

4.2 From the example to general results

For exponential symmetries, Noether flow drives motion along degenerate directions toward a unique noise equilibrium in its setting, independent of the initial position along the orbit. There, noise contributions balance and align. The PRE balance paper obtains stationary distributions for stochastic gradient flow in diagonal linear networks, including phase transitions, broken ergodicity, and fluctuation inversion.[5, 25]

The constraint mechanism connects symmetry to loss of plasticity and collapse.[4] It also motivates exploiting redundancy with *spred* and breaking harmful symmetry-induced traps with *syre*.[8, 9] Symmetry becomes a tool for explaining and controlling learned organization.

5 Entropic forces predict internal decompositions

Neural thermodynamics studies effective forces generated by finite steps and stochastic gradients.[6] For a fixed minibatch, backward-error analysis gives a leading modified-loss correction. Averaging that correction defines the first-order entropic loss

$$F_\eta = L + S, \qquad S = \frac{\eta}{4}\mathbb{E}_{\mathcal{B}}\|\nabla \ell_{\mathcal{B}}\|^2. \tag{7}$$

The correction measures sensitivity to stochastic gradients. For scalar-rate updates, the theory breaks noncompact rescaling symmetries while preserving orthogonal symmetries. Its gradient-balance relations resemble equipartition and connect representational alignment to flattening and sharpening. The description is controlled in the small-step regime.

5.1 The identity task makes selection unusually transparent

Work with Andy Arditi, Weian Xie, and David Bau considers $f(x) = (I + W_2)(I + W_1)x$, $y = x + \varepsilon$, isotropic inputs, and positive label-noise eigenvalues normalized by $\sum_i \lambda_i = d$. Every factorization $\widetilde{W}_2\widetilde{W}_1 = I$, with $\widetilde{W}_\ell = I + W_\ell$, has the same population loss. The entropic term distinguishes them.[7]

For a singular-value mode a_i in the paper's aligned basis, the on-manifold objective is

$$\phi_i(a_i) = \frac{\eta d}{B}(a_i^2 + \lambda_i a_i^{-2}) + \gamma\left[(a_i - 1)^2 + (a_i^{-1} - 1)^2\right]. \tag{8}$$

Without decay, differentiating yields $a_i^4 = \lambda_i$. With decay, the selected value lies between $\lambda_i^{1/4}$ and 1 under the stated conditions.

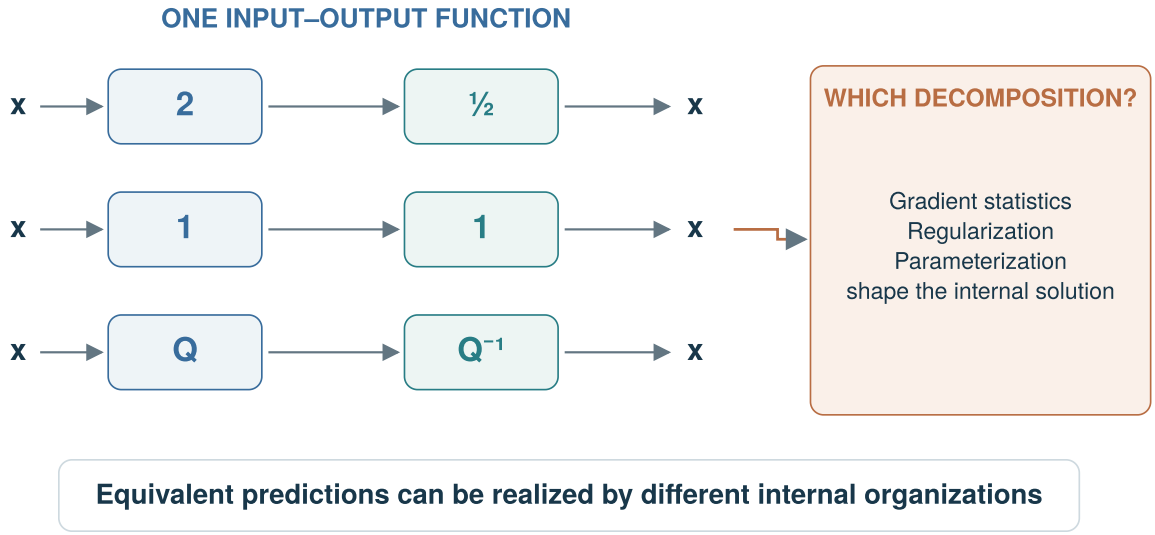


Figure 5. One function, many decompositions.[7] The three paths illustrate equivalent input–output maps, with Q invertible. Entropic selection asks how gradient statistics, regularization, and parameterization distinguish their internal factors.

Reparameterizing residual weights as $W_\ell = A_\ell B_\ell$ favors zero residuals without decay; deeper directly parameterized chains concentrate predicted noise anisotropy at their boundary layers.[7]

5.2 A scalar calculation gives the intuition

For $f = uv$, label $1 + \varepsilon$, and squared loss divided by two, suppose independent label noise has variance σ^2. On the optimal manifold $uv = 1$, the minibatch gradient has second moment $\sigma^2(u^2 + v^2)/B$. Hence $S = \eta\sigma^2(u^2 + v^2)/(4B)$ favors balanced factors, because $u^2 + v^2 \geq 2$ there. Identical functions can have different sensitivity to training noise. Page 9 develops the complementary trajectory-based thermodynamic description.

6 Stochastic thermodynamics and the arrow of learning

Stochastic thermodynamics asks how fluctuations constrain a process away from equilibrium. It is a distinct part of this research, encompassing universal uncertainty relations and the irreversibility of training algorithms. The basic object is a trajectory, rather than just the loss at its endpoint.

6.1 Probability ratios make time direction measurable

Let $P(\Gamma)$ describe forward paths and $Q(\Gamma)$ the corresponding reversed-path law under a specified reverse protocol. With compatible support, the dimensionless path entropy production is

$$s(\Gamma) = \log \frac{P(\Gamma)}{Q(\Gamma)}, \qquad \mathbb{E}_P[s] = D_{\mathrm{KL}}(P\|Q) \geq 0. \tag{9}$$

The mean compares the probabilities of entire histories. Entropy production quantifies time asymmetry; it differs from the entropy measuring the spread of a state distribution.[33]

6.2 A universal uncertainty relation beyond steady states

The work with Masahito Ueda derives a TUR for arbitrary observables in general Markovian dynamics, including time-dependent protocols and absolute irreversibility. Entropy production and nonstationarity both matter for a tight constraint. In the simpler case $Q \ll P$ with finite moments, a change-of-measure bound reads

$$(\mathbb{E}_Q[A] - \mathbb{E}_P[A])^2 \leq \mathrm{Var}_P(A)\big(\mathbb{E}_P[e^{-2s}] - 1\big). \tag{10}$$

Here A is a path observable. The full result includes support corrections when absolute irreversibility is present. Removing an antisymmetry restriction on observables also links thermodynamic uncertainty bounds to a fundamental inequality in theoretical finance.[33]

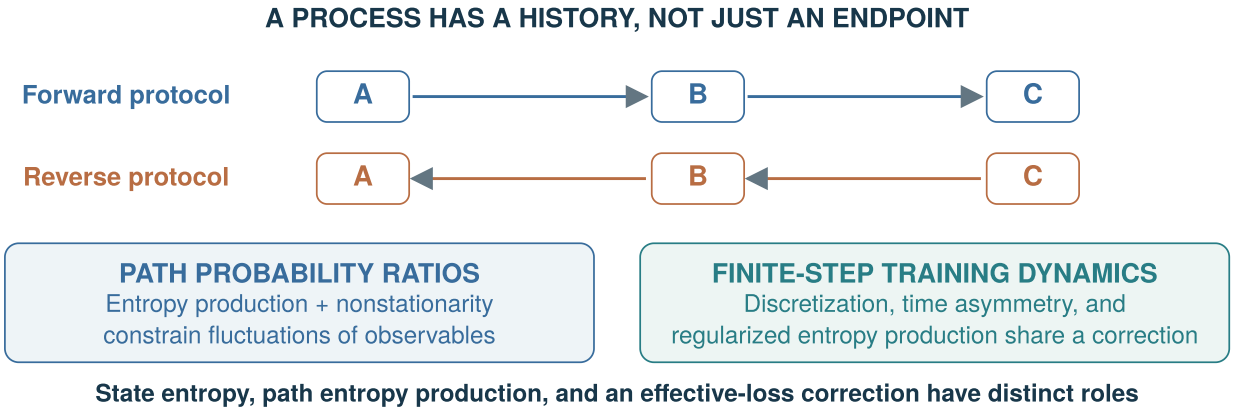


Figure 6. From histories to constraints on dynamics. Comparing forward and reversed path probabilities yields entropy production and uncertainty bounds. For learning algorithms, finite-step analysis supplies a complementary route to effective forces. These are connected levels of description, with assumptions specified separately.

6.3 Four views of algorithmic irreversibility

The 2026 preprint with Yuanjie Ren, Adam Levine, and Isaac Chuang connects numerical discretization error (DE), time renormalization (TR), microscopic time asymmetry (TA), and regularized stochastic entropy production (ST). For $\theta_{t+1} = \theta_t - \eta U(\theta_t)$, smooth update fields with a symmetric Jacobian give

$$\phi_X(\theta) = \tfrac{\eta}{4}\|U(\theta)\|^2 + O(\eta^2), \quad X \in \{\mathrm{DE}, \mathrm{TR}, \mathrm{TA}, \mathrm{ST}\}. \tag{11}$$

The shared leading term connects the arrow of a numerical update to a force that selects learning trajectories. The stochastic construction uses the paper's regularization and normalization.[10] This is a dynamical counterpart to entropic selection: an effective-loss correction, state entropy, and path entropy production play different mathematical roles.[6]

7 Representation learning as matrix alignment

7.1 What is supposed to align?

For a hidden linear map $h_b = Wh_a$, define the representation and neuron-gradient second moments

$$H_c = \mathbb{E}[h_c h_c^\top], \quad G_c = \mathbb{E}[g_c g_c^\top], \quad Z_a = W^\top W, \quad Z_b = WW^\top.$$

The Canonical Representation Hypothesis (CRH) proposes three relations on each side of the map:

$$H_c \propto G_c, \qquad H_c \propto Z_c, \qquad G_c \propto Z_c, \quad c \in \{a, b\}. \tag{12}$$

These are six matrix relations, not merely overlap between a few principal directions. They connect which features are present, which affect the loss, and which the weights transmit.[11]

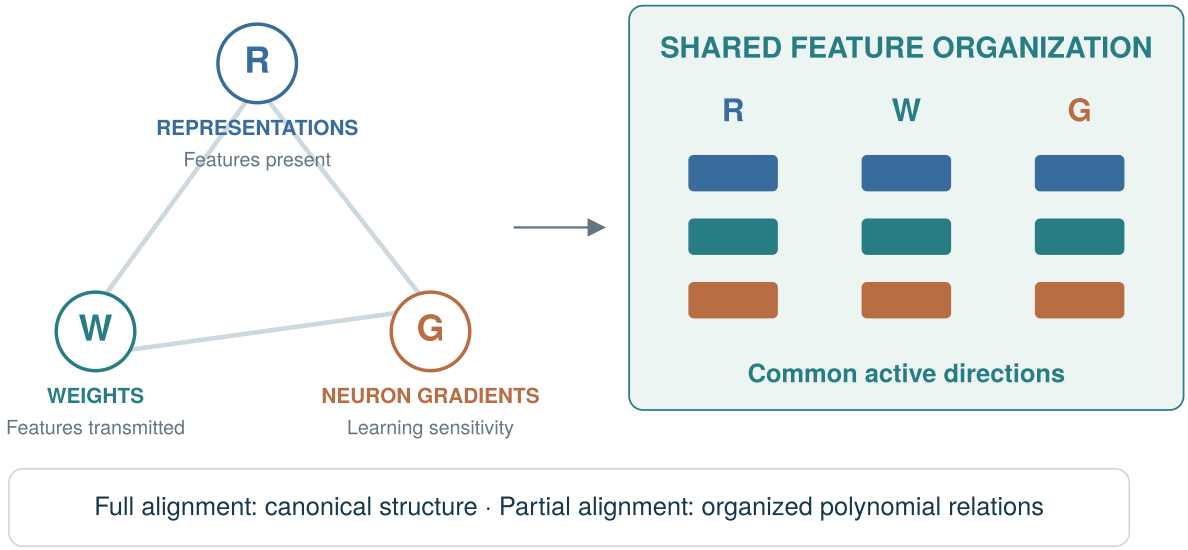


Figure 7. Three views of learned feature organization.[11] Representations, weights, and neuron gradients describe present features, transmitted features, and learning sensitivity. CRH connects their matrix structure on both sides of a linear map. The colored channels illustrate shared directions, not measured spectra.

7.2 A mechanism and a structured way for it to fail

The paper derives alignment from noise–regularization balance under explicit assumptions, including mean-field norm conditions and appropriate stationarity approximations. It also studies finite-time violations and layers where full CRH fails. Polynomial Alignment Hypothesis (PAH) relations involve powers of projected positive-semidefinite matrices, schematically

$$\widetilde{H}_c^{\alpha} \propto \widetilde{G}_c^{\beta} \propto \widetilde{Z}_c^{\delta},$$

with subspaces and exponents specified in the paper. Thus partial failure can have organized spectral structure rather than arbitrary misalignment. The general hypotheses are broader than any one proof.

7.3 Alignment is informative, but task dependent

In its specified embedded deep linear model, the Platonic note proves a selection result: SGD reaches rotation-equivalent representations even though most loss minima are not Platonic.[31] Work with Megan Tjandrasuwita, Chanakya Ekbote, and Paul Pu Liang finds that the benefit of multimodal alignment depends on redundant versus unique task information.[13] Alignment is a task-dependent object of explanation.

In work with Qiyao Liang, Daoyuan Qian, and Ila Fiete, forcing disentangled factors into the output representation improves compositional extrapolation in controlled generation tasks. Later layers can otherwise re-entangle the factors.[14] The result connects generalization to the full organization of the learned computation.

8 Topology, feature learning, and collapse

8.1 Small updates can preserve more than the loss

Topology concerns properties such as connectivity and holes that survive continuous invertible deformation. If an update field u is K-Lipschitz, meaning $\|u(x) - u(y)\| \leq K\|x - y\|$, and $T(x) = x - \eta u(x)$, the triangle inequalities give

$$(1 - \eta K)\|x - y\| \leq \|T(x) - T(y)\| \leq (1 + \eta K)\|x - y\|. \tag{13}$$

For $\eta K < 1$, distinct points cannot merge and the inverse on the image is Lipschitz. For suitable permutation-equivariant rules, including SGD and Adam, the topology paper identifies a small-step regime preserving neuron-distribution topology; larger steps permit simplification that can reduce expressivity.[15]

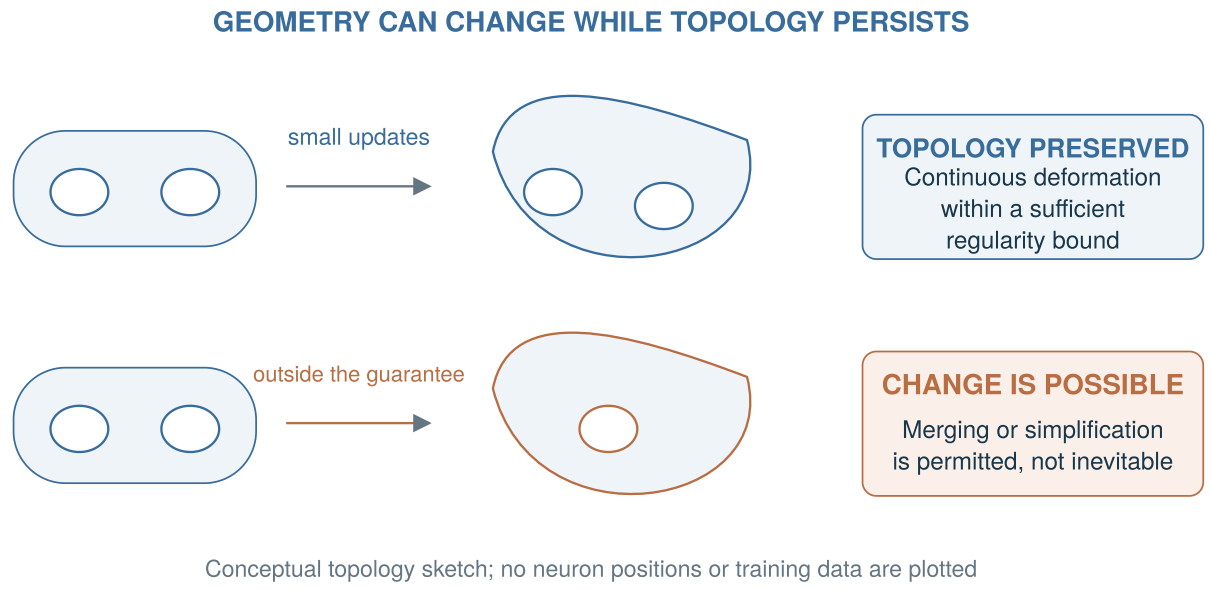


Figure 8. Geometry versus topology.[15] A shape can deform while its connectivity and holes persist. Outside a sufficient regularity guarantee, merging or simplification becomes possible. These shapes explain the distinction; they are not reconstructed neuron distributions.

Crossing the sufficient bound permits topology change; it does not force every larger step to change topology. For example, the scalar map $T(x) = (1 - \eta K)x$ collapses at $\eta K = 1$ but becomes invertible again above it. Distinguishing a sufficient condition from a necessary phase boundary is essential when applying a bound to a real training run.

8.2 Feature learning has multiple dynamical mechanisms

An exactly solved finite-width linear model with Yizhou Xu distinguishes learning by alignment, disalignment, and rescaling, connecting kernel-like and feature-learning regimes.[12] The stepwise SSL study derives sequential eigenmode acquisition in a linearized model and observes related behavior in nonlinear joint-embedding networks.[29] These analyses motivate tracking collective variables instead of only total loss.

8.3 Different collapses require different explanations

The SSL-landscape theory studies complete and dimensional collapse, including effects of normalization, bias, and data structure.[28] The linear latent-variable analysis traces posterior collapse to competition between likelihood fitting and prior-induced regularization of the latent mean.[26] Phase-transition work shows how depth and complexity penalties can change the order of transitions in solvable networks.[27] These are related examples of structure emerging from an objective and dynamics, with distinct assumptions and mechanisms.

9 Neural quadratic forms: from steps to scaling

9.1 A common local description of different architectures

The NQF work with Yizhou Xu derives a quadratic normal form for smooth sums of interchangeable units near zero weights, with a zero-gradient condition at the origin:

$$f_W(x) \simeq \mathrm{Tr}\big(MA(x)\big), \qquad M = WW^{\mathsf{T}}. \tag{14}$$

Perceptrons, attention, mixtures of experts, and convolutions enter through different structure matrices $A(x)$. The matrix M collects outer products of the unit weights, retaining evolving features rather than freezing them in a fixed kernel.[16]

When the data matrices share an eigenbasis, the dynamics become generalized Lotka–Volterra mode equations. In the orthogonal-feature case,

$$\dot{z}_k = z_k(r_k - C_{kk}z_k), \qquad z_k(t) = \frac{r_k z_k(0)}{C_{kk}z_k(0) + (r_k - C_{kk}z_k(0))e^{-r_k t}}. \tag{15}$$

Here z_k is a mode's strength, r_k its initial growth rate, and C_{kk} sets self-limitation. A positive growing mode saturates at r_k/C_{kk}. If $z_k(0) = \epsilon$ is small, its activation time is approximately $r_k^{-1}\log[r_k/(C_{kk}\epsilon)]$. Modes can therefore turn on in well-separated steps despite smooth underlying dynamics.

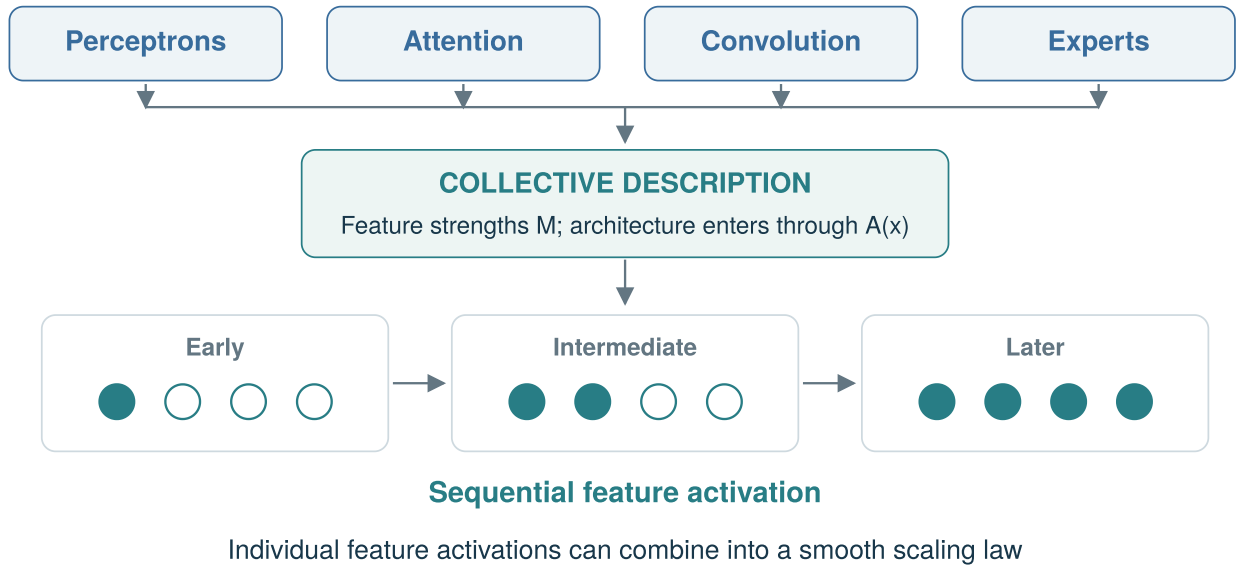


Figure 9. From microscopic architecture to collective learning.[16] Different modules share a local quadratic description under the stated assumptions. Feature modes activate in sequence; their combined contributions yield smooth scaling with the assumed spectral tail. Filled circles schematically indicate mode activation.

9.2 How a spectral tail gives a scaling exponent

Suppose mode growth rates and target strengths obey $r_k \propto k^{-\alpha_2}$ and $V_k \propto k^{-\alpha_1}$, with $\alpha_1 > 1$, $\alpha_2 > 0$. In rescaled time $\tau = t/\log(1/\epsilon)$, the activated index grows as $k_*(\tau) \sim \tau^{1/\alpha_2}$. The unlearned tail then scales as

$$\mathcal{E}(\tau) \sim \sum_{k>k_*} V_k \sim k_*^{1-\alpha_1} \sim \tau^{-(\alpha_1-1)/\alpha_2}. \tag{16}$$

This exposes the mechanism behind the paper's conditional exponent prediction. The spectral power-law tail is an input assumption, not something inferred from permutation symmetry alone. Small initialization, smoothness, and the chosen solvable spectral setting delimit the derivation; higher-order terms can generate additional plateaus. Earlier stepwise SSL results provide a complementary route into this subject.[29]

10 Why symmetric systems can be compressed

The universal compression work with Hong-Yi Wang and Di Luo uses weighted objects and low-order moments to preserve functions of unordered neurons and data points.[17]

10.1 The exact finite idea

For objects $w_i \in \mathbb{R}^m$, define $p_j = d^{-1} \sum_i w_i^{\otimes j}$. Matching moments through degree k preserves every symmetric polynomial determined by them. A positive weighted representation requires at most

$$N_{m,k} = \binom{m+k}{k} \tag{17}$$

support points in the relevant moment-matching construction. For analytic functions, a controlled Taylor remainder turns finite moment matching into approximation.

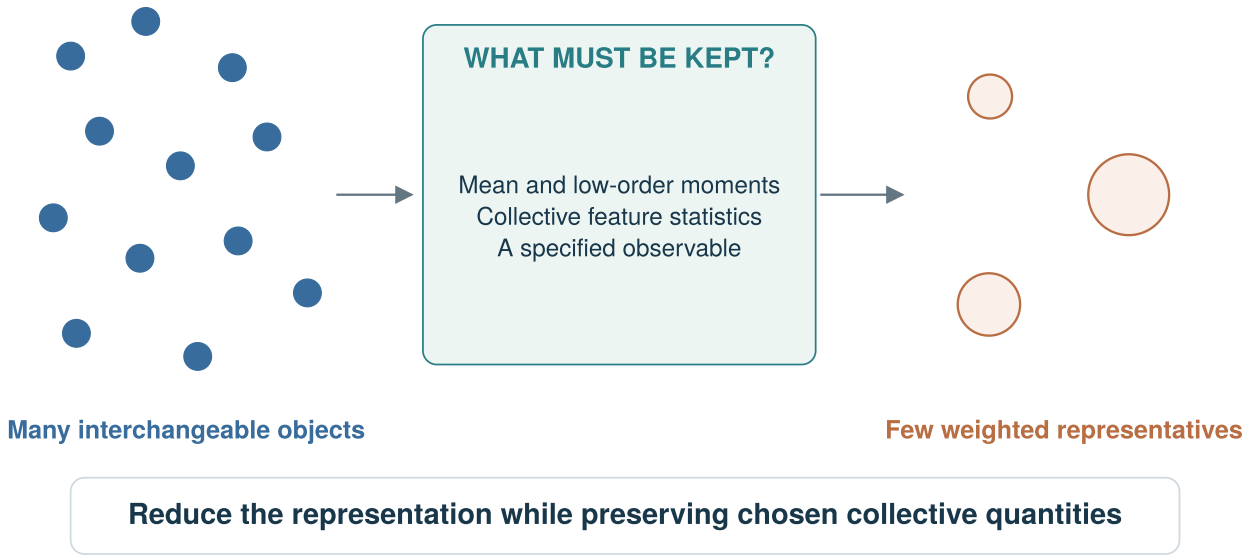


Figure 10. Compression as preservation of collective organization. Many interchangeable objects can be replaced by fewer weighted representatives when the relevant statistics are retained. Moments connect the microscopic collection to the observable being preserved; marker sizes are symbolic.

10.2 From crowded objects to a polylogarithmic bound

Clustering and moment matching give asymptotically vanishing error with a small weighted support. A representative cap is $d' = O([\log(d/\varepsilon)]^m)$, assuming bounded objects and an analytic deep-set representation $h(\sum_i g(w_i))$ with h, g independent of d. Higher moment order improves error but increases cost.

10.3 Preserving learning requires preserving the update

The dynamical lottery-ticket result applies compression to a symmetric trained observable $f \circ T$, with equivariant training and regularity assumptions on that composition. The compressed dynamics use weighted objects: a gradient for retained object w_j is rescaled by c_j^{-1}. Thus the theorem is more specific than pruning a network and rerunning an unchanged optimizer.

Source experiments compare wide, compressed, and random-subnetwork models under matched minibatch choices.[17] Preserving the performance of d objects with d' effective objects can improve loss versus retained size; constructing the compression has its own cost.

11 Compressing a fixed deep network

11.1 Match input derivatives rather than every hidden coordinate

The width-independent theorem with Hong-Yi Wang and Mingze Wang considers a fixed analytic-activation deep MLP and a bounded input domain. A layer contributes $y(x) = \sum_i v_i \sigma(w_i^\top f(x))$. The derivative vector through order k has only

$$D_k = \binom{d_{\text{in}} + k}{d_{\text{in}}} \tag{18}$$

coordinates per output. A rank-reduction procedure retains original neurons and recomputes outgoing weights to match these derivatives. Analyticity controls the remainder by a geometric factor $(R/\rho_0)^{k+1}$, where $R < \rho_0$ relates the real input domain to the complex analyticity radius.[18]

Compressing layers backward controls the propagation of errors, yielding a same-depth approximation with width cap $O((\log(1/\varepsilon))^{d_{\text{in}}})$ for the fixed teacher. Teacher-specific analytic and norm constants remain in the bound; width independence does not eliminate the cost of high effective input dimension.

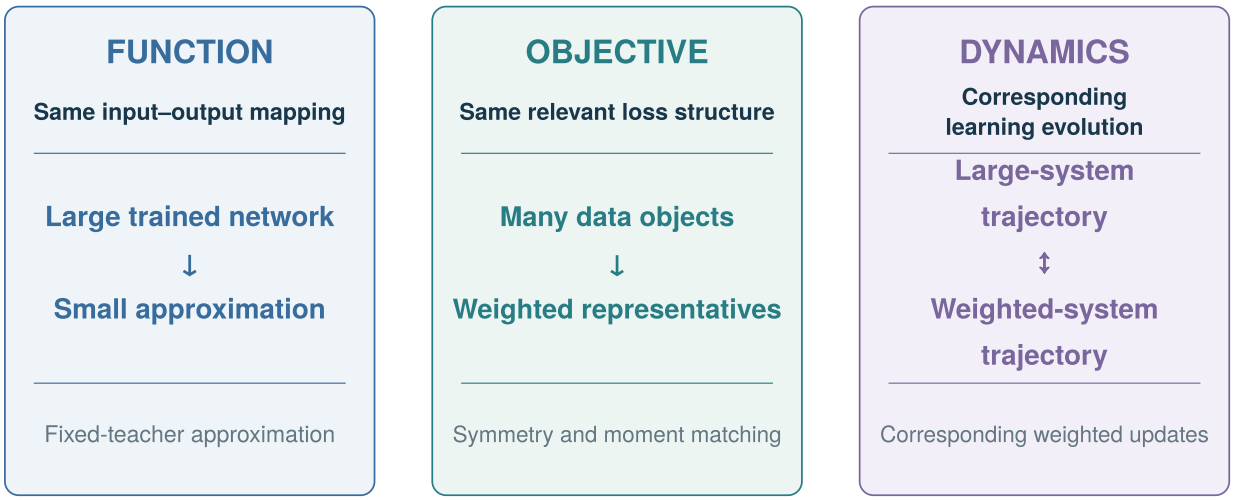


Figure 11. Three levels of preservation.[17, 18] Function approximation, objective compression, and corresponding learning dynamics ask different questions. A compact function representation does not automatically preserve an unchanged optimizer's trajectory. The columns organize the guarantees discussed on these two pages.

11.2 Two guarantees that should remain distinct

This result approximates the input–output function of a fixed deep teacher. The previous page's dynamical result concerns a regular symmetric trained observable and corresponding compressed updates. The fixed-teacher theorem explicitly does not establish that dynamical guarantee. A constructive proof can also be computationally expensive even when its final representation is compact.

11.3 Scaling laws also have invariants

Work with Xing Han, Suchi Saria, and Paul Pu Liang asks whether a scaling law fitted in one domain transfers to another. It examines information-preserving transformations and changes in information resolution, with tests across language, vision, and speech.[19] Alongside NQF and compression, this distinguishes how learning curves arise, which transformations preserve scaling, and how many retained objects achieve a given performance.

12 How a circuit can assemble its learning rule

The PNAS work with Qianli Liao, Yulu Gan, and colleagues asks how a system develops the circuitry that changes its forward weights. Its Self-Assembling Motif (SAM) circuit couples ascending and descending streams through plastic interconnections, allowing different pathway widths.[20]

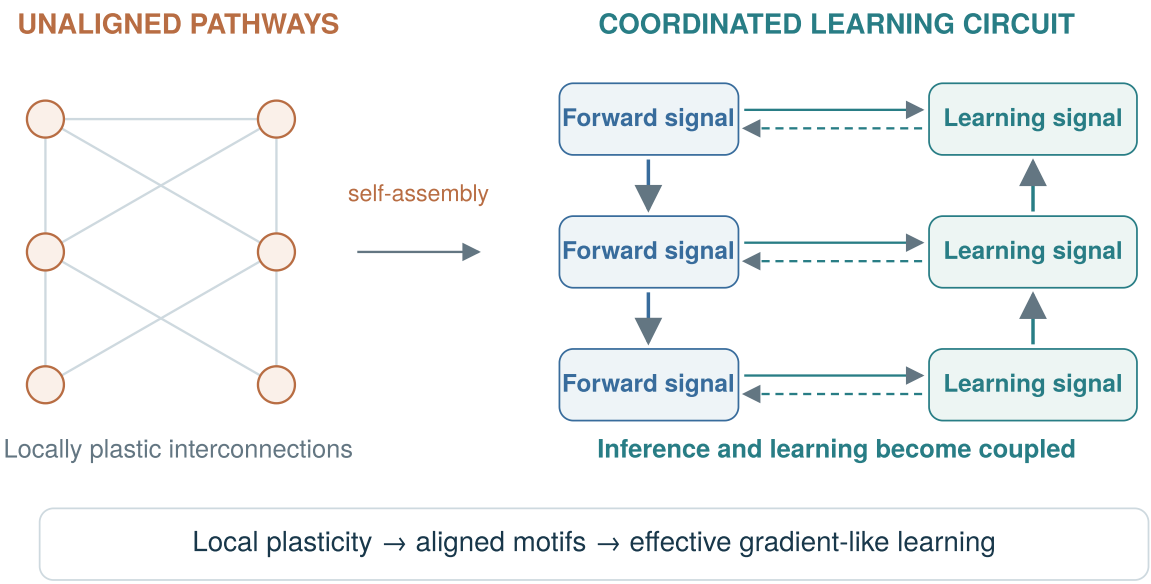


Figure 12. Learning machinery as an emergent organization.[20] Starting with forward and feedback pathways, local plasticity coordinates their connections into a motif supporting gradient-like learning under alignment conditions. The figure shows levels of organization rather than observed cortical wiring.

12.1 What self-assembles, and what the theorem requires

The model has four plastic synapse classes: within the two streams and in both interstream directions. Local rules combine presynaptic activity, a signal delivered by other synapses, and decay. Matched interstream rules drive $E_v = V - \bar{V}^\mathsf{T}$ toward zero. A distinct pathway error E_p compares the forward Jacobian route with the feedback route. Small pathway errors, rank, and boundary conditions give

$$\Delta W^\ell = -\eta_W H_D^\ell \nabla_{W^\ell} F - \gamma W^\ell + R^\ell, \qquad H_D^\ell \succeq 0. \tag{19}$$

The residual R^ℓ is controlled by alignment errors. In the aligned limit, the circuit implements a learned matrix-preconditioned gradient update. Interstream alignment is proved under matched conditions; improvement of pathway alignment in successful runs is separately observed, rather than proved for every initialization.

Selected source results, accuracy (%)	SGD	SAM circuit
MNIST	98.44	98.85
CIFAR-10	50.36	51.91

Selected entries from the paper's Table 1, with five runs per experiment. These concern its model configurations and additional downstream learning machinery, rather than general image-classification records.

13 Local plasticity: mechanism, signature, and time

13.1 Heterosynaptic computation is a design principle

The heterosynaptic-gradient-machine preprint studies how local interactions can implement a broad family of gradient-based meta-learning rules.[21] A synapse receives a forward drive and a modulatory signal associated with other synapses. Their interaction can make the synapse sensitive to an effective learning gradient without requiring each synapse to execute an explicitly programmed global derivative calculation. This is a theoretical construction and a proposed biological principle.

13.2 Why a Hebbian-looking update need not be a Hebbian rule

The ICML study separates a learning signal U from a contractive term: $\Delta W = U - \gamma W$. Near mean stationarity, $\mathbb{E}[U] \approx \gamma W$. For $h_b = W h_a$, a Hebbian matrix is $\mathcal{H} = \mathbb{E}[h_b h_a^\mathsf{T}] = W H_a$, with $H_a \succeq 0$. Therefore,

$$\langle \mathbb{E}[U], \mathcal{H} \rangle_F \approx \gamma \operatorname{Tr}(W^\mathsf{T} W H_a) \geq 0. \tag{20}$$

This simple identity explains how regularization can produce a Hebbian signature in the learning-signal component of quite different update mechanisms. Noise can instead drive anti-Hebbian alignment. The total stationary update and its learning-signal component are different observables.[22]

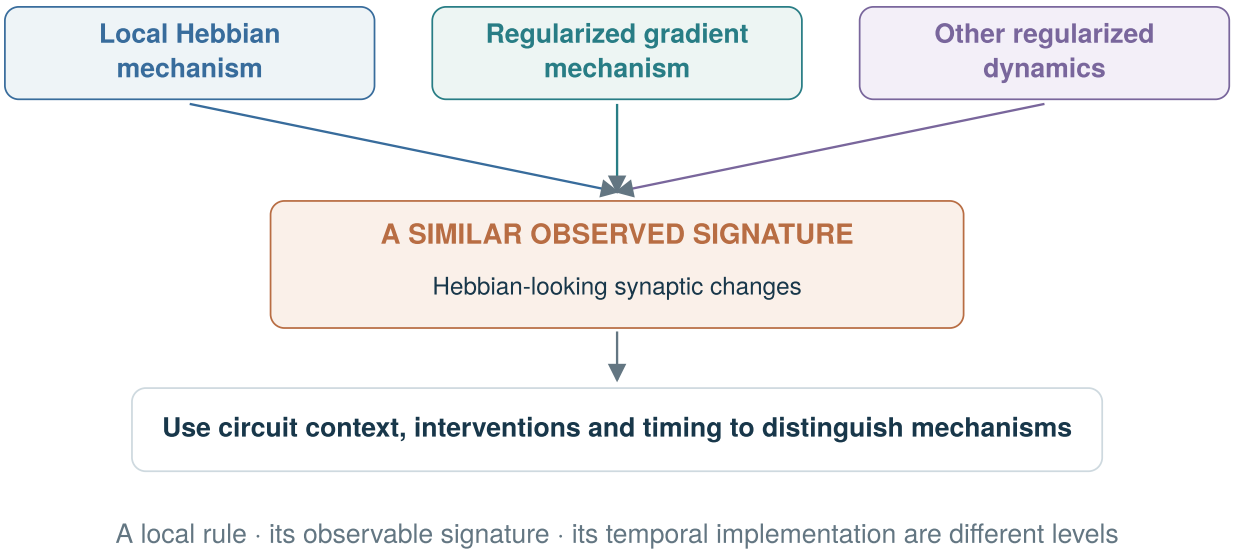


Figure 13. An observed signature can have several mechanisms.[22] Local Hebbian computation and distinct regularized dynamics can produce similar plasticity patterns. Interventions, circuit context, and timing supply additional information for identifying the underlying mechanism.

13.3 A learning rule must also work at the right time

Continuous-time feedback alignment studies coupled propagation, plasticity, and decay, with an effective hierarchy $\tau_{\text{prop}} \ll \tau_{\text{plas}} \ll \tau_{\text{dec}}$. The informative part of a weight update depends on overlap:

$$\Delta W_{\text{informative}} \propto \int_0^T \big(V\varepsilon(t)\big) z(t)^\mathsf{T} k(t)\, dt. \tag{21}$$

Here V is a feedback matrix, $\varepsilon(t)$ an error signal, $z(t)$ presynaptic activity, and $k(t)$ a plasticity kernel. Delays change their temporal match: the model tolerates moderate mismatch but fails when overlap is lost.[23] This connects local implementation, observed plasticity, and timing.

14 From a connectome to relative-motion computation

The 2026 figure–ground preprint combines Drosophila connectomics with detailed biophysical modeling. It investigates a candidate implementation of the Reichardt–Poggio–Hausen architecture: a wide-field pool modulates local motion detectors, helping distinguish object motion from background motion. The proposed circuit involves T4a and the ventral centrifugal horizontal neuron, vCH.[24]

14.1 Different compartments can perform different computations

The reported model assigns a distinction between T4a dendritic motion responses and relative-motion-sensitive axon-terminal signaling. vCH pools broad motion and provides presynaptic shunting inhibition to T4a terminals. Motion-opponent and binocular pathways extend the account. A broader organizational interpretation is that synaptic location helps determine how cells compute together.

14.2 Why shunting can normalize a signal

A generic passive compartment illustrates the logic. Here C_m is membrane capacitance, V the voltage, and g and E are conductances and reversal potentials for leak (L), excitation (E), and inhibition (I):

$$C_m \dot{V} = g_L(E_L - V) + g_E(E_E - V) + g_I(E_I - V). \tag{22}$$

At steady state, $V = (g_L E_L + g_E E_E + g_I E_I)/(g_L + g_E + g_I)$. If $E_I = E_L$, then

$$V - E_L = \frac{g_E(E_E - E_L)}{g_L + g_E + g_I}. \tag{23}$$

An inhibitory conductance near the resting reversal can thus reduce the response to excitation by changing the denominator. It need not be represented as simple subtraction of a fixed voltage.

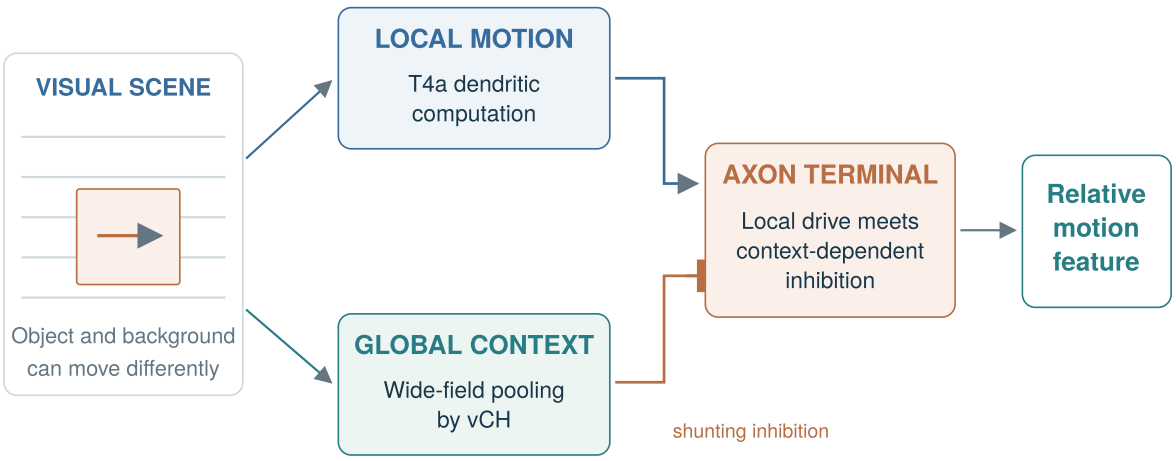


Figure 14. Division of labor across a spatial circuit. Local motion signals and wide-field context meet at the T4a axon terminal in the proposed model. Compartment-specific inhibition changes the computation delivered downstream. This is an organizational cartoon.

14.3 What a connectome contributes to explanation

Anatomy constrains possible interactions; biophysics determines their dynamical consequences. The model's agreement with existing observations supports the proposed mechanism. Perturbations that separate its predictions from alternatives would strengthen the explanation.

Source scope: this section's article-specific statements follow the authors' public bioRxiv abstract. The passive-compartment derivation is an explanatory example, not a reproduction of the paper's quantitative model.

15 Current limitations and open boundaries

The results in this program are strongest when a phenomenon can be isolated, the relevant variables identified, and its dynamics derived. The remaining challenge is to predict which mechanism dominates in a new network or biological circuit. Several boundaries separate an explanatory mechanism from a broadly reliable quantitative theory.

15.1 From solvable regimes to realistic training

Exact conclusions depend on quadratic losses or controlled noise, mean-field conditions, small-weight expansions, or special spectra. Modern training combines normalization, feature evolution, and optimizer state. An effective description's breakdown must become predictable.[1, 11, 16] Global phase boundaries concern optimized values; a finite training path can remain metastable. Likewise, a scalar irreversibility potential requires update-field assumptions, and a nonstationary process requires more than a stationary noise model.[27, 10]

15.2 From learned organization to generalization

Balance, alignment, low rank, and topology characterize internal organization. Their relationship to test performance also depends on which information the task requires. Cross-modal alignment can emphasize shared information while losing useful modality-specific information; disentangled inputs can still be processed through a computation that fails to compose.[13, 14] A complete account needs to predict when a structural statistic improves transfer, robustness, or extrapolation, and when the same statistic is incidental. Interventions that separate these alternatives are more informative than an additional correlation alone.

15.3 From compact representations to practical savings

A small representation may exist while finding it remains expensive. Moment matching and input-derivative constructions involve dimension, approximation order, teacher-dependent constants, and propagation of error through depth. In the fixed-teacher result, width can enter indirectly through those constants even without an additional explicit width factor.[17, 18] Practical compression must therefore account for construction time, memory, calibration, and any subsequent training. Function approximation and preservation of learning dynamics remain distinct targets, with different costs and assumptions.

15.4 From plausible biological rules to identifiable mechanisms

SAM proves interstream alignment under matched conditions, while the pathway alignment needed for a gradient correspondence is a separate condition. Continuous-time models expose another requirement: the correct activity and feedback signals must overlap at the synapse.[20, 23] Establishing a biological implementation requires identifying those signals, their access to plastic synapses, and the circumstances in which the relevant alignment emerges. Hebbian-looking changes alone cannot identify the underlying rule.[22, 24] Likewise, a connectome narrows the wiring possibilities but does not uniquely fix conductances, state, or circuit dynamics; distinguishing candidate mechanisms requires responses or perturbations that separate their predictions.

15.5 From connected mechanisms to an integrated theory

The quantitative bridges between parameter symmetries, feature operators, stochastic path laws, and circuit motifs remain incomplete. Connecting them requires an explicit mapping of variables and observable consequences. Extending connectome models to fish and mammals also requires accounting for reconstruction uncertainty, unmatched physiology, and inputs outside the imaged volume.[34, 35] A useful unification must predict the effect of changing architecture, task, or circuit organization, and specify its domain of validity.

16 Future research: five questions for AI

The ten future questions are divided equally between AI and neuroscience. Each pairs a motivation with an approach that could distinguish mechanisms. These are proposed directions, grounded in the preceding results and in public data resources.

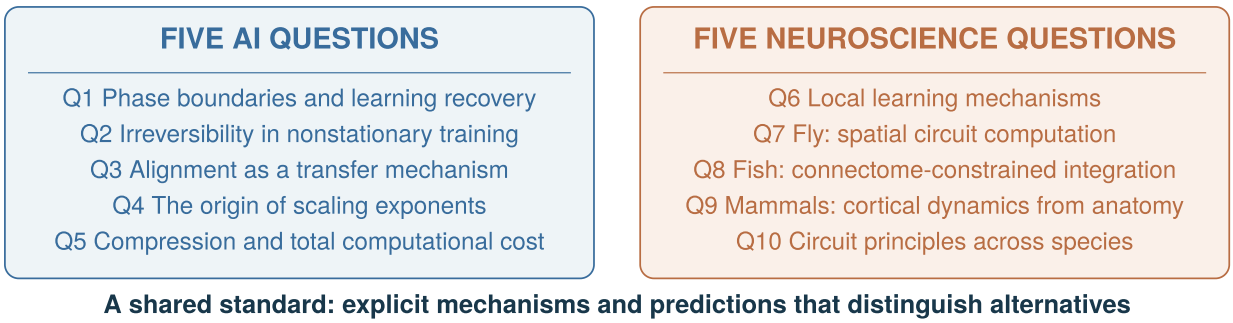


Figure 15. A balanced agenda. Five questions concern AI and five concern neuroscience. Within neuroscience, local learning, a fly circuit, fish, mammals, and comparison across species provide complementary levels of explanation.

Q1. Which phase boundaries predict collapse and recovery of learning in modern networks? The solvable theory links depth and regularization to competing learned and collapsed states. Define order parameters in a normalized residual or attention block, then vary decay, noise, and initialization separately. Predict a boundary and the stability of both states, distinguishing a global optimum switch from optimizer trapping and loss of plasticity.[27, 32, 4, 5]

Q2. Which irreversible quantities constrain nonstationary training with momentum or Adam? An effective potential can miss probability currents and changing state distributions. Include optimizer state, define a forward and reverse process, and derive an uncertainty bound for a specified training observable. A tractable model should identify when a scalar correction suffices and when nonstationarity changes the prediction.[33, 10, 6]

Q3. Which representation-alignment relations causally improve transfer? Alignment can retain shared features while discarding useful unique information. At comparable source-task error, vary alignment and modality-specific information independently in a controlled latent-variable task. Predict performance on a new task or held-out composition, separating a transfer mechanism from a statistic that merely accompanies training.[11, 13, 14]

Q4. Can a scaling exponent be derived without assuming a power-law spectrum? NQF connects a spectral tail to a loss exponent; a deeper explanation would derive that tail. Specify a feature operator $A(x)$ and data distribution, derive their learning spectrum, and predict both an exponent and a case with no power law before fitting a loss curve.[16, 19]

Q5. When does dynamical compression produce a genuine computational saving? Combine a trajectory-preservation guarantee with a construction whose cost is accounted for. Determine which data or representation dimension controls the bound, track errors over a stated horizon, and include construction, memory, and subsequent training. This would connect existence theorems to a regime where the full compressed procedure is cheaper.[17, 18]

17 Future research: five questions for neuroscience

Q6. Which local plasticity mechanisms produce coordinated gradient-like learning? Within the two-stream scaffold, find rules and initial conditions that establish both interstream and pathway alignment despite finite feedback delays. Then predict responses to separate changes in coactivation, modulatory feedback, and decay. The aim is to distinguish a functioning local learning mechanism from a Hebbian-looking correlation shared by different rules.[20, 21, 23, 22]

Q7. Which spatial inhibitory mechanism explains relative-motion computation in the fly circuit? In T4a–vCH candidate models, vary inhibitory location and conductance separately while respecting anatomical constraints. Compare predictions on held-out visual contexts and targeted perturbations. Track uncertainty in synaptic strength and membrane parameters, so a predicted effect can be assigned to a compartment-specific mechanism rather than one fitted parameter choice.[24]

Q8. Which fish connectome constraints make sensory integration robust across conditions? Larval zebrafish resources such as Fish1 offer a starting point for connectome-constrained modeling. Select a hindbrain circuit and test how recurrent excitation, contralateral inhibition, and cellular time constants jointly determine integration and choice. Constrain structure from anatomy, fit remaining physiology on one condition, and predict responses in another. Compare mechanisms while propagating reconstruction and cell-type uncertainty.[34]

Q9. How much mammalian cortical function is predicted by anatomically constrained dynamics? Use matched anatomy and activity, such as MICrONS mouse visual cortex, to compare a connectome-constrained model with degree-, distance-, and cell-type-matched null models. Fit uncertain conductances on training responses, then predict held-out stimuli and contextual effects. Explicitly model inputs crossing the imaged volume. The key test is which anatomical constraints improve predictions beyond functional fitting alone.[35]

Q10. Which effective circuit variables transfer across flies, fish, and mammals? For a matched visual computation, compare anatomically different circuits using effective time constants, recurrent gain, and excitation–inhibition organization. Ask whether a reduced model predicts the same class of perturbation response across species. Separate shared computational principles from species-specific implementation, and test the reduction against held-out responses rather than equating morphological similarity with functional equivalence.[24, 34, 35]

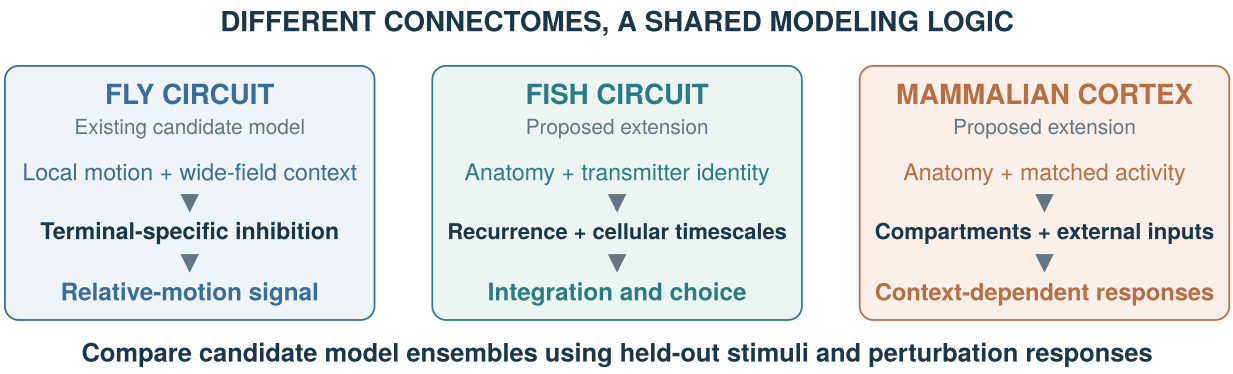


Figure 16. From a fly circuit to vertebrate modeling questions. Different organisms supply different anatomical and physiological constraints. A common test compares candidate dynamics on held-out responses and perturbations. Fish and mammalian modeling are proposed extensions; Fish1 and MICrONS are external resources.

References

Public status: 22 October 2026. Author order and contribution markers follow the website and original papers. * denotes equal contribution or correspondence; see each paper.

[1] Kangqiao Liu*, Liu Ziyin*, Masahito Ueda. *Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent*. ICML 2021. arXiv:2012.03636.

[2] Liu Ziyin*, Kangqiao Liu*, Takashi Mori, Masahito Ueda. *Strength of Minibatch Noise in SGD*. ICLR 2022. arXiv:2102.05375.

[3] Yizhou Xu, Pierfrancesco Beneventano, Isaac Chuang, Liu Ziyin. *Does SGD Seek Flatness or Sharpness? An Exactly Solvable Model*. Preprint, 2026. arXiv:2602.05065.

[4] Liu Ziyin. *Symmetry Induces Structure and Constraint of Learning*. ICML 2024. arXiv:2309.16932.

[5] Liu Ziyin, Mingze Wang, Hongchao Li, Lei Wu. *Parameter Symmetry and Noise Equilibrium of Stochastic Gradient Descent*. NeurIPS 2024. arXiv:2402.07193.

[6] Liu Ziyin*, Yizhou Xu*, Isaac Chuang. *Neural Thermodynamics: Entropic Forces in Deep and Universal Representation Learning*. NeurIPS 2025. arXiv:2505.12387.

[7] Andy Arditi, Weian Xie, David Bau, Liu Ziyin. *Learning the identity: a case study of how SGD selects among functional decompositions*. Preprint, 2026. arXiv:2610.00615.

[8] Liu Ziyin*, Zihao Wang*. *Sparsity by Redundancy: Solving L1 with SGD*. ICML 2023. arXiv:2210.01212.

[9] Liu Ziyin*, Yizhou Xu*, Isaac Chuang. *Remove Symmetries to Control Model Expressivity and Improve Optimization*. ICLR 2025. arXiv:2408.15495.

[10] Liu Ziyin, Yuanjie Ren, Adam Levine, Isaac Chuang. *Thermodynamic Irreversibility of Training Algorithms*. Preprint, 2026. arXiv:2605.21933.

[11] Liu Ziyin, Isaac Chuang, Tomer Galanti, Tomaso Poggio. *Formation of Representations in Neural Networks*. ICLR 2025. arXiv:2410.03006.

[12] Yizhou Xu*, Liu Ziyin*. *Three Mechanisms of Feature Learning in a Linear Network*. ICLR 2025. arXiv:2401.07085.

[13] Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, Paul Pu Liang. *Understanding the Emergence of Multimodal Representation Alignment*. ICML 2025. arXiv:2502.16282.

[14] Qiyao Liang, Daoyuan Qian, Liu Ziyin, Ila Fiete. *Compositional Generalization via Forced Rendering of Disentangled Latents*. ICML 2025. arXiv:2501.18797.

[15] Yongyi Yang, Tomaso Poggio, Isaac Chuang, Liu Ziyin*. *Topological Invariance and Breakdown in Learning Dynamics*. NeurIPS 2026. arXiv:2510.02670.

[16] Liu Ziyin*, Yizhou Xu*, Tomaso Poggio, Isaac Chuang. *Neural Quadratic Forms: A Unified Minimal Model for Sudden Learning and Scaling Laws*. Preprint, 2026. arXiv:2608.13335.

[17] Hong-Yi Wang, Di Luo, Tomaso Poggio, Isaac L. Chuang, Liu Ziyin*. *A universal compression theory for lottery ticket hypothesis and neural scaling laws*. ICLR 2026. arXiv:2510.00504.

[18] Hong-Yi Wang, Mingze Wang, Liu Ziyin. *Width-Independent Compressibility of Deep Neural Networks*. Preprint, 2026. arXiv:2608.21752.

[19] Xing Han, Liu Ziyin, Suchi Saria, Paul Pu Liang. *On the Invariance and Generality of Neural Scaling Laws*. NeurIPS 2026. arXiv:2605.07546.

[20] Qianli Liao*, Liu Ziyin*, Yulu Gan*, Brian Cheung, Mark T. Harnett, Tomaso Poggio. *How Biological Synapses Self-Assemble Gradient Learning*. PNAS 123(41), e2606245123 (2026). doi:10.1073/pnas.2606245123; arXiv:2412.20018v2.

[21] Liu Ziyin, Isaac Chuang, Tomaso Poggio. *Heterosynaptic Circuits Are Universal Gradient Machines*. Preprint, 2025. arXiv:2505.02248.

[22] David Koplow, Tomaso Poggio, Liu Ziyin. *Ubiquity of Emergent Hebbian Dynamics in Regularized Learning*. ICML 2026. arXiv:2505.18069.

[23] Marc Gong Bacvanski, Liu Ziyin, Tomaso Poggio. *Does Feedback Alignment Work at Biological Timescales?* Preprint; revised 2026. arXiv:2510.18808.

[24] Liu Ziyin, Yifei Lu, Alexander Borst, Sven Dorkenwald, Tomaso Poggio. *Connectome-based biophysical modeling of a figure-ground discrimination circuit*. bioRxiv preprint, 2026. doi:10.64898/2026.09.17.752514.

[25] Liu Ziyin*, Hongchao Li*, Masahito Ueda. *Noise Balance and Stationary Distribution of Stochastic Gradient Descent*. Physical Review E 111, 065303 (2025). arXiv:2308.06671.

[26] Zihao Wang*, Liu Ziyin*. *Posterior Collapse of a Linear Latent Variable Model*. NeurIPS 2022. arXiv:2205.04009.

[27] Liu Ziyin, Masahito Ueda. *Zeroth, first, and second-order phase transitions in deep neural networks*. Physical Review Research 5, 043243 (2023). doi:10.1103/PhysRevResearch.5.043243.

[28] Liu Ziyin, Ekdeep Singh Lubana, Masahito Ueda, Hidenori Tanaka. *What shapes the loss landscape of self-supervised learning?*. ICLR 2023. arXiv:2210.00638.

[29] James B. Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J. Fetterman, Joshua Albrecht. *On the stepwise nature of self-supervised learning*. ICML 2023. arXiv:2303.15438.

[30] Liu Ziyin, Tilman Hartwig, Masahito Ueda. *Neural Networks Fail to Learn Periodic Functions and How to Fix It*. NeurIPS 2020. arXiv:2006.08195.

[31] Liu Ziyin, Isaac Chuang. *Proof of a perfect platonic representation hypothesis*. Technical note, 2025. arXiv:2507.01098v2.

[32] Liu Ziyin, Botao Li, Xiangming Meng. *Exact Solutions of a Deep Linear Network*. NeurIPS 2022. arXiv:2202.04777.

[33] Liu Ziyin, Masahito Ueda. *Universal thermodynamic uncertainty relation in nonequilibrium dynamics*. Physical Review Research 5, 013039 (2023); preprint 2022. doi:10.1103/PhysRevResearch.5.013039.

[34] Fish1 collaboration: Lichtman and Engert laboratories with Google Connectomics. *Fish1: A Connectomic Resource for the Larval Zebrafish Brain*. Public dataset and accompanying preprint, 2025. Official resource and manuscript.

[35] MICrONS Consortium. *Functional connectomics spanning multiple areas of mouse visual cortex*. Nature 640, 435–447 (2025). doi:10.1038/s41586-025-08790-w.

[36] Takashi Mori, Liu Ziyin, Kangqiao Liu, Masahito Ueda. *Power-Law Escape Rate of SGD*. ICML 2022, PMLR 162:15959–15975. Conference paper; arXiv:2105.09557.

Figure note. All sixteen figures are original conceptual and organizational diagrams based on the cited research. They depict relationships among questions, mechanisms, frameworks, and levels of neural organization. Diagram geometry and colors are explanatory rather than quantitative. Editable English and Chinese SVGs accompany the source files.