图解研究概述 2026 年 10 月

学习、结构与神经计算

刘子寅及合作者的研究

一个任务通常不能唯一确定神经网络的内部解。隐藏单元可以交换，因子的尺度可以重新分配，同一种计算也可以由不同的逐层分解实现。训练过程会在这些可能性中作出选择。对于生物系统，局部可塑性（local plasticity）还需要组装出学习所需的机制；突触的位置和生物物理性质，则进一步影响回路完成什么计算。

刘子寅的研究结合数学、统计物理、机器学习和神经科学考察这些问题。相变（phase transition）理论解释学得状态的定性改变，随机热力学（stochastic thermodynamics）约束涨落与不可逆动力学。一种反复出现的方法是：在可求解模型中分离机制，找到适当的集体变量（collective variables），再推导可检验预测。长远目标是寻找人工智能与生物智能的科学原理。

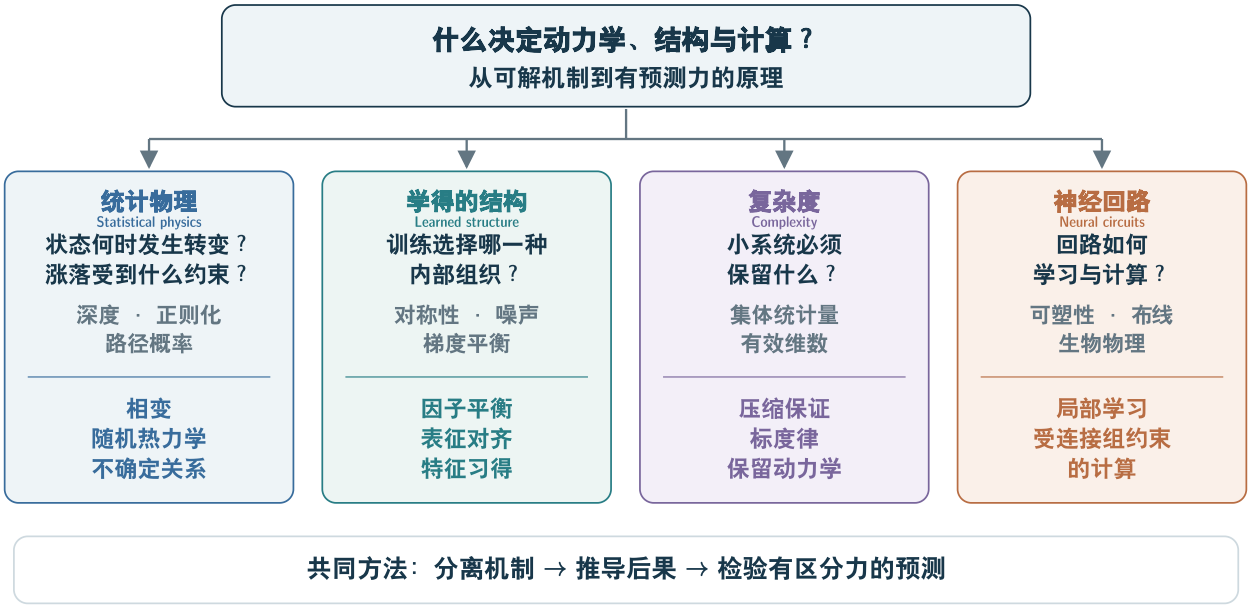


图 1. 研究纲领的组织。统计物理、学习结构、有效复杂度与神经回路，构成四条相连的主线。相变与随机热力学既提供具体成果，也为研究学习过程提供工具。

论述路线

可点击的目录提供各主题的入口。一条路线从有限步长、相变、对称性与热力学展开；另一条路线考察表征、拓扑、标度与压缩。生物章节从局部学习规则走向连接组约束的计算。最后讨论研究局限，并分别提出五个人工智能问题和五个神经科学问题。

图示用于解释概念与组织：哪些问题相互联系，训练选择什么，以及机制如何连接不同描述层次。所有图均依据所引研究原创绘制。中英文版的方程、图号和文献相互对应；广泛假说、条件性定理与生物机制模型，保留各自不同的证据性质。

目录

研究概述 **1**

1 核心概念与阅读方法 **4**
1.1 先分清五组概念 . . . 4
1.2 如何阅读数学记号 . . . 4
1.3 阅读路线与解释标准 . . . 4

2 为什么有限训练步长重要 **5**
2.1 一个精确计算 . . . 5
2.2 结构化噪声选择解，也控制逃逸 . . . 5

3 相变：深度改变学习出现的方式 **6**
3.1 随深度改变的相变分类 . . . 6
3.2 为什么更深的模型可以保留坍缩吸引域? . . . 6
3.3 精确解将分类连接到优化问题 . . . 6

4 对称性使噪声产生有组织的选择 **7**
4.1 标量计算揭示机制 . . . 7
4.2 从示例到一般结果 . . . 7

5 熵力如何预测内部的函数分解 **8**
5.1 恒等函数任务使解选择变得清晰 . . . 8
5.2 一个标量计算给出直觉 . . . 8

6 随机热力学与学习的时间箭头 **9**
6.1 用概率比衡量时间方向 . . . 9
6.2 超越稳态的普适不确定关系 . . . 9
6.3 算法不可逆性的四种视角 . . . 9

7 用矩阵对齐刻画表征学习 **10**
7.1 究竟哪些量应当对齐? . . . 10
7.2 对齐的机制，以及有结构的偏离 . . . 10
7.3 对齐提供信息，但其作用依赖任务 . . . 10

8 拓扑、特征学习与坍缩 **11**
8.1 小更新可以保留损失之外的结构 . . . 11
8.2 特征学习有多种动力学机制 . . . 11
8.3 不同坍缩需要不同解释 . . . 11

9 神经二次型：从突变到标度 **12**
9.1 不同架构共享的局部描述 . . . 12
9.2 谱尾如何给出神经标度律（neural scaling laws）指数 . . . 12

10 为什么对称系统能够压缩 **13**
10.1 一个精确的有限维思路 . . . 13
10.2 从密集对象到多对数（polylogarithmic）上界 . . . 13
10.3 保留学习，还需要保留相应更新 . . . 13

11 压缩一个固定的深层网络 **14**
11.1 匹配输入导数，而非每个隐藏坐标 . . . 14
11.2 两类保证应当分清 . . . 14
11.3 标度律也存在不变量 . . . 14

12 回路如何组装出自己的学习规则 **15**
12.1 哪些结构会自组装，定理又要求什么? . . . 15

13 局部可塑性：机制、现象与时间 **16**
13.1 异突触计算是一种设计原理 16
13.2 为什么像 Hebb 的更新不一定来自 Hebb 规则? 16
13.3 学习规则还必须在恰当的时间工作 16

14 从连接组理解相对运动计算 **17**
14.1 不同区室可以完成不同计算 17
14.2 为什么分流能够归一化信号? 17
14.3 连接组为解释贡献了什么? 17

15 当前研究的局限与边界 **18**
15.1 从可解条件走向真实训练 18
15.2 从内部组织走向泛化能力 18
15.3 从紧凑表示走向实际资源节省 18
15.4 从生物合理规则走向可辨识的真实机制 18
15.5 从相关机制走向整合理论 18

16 未来研究：五个人工智能问题 **19**
Q1. 哪些相边界能够预测现代网络的坍缩与学习恢复? 19
Q2. 哪些不可逆量能够约束带动量（momentum）或 Adam 的非平稳训练? 19
Q3. 哪些表征对齐关系，确实会因果性地改善迁移? 19
Q4. 能否不预设幂律谱，就推导出标度指数? 19
Q5. 保持动力学的压缩（dynamical compression），何时真正节省计算? 19

17 未来研究：五个神经科学问题 **20**
Q6. 哪些局部可塑性机制能够产生协调的类梯度学习? 20
Q7. 果蝇回路中的哪种空间抑制机制解释相对运动计算? 20
Q8. 鱼类连接组中的哪些约束，使感觉整合能跨条件保持稳健? 20
Q9. 受解剖约束的动力学，能够预测多少哺乳动物皮层功能? 20
Q10. 哪些有效回路变量，能够在果蝇、鱼类与哺乳动物之间迁移? 20

参考文献 **21**

1 核心概念与阅读方法

这些研究将科学问题转化为模型、推导和可检验预测。学习理论、物理与神经科学提供互补的切入点。以下五组概念，有助于理解其中的思想，以及不同论文之间的联系。

1.1 先分清五组概念

1. **函数与参数**。预测 $f_\theta(x)$ 是网络计算的结果，权重 θ 是某一种实现方式。多组权重可以实现同一个函数，却对训练更新产生不同响应。

2. **目标函数与学习规则**。损失 $L(\theta)$ 为预测评分；SGD、梯度（gradient）流和自适应优化器，则是不同的权重改变方式。规定评分，并没有规定学习路径或内部解。

3. **表征（representation）与输出**。隐藏活动 $h(x)$ 是输入的中间描述。即使原任务输出已经准确，它仍可能保留、丢弃或重组对其他任务有用的信息。

4. **对称性（symmetry）与被选择的组织**。对称变换保持某个指定的函数或损失不变，训练却仍可能在这些自由度中偏向平衡、稀疏或对齐的状态。阅读时，要先问清楚究竟哪个量保持不变。

5. **连接图与计算**。连接组（connectome）记录连接关系；回路的工作还依赖突触位置、活动、电导和时序。生物机制解释需要把这些因素与可观测响应联系起来。

1.2 如何阅读数学记号

记号	直观含义
$\mathbb{E}[\cdot]$	对指定的数据或训练随机性取平均。
∇L、$\nabla^2 L$	梯度给出局部上升方向；Hessian 描述该方向如何变化，即局部曲率。
$\mathrm{Cov}(g)$	梯度相对均值的涨落；二阶矩（second moment）$\mathbb{E}[hh^\top]$ 还包含均值的贡献。
特征向量（eigenvector）、特征值（eigenvalue）	某个算子的方向或模态，以及相应强度；含义取决于研究的是哪个算子。
$=$、$\simeq$、标度	分别表示精确等式、特定条件下的近似，以及描述增长或衰减的关系。

1.3 阅读路线与解释标准

沿任一条路线，都可以找出要解释的现象、最小模型、假设，以及能够区分机制的预测。要分清精确定理、有效近似、计算观察与生物假说。

切入点	建议路线
物理	相变 → 不确定关系 → 不可逆训练。[27, 33, 10]
学习	有限步长 → 噪声平衡 → 表征 → NQF。[1, 5, 11, 16]
复杂度	对称系统压缩 → 解析固定网络压缩。[17, 18]
神经科学	局部学习 → 果蝇回路 → 鱼类与小鼠资源。[20, 24, 34, 35]

线性代数、微积分、概率、动力系统与突触生理，提供互补的基础。第8页的标量例子，将函数等价与噪声选择联系起来。

2 为什么有限训练步长重要

随机梯度下降（stochastic gradient descent, SGD）用随机小批量（minibatch）的平均梯度 g_t，更新权重 θ：

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

其中 η 为学习率（learning rate），C 描述梯度噪声（gradient noise）的几何结构。梯度流（gradient flow）用连续运动近似这些有限跳跃。有限步长与小批量噪声的一系列工作，直接研究离散更新，并判断连续时间近似在哪些条件下受控。[1, 2]

2.1 一个精确计算

设 $L(\theta) = \frac{1}{2}\theta^\mathsf{T} H\theta$、$H \succ 0$，且 $g_t = H\theta_t + \xi_t$，其中独立加性噪声（additive noise）的均值为零、协方差（covariance）为 C。记 $A = I - \eta H$，得到 $\Sigma_{t+1} = A\Sigma_t A^\mathsf{T} + \eta^2 C$。因此，当 $\rho(A) < 1$ 时，平稳协方差满足

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

这里 Σ 为平稳的参数协方差，$\rho(A)$ 是 A 的最大特征值绝对值；稳定性要求它小于一。连续近似忽略了左侧第三项。在一维情形，

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

当更新接近稳定边界，两种描述的差异迅速增大；不同曲率方向也会得到不同程度的放大。

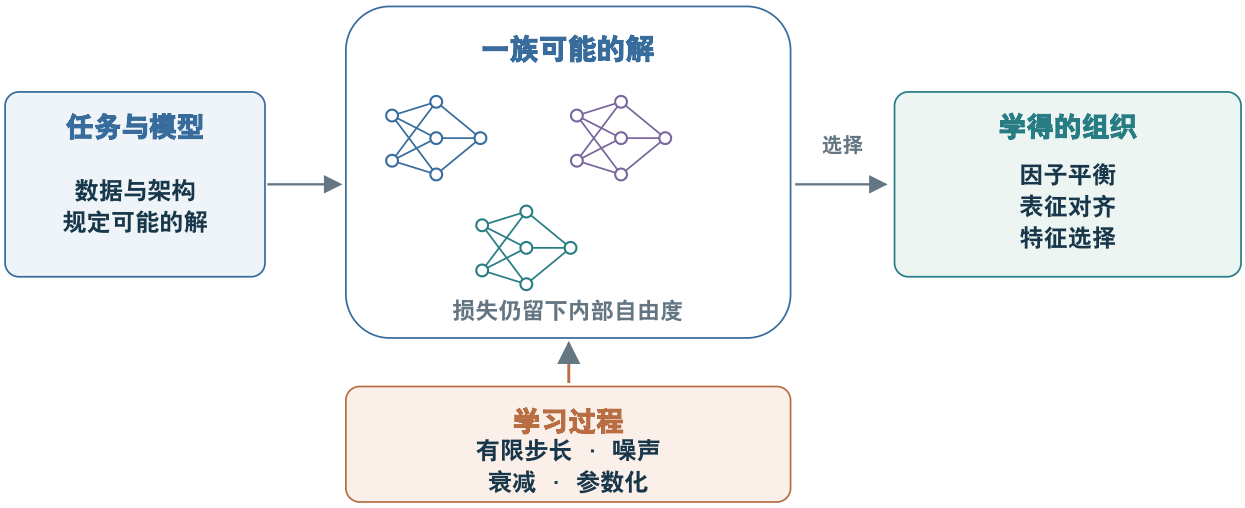


图 2. 目标施加约束，学习过程进行选择。任务与架构允许一族内部解；有限步长、噪声、正则化与参数化决定训练偏向哪些组织。网络图标表示概念上的不同可能。

2.2 结构化噪声选择解，也控制逃逸

小批量噪声依赖参数、数据和残差。在 2026 年的精确可解模型中，各向同性（isotropic）与各向异性（anisotropic）的标签噪声，可以选择更平或更尖的解；起作用的量是梯度涨落。[3]

与 Takashi Mori 等人的合作，分析均方损失下 SGD 的乘性噪声（multiplicative noise）。随机时间变换（random time change）揭示了对数损失势垒（logarithmic loss barrier），在相应条件下给出幂律逃逸（power-law escape）。[36] 因此，噪声结构既改变解的选择，也改变从亚稳态（metastable state）逃出的方式。

3 相变：深度改变学习出现的方式

与 Masahito Ueda 合作的相变研究追问：拟合预测的收益，何时克服保持模型简单的压力？核心对象是全局最小化后的正则化（regularization）目标：

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

序参量（order parameter）区分平凡解与学得的解；$\mathcal{V}$ 或其导数的非解析变化，定义相变阶数。这是一种零温损失景观（zero-temperature loss landscape）理论：即使网络规模有限，最小化也可以产生奇点。[27]

3.1 随深度改变的相变分类

在论文研究的带随机神经元、正则化和有效输入–目标信号的深层线性模型中，岭回归（ridge regression）在正则化强度为正时没有相变；一个隐藏层出现二阶（second-order）的连续起始；多于一个隐藏层，则允许平凡相与学习相之间的一阶（first-order）切换。零阶（zeroth-order）不连续性出现在奇异的无限深度极限中的零正则化处。有限深度时，全局最优值保持连续；训练路径的跳变，则可能来自局部困陷。[27]

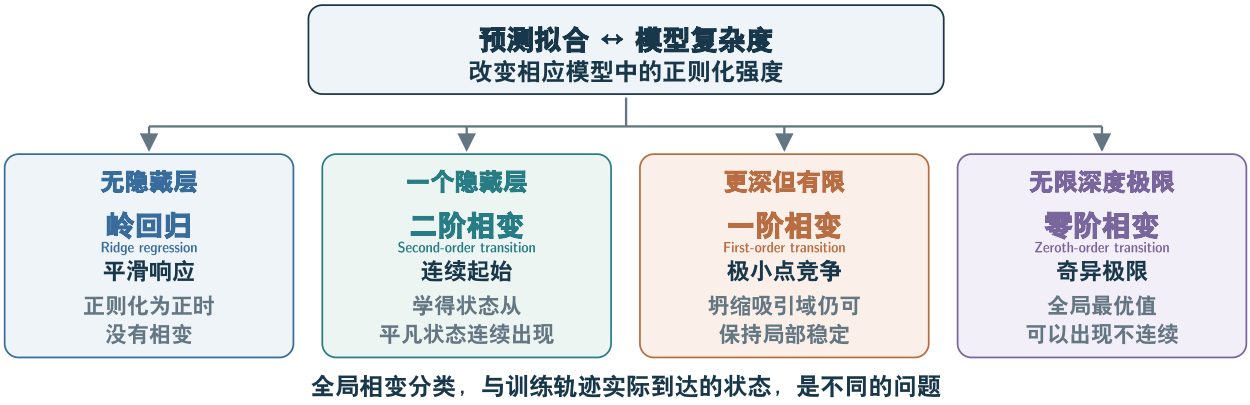


图 3. 深度改变转变的组织方式。[27] 四条分支按正则化变化，对相应模型分类。连续起始、极小点竞争和奇异极限下的最优值跳变，是不同机制。方框组织理论结果，不表示测量曲线或实测阈值。

3.2 为什么更深的模型可以保留坍缩吸引域？

一个平衡的标量链可以揭示机制。设权重层数为 m，目标 $a > 0$，各因子均为 $w \geq 0$，并施加二次衰减：

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

当 $m = 2$ 时，$w_*^2 = \max(a - \gamma, 0)$，学得的因子连续出现。当 $m > 2$ 且 $\gamma > 0$ 时，即使非零解损失更低，正的二次项仍稳定原点；学习因而需要离开局部稳定的坍缩（collapse）状态。这一计算解释损失景观机制，并不把所有突变的学习曲线都视为相变。

3.3 精确解将分类连接到优化问题

与 Botao Li、Xiangming Meng 合作的精确解工作，刻画带衰减和随机神经元的深层线性网络极小点，揭示原点的特殊性。[32] 相变论文的线性原点定理（linear-origin theorem），将这一局部视角推广到满足相应激活函数与原点条件的非线性网络。[27] 互补的 PRE 工作求出对角线性网络中随机梯度流的平稳分布，出现噪声诱导的相变、遍历性破缺与涨落反转。[25] 这些现象，区别于 NQF 研究的特征依次激活动力学。[16]

4 对称性使噪声产生有组织的选择

镜像对称性（mirror symmetry）定理从 $\ell_0(\theta) = \ell_0((I - 2OO^\mathsf{T})\theta)$、$O^\mathsf{T}O = I$ 出发。O 的列是被反射、即符号翻转的正交归一方向。变换后保持不变的固定子空间为

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

梯度保持这一子空间。在 Hessian 有下界的条件下，足够强的权重衰减（weight decay）使每个极小点都位于其中。该框架从重缩放（rescaling）对称性导出稀疏（sparsity），从旋转对称性导出低秩（low rank），从置换对称性导出同质集成（homogeneous ensembling）。[4]

4.1 标量计算揭示机制

令 $\ell_t = \frac{1}{2}(uv - y_t)^2$、$e_t = uv - y_t$。同时更新得到 $u' = u - \eta e_t v$、$v' = v - \eta e_t u$。展开平方后，交叉项抵消，得到精确等式：

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

总体损失上的梯度流保持 D，有限步长却会改变它。当条件二阶矩有限，且 $0 < \eta^2 \mathbb{E}[e_t^2 \mid u, v] < 1$ 时，不平衡量的条件均值收缩。因此，相同的乘积函数可以获得平衡的内部因子。

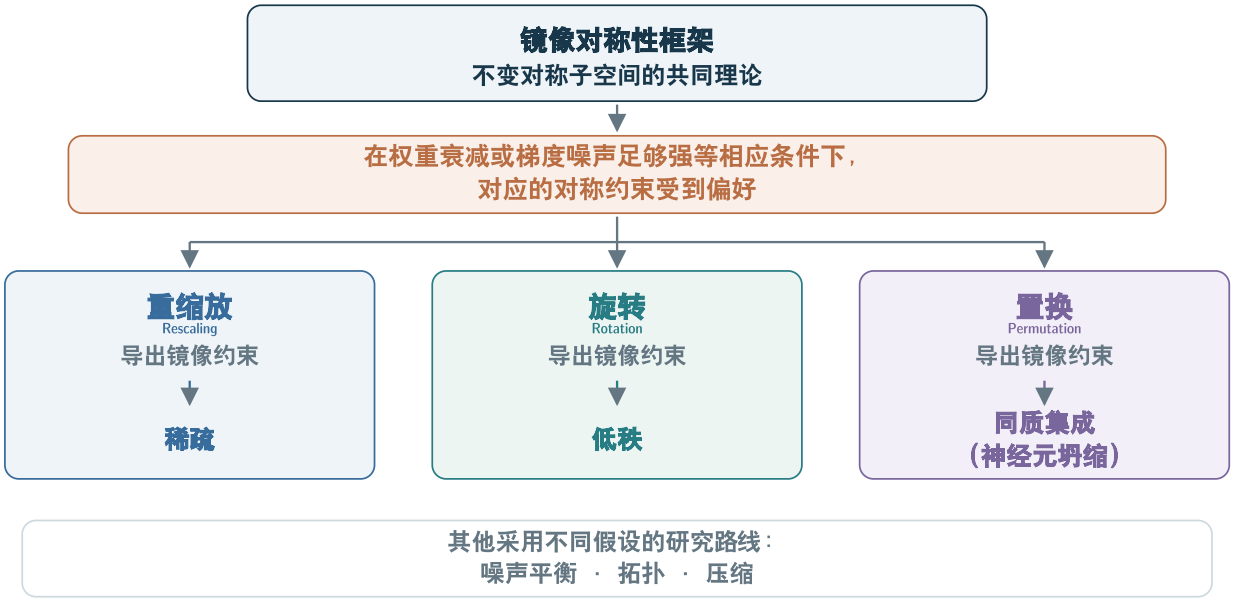


图 4. 镜像对称性组织了约束结果。[4, 5, 15, 17] 重缩放、旋转与置换导出镜像约束，噪声或衰减的共同条件适用于三类推论。噪声平衡、拓扑与压缩是采用各自假设的其他对称性研究路线。

4.2 从示例到一般结果

对于指数型对称性（exponential symmetry），诺特流（Noether flow）推动参数沿退化方向趋向相应条件下的唯一噪声平衡（noise equilibrium）点，不依赖轨道上的初始位置；各噪声贡献在此平衡并对齐。PRE 论文求出对角线性网络中随机梯度流的平稳分布，出现相变、遍历性破缺（ergodicity breaking）和涨落反转（fluctuation inversion）。[5, 25]

约束机制将对称性与可塑性丧失和坍缩联系起来。[4] 它也启发了利用冗余的 *spred*，以及打破有害对称性陷阱的 *syre*。[8, 9] 对称性由此成为解释和控制学习组织的工具。

5 熵力如何预测内部的函数分解

神经热力学（neural thermodynamics）研究有限步长与随机梯度产生的熵力（entropic forces）。[6] 对固定小批量，后向误差分析（backward error analysis）给出修正损失的领先阶项；将其对小批量取平均，定义一阶熵修正损失（entropic loss）：

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

修正项衡量对随机梯度的敏感程度。对于标量学习率，理论破坏非紧（noncompact）的重缩放对称性，同时保持正交对称性（orthogonal symmetry）。由此产生类似能量均分（equipartition）的梯度平衡（gradient balance）关系，联系表征对齐与变平或变尖的现象。有效描述在小步长区域内受控。

5.1 恒等函数任务使解选择变得清晰

与 Andy Arditi、Weian Xie、David Bau 的合作研究考虑 $f(x) = (I + W_2)(I + W_1)x$、$y = x + \varepsilon$。输入各向同性，正的标签噪声特征值归一化为 $\sum_i \lambda_i = d$。记 $\widetilde{W}_\ell = I + W_\ell$，每个满足 $\widetilde{W}_2\widetilde{W}_1 = I$ 的分解都有相同总体损失，但熵项能够区分它们。[7]

在论文的对齐基中，一个奇异值模态（singular-value mode）a_i 对应的流形上目标为

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

无权重衰减时，求导即得 $a_i^4 = \lambda_i$；加入衰减后，在相应条件下，选中的值位于 $\lambda_i^{1/4}$ 与 1 之间。

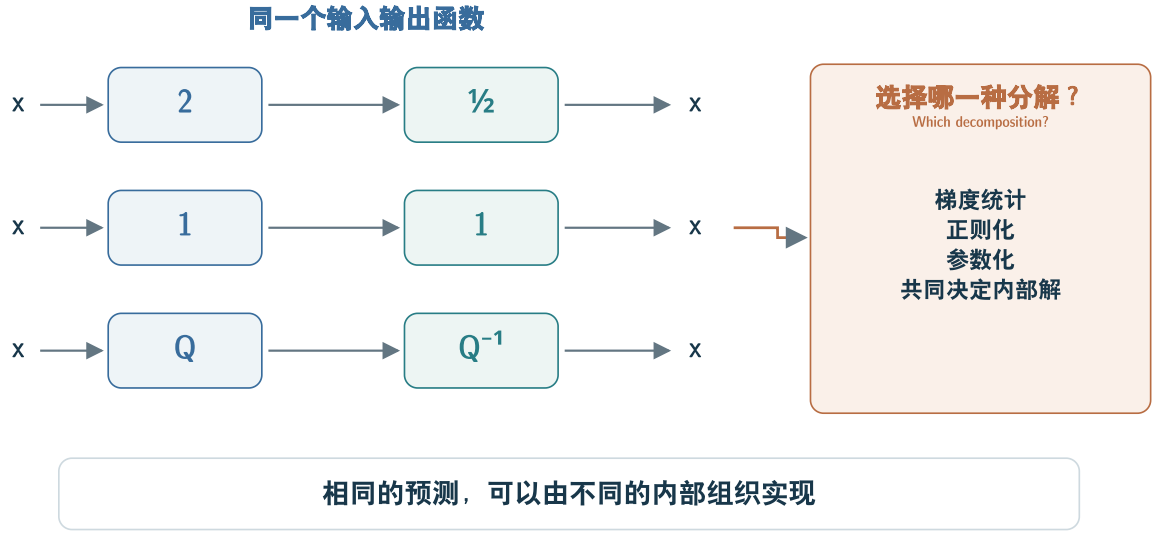


图 5. 一个函数，多种分解。[7] 三条路径表示等价的输入输出映射，其中 Q 可逆。熵驱动的选择研究梯度统计、正则化和参数化如何区分内部因子。

把残差权重重新参数化为 $W_\ell = A_\ell B_\ell$，即使没有衰减，也会偏向零残差；更深的直接参数化链，则将预测的噪声各向异性集中于边界层。[7]

5.2 一个标量计算给出直觉

令 $f = uv$，标签为 $1 + \varepsilon$，平方损失除以二，独立标签噪声的方差为 σ^2。在最优流形 $uv = 1$ 上，小批量梯度的二阶矩为 $\sigma^2(u^2 + v^2)/B$。因此 $S = \eta\sigma^2(u^2 + v^2)/(4B)$ 偏向平衡因子，因为此处 $u^2 + v^2 \geq 2$。同一个函数可以对训练噪声有不同的敏感程度。第9页将介绍互补的、基于轨迹的热力学描述。

6 随机热力学与学习的时间箭头

随机热力学研究，涨落如何约束远离平衡的过程。这是研究中独立的一条主线，涵盖普适热力学不确定关系（universal thermodynamic uncertainty relation, TUR）与训练算法的不可逆性。基本对象是一条轨迹，而不仅是轨迹终点的损失。

6.1 用概率比衡量时间方向

记 $P(\Gamma)$ 为正向路径的分布，$Q(\Gamma)$ 为在明确的反向协议（reverse protocol）下，对应的反向路径分布。当支撑相容时，无量纲的路径熵产生（path entropy production）为

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

其均值比较完整历史的概率。熵产生衡量时间不对称，区别于描述状态分布分散程度的熵。[33]

6.2 超越稳态的普适不确定关系

与 Masahito Ueda 合作的工作，对一般 Markov 动力学中的任意可观测量推导 TUR，允许时变协议和绝对不可逆性（absolute irreversibility）。要得到紧的约束，熵产生与非平稳性（nonstationarity）都很重要。在更简单的 $Q \ll P$、相关矩有限的情形，测度变换（change of measure）给出

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

这里 A 为路径可观测量（path observable）；存在绝对不可逆性时，完整结果还包括支撑修正。取消对可观测量的反对称限制，也将热力学不确定界与理论金融中的一个基本不等式联系起来。[33]

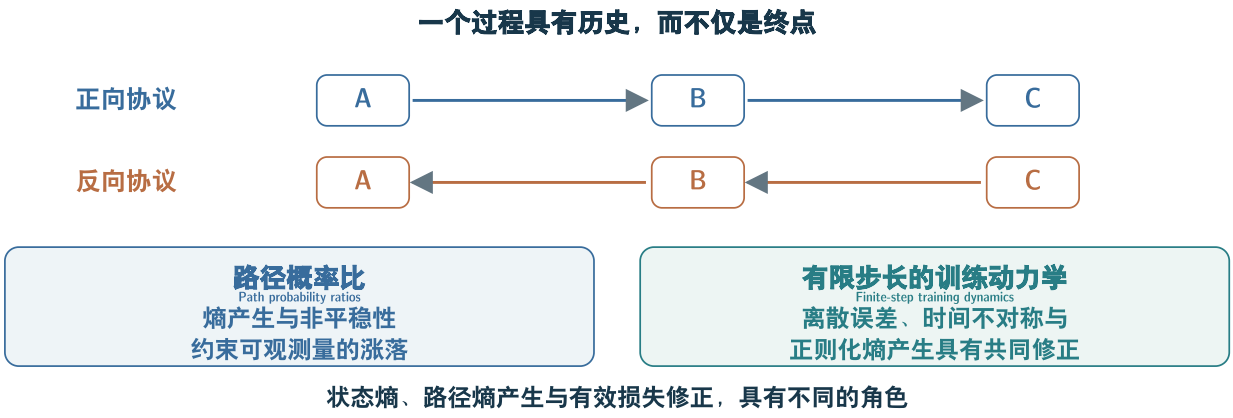


图 6. 从过程历史到动力学约束。正向与反向路径概率的比较，给出熵产生和不确定界。对于学习算法，有限步长分析提供通向有效作用的另一条路线。这些描述层次相互联系，但各自的假设需要分别明确。

6.3 算法不可逆性的四种视角

与 Yuanjie Ren、Adam Levine、Isaac Chuang 合作的 2026 年预印本，连接数值离散误差（numerical discretization error, DE）、时间重整化（time renormalization, TR）、微观时间不对称（microscopic time-reversal asymmetry, TA）和正则化的随机熵产生（regularized stochastic entropy production, ST）。对更新 $\theta_{t+1} = \theta_t - \eta U(\theta_t)$，在更新场平滑且 Jacobian 对称的条件下，

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

共同的领先阶项，将数值更新的时间箭头，连接到选择学习轨迹的有效作用。随机热力学构造使用论文规定的正则化与归一化。[10] 这为熵驱动的选择提供动力学视角：有效损失修正、状态熵（state entropy）和路径熵产生，具有不同的数学角色。[6]

7 用矩阵对齐刻画表征学习

7.1 究竟哪些量应当对齐？

对于隐藏线性映射 $h_b = Wh_a$，定义表征与神经元梯度的二阶矩：

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

典范表征假说（Canonical Representation Hypothesis, CRH）在线性映射的两侧各提出三种关系：

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

这六个关系针对完整矩阵，不只是少数主方向之间的重叠。它们联系了哪些特征存在、哪些特征影响损失，以及权重传递哪些特征。[11]

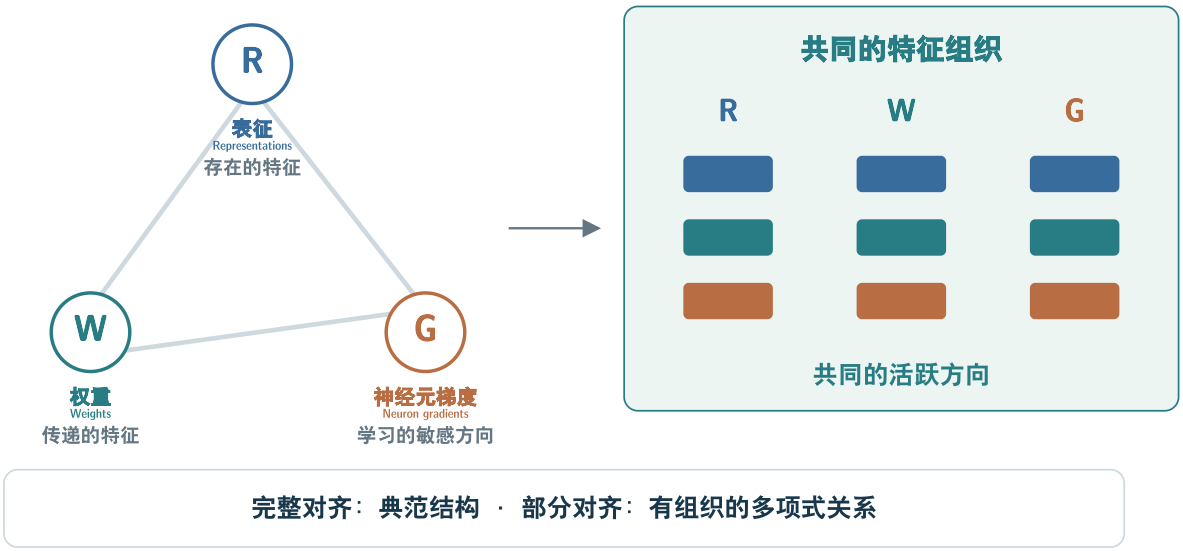


图 7. 学得特征组织的三个视角。[11] 表征、权重和神经元梯度，分别描述存在的特征、传递的特征及学习敏感方向。CRH 联系线性映射两侧的矩阵结构。彩色通道示意共同方向，而非实测谱。

7.2 对齐的机制，以及有结构的偏离

论文在明确假设下，从噪声–正则化平衡（noise–regularization balance）推导对齐，其中包括平均场范数条件（mean-field norm conditions）和适当的平稳近似。它也分析有限时间造成的偏离，以及不满足完整 CRH 的层。多项式对齐假说（Polynomial Alignment Hypothesis, PAH）研究投影后半正定矩阵（positive-semidefinite matrix）的幂次关系，可示意为

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

其中子空间和指数由原论文具体给定。因此，部分关系失效时，谱结构仍可能具有可分析的组织方式。假说的广泛适用性，与某个具体证明覆盖的范围，是不同的问题。

7.3 对齐提供信息，但其作用依赖任务

该技术笔记在指定的嵌入式深线性模型中证明一种解选择结果：尽管多数损失极小点不具柏拉图式表征（Platonic representation），SGD 却选择相差旋转的等价表征。[31] 与 Megan Tjandrasuwita、Chanakya Ekbote、Paul Pu Liang 的合作发现，跨模态对齐（multimodal alignment）是否有益，取决于任务中的冗余和独有信息。[13] 因此，对齐本身也是需要结合任务解释的对象。

与 Qiyao Liang、Daoyuan Qian、Ila Fiete 的合作表明，迫使解耦因子（disentangled factors）进入输出表征，可以改善受控生成任务中的组合外推（compositional extrapolation）；否则后续层仍可能重新纠缠这些因子。[14] 这一结果将泛化与学得计算的完整组织联系起来。

8 拓扑、特征学习与坍缩

8.1 小更新可以保留损失之外的结构

拓扑（topology）关注连通关系、孔洞等在连续可逆变形下保留的性质。若更新场 u 为 K-Lipschitz，即 $\|u(x)-u(y)\| \le K\|x-y\|$，且 $T(x)=x-\eta u(x)$，由三角不等式可得

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

当 $\eta K < 1$ 时，不同的点不能合并，且像集上的逆映射也是 Lipschitz 的。对于满足条件的置换等变（permutation equivariance）规则，包括 SGD 和 Adam，拓扑论文识别出保持神经元分布拓扑的小步长区域；较大步长则允许简化，并可能降低表达能力（expressivity）。[15]

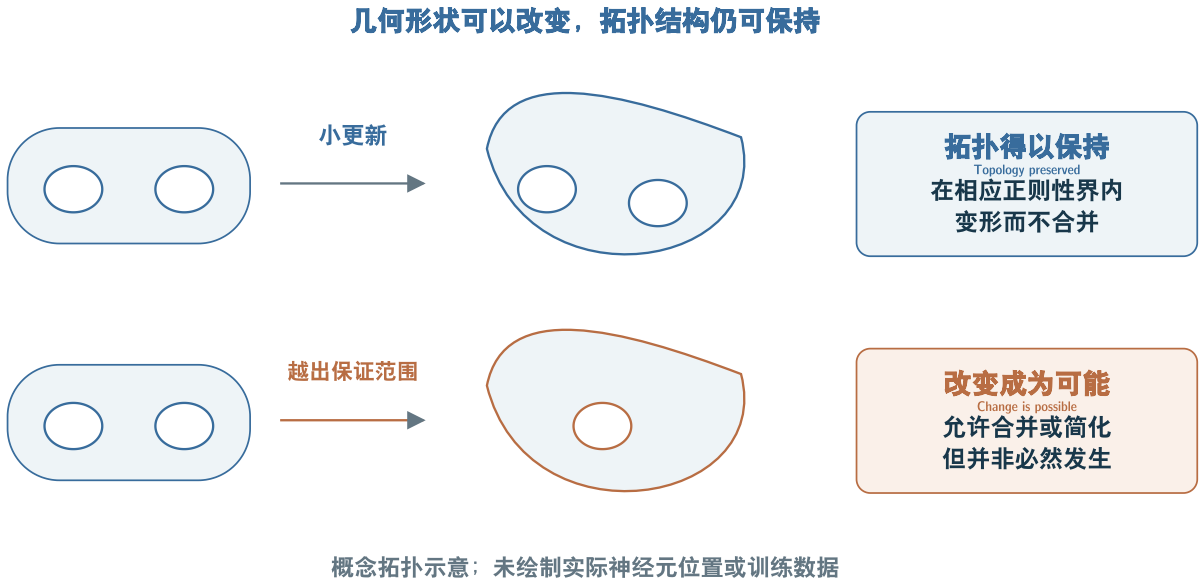


图 8. 几何与拓扑。[15] 形状可以变形，同时保持连通关系和孔洞。越出某个充分的正则性保证后，合并或简化成为可能。这些形状用于解释概念，并非重建的神经元分布。

越过上述充分条件的界限，意味着拓扑可以改变，并不意味着每个更大的步长都必然改变拓扑。例如，标量映射 $T(x)=(1-\eta K)x$ 在 $\eta K=1$ 时坍缩，再增大步长却又恢复可逆性。将界限应用于实际训练时，需要区分充分条件与必要的相边界。

8.2 特征学习有多种动力学机制

与 Yizhou Xu 合作的有限宽线性模型精确解，区分了对齐、去对齐（disalignment）和重缩放三种机制，联系了近似核方法与特征学习区域。[12] 自监督学习（self-supervised learning, SSL）的逐步习得研究在线性化模型中导出特征模态的依次习得，并在非线性联合嵌入网络中观察到类似行为。[29] 这些分析支持追踪集体变量，而不只是总损失。

8.3 不同坍缩需要不同解释

自监督损失景观理论分析完全坍缩（complete collapse）与维度坍缩（dimensional collapse），并研究归一化（normalization）、偏置及数据结构的作用。[28] 线性潜变量（latent variable）模型则将后验坍缩（posterior collapse）归因于似然拟合与先验引入的潜变量均值正则化之间的竞争。[26] 相变工作说明，深度与复杂度惩罚能够改变可解网络中的相变阶数。[27] 这些都是结构由目标和动力学共同产生的例子，但具体机制与假设各不相同。

9 神经二次型：从突变到标度

9.1 不同架构共享的局部描述

与 Yizhou Xu 合作的神经二次型（Neural Quadratic Forms, NQF）工作，针对平滑的可交换单元求和模块，在小权重附近、原点满足零梯度条件时，推导二次型正规形式（normal form）：

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

感知机（perceptron）、注意力（attention）、专家混合（mixture of experts）与卷积（convolution），对应不同的结构矩阵 $A(x)$。M 汇集单元权重的外积（outer product），保留特征演化，而没有将其冻结为固定核。[16]

当数据矩阵共享特征基时，动力学化为广义 Lotka–Volterra 模态方程。在特征正交的情形，

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

这里 z_k 是模态强度，r_k 是初始增长率，C_{kk} 控制自限制。正增长模态的饱和值为 r_k/C_{kk}。若 $z_k(0) = \epsilon$ 很小，激活时间近似为 $r_k^{-1}\log[r_k/(C_{kk}\epsilon)]$。因此，平滑动力学中也可以出现时间上彼此分离的特征激活。

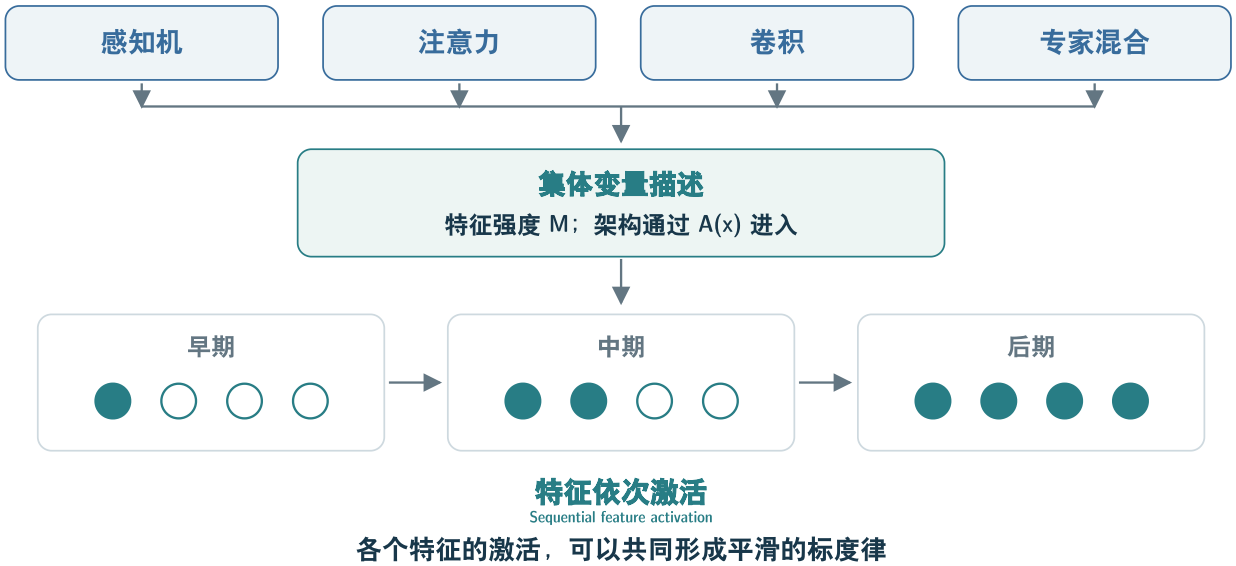


图 9. 从微观架构到集体学习。[16] 在相应假设下，不同模块共享局部二次描述。特征模态依次激活，其贡献结合后，在假定谱尾下产生平滑标度。实心圆示意模态激活。

9.2 谱尾如何给出神经标度律（neural scaling laws）指数

设模态增长率与目标强度分别满足 $r_k \propto k^{-\alpha_2}$、$V_k \propto k^{-\alpha_1}$，其中 $\alpha_1 > 1$、$\alpha_2 > 0$。在重标度时间（rescaled time）$\tau = t/\log(1/\epsilon)$ 中，已激活的模态序号增长为 $k_*(\tau) \sim \tau^{1/\alpha_2}$。尚未学得的尾部贡献于是满足

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

这直接展示了论文中条件性指数预测的机制。谱的幂律尾部（power-law spectral tail）是输入假设，不能仅凭置换对称性推出。小初始化、平滑性和选定的可解谱条件限定了推导范围；高阶项还可能产生其他平台。较早的逐步自监督学习工作提供了另一条理解这一问题的路线。[29]

10 为什么对称系统能够压缩

与 Hong-Yi Wang、Di Luo 合作的普适压缩理论（universal compression theory），利用带权对象与低阶矩，保留神经元、数据点等无序集合所决定的函数。[17]

10.1 一个精确的有限维思路

对于 $w_i \in \mathbb{R}^m$，定义 $p_j = d^{-1}\sum_i w_i^{\otimes j}$。匹配到 k 阶的矩，能够保留由这些矩决定的对称多项式。在相应的矩匹配（moment matching）构造中，一个正权表示至多需要

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

个支撑点。对于解析函数（analytic function），可控的泰勒余项进一步把有限阶矩匹配转化为近似保证。

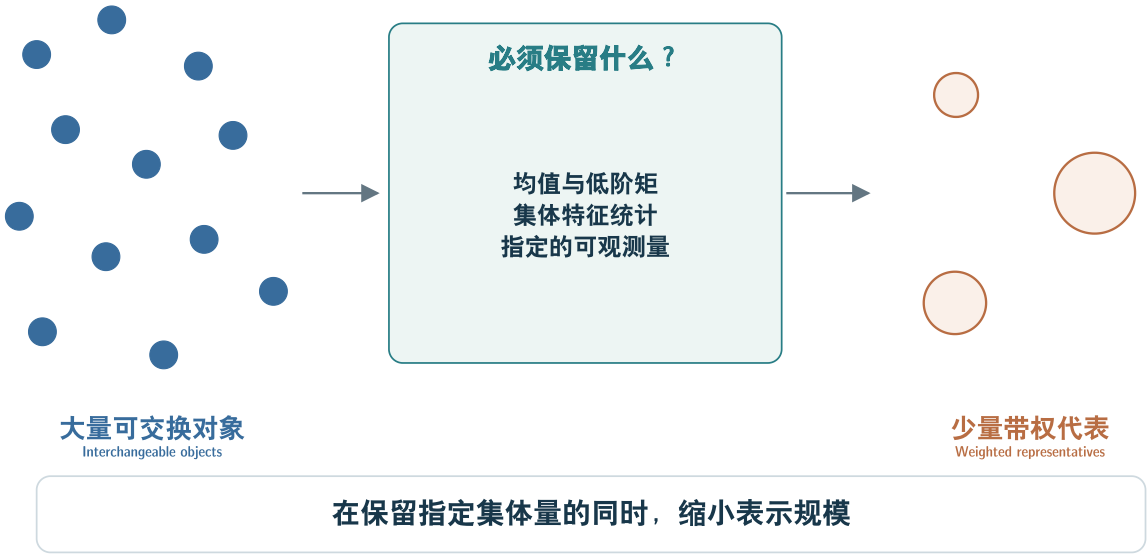


图 10. 把压缩理解为保留集体组织。当相关统计量得以保留时，大量可交换对象可以由少量带权代表（weighted representatives）替代。矩联系微观集合与需要保留的可观测量；图中点的大小是符号示意。

10.2 从密集对象到多对数（polylogarithmic）上界

聚类与矩匹配结合，使小规模带权支撑下的误差渐近消失。代表性上界为 $d' = O([\log(d/\varepsilon)]^m)$，要求对象有界，且具有 $h(\sum_i g(w_i))$ 的解析深集合表示（deep-set representation），其中 h, g 不依赖 d。更高矩阶数改善误差，也增加成本。

10.3 保留学习，还需要保留相应更新

动力学彩票（dynamical lottery tickets）结果将压缩应用于对称的受训可观测量 $f \circ T$，要求训练等变（equivariant training），且复合映射满足相应正则性（regularity）。压缩动力学作用于带权对象，保留对象 w_j 的梯度要按 c_j^{-1} 重缩放。因此，这比剪去部分网络后直接使用不变的优化器更为具体。

原论文在相同小批量选择下，比较宽网络、压缩网络与随机子网络。[17] 用 d' 个有效对象保留 d 个对象的性能，可以改善以保留规模衡量的损失；但压缩构造本身仍有成本。

11 压缩一个固定的深层网络

11.1 匹配输入导数，而非每个隐藏坐标

与 Hong-Yi Wang、Mingze Wang 合作的宽度无关定理（width-independent compression theorem），考虑固定的解析激活深层多层感知机（multilayer perceptron, MLP）及有界输入域。某层贡献为 $y(x) = \sum_i v_i \sigma(w_i^\top f(x))$。对每个输出，到 k 阶的输入导数向量只有

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

个坐标。秩约化（rank reduction）过程保留部分原神经元，重新计算输出权重以匹配这些导数。解析性通过 $(R/\rho_0)^{k+1}$ 的几何因子控制余项，其中 $R < \rho_0$ 联系了实输入域和复解析半径。[18]

从后向前压缩各层，可控制误差的逐层传播，构造深度相同的近似网络。对于固定教师网络（teacher network），宽度上界为 $O((\log(1/\varepsilon))^{d_{\text{in}}})$。教师网络的解析性与范数常数仍保留在界内；宽度无关，不代表高有效输入维数的代价也消失。



图 11. 三种保留层次。[17, 18] 函数近似、目标函数压缩和相对应的学习动力学，提出不同的问题。函数的紧凑表示，不会自动保留不变优化器的轨迹。三列组织了这两页讨论的不同保证。

11.2 两类保证应当分清

这项结果近似一个固定深层教师的输入输出函数。上一页的动力学结果则针对正则的对称受训可观测量，以及相应的压缩更新。固定教师定理明确没有建立动力学彩票保证。即使最终表示很紧凑，构造性证明对应的计算也可能昂贵。

11.3 标度律也存在不变量

与 Xing Han、Suchi Saria、Paul Pu Liang 的合作追问：在一个领域拟合的标度律，能否迁移到另一个领域？它考察信息保持变换（information-preserving transformation）与信息分辨率（information resolution）变化，在语言、视觉和语音场景检验。[19] 连同 NQF 与压缩理论，这区分了学习曲线怎样产生、哪些变换保持标度，以及给定性能需要保留多少对象。

12 回路如何组装出自己的学习规则

与 Qianli Liao、Yulu Gan 等合作的 PNAS 论文追问，系统能否形成改变前向权重所需的回路？自组装基元（Self-Assembling Motif, SAM）通过可塑的交叉连接耦合上行与下行流，允许两者宽度不同。[20]

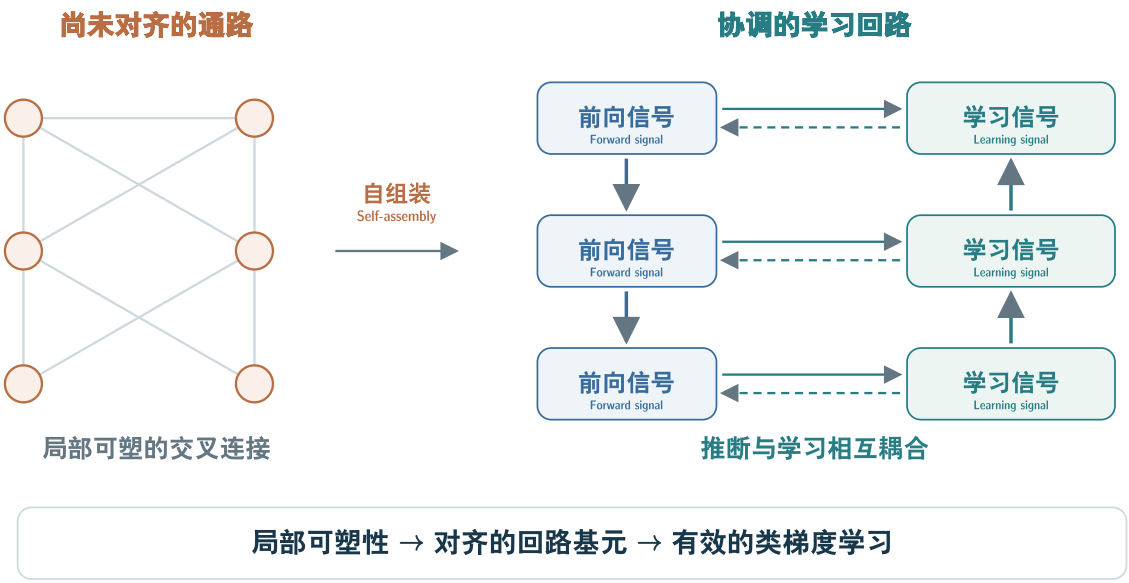


图 12. 学习机制作为涌现的组织。[20] 从已有的前向与反馈通路出发，局部可塑性协调其连接，在对齐条件下形成支持类梯度学习的回路基元。此图展示组织层次，并非实测皮层连接图。

12.1 哪些结构会自组装，定理又要求什么？

模型有四类可塑突触（plastic synapses）：两条流内部，以及流之间的两个方向。局部规则结合突触前活动（presynaptic activity）、其他突触送达的信号与衰减。匹配规则使上下行流之间的对齐（interstream alignment）误差 $E_v = V - \bar{V}^{\mathsf{T}}$ 趋向零；通路对齐（pathway alignment）误差 E_p 则比较前向 Jacobian 与反馈路径。当通路误差较小，并满足秩与边界条件时，

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

余项 R^{ℓ} 由对齐误差控制。在对齐极限，回路实现由学得矩阵预条件的梯度更新（matrix-preconditioned gradient update）。匹配条件下，流之间的对齐已有证明；通路对齐在成功运行中的改善另有观察，其对任意初始化的收敛仍是开放问题。

原论文部分结果，准确率（%）	SGD	自组装回路
MNIST	98.44	98.85
CIFAR-10	50.36	51.91

取自论文表 1，各实验进行五次运行。这些数值对应其具体模型配置及额外的下行学习机制，并非通用图像分类性能纪录。

13 局部可塑性：机制、现象与时间

13.1 异突触计算是一种设计原理

异突触梯度机器（heterosynaptic gradient machine）预印本研究，局部相互作用如何实现广泛的梯度元学习（gradient-based meta-learning）规则。[21] 突触同时接收前向驱动，以及与其他突触有关的调制信号；二者相互作用，可使突触对有效学习梯度敏感，而不必让每个突触执行显式编程的全局导数计算。这是一类理论构造，也是一种提出的生物设计原理。

13.2 为什么像 Hebb 的更新不一定来自 Hebb 规则？

ICML 研究将学习信号（learning signal）U 与收缩项分开：$\Delta W = U - \gamma W$。接近均值平稳时，$\mathbb{E}[U] \approx \gamma W$。对 $h_b = W h_a$，Hebb 矩阵为 $\mathcal{H} = \mathbb{E}[h_b h_a^\top] = W H_a$，且 $H_a \succeq 0$。于是

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

这个简单等式说明，正则化可以使相当不同的更新机制，其学习信号分量都呈现 Hebb 型特征（Hebbian signature）；噪声则可能推动反 Hebb 对齐（anti-Hebbian alignment）。平稳时的总更新，与其中的学习信号分量，是不同的可观测量。[22]

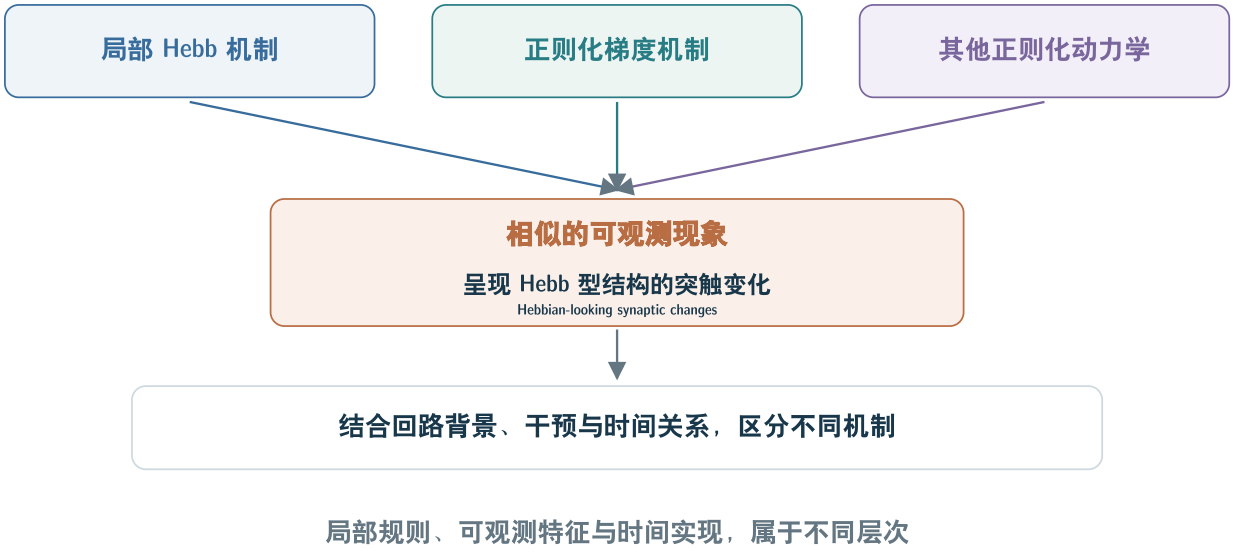


图 13. 同一种可观测特征，可以来自多种机制。[22] 局部 Hebb 计算和不同的正则化动力学，可以产生相似的可塑性模式。干预、回路背景和时间关系，为辨识底层机制提供额外信息。

13.3 学习规则还必须在恰当的时间工作

连续时间反馈对齐（feedback alignment, FA）研究，把传播、可塑性与衰减写成耦合过程，并研究 $\tau_{\text{prop}} \ll \tau_{\text{plas}} \ll \tau_{\text{dec}}$ 的有效时间尺度层次。权重更新中有信息的部分取决于重叠：

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

这里 V 为反馈矩阵，$\varepsilon(t)$ 为误差信号，$z(t)$ 为突触前活动，$k(t)$ 为可塑性核（plasticity kernel）。延迟改变时间上的匹配：模型容忍适度错位，但重叠消失时学习失败。[23] 这联系了局部实现、可塑性现象与时间要求。

14 从连接组理解相对运动计算

2026 年的图形–背景辨别（figure–ground discrimination）预印本，将果蝇连接组与细致的生物物理建模（biophysical modeling）结合，研究 Reichardt–Poggio–Hausen 架构的一种候选实现：大视野运动汇集单元（wide-field pool）调制局部运动检测器，从而帮助区分物体运动与背景运动。提出的回路涉及 T4a 和腹侧离心水平神经元（ventral centrifugal horizontal neuron, vCH）。[24]

14.1 不同区室可以完成不同计算

论文报告的模型区分了 T4a 树突（dendrite）的运动响应，以及轴突末梢（axon terminal）对相对运动（relative motion）敏感的信号。vCH 汇集大范围运动，并向 T4a 末梢施加突触前分流抑制（presynaptic shunting inhibition）；运动拮抗（motion opponency）和双眼通路（binocular pathways）进一步扩展了这一解释。由此可提出更一般的组织层次理解：突触作用的位置，会影响细胞如何共同完成计算。

14.2 为什么分流能够归一化信号？

一个一般的被动区室（passive compartment）揭示了生物物理逻辑。C_m 为膜电容（membrane capacitance），V 为膜电位，g 与 E 分别为电导（conductance）和反转电位（reversal potential）；下标 L、E、I 对应漏电、兴奋和抑制：

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

平稳时，$V = (g_L E_L + g_E E_E + g_I E_I)/(g_L + g_E + g_I)$。若 $E_I = E_L$，则

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

因此，反转电位接近静息电位的抑制电导，可以通过改变分母来减弱兴奋响应；不能总将其等同于减去一个固定电压。

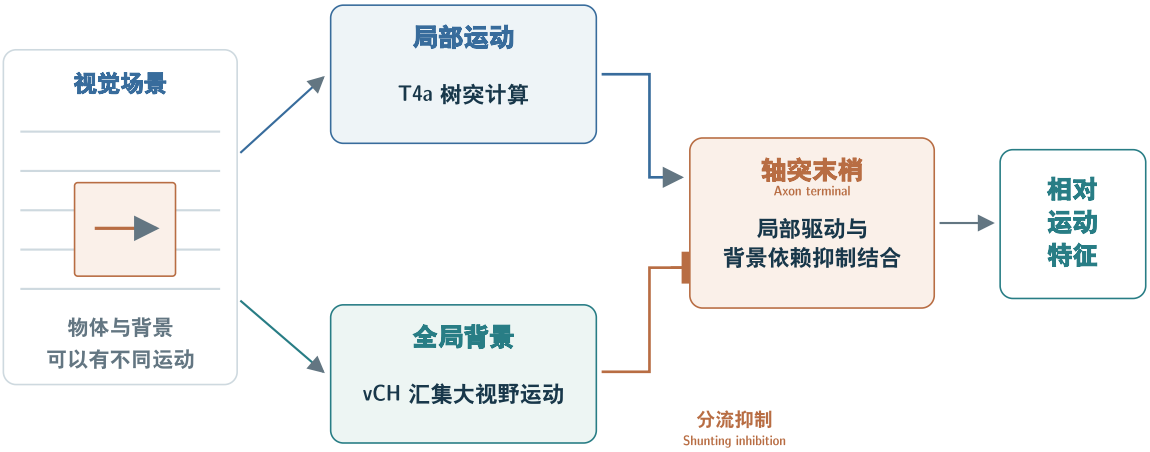


图 14. 空间回路中的分工。在所提模型中，局部运动与大视野背景在 T4a 轴突末梢结合。区室特异的抑制改变向下游传递的计算。此图展示组织关系。

14.3 连接组为解释贡献了什么？

解剖结构约束可能的相互作用，生物物理性质决定其动力学后果。模型与已有观察的一致性，支持所提出的机制。如果干预能够将其预测与替代解释区分开，就会进一步增强这一解释。

来源范围：本节关于该论文的具体陈述依据作者公开的 bioRxiv 摘要。被动区室推导用于解释机制，并非复现论文的定量模型。

15 当前研究的局限与边界

这些成果最有力的地方，是分离具体的现象，找出相关变量，并推导其动力学。接下来的难点，是预测在一个新的网络或生物回路中，究竟哪种机制占主导。从解释某个机制，到建立广泛可靠的定量理论，中间仍有几类明确的边界。

15.1 从可解条件走向真实训练

精确结论依赖二次损失或受控噪声、平均场条件、小权重展开，以及特殊的谱结构。真实训练则包含归一化、特征演化与优化器状态，有效描述的失效应当变得可预测。[1, 11, 16] 全局相边界描述优化后的值，而有限训练路径仍可能停留在亚稳态。同样，标量不可逆势依赖更新场假设，非平稳过程也需要超出稳态噪声模型的描述。[27, 10]

15.2 从内部组织走向泛化能力

平衡、对齐、低秩与拓扑刻画内部组织，而它们与测试性能的关系，还取决于任务需要哪些信息。跨模态对齐可能强调共享信息，同时丢失有用的模态独有信息；即使输入因子解耦，后续计算仍可能无法组合泛化。[13, 14] 完整理论需要预测，某个结构统计量何时改善迁移、鲁棒性（robustness）或外推，何时只是伴随现象。能够区分这些可能性的干预，比再观察一组相关性更有解释力。

15.3 从紧凑表示走向实际资源节省

小表示可以存在，但寻找它仍可能昂贵。矩匹配与输入导数构造涉及维数、近似阶数、依赖教师网络的常数，以及误差随深度的传播。固定教师结果虽然没有额外显式的宽度因子，宽度仍可能通过这些常数间接进入。[17, 18] 因此，实用压缩必须同时计算构造时间、内存、校准（calibration）及后续训练成本；近似函数与保留学习动力学，仍是具有不同成本和假设的目标。

15.4 从生物合理规则走向可辨识的真实机制

SAM 在匹配条件下证明了流之间的对齐，而梯度对应所需的通路对齐仍是另一项条件。连续时间模型还指出，正确的活动与反馈信号必须在突触处具有时间重叠。[20, 23] 建立生物实现，需要识别这些信号、它们如何到达可塑突触，以及相关对齐何时形成。仅凭 Hebb 型变化，不能唯一确定底层规则。[22, 24] 同样，连接组缩小了布线范围，却不能独自确定电导、状态与回路动力学；要区分候选机制，需要能够使其预测分离的响应或干预。

15.5 从相关机制走向整合理论

参数对称性、特征算子、随机路径分布与回路基元之间的定量桥梁，仍未完整。连接它们，需要明确变量与可观测后果的对应。将连接组模型拓展到鱼类和哺乳动物，还需要处理重建不确定性、未配准的生理，以及成像体积之外的输入。[34, 35] 有用的统一理论，必须预测架构、任务或回路组织改变的后果，并明确适用边界。

16 未来研究：五个人工智能问题

未来十个问题，在人工智能与神经科学之间各占一半。每个问题都配以动机，以及能够区分机制的研究方法。这些是由前述成果与公开数据资源支持的未来方向。

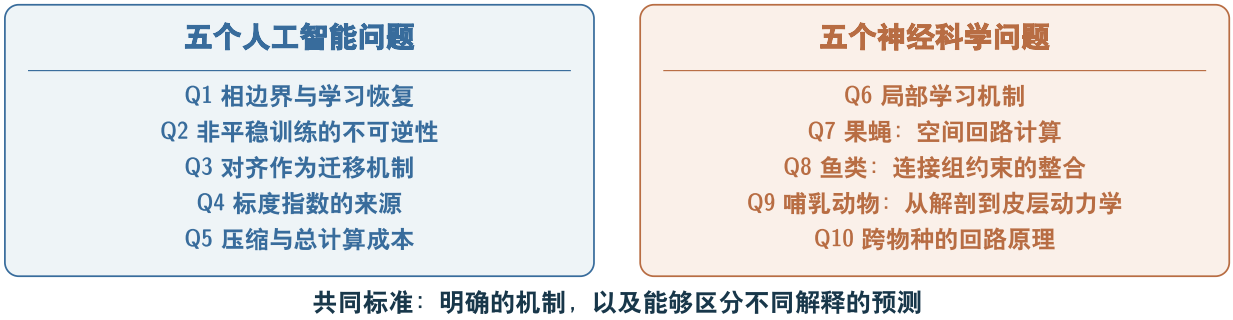


图 15. 均衡的研究议程。人工智能与神经科学各五题。神经科学部分从局部学习、果蝇回路、鱼类、哺乳动物与跨物种比较，提供互补的解释层次。

Q1. 哪些相边界能够预测现代网络的坍缩与学习恢复？可解理论将深度和正则化连接到学得状态与坍缩状态的竞争。可以在带归一化的残差模块（residual block）或注意力模块中定义序参量，再分别改变衰减、噪声与初始化，预测相边界及两种状态的稳定性，并区分全局最优解切换、优化器困陷和可塑性丧失。[27, 32, 4, 5]

Q2. 哪些不可逆量能够约束带动量（momentum）或 Adam 的非平稳训练？有效势可能遗漏概率流（probability current）与不断变化的状态分布。可以纳入优化器状态，定义正向和反向过程，再对指定的训练可观测量推导不确定界。可解模型应当区分，何时一个标量修正已经足够，何时非平稳性会改变预测。[33, 10, 6]

Q3. 哪些表征对齐关系，确实会因果性地改善迁移？对齐可能保留共享特征，也可能丢弃有用的独有信息。可以在受控隐变量任务中，保持原任务误差相当，独立改变对齐与模态独有信息，再预测新任务或未见组合的表现，区分迁移机制与仅仅伴随训练出现的统计量。[11, 13, 14]

Q4. 能否不预设幂律谱，就推导出标度指数？ NQF 把谱尾与损失指数联系起来；更深的解释需要推导谱本身。可以选定特征算子 $A(x)$ 与数据分布，推导学习谱，在拟合损失曲线之前，同时预测一个指数，以及一种不应出现幂律的情形。[16, 19]

Q5. 保持动力学的压缩（dynamical compression），何时真正节省计算？需要把轨迹保留保证，与完整计算构造成本的算法结合起来：确定数据或表征的哪种维数控制压缩界，在指定时间范围内追踪误差，并计入构造、内存与后续训练。这将存在性定理，连接到整个压缩流程确实更便宜的条件范围。[17, 18]

17 未来研究：五个神经科学问题

Q6. 哪些局部可塑性机制能够产生协调的类梯度学习？在双流框架内，寻找能在有限反馈延迟下，同时建立流之间对齐与通路对齐的规则和初始条件。再分别改变共同激活、调制反馈与衰减，预测相应响应。目标是将真正发挥作用的局部学习机制，与不同规则都可能产生的 Hebb 型相关性区分开。[20, 21, 23, 22]

Q7. 果蝇回路中的哪种空间抑制机制解释相对运动计算？在 T4a–vCH 候选模型中，遵守解剖约束，分别改变抑制位置与电导强度，在留出的视觉背景和定向干预上比较预测。还需要追踪突触强度与膜参数的不确定性，使预测效应能够归因于区室特异的机制，而非某一组拟合参数。[24]

Q8. 鱼类连接组中的哪些约束，使感觉整合能跨条件保持稳健？ Fish1 等幼体斑马鱼资源，为受连接组约束的建模提供起点。可以选择一个后脑回路，研究递归兴奋（recurrent excitation）、对侧抑制（contralateral inhibition）与细胞时间常数，如何共同决定整合与选择。用解剖约束结构，在一种条件下拟合其余生理参数，再预测另一种条件的响应；比较机制时，需要传播重建与细胞类型的不确定性。[34]

Q9. 受解剖约束的动力学，能够预测多少哺乳动物皮层功能？利用 MICrONS 小鼠视觉皮层等配准的解剖与活动数据，将连接组约束模型，与匹配度数、距离和细胞类型的空模型（null model）比较。用训练响应拟合不确定电导，再预测留出刺激与背景效应，同时显式建模跨越成像体积的输入。关键检验是：哪些解剖约束，能够在仅拟合功能之外提高预测力？[35]

Q10. 哪些有效回路变量，能够在果蝇、鱼类与哺乳动物之间迁移？对于匹配的视觉计算，可以用有效时间常数（effective time constant）、递归增益（recurrent gain）与兴奋–抑制组织，比较解剖不同的回路，检验约化模型能否跨物种预测同类干预响应。需要分开共享的计算原理与物种特异的实现，并用留出响应检验约化，避免将形态相似直接视为功能等价。[24, 34, 35]

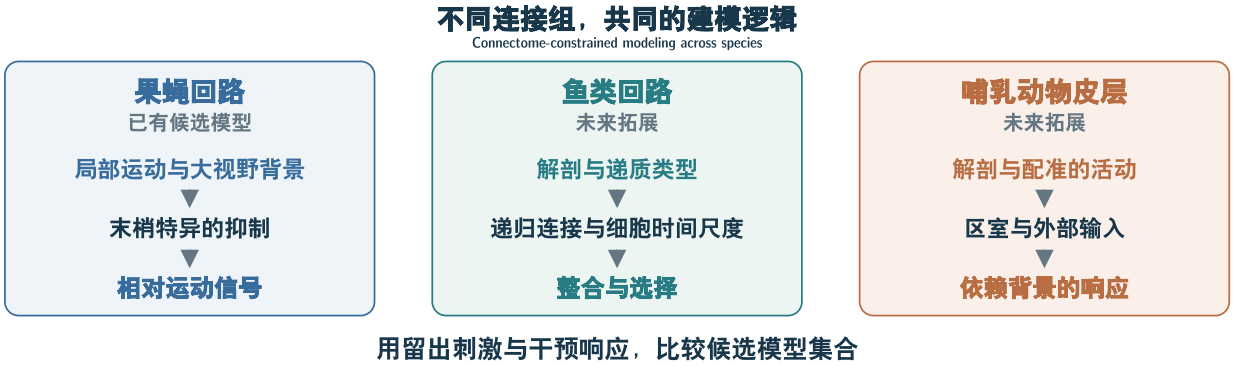


图 16. 从果蝇回路到脊椎动物建模问题。不同物种提供不同的解剖与生理约束，共同的检验方式是利用留出响应与干预比较候选动力学。鱼类与哺乳动物建模属于未来拓展，Fish1 与 MICrONS 是外部资源。

参考文献

公开状态截至 2026 年 10 月 22 日。作者顺序及贡献标记依据当前网站与原始论文；* 表示共同贡献或通讯作者，具体以各论文为准。

[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.

图示说明：十六幅图均为依据所引研究原创的概念与组织关系图，展示问题、机制、理论框架及神经组织层次之间的联系。图形几何与配色用于解释，不表示定量数据。可编辑的中英文 SVG 随源文件提供。