DISCOVERING COMMON CONSERVATION LAWS ACROSS GRADIENT-FLOW ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Common conservation laws remain invariant across a family of neural-network training problems. Tangency on finitely sampled fields is ambiguous: nearly parallel fields create false near-invariants, while functionally independent candidate rows can exceed the normal-space dimension available to common laws. We formulate this finite identification problem as *discover–certify–preserve*. Multi-environment neural deflation (MEND) jointly learns a vector-valued candidate from training fields. A held-out certificate then tests field tangency, pointwise rank, first-bracket compatibility, and codimension capacity on frozen candidates. Conditional matrix bounds connect residuals to normal-space error when candidate and direction singular margins, derivative accuracy, and support coverage are resolved. On shallow-network controls, the full certificate resolves all 40 predeclared boundary/null decisions on fresh tasks; matched stages show that brackets reject task-local candidates and capacity rejects the remaining overcounts. Preservation is a separate deployment question. The learned-law Conservation-Constrained Adam evaluation fails its formal utility gate because of pathwise conditioning events, whereas supplied exact attention laws are preserved at Adam-comparable loss. The evidence supports a finite, boundary-conditioned certification framework. Blind law-count recovery, nonquadratic discovery, and robust learned-law optimizer utility remain open evaluation problems.

1 INTRODUCTION

Conservation laws expose the geometry of neural-network training and provide constraints for optimization. A useful law remains constant for a declared family of losses that share an architecture but differ in data or targets. Slow variation along one observed trajectory supplies only a task-local clue. Given finitely many parameter points and loss environments, the central question is therefore: when does a learned candidate provide defensible evidence for a *common* conservation law?

For Euclidean gradient flow,

$$\dot{\boldsymbol{\theta}}(t) = -\nabla L(\boldsymbol{\theta}(t)), \qquad \frac{\mathrm{d}}{\mathrm{d}t} C(\boldsymbol{\theta}(t)) = -\nabla C(\boldsymbol{\theta}(t))^{\top} \nabla L(\boldsymbol{\theta}(t)), \tag{1}$$

so a law is represented locally by the normal vector ∇C. In the factorized model

$$L_y(u, v) = \tfrac{1}{2}(uv - y)^2, \qquad C(u, v) = u^2 - v^2, \tag{2}$$

the trajectory follows a hyperbola $C = \text{constant}$. Any regular reparameterization $h \circ C$ describes the same foliation. The normal space is therefore the identifiable object. It supports both a geometric description of training and a projected optimizer that preserves the learned quantities.

For a fixed target y, all zero-loss factorizations satisfy $uv = y$. An initialization with $u(0)^2 - v(0)^2 = c$ restricts any zero-loss limit to the intersection of this solution curve with $u^2 - v^2 = c$. The conserved value thus distinguishes parameter configurations that realize the same predictor. Changing the target moves the solution curve while retaining the law, since the two contributions to $\dot{C}$ cancel for every y. This is the form of shared structure sought here: a function of the parameters whose level sets constrain training across the declared loss family.

Small tangency error admits several explanations. Near-parallel fields can make a false candidate appear invariant, while a globally well-conditioned collection of candidate functions can still be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

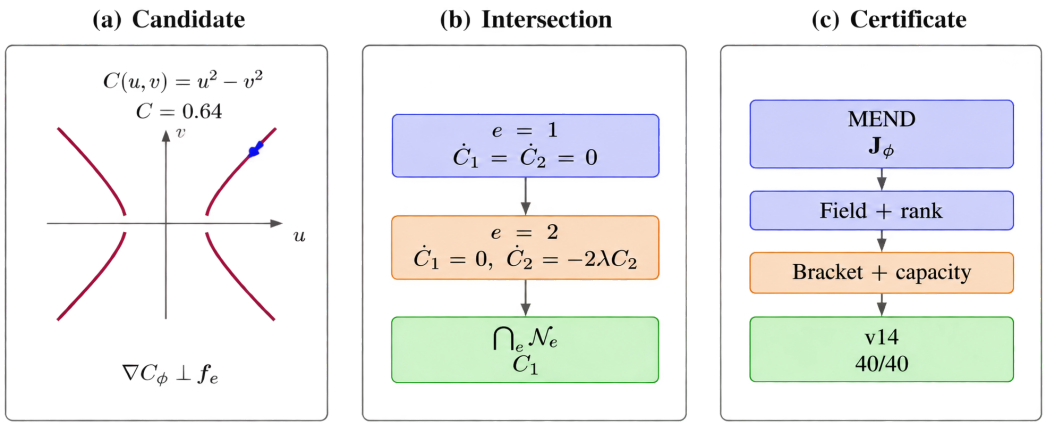


Figure 1: The finite common-law identification problem. (a) For $L_y(u,v) = \frac{1}{2}(uv - y)^2$, gradient flow follows a level set of $C = u^2 - v^2$; MEND estimates this leaf geometry rather than a preferred scalar formula. (b) Let $L_1 = \frac{1}{2}\sum_{j=1}^{2}(u_j v_j - y_j)^2$ and $C_j = u_j^2 - v_j^2$. Both candidates are conserved by L_1, but $L_2 = L_1 + \frac{\lambda}{2}(u_2^2 + v_2^2)$ preserves only C_1. The second environment separates the shared law from the task-local law. (c) MEND proposes a candidate normal space; the held-out certificate tests local tangency and independence, then differential compatibility and available codimension. The frozen v14 count is empirical; panels (a)–(b) are analytic.

pointwise redundant. More fundamentally, the flow-box theorem supplies $d - 1$ local first integrals near any regular point of a single nonvanishing field (Arnold, 1989). A flexible network can therefore learn a task-local coordinate that is exact for one loss and changes meaning across losses. Figure 1b gives a two-channel example in which one environment preserves two quantities and a second environment preserves only their intersection.

Analytic work derives architecture-dependent laws from parameter symmetries and Lie algebras (Du et al., 2018; Kunin et al., 2021; Zhao et al., 2023; Marcotte et al., 2023; 2024; 2025; Tran et al., 2026). In particular, Marcotte et al. (2023) characterize common laws through the Lie algebra generated by the loss fields. Finite inference additionally requires a decision rule for fitted candidates, sampled environments, and numerical tolerances. Data-driven invariant discovery typically studies one physical system using trajectories, derivatives, or a fixed function dictionary (Ha and Jeong, 2021; Liu et al., 2022; Lu et al., 2023; Liu et al., 2024). Neural Deflation learns scalar integrals sequentially for a Hamiltonian field, using Poisson involution and gradient-space deflation to suppress duplicates (Zhu et al., 2023; Chen et al., 2025). Gradient flow instead calls for cross-environment compatibility and a pointwise rank test. Sparse or neural-symbolic generators could nevertheless provide candidates to the certificate developed here (Liu et al., 2024; Ray, 2026). A matched comparison among candidate generators remains a separate empirical question.

Our workflow separates discovery from certification. Multi-environment neural deflation (MEND) jointly fits a vector-valued candidate across training losses. A held-out certificate then tests four nested conditions on the same frozen rows and outer tensors: field tangency and noncollapse, pointwise row rank, first-bracket compatibility, and the codimension left by the observed field–bracket distribution. The rank test removes duplicate coordinates; brackets test differential compatibility across environments; capacity prevents more independent candidates than the observed geometry can accommodate. Conditional matrix bounds show why a small residual becomes informative only when both the candidate rows and generated directions have resolved singular margins and the validation support covers the region of interest.

Preservation is a separate decision. Conservation-Constrained Adam (CCA) projects an Adam direction into the null space of a supplied constraint Jacobian. Riemannian adaptive methods use a manifold atlas and retraction (Kasai et al., 2019); symmetry teleportation moves along a known symmetry orbit (Zhao et al., 2022; 2024); and AdamP applies a per-tensor scale correction (Heo et al., 2021). CCA directly encodes Jacobian rows that couple several tensors. Deployment additionally

requires the learned Jacobian to remain conditioned along the optimizer path, so the deployment gate is reported independently.

We evaluate frozen shallow-network boundary and null controls. The complete certificate makes all 40 declared decisions on fresh tasks; matched component changes show that brackets remove the task-local null candidates and capacity removes the remaining positive overcounts. Accepted positive candidates align with the analytic normal spaces in a post-freeze audit. A later learned-law CCA evaluation crosses its operational conditioning safeguard. With supplied exact attention laws, projection reduces conservation drift at comparable loss and isolates the preservation mechanism. Together these results establish a finite, boundary-conditioned certificate and an explicit separation between *discover*, *certify*, and *preserve*. Future evaluations can extend this framework to blind law-count recovery, global completeness, estimator comparisons, and learned-law optimizer utility.

2 DISCOVER, CERTIFY, AND PRESERVE

Let $\mathcal{P} \subseteq \mathbb{R}^d$ be open and let $\mathcal{E}$ index loss environments sharing an architecture. Environment e induces

$$\dot{\boldsymbol{\theta}} = \mathbf{f}_e(\boldsymbol{\theta}) := -\nabla_{\boldsymbol{\theta}} L_e(\boldsymbol{\theta}). \tag{3}$$

For $C_{\boldsymbol{\phi}} \colon \mathcal{P} \to \mathbb{R}^K$, write $\mathbf{J}_{\boldsymbol{\phi}} := \nabla_{\boldsymbol{\theta}} C_{\boldsymbol{\phi}} \in \mathbb{R}^{K \times d}$. Its components are common first integrals on an open region $\mathcal{O}$ when

$$\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\mathbf{f}_e(\boldsymbol{\theta}) = \mathbf{0}_K \quad \text{for all } \boldsymbol{\theta} \in \mathcal{O} \text{ and } e \in \mathcal{E}. \tag{4}$$

Finite data support only a distribution-conditioned analogue of this identity. We therefore call an estimator output a *candidate* until it passes the prespecified outer certificate, and even a pass refers only to the declared support and environment family. Discovery uses field evaluations; certification also requires field derivatives. For twice differentiable losses,

$$[\mathbf{f}_e, \mathbf{f}_{e'}] = \nabla^2 L_{e'} \nabla L_e - \nabla^2 L_e \nabla L_{e'}, \tag{5}$$

so Hessian–vector access or an equivalent differentiable loss evaluator is required.

2.1 DISCOVER: A JOINT CANDIDATE NORMAL SPACE

For each declared K, MEND fits $C_{\boldsymbol{\phi}} = q_{\boldsymbol{\beta}} + r_{\boldsymbol{\psi}}$, where $q_{\boldsymbol{\beta}}$ is a generic degree-two map in all parameter coordinates and $r_{\boldsymbol{\psi}}$ is an MLP with a zero-initialized final readout. The map takes parameter coordinates as input; sampled fields constrain its Jacobian during fitting. Analytic formulas are reserved for the post-freeze audit. Each training observation pairs a point $\boldsymbol{\theta}$ with the fields of several losses evaluated at that same point. The candidate values are learned without conservation-value labels. Sharing one map across these losses makes its level sets a common geometric representation of their dynamics. For nonzero fields, define $\widehat{\mathbf{f}}_e = \mathbf{f}_e / \|\mathbf{f}_e\|_2$ and

$$\ell_{e,k}(\boldsymbol{\theta}; \boldsymbol{\phi}) := \big([\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\widehat{\mathbf{f}}_e(\boldsymbol{\theta})]_k\big)^2, \qquad \mathbf{G}^{\mathrm{tr}}_{\boldsymbol{\phi}} := \mathbb{E}_{\mu_{\mathrm{tr}}}[\mathbf{J}_{\boldsymbol{\phi}} \mathbf{J}_{\boldsymbol{\phi}}^{\top}]. \tag{6}$$

The training objective is

$$\begin{aligned} L_{\mathrm{MEND}}(\boldsymbol{\phi}) = (1-\omega)\mathbb{E}_{\boldsymbol{\theta},e,k}[\ell_{e,k}] + \omega \mathbb{E}_{\boldsymbol{\theta},k}\big[\max_{e \in \mathcal{E}_{\mathrm{tr}}} \ell_{e,k}\big] \\ + \frac{\lambda_{\mathrm{frame}}}{K^2} \|\mathbf{G}^{\mathrm{tr}}_{\boldsymbol{\phi}} - \mathbf{I}_K\|_F^2. \end{aligned} \tag{7}$$

The normalized residual prevents field magnitude from dominating; the worst-environment term prevents an easy loss from masking another; and the frame term controls global scale and discourages redundant rows. The controlled suite uses $\omega = 1$. The held-out certificate determines the supported law count, while pointwise rank resolves candidate dependence that a global frame can conceal.

The output coordinates have a natural equivalence. For a smooth local reparameterization $h \colon \mathbb{R}^K \to \mathbb{R}^K$ with invertible derivative,

$$\mathbf{J}_{h \circ C}(\boldsymbol{\theta}) = \mathbf{D}h(C(\boldsymbol{\theta}))\mathbf{J}_C(\boldsymbol{\theta}), \qquad \ker \mathbf{J}_{h \circ C}(\boldsymbol{\theta}) = \ker \mathbf{J}_C(\boldsymbol{\theta}). \tag{8}$$

Thus several fitted coordinate systems can describe the same local leaves. The post-freeze analytic comparison evaluates their normal spaces, while the rank and conditioning checks concern the particular coordinates returned by the estimator. This distinction separates geometric recovery from the numerical quality of a representation used in optimization.

2.2 CERTIFY: LOCAL RANK, LIE COMPATIBILITY, AND CAPACITY

At each held-out point $\boldsymbol{\theta}_b$, let a_{bk} be the norm of row k of $\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_b)$ and normalize that row by $\max(a_{bk}, \epsilon)$. If the resulting matrix has singular values $\sigma_{b,1} \geq \cdots \geq \sigma_{b,K}$, define

$$m_{\mathrm{row}} := \min_{b,k} a_{bk}, \qquad \widehat{p}_{\mathrm{full}} := \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{\sigma_{b,K} > \tau_{\mathrm{rank}} \max(\sigma_{b,1}, \epsilon)\}\,. \tag{9}$$

Noncollapse excludes nearly constant rows; pointwise rank excludes candidates such as C and C^3 that can appear independent after global whitening. For example, take $C_1(x, y) = x$ and $C_2(x, y) = x^3$. For an empirical support measure μ, their joint Jacobian and global gradient Gram matrix satisfy

$$\mathbf{J}_C(x, y) = \begin{pmatrix} 1 & 0 \\ 3x^2 & 0 \end{pmatrix}, \qquad \det \mathbb{E}_\mu[\mathbf{J}_C \mathbf{J}_C^\top] = 9 \operatorname{Var}_\mu(x^2). \tag{10}$$

Whenever x^2 varies over the support, the global matrix has full rank even though the pointwise Jacobian has rank one everywhere. Whitening this global matrix preserves that pointwise dependence. Both outputs encode the same local constraint, and the second output adds a coordinate rather than a law.

The first-order generated distribution is

$$\mathcal{D}^{(1)}_{\boldsymbol{\theta}} := \operatorname{span}(\{\mathbf{f}_e(\boldsymbol{\theta})\}_{e \in \mathcal{E}} \cup \{[\mathbf{f}_e, \mathbf{f}_{e'}](\boldsymbol{\theta})\}_{e < e'})\,. \tag{11}$$

A necessary condition connects the finite screen to the population geometry (Marcotte et al., 2023). The bracket records the leading difference between applying two small flow steps in opposite orders. A common level set must accommodate this displacement as well as each individual field. Bracket tangency therefore checks how the fields fit together as the parameters vary, using differential information beyond their directions at an isolated point.

Theorem 1 (Lie-codimension principle)**.** *Let $C\colon \mathcal{O} \to \mathbb{R}^K$ and $\{\mathbf{f}_e\}$ be smooth, suppose $\operatorname{rank} \mathbf{J}_C = K$ on $\mathcal{O}$, and assume $\mathbf{J}_C \mathbf{f}_e = \mathbf{0}_K$ for every e. Then $\mathbf{J}_C$ annihilates the full Lie distribution. Consequently,*

$$K + \operatorname{rank} \mathcal{D}^{(\infty)}_{\mathcal{E}}(\boldsymbol{\theta}) \leq d, \qquad K + \operatorname{rank} \mathcal{D}^{(1)}_{\boldsymbol{\theta}} \leq d. \tag{12}$$

For a held-out direction collection $\mathcal{V}_b$, define the worst row-normalized residual

$$\rho_{\mathcal{V}}(\boldsymbol{\phi}) := \max_k \frac{1}{|\mathcal{B}_{\mathcal{V}}|} \sum_{b \in \mathcal{B}_{\mathcal{V}}} \max_{\mathbf{v} \in \mathcal{V}_b} \left(\frac{[\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_b)]_{k,:}\mathbf{v}}{\max(a_{bk}, \epsilon)\|\mathbf{v}\|_2} \right)^2, \tag{13}$$

using the nonzero directions. Let $\widehat{r}_b^{(1)}$ be the numerical rank of the normalized held-out fields and first brackets, and set

$$\widehat{p}_{\mathrm{cap}}(K) := B^{-1} \sum_{b=1}^{B} \mathbf{1}\Big\{K + \widehat{r}_b^{(1)} \leq d\Big\}\,. \tag{14}$$

The four nested rules use identical frozen candidates and outer tensors:

$$\begin{aligned} \textsc{Field} &:= \{\rho_{\mathrm{fld}} \leq \tau_{\tan}\} \cap \{m_{\mathrm{row}} > \tau_{\mathrm{row}}\}, \\ \textsc{Rank} &:= \textsc{Field} \cap \{\widehat{p}_{\mathrm{full}} \geq \tau_{\mathrm{full}}\}, \\ \textsc{Bracket} &:= \textsc{Rank} \cap \{\rho_{\mathrm{br}} \leq \tau_{\mathrm{br}}\}, \\ \textsc{Full} &:= \textsc{Bracket} \cap \{\widehat{p}_{\mathrm{cap}}(K) \geq \tau_{\mathrm{cap}}\}. \end{aligned} \tag{15}$$

The nominal K enters the capacity test. First brackets cover the first layer of the full Lie closure, while numerical ranks and the raw-direction floor depend on tolerance and loss units. A pass is a finite certificate on the declared support; population completeness follows under the conditions in Appendix B.

With exact ranks and zero residual, full-row-rank tangency already implies the codimension inequality: $\operatorname{row}(\mathbf{J}_C)$ lies in the orthogonal complement of the tested distribution. The explicit capacity screen addresses the finite-tolerance setting. A candidate can satisfy the separate residual and row-rank cutoffs while its nominal dimension exceeds the measured codimension. The capacity test

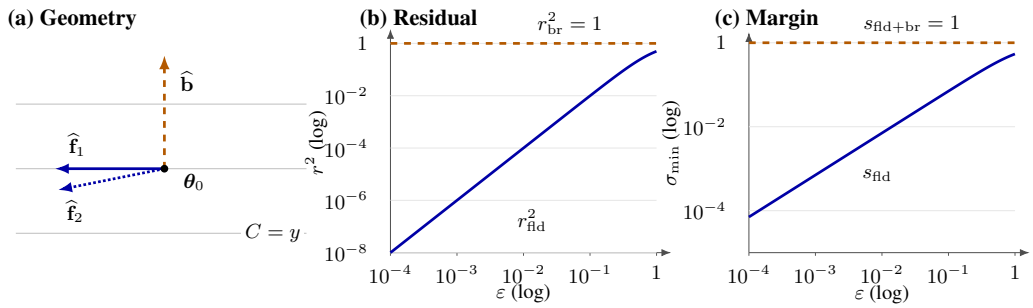


Figure 2: Near tangency can hide both a false candidate and an ill-conditioned direction span. Let $L_1 = x^2/2$, $L_2 = (x + \varepsilon y)^2/2$, $C = y$, and $\boldsymbol{\theta}_0 = (1, 0)$. Write $\widehat{\mathbf{f}}_1$ and $\widehat{\mathbf{f}}_2$ for the unit gradient-flow directions and $\widehat{\mathbf{b}}$ for the normalized bracket direction. (a) The two fields approach the horizontal level set of C, whereas the bracket is transverse. (b) The exact squared scores satisfy $r^2_{\mathrm{fld}} = \varepsilon^2/(1+\varepsilon^2)$ and $r^2_{\mathrm{br}} = 1$. (c) For $\mathbf{B}_{\mathrm{fld}} = [\widehat{\mathbf{f}}_1, \widehat{\mathbf{f}}_2]$ and $\mathbf{B}_{\mathrm{fld+br}} = [\mathbf{B}_{\mathrm{fld}}, \widehat{\mathbf{b}}]$, $s_{\mathrm{fld}} = \sigma_{\min}(\mathbf{B}_{\mathrm{fld}})$ tends to zero while $s_{\mathrm{fld+br}} = 1$. Thus bracket normalization supplies an order-one transverse direction even as the observed fields become nearly parallel. The display covers $\varepsilon > 0$; at $\varepsilon = 0$, the bracket vanishes and C is exact.

checks this dimension relation directly under the frozen numerical-rank convention. Its contribution is thus a consistency check among approximate decisions. The bounds below give the corresponding exact and margin-conditioned geometric statements.

At a fixed point, singular margins quantify when the residual is informative.

Theorem 2 (Residual, capacity, and normal-space error)**.** *Let $\mathbf{A} \in \mathbb{R}^{K\times d}$ have rank K and let $\mathbf{B} \in \mathbb{R}^{d\times m}$ have rank r, with $a = \sigma_K(\mathbf{A}) > 0$ and $b = \sigma_r(\mathbf{B}) > 0$. Then*

$$\|\mathbf{AB}\|_F^2 \geq (K + r - d)_+ a^2 b^2. \tag{16}$$

If $K = d - r$, then

$$\left\|\mathbf{P}_{\mathrm{row}(\mathbf{A})} - \mathbf{P}_{\ker(\mathbf{B}^\top)}\right\|_2 \leq \min\left\{1, \frac{\|\mathbf{AB}\|_2}{ab}\right\}. \tag{17}$$

If $K + r > d$, then $\|\mathbf{AB}\|_2 \geq ab$.

Thus every well-conditioned overcomplete candidate obeys a positive residual lower bound against a well-resolved distribution. A small residual becomes informative when both singular margins are resolved. Taking $\mathbf{A}$ to be the candidate Jacobian and $\mathbf{B}$ to contain the tested directions makes the two margins explicit: a measures separation among candidate rows, and b measures separation among the generated directions. The residual alone mixes geometric agreement with these two conditioning effects. The example below isolates the second effect. For

$$L_1(x, y) = \tfrac{1}{2}x^2, \quad L_2(x, y) = \tfrac{1}{2}(x + \varepsilon y)^2, \quad C(x, y) = y, \tag{18}$$

fix $\varepsilon > 0$ and a connected open set $\mathcal{O}_0 \subset \{(x, y) : x(x + \varepsilon y) \neq 0\}$. The two fields are independent on $\mathcal{O}_0$, so every common law there is constant, yet the normalized field residual at $(1, 0)$ is $\varepsilon^2/(1 + \varepsilon^2)$. Their bracket is transverse and has unit residual at that point. The smallest singular value of the two normalized field directions vanishes with ε, whereas augmenting them by the normalized bracket fixes that margin at one (Fig. 2).

Here the raw bracket magnitude at $(1, 0)$ is ε. Normalization makes its transverse geometry visible, while accurate evaluation still requires resolving a direction of that scale. The raw-direction floor in the finite certificate makes this numerical requirement explicit.

Appendix C proves the bounds and states their perturbation, support-coverage, and frozen-score variants. The v14 experiment reports finite sampled margins. Uniform coverage, exact population rank, and derivative-error bounds remain assumptions for numerical confidence guarantees.

2.3 PRESERVE: THE PATHWISE CONDITION

Let $\mathbf{u}_t$ be the bias-corrected Adam direction and let $\widetilde{\mathbf{J}}_t \in \mathbb{R}^{K\times d}$ be a supplied constraint Jacobian. CCA uses

$$\begin{aligned}\mathbf{P}_t &:= \mathbf{I}_d - \widetilde{\mathbf{J}}_t^\top(\widetilde{\mathbf{J}}_t\widetilde{\mathbf{J}}_t^\top + \delta\mathbf{I}_K)^{-1}\widetilde{\mathbf{J}}_t,\\ \mathbf{d}_t &:= \mathbf{P}_t\mathbf{u}_t, \qquad \boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t - \eta_t\mathbf{d}_t.\end{aligned} \tag{19}$$

With $\delta = 0$ and full row rank, this is orthogonal projection onto $\ker\widetilde{\mathbf{J}}_t$. A positive ridge leaves the first-order residual

$$\widetilde{\mathbf{J}}_t\mathbf{d}_t = \delta(\widetilde{\mathbf{J}}_t\widetilde{\mathbf{J}}_t^\top + \delta\mathbf{I}_K)^{-1}\widetilde{\mathbf{J}}_t\mathbf{u}_t. \tag{20}$$

Theorem 3 (Local discrete conservation error)**.** *Let C_k be a three-times continuously differentiable exact law, use its fresh Jacobian in Eq.* (19)*, and hold $\boldsymbol{\theta}_t, \mathbf{d}_t$ fixed as $\eta_t \to 0$. Then*

$$\begin{aligned}C_k(\boldsymbol{\theta}_{t+1}) - C_k(\boldsymbol{\theta}_t) = &-\eta_t\delta\left[(\mathbf{J}_t\mathbf{J}_t^\top + \delta\mathbf{I}_K)^{-1}\mathbf{J}_t\mathbf{u}_t\right]_k\\ &+ \tfrac{1}{2}\eta_t^2\mathbf{d}_t^\top\nabla^2 C_k(\boldsymbol{\theta}_t)\mathbf{d}_t + O(\eta_t^3).\end{aligned} \tag{21}$$

Zero ridge removes the first-order term for exact, fresh rows. Learned rows require an additional pathwise normal-space and conditioning margin. To see the role of geometric accuracy, let $\mathbf{J}_\star$ be the Jacobian of an exact law map, and let $\widehat{\mathbf{P}}_t$ and $\mathbf{P}_{\star,t}$ be the orthogonal projectors onto the learned and exact normal spaces. For a fresh, full-row-rank learned Jacobian and zero ridge, $\mathbf{d}_t = (\mathbf{I}_d - \widehat{\mathbf{P}}_t)\mathbf{u}_t$, and

$$\|\mathbf{J}_\star(\boldsymbol{\theta}_t)\mathbf{d}_t\|_2 \le \|\mathbf{J}_\star(\boldsymbol{\theta}_t)\|_2\|\widehat{\mathbf{P}}_t - \mathbf{P}_{\star,t}\|_2\|\mathbf{u}_t\|_2. \tag{22}$$

Normal-space error therefore enters the exact-law drift bound linearly in the step size, alongside the quadratic curvature term. Proposition 13 states the resulting bound under explicit smoothness and path-coverage assumptions. The certificate and CCA consequently require accuracy on different supports: sampled outer points for the former, and the evolving optimizer path for the latter. Freshness and conditioning must also hold along that path; stale Jacobians add another residual. Appendix B gives the full update and safeguards. The learned-law gate evaluates these deployment conditions separately.

3 EVIDENCE

The experiments separate the three claims in the workflow. Discovery and certification use bias-free $(2, 3, 2)$ networks with independently sampled 50-example MSE environments. Linear and ReLU models provide positive controls with six and three generically independent laws; GELU and SiLU are null controls only for this homogeneous-rescaling family. MEND sees gradient fields and a generic quadratic–neural candidate map, while analytic formulas are withheld until a post-freeze row-space audit. The decisive v14 protocol fixes ten seeds, disjoint train/development/outer environments, 96 outer support points, all 66 first brackets among 12 outer fields, and the boundary pairs $K = 6, 7$ (linear) and $K = 3, 4$ (ReLU), with $K = 1$ for each null control. All 60 candidates are serialized before outer tensors are constructed. Exact settings, thresholds, hardware, hashes, and the ReLU stratumwise regularity qualification appear in Appendix G.

The positive laws extend the factorized example in Eq. (2). Write the network as $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$, with $\mathbf{W}_1 \in \mathbb{R}^{3\times 2}$ and $\mathbf{W}_2 \in \mathbb{R}^{2\times 3}$. For the linear activation, gradient flow conserves

$$\mathbf{Q}_{\text{lin}} = \mathbf{W}_2^\top\mathbf{W}_2 - \mathbf{W}_1\mathbf{W}_1^\top. \tag{23}$$

The six symmetric entries supply six generically independent scalar laws. For ReLU, the retained balances are its three diagonal entries, $q_i = \|\mathbf{W}_{2,:,i}\|_2^2 - \|\mathbf{W}_{1,i,:}\|_2^2$. Positive rescaling of a hidden unit's incoming weights, paired with inverse rescaling of its outgoing weights, preserves the represented function. Differentiating this symmetry gives cancellation of the two contributions to $\dot{q}_i$ on each differentiability stratum. Proposition 4 proves the conservation and generic independence statements. These controls test recovery of coupled layer balances through gradient-field observations, with the formulas available only to the evaluation audit.

The declared boundary pairs isolate a concrete counting question: can a certificate retain a candidate at the analytic law count and reject an independently fitted candidate with one extra output? Joint success requires both decisions. The null controls test the complementary case in which even one

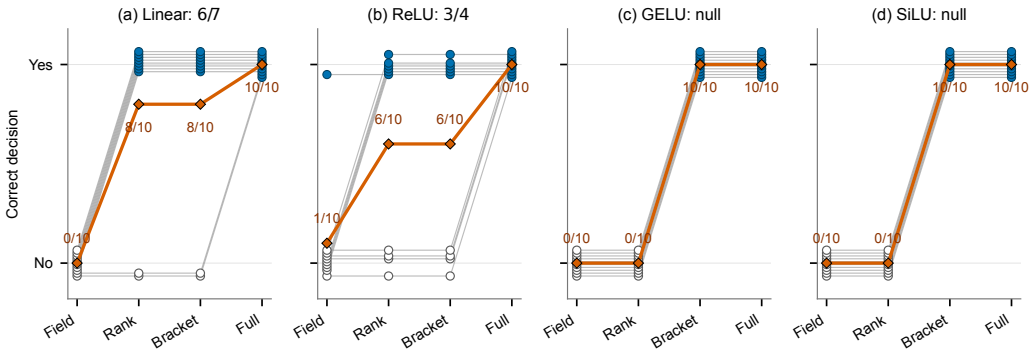


Figure 3: Matched seed trajectories through the certificate staircase. Thin gray paths track the ten frozen seeds; filled and open circles mark correct and incorrect declared decisions. Orange diamonds give the fraction correct, with the count printed at each stage. Linear and ReLU require accepting K_0 and rejecting the independently fitted $K_0 + 1$ candidate; GELU and SiLU require rejecting $K = 1$ for the declared homogeneous-rescaling family. Every stage uses the same candidates and outer tensors. Rank removes some positive overcounts, brackets remove the task-local null candidates, and capacity removes the remaining positive overcounts. The Wilson interval for each 10/10 stage is $[0.72, 1.00]$; the experiment evaluates the declared count boundaries.

Table 1: Frozen v14 component removal on identical candidates and outer tensors. Unnamed conditions remain active. Every row reuses the same fitted candidates and thresholds.

Rule	Linear	ReLU	GELU	SiLU	Total
Full	10/10	10/10	10/10	10/10	40/40
− field	10/10	10/10	10/10	10/10	40/40
− rank	10/10	10/10	10/10	10/10	40/40
− bracket	10/10	10/10	0/10	0/10	20/40
− capacity	8/10	6/10	10/10	10/10	34/40
Capacity only	10/10	10/10	0/10	0/10	20/40

proposed output requires rejection for the specified family. This design separates acceptance at a known positive boundary from rejection of excess or task-local structure. Its information order also separates fitting from evaluation: analytic laws define the control labels and the later span audit, while the certificate acts on the frozen candidates and held-out differential information.

3.1 The certificate resolves the declared boundary

The complete rule alone makes every declared boundary/null decision (Fig. 3). Because each stage reuses the same candidates and outer tensors, the transitions localize the mechanism rather than confounding it with retraining. Pointwise rank corrects some positive overcounts and leaves the null candidates for the next stage. First brackets then reject all task-local GELU/SiLU candidates, as expected when environment variation creates transverse directions. Bracket compatibility admits six independently fitted above-boundary positive candidates; the codimension condition removes those and retains every analytic-boundary candidate.

A retrospective removal analysis on the same saved component flags confirms this division of labor: deleting bracket tangency restores all 20 null errors, whereas deleting capacity restores the six positive overcounts (Table 1). Capacity measures the available codimension, while brackets measure differential compatibility. Removing field tangency or pointwise rank leaves the final v14 decisions unchanged, though separate executable controls show that the rank test rejects a globally whitened pair $C, h(C)$ with pointwise rank one. The ablation describes this frozen cohort. The Wilson 95% interval for a 10/10 family decision is $[0.72, 1.00]$, which gives the corresponding sampling uncertainty.

The method-free premise audit checks that the benchmark quantities are laws before they are used as labels. Figure 4 shows all five matched seeds. Linear and ReLU tangency is near numerical preci-

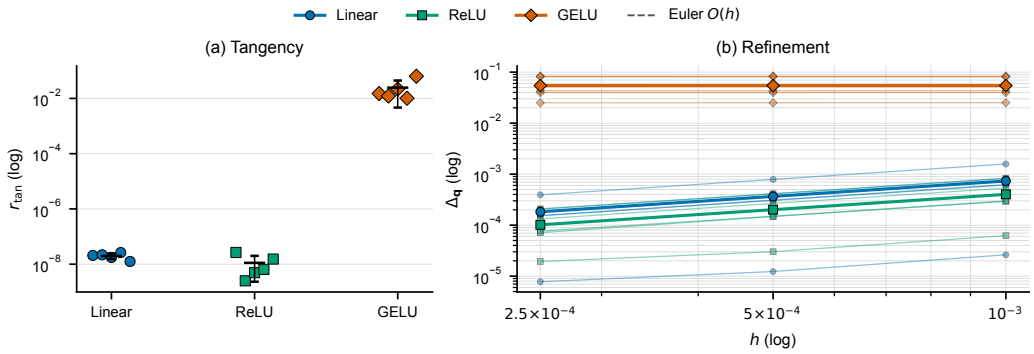


Figure 4: Seed-resolved premise audit for $q_i = \|\mathbf{W}_{2,:,i}\|_2^2 - \|\mathbf{W}_{1,i,:}\|_2^2$. (a) Relative tangency at matched initial conditions; small markers show all five fixed seeds and black summaries show mean $\pm$ one population standard deviation. (b) Relative $\mathbf{q}$ drift at physical horizon 0.5 for all seed trajectories under Euler refinement; thin lines are seeds and thick lines are their means. Both axes are logarithmic where marked. Linear and ReLU follow the expected $O(h)$ global discretization trend, whereas GELU remains approximately constant under refinement. Every seed is shown. This exhibit audits the law premise used by the control suite.

sion and their fixed-horizon drift contracts under Euler refinement, whereas GELU retains nonzero tangency and nonvanishing drift. Positive homogeneity accounts for the difference by cancelling the two layer contributions to $\dot{q}_i$. This audit validates the declared control family. Discovery thresholds and CCA are evaluated in their respective experiments.

The one-environment sequential diagnostic reinforces the same distinction: locally noncollapsed candidates receive zero fresh 12-environment certifications across the five seeds (Table 4). Its candidate class and budget differ from MEND, so the comparison diagnoses task-locality. The post-freeze analytic audit separately confirms that every accepted linear/ReLU boundary candidate recovers the intended normal space; the data-only certificate uses field information exclusively.

3.2 Preservation with supplied and learned rows

The prespecified v15 learned-law deployment gate has formal status FAIL. It consumes all 20 certified positive-control candidates in three new realizable tasks per model and compares five paired optimizer arms. The gate requires every arm and all 20 model-level clusters. On one linear task, both oracle CCA and MEND-CCA cross the frozen relative-rank safeguard; two scrambled-row arms also cross it, and three random-rank arms miss the matched-loss target. These outcomes remain in the record. Complete-cluster drift ratios favor MEND-CCA. The prespecified estimand also includes method-triggered safeguard events, which makes pathwise rank maintenance an additional deployment requirement. Full arm definitions, conditional summaries, and timing ratios appear in Appendix G.

The attention control isolates what projection can do when exact rows are supplied. Its constraints are the cross-tensor laws $q_{\mathrm{QK}} = \|\mathbf{W}_Q\|_F^2 - \|\mathbf{W}_K\|_F^2$ and $q_{\mathrm{VO}} = \|\mathbf{W}_V\|_F^2 - \|\mathbf{W}_O\|_F^2$ (Proposition 16). At an Adam-comparable recorded loss, CCA and the pre-/post-projection control have nearly identical drift, while Adam and AdamP accumulate large drift in both cross-tensor quantities (Table 2). This pattern is the one predicted by Eq. (19): both explicit implementations remove the component of the preconditioned update in the supplied normal space, whereas AdamP applies a per-tensor correction rather than the coupled Jacobian rows.

Table 2: Supplied-law attention control (three seeds). Recorded loss is the cached value before the final update, reported as mean ± sample standard deviation; drift is the seed mean at final parameters. CCA and Pre/post Adam explicitly project using the analytic cross-tensor Jacobian. GD provides a discretization reference at its recorded loss. Absolute drifts and full protocol details are in Appendix E.

Method	**Recorded loss**	rel. drift q_{QK}	rel. drift q_{VO}
GD	0.2894 ± 0.0131	0.0059%	0.0024%
Adam	0.0221 ± 0.0009	1841%	878%
CCA	0.0219 ± 0.0005	1.134%	0.257%
AdamP	0.0221 ± 0.0009	1841%	878%
Pre/post Adam	0.0219 ± 0.0005	1.134%	0.257%

3.3 SUPPORTED SCOPE

The certificate addresses declared $(K_0, K_0 + 1)$ boundaries in shallow controls using first brackets, fixed numerical tolerances, and a loss-unit-dependent raw-direction floor. Its quantitative interpretation is conditioned on singular margins, derivative accuracy, and support coverage. Both the positive laws and the generic candidate features are quadratic. The next evaluation tier consists of nonquadratic law families, blind count recovery, full Lie-closure diagnostics, and distribution-shift transfer. V15 adds pathwise coverage and conditioning to the requirements for optimizer use.

4 CONCLUSION

Common-law identification from finite gradient-field observations requires a claim layer beyond tangency fitting. MEND supplies joint candidate normal spaces; the held-out rank, bracket, and capacity screens test functional independence, differential compatibility, and geometric capacity. On the frozen controls, the full certificate resolves every declared boundary/null case, while the quantitative analysis shows why residuals require singular and coverage margins. CCA then exposes a separate pathwise requirement. Supplied exact attention laws are preserved at comparable loss, while the learned-law gate records FAIL at its conditioning criterion. The present result is a finite certificate with explicit abstention and scope. It provides a foundation for nonquadratic recovery, adaptive numerical resolution, blind count estimation, and certification along paths near singular strata.

Reproducibility statement. The anonymized supplement contains the original runners, dependency file, frozen protocols, models, source snapshots, seed-level records, safeguard outcomes, and complete manifests. A read-only notebook verifies the v14 and v15 hashes and reconstructs every reported table and figure, keeps retrospective calculations separate from formal decisions, and rebuilds the paper. Appendix G records the information order, thresholds, candidate and task barriers, and hardware; Appendix H gives executable pseudocode. The replay establishes package and report consistency; the frozen protocols also specify the inputs required for independent training replication.

Ethics statement. The experiments use synthetic computational controls exclusively. The conditioning safeguard records every rank-boundary event and abstention, and the learned-law gate retains its formal status. The study concerns optimizer geometry in controlled models.

AI use statement. Generative AI tools assisted with polishing the paper only. The authors take responsibility for the paper's text, claims, and code.

REFERENCES

Vladimir I. Arnold. *Mathematical Methods of Classical Mechanics*, volume 60 of *Graduate Texts in Mathematics*. Springer-Verlag, 2 edition, 1989. doi: 10.1007/978-1-4757-2063-1.

Shaoxuan Chen, Panayotis G. Kevrekidis, Hong-Kun Zhang, and Wei Zhu. Data-driven discovery of conservation laws from trajectories via neural deflation. *Communications in Nonlinear Science and Numerical Simulation*, 143:108563, 2025. doi: 10.1016/j.cnsns.2024.108563.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Seungwoong Ha and Hawoong Jeong. Discovering invariants via machine learning. *Physical Review Research*, 3:L042035, 2021. doi: 10.1103/PhysRevResearch.3.L042035.

Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In *International Conference on Learning Representations*, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Hiroyuki Kasai, Pratik Jawanpuria, and Bamdev Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3262–3271. PMLR, 2019.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Ziming Liu, Varun Madhavan, and Max Tegmark. Machine learning conservation laws from differential equations. *Physical Review E*, 106:045307, 2022. doi: 10.1103/PhysRevE.106.045307.

Ziming Liu, Patrick Obin Sturm, Saketh Bharadwaj, Sam J. Silva, and Max Tegmark. Interpretable conservation laws as sparse invariants. *Physical Review E*, 109:L023301, 2024. doi: 10.1103/PhysRevE.109.L023301.

Peter Y. Lu, Rumen Dangovski, and Marin Soljačić. Discovering conservation laws using optimal transport and manifold learning. *Nature Communications*, 14:4744, 2023. doi: 10.1038/s41467-023-40325-7.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2760.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Keep the momentum: Conservation laws beyond Euclidean gradient flows. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34790–34821. PMLR, 2024.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Transformative or conservative? conservation laws for ResNets and Transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 43140–43176. PMLR, 2025.

Rahul D. Ray. From data to laws: Neural discovery of conservation laws without false positives, 2026.

Viet-Hoang Tran, Vinh Khanh Bui, Tan Lai Ngoc, Nam Nguyen, Tuan Dam, and Tan M. Nguyen. Conservation laws for modern neural architectures. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. doi: 10.48550/arXiv.2606.17816. Also available as arXiv:2606.17816.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1213.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. In *International Conference on Learning Representations*, 2023.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries. In *International Conference on Learning Representations*, 2024.

Wei Zhu, Hong-Kun Zhang, and P. G. Kevrekidis. Machine learning of independent conservation laws through neural deflation. *Physical Review E*, 108:L022301, 2023. doi: 10.1103/PhysRevE.108.L022301.

A Complete decision and control tables

The main text uses seed-resolved figures, the component-removal audit, and a compact preservation table. The tables below retain the corresponding aggregate counts, secondary sequential diagnostic, and absolute attention-law drifts.

Table 3: Aggregate form of the frozen v14 staircase. Positive entries require accepting K_0 and rejecting the independently fitted $K_0 + 1$ candidate; null entries require rejecting $K = 1$. "Span" is a post-freeze analytic audit reported alongside the data-only decision.

Architecture	Field	Rank	Bracket	Full	Span
Linear, target $K = 6$	0/10	8/10	8/10	**10/10**	10/10
ReLU, target $K = 3$	1/10	6/10	6/10	**10/10**	10/10
GELU, null control	0/10	0/10	**10/10**	**10/10**	N/A
SiLU, null control	0/10	0/10	**10/10**	**10/10**	N/A

Table 4: Prespecified one-environment GELU diagnostic. K is selected with one shared development environment; "outer" applies the fresh 12-environment common-law gate. ρ_{out} is mean [min, max] over five seeds. This diagnostic uses its own objective and budget.

Objective	K by seed	Local	Outer	ρ_{out}
Asymmetric + amp.	3,1,3,3,2	5/5	0/5	$8.81\ [3.64, 26.0]\times 10^{-2}$
Symmetric normalized	1,3,3,1,3	3/5	0/5	$1.75\ [0.441, 3.16]\times 10^{-2}$

Table 5: Complete supplied-law attention control. Loss is the cached value before the final update, reported as mean $\pm$ sample standard deviation. Each drift entry is a seed mean at final parameters. The projection control projects before moment accumulation and again after the Adam preconditioner, with ridge 10^{-10}; GD provides the unprojected gradient-flow reference.

Method	Recorded loss	$\lvert\Delta q_{\mathrm{QK}}\rvert$	rel. drift q_{QK}	$\lvert\Delta q_{\mathrm{VO}}\rvert$	rel. drift q_{VO}
GD	0.2894 ± 0.0131	1.26×10^{-5}	0.0059%	8.86×10^{-6}	0.0024%
Adam	0.0221 ± 0.0009	4.109	1841%	3.565	878%
CCA	0.0219 ± 0.0005	2.221×10^{-3}	1.134%	1.058×10^{-3}	0.257%
AdamP	0.0221 ± 0.0009	4.109	1841%	3.565	878%
Pre/post Adam	0.0219 ± 0.0005	2.219×10^{-3}	1.134%	1.057×10^{-3}	0.257%

B Additional method details

B.1 Relation to Hamiltonian Neural Deflation

Table 6 isolates which parts of Neural Deflation are retained and which are replaced by the common-law formulation.

B.2 Analytic laws in the positive controls

The benchmark boundaries use conserved quantities that follow directly from the two-layer parameterization. The statement below establishes the analytic generators used in the experiment. Population maximality is a separate architecture-level property.

Proposition 4 (Balance laws in the positive controls)**.** *Consider a bias-free two-layer network with $\mathbf{W}_1 \in \mathbb{R}^{3\times 2}$ and $\mathbf{W}_2 \in \mathbb{R}^{2\times 3}$, trained by Euclidean gradient flow for a differentiable loss that depends on the parameters through the network output. For the linear activation,*

$$\mathbf{Q}_{\mathrm{lin}} := \mathbf{W}_2^\top \mathbf{W}_2 - \mathbf{W}_1 \mathbf{W}_1^\top \tag{24}$$

Table 6: Neural Deflation and MEND share a tangency-based candidate-learning principle but solve different identification problems. The Hamiltonian integrability criterion is replaced by cross-environment compatibility for gradient flow.

Property	**Neural Deflation (Zhu et al., 2023)**	**MEND**
Dynamics	One known Hamiltonian field $\mathbf{f} = \mathbf{J}\nabla H$.	Several Euclidean gradient fields $\mathbf{f}_e = -\nabla L_e$ from a declared environment family.
Candidate	Scalar networks fitted sequentially.	One vector-valued quadratic-neural map fitted jointly for each proposed K.
Duplicate control	Poisson involution and a deflation factor based on the gradient component outside the preceding span.	A global gradient frame, held-out pointwise row rank, and a separate Lie-capacity screen.
Meaning of success	Independent, Poisson-commuting integrals for the sampled Hamiltonian system.	Common-law behavior on the declared support and environment family through a finite certificate.
Law count	Stop a sequential prefix by validation loss.	Evaluate a frozen K grid in general; the decisive experiment tests predeclared positive-control boundaries K_0 versus $K_0 + 1$.

is conserved, and its six symmetric entries have generically independent differentials. For ReLU, define

$$q_i := \|\mathbf{W}_{2,:,i}\|_2^2 - \|\mathbf{W}_{1,i,:}\|_2^2, \qquad i = 1, 2, 3. \tag{25}$$

Their time derivatives vanish at every differentiability point, and the three differentials are generically independent. Consequently, the q_i are conserved along any absolutely continuous gradient-flow solution for which the chain rule holds almost everywhere.

Proof. For the linear model, write $\mathbf{M} = \mathbf{W}_2\mathbf{W}_1$ and $\mathbf{G} = \nabla_{\mathbf{M}} L$. The chain rule gives $\nabla_{\mathbf{W}_2} L = \mathbf{G}\mathbf{W}_1^\top$ and $\nabla_{\mathbf{W}_1} L = \mathbf{W}_2^\top \mathbf{G}$. Substitution into the time derivative of Eq. (24) yields

$$\begin{aligned} \dot{\mathbf{Q}}_{\mathrm{lin}} = & -\mathbf{W}_1\mathbf{G}^\top\mathbf{W}_2 - \mathbf{W}_2^\top\mathbf{G}\mathbf{W}_1^\top \\ & + \mathbf{W}_2^\top\mathbf{G}\mathbf{W}_1^\top + \mathbf{W}_1\mathbf{G}^\top\mathbf{W}_2 = \mathbf{0}_{3\times 3}. \end{aligned} \tag{26}$$

To verify generic independence, set $\mathbf{A} = \mathbf{W}_2^\top$ and $\mathbf{B} = \mathbf{W}_1$. At $\mathbf{A} = [\mathbf{e}_1, \mathbf{e}_2]$ and $\mathbf{B} = [\mathbf{e}_2, \mathbf{e}_3]$, the differentials of $\mathbf{A}\mathbf{A}^\top$ span all symmetric coordinate directions except $(3, 3)$, while those of $\mathbf{B}\mathbf{B}^\top$ include $(3, 3)$. Hence the differential of $\mathbf{Q}_{\mathrm{lin}}$ is surjective onto the six-dimensional space of symmetric 3×3 matrices at this point. The rank is therefore six on an open dense set.

For ReLU, the positive rescaling $(\mathbf{W}_{1,i,:}, \mathbf{W}_{2,:,i}) \mapsto (a\mathbf{W}_{1,i,:}, a^{-1}\mathbf{W}_{2,:,i})$ leaves the network output unchanged for $a > 0$. Differentiating at $a = 1$ and using Euclidean gradient flow gives $\dot{q}_i = 0$ wherever the loss field is differentiable. Absolute continuity then gives pathwise conservation under the stated chain-rule condition. The three gradients have disjoint parameter blocks and are independent whenever no corresponding pair $(\mathbf{W}_{1,i,:}, \mathbf{W}_{2,:,i})$ vanishes simultaneously. This condition holds generically. □

B.3 Population statements behind the finite certificate

Theorem 1 gives a necessary codimension condition. The next statements give conditions under which environment restriction preserves the population law set and the full codimension becomes locally exact. The finite certificate treats these conditions as explicit assumptions.

Proposition 5 (Environment sufficiency)**.** *Let $\mathcal{O} \subseteq \mathbb{R}^d$ be open, let $\mathcal{E}_0 \subseteq \mathcal{E}_\star$, and suppose all corresponding fields are smooth on $\mathcal{O}$. If*

$$\mathcal{D}_{\mathcal{E}_0}^{(\infty)}(\boldsymbol{\theta}) = \mathcal{D}_{\mathcal{E}_\star}^{(\infty)}(\boldsymbol{\theta}) \quad \textit{for every } \boldsymbol{\theta} \in \mathcal{O}, \tag{27}$$

then a smooth function is an exact common first integral on $\mathcal{O}$ for $\mathcal{E}_0$ if and only if it is an exact common first integral there for $\mathcal{E}_\star$.

The next consequence of Frobenius theory is the local completeness characterization used by Marcotte et al. (2023, Theorem 3.3).

Corollary 6 (Local completeness under regular Lie closure)**.** *Suppose the full Lie closure defines a smooth involutive distribution $\mathcal{D}$ of constant rank r on a neighborhood $\mathcal{U}$. Every point then has a smaller neighborhood with exactly $d - r$ functionally independent common first integrals.*

For a noncollapsed candidate and step size h, a related one-step diagnostic is

$$D_h^2(C_{\boldsymbol{\phi}}; e) := \frac{\mathbb{E}_\mu \|C_{\boldsymbol{\phi}}(\boldsymbol{\theta} + h\mathbf{f}_e(\boldsymbol{\theta})) - C_{\boldsymbol{\phi}}(\boldsymbol{\theta})\|_2^2}{\mathbb{E}_\mu \|C_{\boldsymbol{\phi}}(\boldsymbol{\theta}) - \mathbb{E}_\mu C_{\boldsymbol{\phi}}\|_2^2 + \epsilon}. \tag{28}$$

Define $D_h := \sqrt{D_h^2} \geq 0$. If the denominator is bounded below by a positive constant independently of h, the first-order term has a finite second moment, and the second-order Taylor remainder is bounded in $L^2(\mu)$ by a constant times h^2, then $D_h = O(h^2)$ for exact infinitesimal tangency. If the first-order term has positive second moment, D_h is instead of order h. At fixed physical horizon, an exact law exhibits the global error of the numerical integrator under the usual stability assumptions. The reported premise audit uses relative $\mathbf{q}$ drift at a fixed horizon, whereas Eq. (28) defines a local diagnostic. The certificate in Eq. (15) uses its own directional statistics.

B.4 Single-environment objective

For one field, the legacy asymmetric and symmetric residuals are

$$\begin{aligned} \ell_{\mathrm{asym}}(C) &:= \mathbb{E}_\mu[\big(\widehat{\mathbf{f}}^\top \nabla C\big)^2], \\ \ell_{\mathrm{sym}}(C) &:= \mathbb{E}_\mu[\big(\widehat{\mathbf{f}}^\top \widehat{\nabla C}\big)^2]. \end{aligned} \tag{29}$$

The first requires an explicit amplitude floor to exclude shrinking; the second removes scale but is undefined at a constant candidate. Both remain one-field objectives.

Theorem 7 (Scale identities for one-field tangency)**.** *Let $C \in \mathcal{C}^1(\mathcal{O})$, let μ be a probability measure on $\mathcal{O}$, assume $\|\mathbf{f}(\boldsymbol{\theta})\|_2 > 0$ μ-almost everywhere, and let $t \neq 0$. For the symmetric residual, also assume $\|\nabla C(\boldsymbol{\theta})\|_2 > 0$ μ-almost everywhere. Then*

$$\ell_{\mathrm{sym}}(tC) = \ell_{\mathrm{sym}}(C), \qquad \ell_{\mathrm{asym}}(tC) = t^2 \ell_{\mathrm{asym}}(C). \tag{30}$$

Both residuals vanish for an exact first integral. If g is random and, jointly with $\boldsymbol{\theta} \sim \mu$, the direction $\widehat{\nabla g(\boldsymbol{\theta})}$ is independent of $\widehat{\mathbf{f}}(\boldsymbol{\theta})$ and uniform on $\mathbb{S}^{d-1}$, then $\mathbb{E}_g[\ell_{\mathrm{sym}}(g)] = 1/d$.

B.5 CCA update and pathwise regularity

For completeness, the standard Adam direction used in CCA is

$$\begin{aligned} \mathbf{m}_{t+1} &= \beta_1 \mathbf{m}_t + (1-\beta_1)\mathbf{g}_t, & \mathbf{v}_{t+1} &= \beta_2 \mathbf{v}_t + (1-\beta_2)(\mathbf{g}_t \odot \mathbf{g}_t), \\ \widehat{\mathbf{m}}_{t+1} &= \mathbf{m}_{t+1}/(1-\beta_1^{t+1}), & \widehat{\mathbf{v}}_{t+1} &= \mathbf{v}_{t+1}/(1-\beta_2^{t+1}), \\ \mathbf{u}_t &= \widehat{\mathbf{m}}_{t+1} \oslash (\sqrt{\widehat{\mathbf{v}}_{t+1}} + \epsilon_{\mathrm{Adam}} \mathbf{1}_d). \end{aligned} \tag{31}$$

At zero ridge and full row rank, Eq. (19) gives the orthogonal decomposition

$$\mathbf{u}_t = \mathbf{d}_t + \widetilde{\mathbf{J}}_t^\top (\widetilde{\mathbf{J}}_t \widetilde{\mathbf{J}}_t^\top)^{-1} \widetilde{\mathbf{J}}_t \mathbf{u}_t. \tag{32}$$

For a compact path $\Theta \colon [0,T] \to \mathcal{O}$ on which the largest singular value is nonzero, define

$$\begin{aligned} \gamma_T(\widetilde{C}; \Theta) &:= \inf_{s \in [0,T]} \sigma_K(\mathbf{J}_{\widetilde{C}}(\Theta(s))), \\ \overline{\gamma}_T(\widetilde{C}; \Theta) &:= \inf_{s \in [0,T]} \frac{\sigma_K(\mathbf{J}_{\widetilde{C}}(\Theta(s)))}{\sigma_1(\mathbf{J}_{\widetilde{C}}(\Theta(s)))}. \end{aligned} \tag{33}$$

The absolute margin γ_T controls invertibility of the zero-ridge projection, whereas the relative margin $\overline{\gamma}_T$ is invariant to uniform rescaling but remains coordinate dependent; it is the conditioning safeguard used in the preservation audit.

Proposition 8 (Pathwise deployability of zero-ridge CCA)**.** *Let $\widetilde{C} \in \mathcal{C}^2(\mathcal{O}; \mathbb{R}^K)$ and let $\Theta\colon [0,T] \to \mathcal{O}$ be continuous. If $\gamma_T(\widetilde{C}; \Theta) \geq \gamma > 0$, then the zero-ridge CCA projector is well defined and continuous along the path, with*

$$\sup_{s\in[0,T]} \left\| \left(\mathbf{J}_{\widetilde{C}}(\Theta(s)) \mathbf{J}_{\widetilde{C}}(\Theta(s))^\top \right)^{-1} \right\|_2 \leq \gamma^{-2}. \tag{34}$$

The projector is locally Lipschitz on a neighborhood of the path. At a rank-deficient Jacobian, direction-dependent limits produce distinct zero-ridge projectors.

Proposition 9 (Local foliation by conservation values)**.** *Let $L\colon \mathcal{P} \to \mathbb{R}$ be smooth and let $\mathcal{O} \subseteq \mathcal{P}$ be open with $\nabla L(\boldsymbol{\theta}) \neq \mathbf{0}_d$ on $\mathcal{O}$. Every $\boldsymbol{\theta}_0 \in \mathcal{O}$ has a neighborhood supporting $d-1$ functionally independent smooth first integrals. Their common level sets are local one-dimensional flow leaves.*

C Quantitative certificate analysis

The analysis below bounds approximate normal-space identification from finite tolerances. All matrix norms without subscripts in inequalities are operator norms. The direction matrix may contain any declared finite collection of fields and brackets. Its orthogonal complement represents the normal space of common-law leaves under the full-Lie-span and regularity assumptions stated in the main text.

C.1 Pointwise geometry and perturbations

Proof of Theorem 2. Set $\mathcal{U} = \mathrm{row}(\mathbf{A})$ and $\mathcal{V} = \mathrm{col}(\mathbf{B})$. The singular-value assumptions give $\mathbf{A}^\top \mathbf{A} \succeq a^2 \mathbf{P}_{\mathcal{U}}$ and $\mathbf{B}\mathbf{B}^\top \succeq b^2 \mathbf{P}_{\mathcal{V}}$. The trace of the product of two positive semidefinite matrices is nonnegative. Applying this observation twice yields

$$\|\mathbf{AB}\|_F^2 = \mathrm{Tr}(\mathbf{A}^\top \mathbf{A}\mathbf{B}\mathbf{B}^\top) \geq a^2 b^2 \,\mathrm{Tr}(\mathbf{P}_{\mathcal{U}} \mathbf{P}_{\mathcal{V}}) \geq a^2 b^2 \dim(\mathcal{U} \cap \mathcal{V}). \tag{35}$$

Since $\dim(\mathcal{U} \cap \mathcal{V}) \geq (K + r - d)_+$, this proves Eq. (16).

Choose column-orthonormal bases $\mathbf{Q}_{\mathcal{U}}$ and $\mathbf{Q}_{\mathcal{V}}$, and write $\mathbf{A} = \mathbf{R}\mathbf{Q}_{\mathcal{U}}^\top$ and $\mathbf{B} = \mathbf{Q}_{\mathcal{V}}\mathbf{S}$, where $\mathbf{R}$ is invertible and $\mathbf{S}$ has full row rank. Their inverse and right pseudoinverse satisfy $\|\mathbf{R}^{-1}\|_2 = a^{-1}$ and $\|\mathbf{S}^\dagger\|_2 = b^{-1}$. Consequently,

$$\|\mathbf{Q}_{\mathcal{U}}^\top \mathbf{Q}_{\mathcal{V}}\|_2 = \|\mathbf{R}^{-1}\mathbf{A}\mathbf{B}\mathbf{S}^\dagger\|_2 \leq \frac{\|\mathbf{AB}\|_2}{ab}. \tag{36}$$

If $K = d - r$, the subspaces $\mathcal{U}$ and $\mathcal{V}^\perp$ have equal dimension. The principal-angle identity gives $\|\mathbf{P}_{\mathcal{U}} - \mathbf{P}_{\mathcal{V}^\perp}\|_2 = \|\mathbf{P}_{\mathcal{U}}\mathbf{P}_{\mathcal{V}}\|_2 = \|\mathbf{Q}_{\mathcal{U}}^\top \mathbf{Q}_{\mathcal{V}}\|_2$. Orthogonal-projector differences have norm at most one, proving Eq. (17). If $K + r > d$, the intersection $\mathcal{U} \cap \mathcal{V}$ contains a unit vector, so $\|\mathbf{P}_{\mathcal{U}}\mathbf{P}_{\mathcal{V}}\|_2 = 1$. Equation (36) then gives the final claim. □

Observed fields and their derivatives may be approximate. A perturbation bound separates candidate residual, direction error, and the retained singular margin. Suppose the exact direction matrix $\mathbf{B}_\star$ has rank r and $\|\widehat{\mathbf{B}} - \mathbf{B}_\star\|_2 \leq \nu$. If $\widehat{b} := \sigma_r(\widehat{\mathbf{B}}) > \nu$ and $K = d - r$, then

$$\left\| \mathbf{P}_{\mathrm{row}(\mathbf{A})} - \mathbf{P}_{\ker(\mathbf{B}_\star^\top)} \right\|_2 \leq \min\left\{ 1, \frac{\|\mathbf{A}\widehat{\mathbf{B}}\|_2 + \|\mathbf{A}\|_2 \nu}{a(\widehat{b} - \nu)} \right\}. \tag{37}$$

The bound assumes $\mathrm{rank}\,\mathbf{B}_\star = r$ and becomes quantitative when the retained margin $\widehat{b} - \nu$ is resolved.

Proof of Eq. (37). The triangle inequality implies

$$\|\mathbf{AB}_\star\|_2 \leq \|\mathbf{A}\widehat{\mathbf{B}}\|_2 + \|\mathbf{A}\|_2 \nu. \tag{38}$$

The singular-value perturbation inequality gives $\sigma_r(\mathbf{B}_\star) \geq \widehat{b} - \nu > 0$. Applying Eq. (17) to $\mathbf{B}_\star$ proves the stated bound. □

Whitening the candidate rows isolates the exact geometric object from its coordinate conditioning. Define $\mathbf{Q} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1/2}$, whose columns are an orthonormal basis for $\mathrm{row}(\mathbf{A})$. Let $\widehat{\mathbf{V}} \in \mathbb{R}^{d\times r}$ be an orthonormal basis for any selected leading r-dimensional left-singular subspace of $\widehat{\mathbf{B}}$. Its retained normal space is $\mathrm{col}(\widehat{\mathbf{V}})^\perp$. The next bound uses these retained singular directions rather than the algebraic rank of a noisy matrix.

Proposition 10 (Truncated direction-space bound)**.** *Suppose* $\mathrm{rank}\,\mathbf{A} = K = d - r$, $r \geq 1$, $\mathrm{rank}\,\mathbf{B}_\star = r$, *and* $\|\widehat{\mathbf{B}} - \mathbf{B}_\star\|_2 \leq \nu$. *If* $\widehat{b} = \sigma_r(\widehat{\mathbf{B}}) > 0$, *then*

$$\left\|\mathbf{P}_{\mathrm{row}(\mathbf{A})} - \mathbf{P}_{\ker(\mathbf{B}_\star^\top)}\right\|_2 \leq \min\left\{1, \frac{\|\mathbf{Q}^\top\widehat{\mathbf{B}}\|_2 + \nu}{\widehat{b}}\right\}. \tag{39}$$

Proof. Write the leading singular triplets as $\widehat{\mathbf{B}}\widehat{\mathbf{W}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}$, with $\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}} = \mathbf{I}_r$ and $\sigma_r(\widehat{\mathbf{\Sigma}}) = \widehat{b}$. Let $\mathcal{V}_\star = \mathrm{col}(\mathbf{B}_\star)$. Then

$$\|(\mathbf{I}_d - \mathbf{P}_{\mathcal{V}_\star})\widehat{\mathbf{V}}\|_2 \leq \frac{\nu}{\widehat{b}}, \qquad \|\mathbf{Q}^\top\widehat{\mathbf{V}}\|_2 \leq \frac{\|\mathbf{Q}^\top\widehat{\mathbf{B}}\|_2}{\widehat{b}}. \tag{40}$$

The first inequality follows by multiplying $(\mathbf{I}_d - \mathbf{P}_{\mathcal{V}_\star})(\widehat{\mathbf{B}} - \mathbf{B}_\star)$ by $\widehat{\mathbf{W}}\widehat{\mathbf{\Sigma}}^{-1}$; the second follows from the same right multiplication. Equal dimensions convert both inequalities into projector-distance bounds. The triangle inequality through $\mathrm{col}(\widehat{\mathbf{V}})^\perp$ proves Eq. (39). □

Exact whitening is invariant, up to an orthogonal basis change, under invertible mixing of the law coordinates. Its numerical computation requires a resolved candidate margin. A small unretained singular value is compatible with either a weak direction or a zero direction perturbed by noise. Structural information therefore supplies the assumption $\mathrm{rank}\,\mathbf{B}_\star = r$ beyond a fixed relative cutoff. Equations (37) and (39) state this dependence explicitly.

The field-error budget must include differentiation. Suppose, at a fixed point, two exact fields have norms at most m_f and derivative norms at most m_d. Suppose their approximations have field errors at most ε_f and derivative errors at most ε_d. Expanding the two products in the bracket gives

$$\|[\widehat{\mathbf{f}}, \widehat{\mathbf{g}}] - [\mathbf{f}, \mathbf{g}]\|_2 \leq 2m_d\varepsilon_f + 2m_f\varepsilon_d + 2\varepsilon_f\varepsilon_d. \tag{41}$$

For example, $\mathbf{D}\widehat{\mathbf{g}}\widehat{\mathbf{f}} - \mathbf{D}\mathbf{g}\mathbf{f} = \mathbf{D}\mathbf{g}(\widehat{\mathbf{f}} - \mathbf{f}) + (\mathbf{D}\widehat{\mathbf{g}} - \mathbf{D}\mathbf{g})\mathbf{f} + (\mathbf{D}\widehat{\mathbf{g}} - \mathbf{D}\mathbf{g})(\widehat{\mathbf{f}} - \mathbf{f})$; the other term is bounded identically. For a matrix with column-error bounds ν_j, one can take $\nu \leq (\sum_j \nu_j^2)^{1/2}$. After direction normalization, error amplification also depends on the raw norm. If $\|\widehat{\mathbf{v}} - \mathbf{v}\|_2 \leq \xi < \|\mathbf{v}\|_2$, then

$$\left\|\frac{\widehat{\mathbf{v}}}{\|\widehat{\mathbf{v}}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\|_2 \leq \frac{2\xi}{\|\mathbf{v}\|_2}. \tag{42}$$

Thus a bound for normalized brackets requires a raw resolvability margin.

C.2 From sampled points to support coverage

Pointwise control becomes uniform only when the validation points cover the region of interest. The next result makes that dependence explicit. An h-net of a compact set $\mathcal{S} \subset \mathcal{O}$ places a validation point within Euclidean distance h of every point in $\mathcal{S}$.

Theorem 11 (Finite-support normal-space bound)**.** *Let* $\mathbf{A}\colon \mathcal{S} \to \mathbb{R}^{K\times d}$ *and* $\mathbf{B}_\star\colon \mathcal{S} \to \mathbb{R}^{d\times m}$ *be Lipschitz in operator norm with constants* ℓ_a, ℓ_b *and bounded by* m_a, m_b. *Suppose* $\mathrm{rank}\,\mathbf{B}_\star(\boldsymbol{\theta}) = r \geq 1$ *on* $\mathcal{S}$. *Let* $\{\boldsymbol{\theta}_i\}_{i=1}^n$ *be an* h*-net and suppose at every validation point*

$$\|\widehat{\mathbf{B}}_i - \mathbf{B}_\star(\boldsymbol{\theta}_i)\|_2 \leq \nu, \quad \sigma_K(\mathbf{A}(\boldsymbol{\theta}_i)) \geq a_0, \quad \sigma_r(\widehat{\mathbf{B}}_i) \geq b_0, \quad \|\mathbf{A}(\boldsymbol{\theta}_i)\widehat{\mathbf{B}}_i\|_2 \leq \varepsilon_0. \tag{43}$$

Define

$$\begin{aligned} a_h &:= a_0 - \ell_a h > 0, \qquad b_h := b_0 - \nu - \ell_b h > 0, \\ \varepsilon_h &:= \varepsilon_0 + m_a\nu + (\ell_a m_b + m_a\ell_b)h. \end{aligned} \tag{44}$$

If $K = d - r$, *then*

$$\sup_{\boldsymbol{\theta}\in\mathcal{S}} \left\|\mathbf{P}_{\mathrm{row}(\mathbf{A}(\boldsymbol{\theta}))} - \mathbf{P}_{\ker(\mathbf{B}_\star(\boldsymbol{\theta})^\top)}\right\|_2 \leq \min\{1, \varepsilon_h/(a_h b_h)\}. \tag{45}$$

If $K + r > d$, *the same conditions imply* $\varepsilon_h \geq a_h b_h$.

Proof of Theorem 11. Fix $\boldsymbol{\theta} \in \mathcal{S}$ and choose a validation point $\boldsymbol{\theta}_i$ within distance h. Lipschitz continuity and the singular-value perturbation inequality give

$$\sigma_K(\mathbf{A}(\boldsymbol{\theta})) \geq a_0 - \ell_a h = a_h, \qquad \sigma_r(\mathbf{B}_\star(\boldsymbol{\theta})) \geq b_0 - \nu - \ell_b h = b_h. \tag{46}$$

In particular, $\mathbf{A}(\boldsymbol{\theta})$ has full row rank. At the validation point the exact residual is at most $\varepsilon_0 + m_a\nu$. Adding and subtracting $\mathbf{A}(\boldsymbol{\theta}_i)\mathbf{B}_\star(\boldsymbol{\theta})$ yields

$$\begin{aligned}\|\mathbf{A}(\boldsymbol{\theta})\mathbf{B}_\star(\boldsymbol{\theta})\|_2 \leq \varepsilon_0 + m_a\nu + \|\mathbf{A}(\boldsymbol{\theta}) - \mathbf{A}(\boldsymbol{\theta}_i)\|_2\|\mathbf{B}_\star(\boldsymbol{\theta})\|_2 \\ + \|\mathbf{A}(\boldsymbol{\theta}_i)\|_2\|\mathbf{B}_\star(\boldsymbol{\theta}) - \mathbf{B}_\star(\boldsymbol{\theta}_i)\|_2 \leq \varepsilon_h.\end{aligned} \tag{47}$$

Theorem 2 now applies at every point. Its normal-space bound proves Eq. (45); its operator-norm overcount bound gives $\varepsilon_h \geq a_h b_h$ when $K + r > d$. □

To justify the sampling statement, let $\{\mathbf{z}_j\}_{j=1}^N$ be an $h/2$-cover with centers in $\mathcal{S}$ and assume $\mu(\mathcal{S} \cap \mathcal{B}(\mathbf{z}_j, h/2)) \geq p$ for each j. For independent and identically distributed draws from μ, the probability that any one ball remains uncovered is at most $(1-p)^n \leq e^{-np}$. A union bound over the N balls gives an uncovered probability at most Ne^{-np}. If every ball is occupied, each point of $\mathcal{S}$ is within h of a sample, by the triangle inequality. Hence $n \geq p^{-1}\log(N/\alpha)$ suffices for coverage probability at least $1 - \alpha$. The cover and sampling law must be fixed independently of the validation draw. This argument concerns support points conditional on the environments; environment sufficiency is the separate premise in Proposition 5.

For intuition about the constants, if the dimensionless margins in Eq. (44) were $a_h = b_h = 1/2$ and $\varepsilon_h = 10^{-2}$, the uniform projector bound would be 0.04. If b_h instead fell to 10^{-3} with the other values unchanged, the bound would saturate at one. These illustrative substitutions show how the residual, singular-value margins, and coverage radius jointly determine the guarantee.

C.3 Concentration of a frozen population screen

A support-average decision is weaker than a uniform geometric bound. To state its statistical meaning, condition on the trained candidates, numerical evaluation rules, and validation environments. Let $g_1, \ldots, g_q \colon \mathcal{S} \to [0,1]$ denote every scalar score that is to be inspected, including per-row squared directional residuals and pointwise rank/capacity indicators for all frozen candidates. Dependence between different scores at the same point is allowed.

Proposition 12 (Simultaneous frozen-score bound)**.** *Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ be independent draws from μ, independent of the frozen score functions. Define $p_j = \mathbb{E}_\mu[g_j]$ and $\widehat{p}_j = n^{-1}\sum_i g_j(\boldsymbol{\theta}_i)$. For $0 < \alpha < 1$, with probability at least $1 - \alpha$,*

$$\max_{1 \leq j \leq q} |\widehat{p}_j - p_j| \leq t_n, \qquad t_n := \sqrt{\frac{\log(2q/\alpha)}{2n}}. \tag{48}$$

The same statement holds for independent, nonidentically distributed draws when p_j is replaced by $n^{-1}\sum_i \mathbb{E}[g_j(\boldsymbol{\theta}_i)]$.

Proof. For each fixed j, Hoeffding's inequality (Hoeffding, 1963) gives $\Pr(|\widehat{p}_j - p_j| > t) \leq 2e^{-2nt^2}$. A union bound over the q scores and substitution of $t = t_n$ yield Eq. (48). The bounded-variable inequality requires independence across i and allows heterogeneous distributions and dependence across j, so the second statement follows in the same way. □

The proposition is the standard Hoeffding union bound for bounded scores. On the stated event, an empirical upper-threshold pass $\widehat{p}_j \leq \tau_j$ implies $p_j \leq \tau_j + t_n$. A lower-threshold pass implies $p_j \geq \tau_j - t_n$. Simultaneous empirical and population decisions agree if every population score is separated from its cutoff by more than t_n. A rule that requires $\widehat{p}_j + t_n \leq \tau_j$ (or $\widehat{p}_j - t_n \geq \tau_j$) instead certifies the corresponding population inequality. The v14 protocol uses the empirical rule. A confidence-adjusted rule certifies a positive-tolerance population screen; exact conservation and uniform discrimination among arbitrarily close laws add a separation assumption, as Eq. (18) illustrates.

The implemented bracket statistic averages only over points with a retained nonzero bracket. Conditional on a fixed positive number of such points, independent and identically distributed support

draws give independent draws from the corresponding conditional support law. The bound then uses that number as the sample size and targets a conditional mean. An empty retained collection triggers the implementation's vacuous convention. A prospective screen can instead estimate the retention probability and an unconditional zero-extended score. A minimum row norm can be represented by its threshold-violation indicator, to which the proposition applies. An observed zero count then bounds the population violation probability.

The frozen support generator fixes stratum counts before permutation: 38 Gaussian points, 38 shell points, and 20 short-trajectory endpoints when $n = 96$. Each endpoint has a separate random initialization and integration length. Conditional on the calibration environment, the pre-permutation draws are independent within and across strata in the ideal sampling model; their average targets the rounded $38/38/20$ mixture. The nonidentically distributed form of Proposition 12 therefore describes unmasked score averages for this design. The shared permutation changes no average, but valid-point masking still requires the conditional or ratio analysis described above. For one unmasked score, $n = 96$ and $\alpha = 0.05$ give $t_n \approx 0.139$; the corresponding population interval is wider than the empirical 0.02 tangency threshold. Simultaneous-score bounds are wider. The existing gate therefore reports its empirical decision. The 66 brackets share 12 sampled environments and are treated as dependent pairs rather than independent environment replicates.

C.4 Numerical margins on the frozen outer support

This retrospective calculation reuses the saved outer tensors of all 20 accepted v14 positive-control candidates, with 96 points per candidate and the original formal decisions. At each point, $\mathbf{A}$ contains the candidate rows normalized by the recorded row-norm floor 10^{-12}. The columns of $\mathbf{B}$ are the held-out fields and first brackets whose raw norms exceed 10^{-10}, each normalized to unit length. Let $\widehat{r}$ count singular values larger than $0.05\sigma_1(\mathbf{B})$, exactly as in the frozen capacity statistic, and let $\mathbf{B}_{\widehat{r}}$ be the corresponding truncated SVD. This rank refers to the retained numerical direction space.

Where $K + \widehat{r} = d$, $\widehat{r} \geq 1$, and $\mathbf{A}$ has full row rank, we evaluate Eq. (17) with $\mathbf{B}_{\widehat{r}}$ and its row-whitened counterpart. For an orthonormal basis $\mathbf{Q}$ of $\mathrm{row}(\mathbf{A})$, define

$$\begin{aligned} u &:= \frac{\|\mathbf{A}\mathbf{B}_{\widehat{r}}\|_2}{\sigma_K(\mathbf{A})\sigma_{\widehat{r}}(\mathbf{B})}, \qquad w := \frac{\|\mathbf{Q}^\top \mathbf{B}_{\widehat{r}}\|_2}{\sigma_{\widehat{r}}(\mathbf{B})}, \\ g &:= \left\| \mathbf{Q}\mathbf{Q}^\top - \mathbf{P}_{\ker(\mathbf{B}_{\widehat{r}}^\top)} \right\|_2, \quad g \leq \min\{1, u, w\}. \end{aligned} \tag{49}$$

These are exact matrix inequalities for the truncated numerical object. Under the rank and perturbation conditions above, g also bounds error against the common-law normal space.

The numerical dimension equality holds at $622/960$ linear points and $477/960$ ReLU points. On these respective subsets, the median values of (u, w, g) are $(0.347, 0.207, 0.0311)$ and $(0.159, 0.126, 0.0246)$. The maximum whitened bounds are 0.9875 and 0.7278, and the maximum measured gaps are 0.2151 and 0.0873. Fifty linear points have $u \geq 1$; all reported w values are below one. Whitening measures candidate-span alignment independently of coordinate conditioning within that span. It preserves the learned functions and the frozen certificate. These pooled pointwise summaries describe the saved outer support; seed-level uncertainty is reported in the staircase.

The other 338 linear and 483 ReLU points carry a dimension-mismatch flag. Their capacity inequality holds, and the retained directions leave more than K numerical normal dimensions; Eq. (49) then has several admissible target subspaces. The replay writes complete point-level and candidate-level records, including singular margins, measured gaps, both bounds, and raw bracket-retention diagnostics. All points retain their original v14 labels and cohort membership.

C.5 Exact examples and numerical scope

For Eq. (18), direct differentiation gives $\mathbf{f}_1 = (-x, 0)^\top$ and $\mathbf{f}_2 = (-(x + \varepsilon y), -\varepsilon(x + \varepsilon y))^\top$. Their determinant and bracket are

$$\det[\mathbf{f}_1, \mathbf{f}_2] = \varepsilon x(x + \varepsilon y), \qquad [\mathbf{f}_1, \mathbf{f}_2] = \varepsilon(-y, x)^\top. \tag{50}$$

Any common first integral has gradient orthogonal to two independent vectors on $\mathcal{O}_0$, hence zero gradient. It is constant because $\mathcal{O}_0$ is connected. At $(1, 0)$ the bracket residual is exactly one for every $\varepsilon > 0$.

The numerical implementation has a separate resolution limit. Fix $\varepsilon = 0.01$ and evaluate at $(1, 0)$. With unit loss scaling, the fields are $(-1, 0)^\top$ and $(-1, -0.01)^\top$ and their bracket is $(0, 0.01)^\top$. Normalized field tangency of $C = y$ is approximately 10^{-4}, while normalized bracket tangency is one. At the relative 0.05 cutoff, the field–bracket matrix has numerical rank two and capacity zero for $K = 1$. Multiplying both losses by 10^{-6} preserves the exact spans and places the bracket norm at 10^{-14}, below the implemented 10^{-10} floor. The retained field directions then have numerical rank one, the assigned bracket score is zero, and the corresponding capacity is one. This calculation specifies the role of the recorded loss units in the v14 gate.

A nonquadratic construction extends the common-law target beyond polynomial generators. For $a_e > 0$, define on $(-\pi/2, \pi/2) \times \mathbb{R}$

$$L_e(x, y) = \tfrac{1}{2} a_e(\sin x + y - b_e)^2, \qquad C(x, y) = y - \operatorname{atanh}(\sin x). \tag{51}$$

Here $\nabla C = (-\sec x, 1)^\top$ and each loss gradient is a scalar multiple of $(\cos x, 1)^\top$. Their inner product is zero for every choice of a_e, b_e. The regular level sets $y = \operatorname{atanh}(\sin x) + c$ replace the hyperbolas of the factorized example. On $|x| \le 0.9$, $|y| \le 1.5$, the law derivatives are bounded and $\|\nabla C\|_2 \ge 1$. The supplement includes an executable check of this construction. The example establishes analytic feasibility and defines a future formal test of nonquadratic recovery and the neural residual.

C.6 A CONDITIONAL BRIDGE FROM NORMAL-SPACE ERROR TO PRESERVATION

The preceding bounds concern a static region. To control optimizer drift, normal-space accuracy must also hold at the points visited by the optimizer. The next proposition states this requirement for zero ridge and a fresh Jacobian.

Proposition 13 (Drift under an approximate normal space)**.** *Let $C_\star \colon \mathcal{O} \to \mathbb{R}^K$ be twice continuously differentiable, with full-row-rank Jacobian $\mathbf{J}_\star(\boldsymbol{\theta})$. Let $\widehat{\mathbf{P}}_t$ and $\mathbf{P}_{\star,t}$ project onto the learned and exact normal spaces at $\boldsymbol{\theta}_t$, both orthogonally, and suppose $\|\widehat{\mathbf{P}}_t - \mathbf{P}_{\star,t}\|_2 \le \zeta_t$. For the update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t(\mathbf{I}_d - \widehat{\mathbf{P}}_t)\mathbf{u}_t$, assume its line segment lies in $\mathcal{O}$, $\eta_t \ge 0$, $\|\mathbf{J}_\star(\boldsymbol{\theta}_t)\|_2 \le m_1$, and $\|\mathbf{D}^2 C_\star(\boldsymbol{\theta})[\mathbf{v}, \mathbf{v}]\|_2 \le m_2\|\mathbf{v}\|_2^2$ on that segment. Then*

$$\|C_\star(\boldsymbol{\theta}_{t+1}) - C_\star(\boldsymbol{\theta}_t)\|_2 \le \eta_t m_1 \zeta_t \|\mathbf{u}_t\|_2 + \tfrac{1}{2}\eta_t^2 m_2 \|\mathbf{u}_t\|_2^2. \tag{52}$$

Proof. Let $\mathbf{d}_t = (\mathbf{I}_d - \widehat{\mathbf{P}}_t)\mathbf{u}_t$. Since $\mathbf{J}_\star(\mathbf{I}_d - \mathbf{P}_{\star,t}) = \mathbf{0}_{K \times d}$,

$$\|\mathbf{J}_\star(\boldsymbol{\theta}_t)\mathbf{d}_t\|_2 = \|\mathbf{J}_\star(\boldsymbol{\theta}_t)(\mathbf{P}_{\star,t} - \widehat{\mathbf{P}}_t)\mathbf{u}_t\|_2 \le m_1 \zeta_t \|\mathbf{u}_t\|_2. \tag{53}$$

Taylor's formula with integral remainder bounds the second-order term by $\eta_t^2 m_2 \|\mathbf{d}_t\|_2^2/2$. Orthogonal projection gives $\|\mathbf{d}_t\|_2 \le \|\mathbf{u}_t\|_2$, proving the result. □

Summing Eq. (52) bounds cumulative drift when its hypotheses hold at every step. A uniform support theorem supplies ζ_t while the path remains in its certified region. Static averages describe the sampled support, whereas this proposition addresses pathwise drift. Objective decrease and generalization require their corresponding optimization and statistical analyses. The v15 all-task decision therefore governs the learned-law utility result.

D REMAINING PROOFS

Proof of Theorem 7. We first observe that, whenever C is a first integral and the normalizations are defined,

$$\widehat{\mathbf{f}}(\boldsymbol{\theta})^\top \nabla C(\boldsymbol{\theta}) = 0$$

almost everywhere. Hence both tangency residuals vanish. For $t \neq 0$,

$$\widehat{\nabla(tC)} = \operatorname{sign}(t)\widehat{\nabla C},$$

which implies

$$\ell_{\mathrm{sym}}(tC) = \ell_{\mathrm{sym}}(C).$$

Since $\nabla(tC) = t\nabla C$, direct expansion also gives

$$\ell_{\mathrm{asym}}(tC) = t^2 \ell_{\mathrm{asym}}(C).$$

Finally, condition on $\widehat{\mathbf{f}}(\boldsymbol{\theta})$. Rotational invariance of a direction uniformly distributed on $\mathbb{S}^{d-1}$ yields

$$\mathbb{E}\left[\left(\widehat{\mathbf{f}}(\boldsymbol{\theta})^\top \widehat{\nabla g}(\boldsymbol{\theta})\right)^2 \,\middle|\, \widehat{\mathbf{f}}(\boldsymbol{\theta})\right] = \frac{1}{d}.$$

Taking expectations establishes the remaining claim. □

Proof of Theorem 1. For a smooth vector field $\mathbf{f}$, write $\mathbf{f}C_k := \nabla C_k^\top \mathbf{f}$. By assumption, $\mathbf{f}_e C_k = 0$ for every environment e and law k. Hence, for any pair e, e',

$$[\mathbf{f}_e, \mathbf{f}_{e'}]C_k = \mathbf{f}_e(\mathbf{f}_{e'} C_k) - \mathbf{f}_{e'}(\mathbf{f}_e C_k) = 0. \tag{54}$$

Repeating the argument shows that every iterated bracket in $\mathrm{Lie}(\mathcal{F}_\mathcal{E})$ is annihilated by every row of $\mathbf{J}_C$. Therefore $\mathcal{D}_\mathcal{E}^{(\infty)}(\boldsymbol{\theta})$ is contained in $\ker \mathbf{J}_C(\boldsymbol{\theta})$. Since $\operatorname{rank} \mathbf{J}_C(\boldsymbol{\theta}) = K$, the rank-nullity theorem gives

$$\operatorname{rank} \mathcal{D}_\mathcal{E}^{(\infty)}(\boldsymbol{\theta}) \le \dim \ker \mathbf{J}_C(\boldsymbol{\theta}) = d - K, \tag{55}$$

which proves the first inequality in Eq. (12). The fields and their first brackets belong to the generated Lie algebra, so $\mathcal{D}_{\boldsymbol{\theta}}^{(1)} \subseteq \mathcal{D}_\mathcal{E}^{(\infty)}(\boldsymbol{\theta})$; the second inequality follows. □

Proof of Proposition 5. Every common first integral for $\mathcal{E}_\star$ is also common for $\mathcal{E}_0$ because $\mathcal{E}_0 \subseteq \mathcal{E}_\star$. Conversely, let C be common for $\mathcal{E}_0$. The bracket identity used in the proof of Theorem 1 implies that $\mathrm{d}C$ annihilates $\mathcal{D}_{\mathcal{E}_0}^{(\infty)}(\boldsymbol{\theta})$ at every $\boldsymbol{\theta} \in \mathcal{O}$. Under Eq. (27), each $\mathbf{f}_e(\boldsymbol{\theta})$ with $e \in \mathcal{E}_\star$ belongs to this same distribution. Consequently, $\mathrm{d}C(\boldsymbol{\theta})[\mathbf{f}_e(\boldsymbol{\theta})] = 0$ for every $e \in \mathcal{E}_\star$, proving that C is common for the larger family. □

Proof of Corollary 6. The Frobenius theorem gives local coordinates $(z_1, \ldots, z_d)$ in which $\mathcal{D} = \mathrm{span}\{\partial_{z_1}, \ldots, \partial_{z_r}\}$. Therefore $z_{r+1}, \ldots, z_d$ have linearly independent differentials and are annihilated by every section of $\mathcal{D}$, hence by every field in $\mathcal{F}_\mathcal{E}$. They are $d - r$ independent common first integrals. Conversely, Theorem 1 bounds any independent common family by $d - r$, so the count is exact. □

Theorem 14 (Discrete conservation error of gradient descent)**.** *Let $\mathcal{O} \subseteq \mathbb{R}^d$ be open, let $C \in \mathcal{C}^3(\mathcal{O})$ be an exact first integral of* (1) *on $\mathcal{O}$, and suppose the update remains in $\mathcal{O}$. Under gradient descent $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \epsilon \nabla L(\boldsymbol{\theta}_k)$,*

$$C(\boldsymbol{\theta}_{k+1}) - C(\boldsymbol{\theta}_k) = \frac{\epsilon^2}{2} \nabla L(\boldsymbol{\theta}_k)^\top \nabla^2 C(\boldsymbol{\theta}_k) \nabla L(\boldsymbol{\theta}_k) + O(\epsilon^3)$$

as $\epsilon \to 0$ with $\boldsymbol{\theta}_k$ fixed.

Proof. Taylor's theorem at $\boldsymbol{\theta}_k$ in the direction $-\epsilon \nabla L(\boldsymbol{\theta}_k)$ gives

$$\begin{aligned} C(\boldsymbol{\theta}_{k+1}) - C(\boldsymbol{\theta}_k) = & -\epsilon \nabla C(\boldsymbol{\theta}_k)^\top \nabla L(\boldsymbol{\theta}_k) \\ & + \frac{\epsilon^2}{2} \nabla L(\boldsymbol{\theta}_k)^\top \nabla^2 C(\boldsymbol{\theta}_k) \nabla L(\boldsymbol{\theta}_k) + O(\epsilon^3). \end{aligned}$$

The first-order term vanishes because C is a first integral. Consequently, the displayed expansion follows. □

Proposition 15 (Conservation normal is a Hessian null vector)**.** *Let $\mathcal{O} \subseteq \mathbb{R}^d$ be open, let $C, L \in \mathcal{C}^2(\mathcal{O})$, and assume $\nabla C(\boldsymbol{\theta})^\top \nabla L(\boldsymbol{\theta}) = 0$ throughout $\mathcal{O}$. If $\boldsymbol{\theta}^\star \in \mathcal{O}$ and $\nabla L(\boldsymbol{\theta}^\star) = 0$, then $\nabla^2 L(\boldsymbol{\theta}^\star) \nabla C(\boldsymbol{\theta}^\star) = 0$. If also $\nabla C(\boldsymbol{\theta}^\star) \neq \mathbf{0}_d$, then $\nabla^2 L(\boldsymbol{\theta}^\star)$ is singular.*

Proof. Differentiating the assumed identity with respect to $\boldsymbol{\theta}$ yields

$$\nabla^2 C(\boldsymbol{\theta}) \nabla L(\boldsymbol{\theta}) + \nabla^2 L(\boldsymbol{\theta}) \nabla C(\boldsymbol{\theta}) = \mathbf{0}_d.$$

At $\boldsymbol{\theta}^\star$, the first term vanishes by $\nabla L(\boldsymbol{\theta}^\star) = \mathbf{0}_d$, which proves the null-vector identity. If $\nabla C(\boldsymbol{\theta}^\star)$ is nonzero, the Hessian has a nontrivial kernel and is therefore singular. □

Proof of Proposition 8. Write $\mathbf{J}(s) := \mathbf{J}_{\widetilde{C}}(\Theta(s))$. The eigenvalues of $\mathbf{J}(s)\mathbf{J}(s)^\top$ are the squared singular values of $\mathbf{J}(s)$. Hence the assumed margin gives

$$\lambda_{\min}\big(\mathbf{J}(s)\mathbf{J}(s)^\top\big) = \sigma_K(\mathbf{J}(s))^2 \geq \gamma^2 \tag{56}$$

uniformly in s, which proves Eq. (34). Matrix multiplication and inversion are continuous on the set of matrices whose smallest singular value is bounded below by γ. Since $\widetilde{C} \in \mathcal{C}^2$, its Jacobian is locally Lipschitz; compactness of $\Theta([0,T])$ and the uniform spectral gap therefore make

$$\mathbf{P}(s) = \mathbf{I}_d - \mathbf{J}(s)^\top \left(\mathbf{J}(s)\mathbf{J}(s)^\top\right)^{-1} \mathbf{J}(s) \tag{57}$$

continuous along the path and locally Lipschitz on a neighborhood of it.

The necessity of the rank assumption is illustrated by $K = 1$, $d = 2$, and

$$\mathbf{J}_\varepsilon = [\varepsilon \quad 0], \qquad \mathbf{J}'_\varepsilon = [0 \quad \varepsilon], \qquad \varepsilon > 0. \tag{58}$$

Both sequences converge to the same rank-zero matrix, while their zero-ridge projectors are respectively $\mathrm{diag}(0,1)$ and $\mathrm{diag}(1,0)$. Their operator-norm difference is one, which establishes direction dependence at the common limit. □

Proof of Theorem 3. Set

$$\mathbf{G}_t := \mathbf{J}_t\mathbf{J}_t^\top + \delta\mathbf{I}_K.$$

The identity

$$\mathbf{J}_t\mathbf{J}_t^\top\mathbf{G}_t^{-1} = \mathbf{I}_K - \delta\mathbf{G}_t^{-1}$$

and the CCA update in Eq. (19) imply

$$\mathbf{J}_t\mathbf{d}_t = \mathbf{J}_t\mathbf{u}_t - \mathbf{J}_t\mathbf{J}_t^\top\mathbf{G}_t^{-1}\mathbf{J}_t\mathbf{u}_t = \delta\mathbf{G}_t^{-1}\mathbf{J}_t\mathbf{u}_t.$$

Consequently, Taylor's theorem applied to $C_k(\boldsymbol{\theta}_t - \eta_t\mathbf{d}_t)$ gives

$$\begin{aligned} C_k(\boldsymbol{\theta}_{t+1}) - C_k(\boldsymbol{\theta}_t) = & -\eta_t\delta\left[\mathbf{G}_t^{-1}\mathbf{J}_t\mathbf{u}_t\right]_k \\ & + \frac{\eta_t^2}{2}\mathbf{d}_t^\top\nabla^2 C_k(\boldsymbol{\theta}_t)\mathbf{d}_t + O(\eta_t^3), \end{aligned}$$

which is the stated formula. If $\delta = 0$, the first-order term vanishes identically. □

Proof of Proposition 9. Let $\mathbf{X} := -\nabla L$. Since ∇L is nonzero on $\mathcal{O}$, $\mathbf{X}$ is a nonvanishing smooth vector field there. The flow-box theorem therefore yields, for each $\boldsymbol{\theta}_0 \in \mathcal{O}$, a neighborhood $\mathcal{U} \subseteq \mathcal{O}$ with smooth coordinates $(s, y_1, \ldots, y_{d-1})$ in which $\mathbf{X} = \partial_s$. Define $C_i := y_i$ on $\mathcal{U}$. Then

$$\mathbf{X}C_i = 0,$$

and hence

$$\nabla C_i(\boldsymbol{\theta})^\top\nabla L(\boldsymbol{\theta}) = 0$$

on $\mathcal{U}$. Since $(s, y_1, \ldots, y_{d-1})$ is a coordinate chart, the functions $C_1, \ldots, C_{d-1}$ are functionally independent. Fixing their values fixes $(y_1, \ldots, y_{d-1})$ and leaves only the s-coordinate free; the resulting common level set is therefore a local one-dimensional flow leaf. □

E EXACT CONSERVATION LAWS IN THE ATTENTION CONTROL

For a bias-free single attention head, let $\mathbf{X} \in \mathbb{R}^{n\times d_m}$, $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_m\times d_h}$, and $\mathbf{W}_O \in \mathbb{R}^{d_h\times d_o}$, and define

$$\mathbf{A}_{\mathbf{X}}(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O) := \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top}{\sqrt{d_h}}\right)\mathbf{X}\mathbf{W}_V\mathbf{W}_O. \tag{59}$$

This symmetry gives the exact premise used by the optimizer control.

Proposition 16 (Cross-tensor conservation in softmax attention)**.** *Let L be any differentiable scalar loss whose parameter dependence is only through one or more maps $\mathbf{A}_{\mathbf{X}}$. Along Euclidean gradient flow in the four projection matrices,*

$$q_{\mathrm{QK}} := \|\mathbf{W}_Q\|_F^2 - \|\mathbf{W}_K\|_F^2, \qquad q_{\mathrm{VO}} := \|\mathbf{W}_V\|_F^2 - \|\mathbf{W}_O\|_F^2 \tag{60}$$

are first integrals.

Proof. For every $a > 0$, the transformations $(\mathbf{W}_Q, \mathbf{W}_K) \mapsto (a\mathbf{W}_Q, a^{-1}\mathbf{W}_K)$ and $(\mathbf{W}_V, \mathbf{W}_O) \mapsto (a\mathbf{W}_V, a^{-1}\mathbf{W}_O)$ leave $\mathbf{A}_\mathbf{X}$ unchanged for every input $\mathbf{X}$. Differentiating the first invariance identity at $a = 1$ gives

$$\left\langle \nabla_{\mathbf{W}_Q} L, \mathbf{W}_Q \right\rangle_F - \left\langle \nabla_{\mathbf{W}_K} L, \mathbf{W}_K \right\rangle_F = 0. \tag{61}$$

Consequently, under $\dot{\mathbf{W}} = -\nabla_\mathbf{W} L$,

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{\mathrm{QK}} = -2 \left(\left\langle \nabla_{\mathbf{W}_Q} L, \mathbf{W}_Q \right\rangle_F - \left\langle \nabla_{\mathbf{W}_K} L, \mathbf{W}_K \right\rangle_F \right) = 0. \tag{62}$$

Applying the same argument to the second transformation yields $\mathrm{d}q_{\mathrm{VO}}/\mathrm{d}t = 0$. □

The proposition applies to objectives that preserve both rescaling symmetries; the reported attention control uses such an objective. Its laws are supplied analytically.

F ANALYTIC-PREMISE AUDIT

The audit in Fig. 4 tests the law premise before any optimizer is evaluated. For each of five fixed seeds, a bias-free $(2, 3, 2)$ network is paired with an independently drawn 50-example MSE environment and an initial parameter vector sampled uniformly from $[-2, 2]^{12}$. We evaluate the three balance candidates collected as $\mathbf{q} = (q_1, q_2, q_3)^\top$ and write $\mathbf{J}_\mathbf{q} := \nabla_{\boldsymbol{\theta}} \mathbf{q}$. Their normalized infinitesimal residual is

$$\rho_\mathbf{q}(\boldsymbol{\theta}) := \frac{\|\mathbf{J}_\mathbf{q}(\boldsymbol{\theta}) \nabla L(\boldsymbol{\theta})\|_2}{\max\{\|\mathbf{J}_\mathbf{q}(\boldsymbol{\theta})\|_F \|\nabla L(\boldsymbol{\theta})\|_2, 10^{-30}\}} \tag{63}$$

and the relative change in the diagonal q_i values after integrating to the fixed horizon $T = 0.5$. The denominator floor matches the implementation. The three Euler step sizes are specified before the trajectories are evaluated. The linear benchmark's full six-dimensional Gram law follows from the corresponding matrix identity; this numerical audit uses its three diagonal members so that the same quantities can be compared across linear, ReLU, and GELU networks.

Table 7: Balance-candidate premise audit. Entries are mean ± one population standard deviation (divisor five) across five matched seeds. The last three columns report fixed-horizon relative drift. Refinement removes Euler error for the exact linear/ReLU laws, while GELU drift remains stable.

Activation	$\rho_\mathbf{q}$	$h = 10^{-3}$	$h = 5\times10^{-4}$	$h = 2.5\times10^{-4}$
Linear	$(1.98\pm0.46)\times10^{-8}$	$(7.35\pm5.01)\times10^{-4}$	$(3.66\pm2.50)\times10^{-4}$	$(1.83\pm1.24)\times10^{-4}$
ReLU	$(1.12\pm0.89)\times10^{-8}$	$(4.02\pm2.57)\times10^{-4}$	$(2.01\pm1.29)\times10^{-4}$	$(1.02\pm0.64)\times10^{-4}$
GELU	$(2.46\pm1.99)\times10^{-2}$	$(5.45\pm2.34)\times10^{-2}$	$(5.45\pm2.35)\times10^{-2}$	$(5.45\pm2.36)\times10^{-2}$

The linear and ReLU trajectories display the $O(h)$ fixed-horizon global error of the Euler scheme. The GELU curve is flat under refinement because its continuous-time tangency residual is nonzero. A single finite-step drift value could confound these two effects; the refinement sequence separates them. The result identifies GELU as the null control for this homogeneous-rescaling family. Other law families and optimizer utility have their own evaluations.

The supplement provides the executable audit, all five seed records, and the deterministic figure renderer. The reproduction notebook checks the stored records before reconstructing the table and figure.

G FROZEN CERTIFICATE AND PRESERVATION AUDITS

Table 8: Evidence map. Each question uses its corresponding frozen evidence and decision rule.

Question	Admissible evidence	Decision
Are the supplied balance quantities exact laws?	Tangency and Euler refinement	Linear/ReLU: yes; GELU: null control for this family
Does the FULL certificate separate the declared boundary?	Held-out certificate cohort	40/40 seed-family decisions; pass
Which components correct recorded errors?	Matched staircase and retrospective removal	Bracket and capacity remove distinct errors
Does learned-law CCA pass its utility gate?	Frozen deployment audit, 60 tasks and 20 clusters	Formal status: FAIL; four arm abstentions and three random-arm target misses
Does CCA preserve a supplied exact law?	Exact-law attention control	Supported in that control

Table 9: Protocol chronology. Each row answers a distinct frozen question; development, stopped, and formal cohorts retain separate provenance.

Stage	Question	Outcome and evidentiary role
v1–v2	Can prespecified candidates recover known linear/ReLU row spaces?	Conditional recovery on early confirmation cohorts; historical only.
v3	Can a blind overcomplete-K grid recover the law count?	Formal status: FAIL; learned-law CCA remained locked.
v5	Does first-bracket closure repair the blind count gate?	Formal status: FAIL at the linear stop rule.
v10–v13	Which fitting rules, thresholds, and candidate artifacts enter the successor test?	Development and freeze stages.
v14	Does the full certificate separate each declared K_0 versus $K_0 + 1$ boundary and reject null controls?	Formal pass: 20 positive-family boundary decisions and 20 null-control rejections.
v15	Do v14-certified learned laws pass the pathwise CCA utility gate?	Formal status: FAIL; complete-cluster ratios remain descriptive.

G.1 HELD-OUT CERTIFICATE COHORT

The v14 protocol inherits the v13 target architectures, candidate class, training schedule, and numerical thresholds. Its sole methodological addition is the first-bracket capacity statistic in Eq. (14). The formal run uses ten seeds $\{22003, 22109, 22229, 22307, 22409, 22511, 22613, 22717, 22807, 22901\}$. For each seed, all six candidate artifacts (linear $K = 6, 7$, ReLU $K = 3, 4$, GELU $K = 1$, and SiLU $K = 1$) are fitted and hashed before the outer support or any of the 66 first brackets is constructed. The global freeze contains 40 seed-family records and 60 model files; all hashes remain unchanged after outer evaluation.

ReLU fields are piecewise smooth. Automatic differentiation evaluates the classical bracket within an activation-pattern cell. Exact kinks have probability zero under continuous sampling. The ReLU result is stratumwise and assumes differentiability at the sampled points; its balance laws hold almost everywhere under Euclidean gradient flow.

Table 10: Complete fitting and decision specification for the frozen v14 cohort. The table reports the values fixed before outer evaluation.

Quantity	Frozen value
Target and environments	Bias-free $(2, 3, 2)$ network ($d = 12$); each task has 50 independently sampled standard-normal inputs and regression targets.
Candidate map	$\mathbf{A}[\boldsymbol{\theta}; (\theta_i\theta_j)_{i\le j}] + \mathbf{b} + r_{\boldsymbol{\psi}}(\boldsymbol{\theta})$; $r_{\boldsymbol{\psi}}$ has three width-128 SiLU hidden layers and a zero-initialized final affine layer.
Candidate optimization	Adam, learning rate 10^{-3}, batch size 256, at most 5000 updates, evaluation every 100 updates, patience 50, $\omega = \lambda_{\mathrm{frame}} = 1$, CPU float32.
Development rule	Retain the checkpoint minimizing development tangency plus frame error, then freeze every declared K.
Support	1024/512/96 train/development/outer points. The generator uses nominal weights $0.4/0.4/0.2$ for $\mathcal{N}(\mathbf{0}, 0.5^2\mathbf{I})$, a random-direction shell with radius $\mathrm{Unif}[0.5, 2]$, and calibration-flow states from 1–8 Euler steps of size 10^{-3}. Integer allocation uses rounding; the 96-point outer split is 38/38/20.
Outer directions	12 fresh environments and all 66 unordered first brackets. Fields, brackets, and decision linear algebra use float64; Jacobians from the frozen float32 candidates are cast to float64 before certification.
Normalization and zero rule	Unit fields $\mathbf{f}/\|\mathbf{f}\|_2$; row-normalization floor $\epsilon = 10^{-12}$; the construction retains field and bracket directions with norm above 10^{-10}.
Field rule	$\rho_{\mathrm{fld}} \le 0.02$ and $m_{\mathrm{row}} > \tau_{\mathrm{row}} = 10^{-8}$.
Rank rule	Relative singular-value tolerance $\tau_{\mathrm{rank}} = 0.05$ and $\widehat{p}_{\mathrm{full}} \ge 0.80$.
Bracket and capacity rules	$\rho_{\mathrm{br}} \le 0.05$ and $\widehat{p}_{\mathrm{cap}}(K) \ge 0.80$; capacity uses the same 0.05 relative-rank tolerance.
Analytic span audit	Pointwise SVD rank tolerance 10^{-6}; maximum allowed mean row-space error 0.02.
Cohort gate	At least $9/10$ declared-boundary decisions per positive family and at most $1/10$ false positives per null family, together with the paired and integrity conditions in the frozen protocol.
Threshold provenance	Tangency, bracket, rank, and support-fraction thresholds were fixed in the v10 development protocol; the row floor was fixed in v11. They were carried unchanged through v13/v14. The v14 capacity rule reused the existing rank tolerance and support fraction.

For completeness, the analytic span diagnostic is defined pointwise. Let $\mathbf{U}_b \in \mathbb{R}^{d\times r_c}$ and $\mathbf{V}_b \in \mathbb{R}^{d\times r_a}$ be column-orthonormal bases for the candidate and analytic row spaces, and let $r = \min(r_c, r_a)$. If $r = 0$, set $e_b = 1$; otherwise the implemented error is

$$e_b = \min\left\{1, 1 - \frac{\|\mathbf{U}_b^\top \mathbf{V}_b\|_F^2}{r} + \frac{|r_c - r_a|}{\max(r_c, r_a)}\right\}, \qquad e_{\mathrm{span}} = \frac{1}{B}\sum_{b=1}^{B} e_b. \tag{64}$$

When both row spaces have dimension K, this is the mean squared sine of their principal angles. With $\zeta_b = \|\mathbf{U}_b\mathbf{U}_b^\top - \mathbf{V}_b\mathbf{V}_b^\top\|_2$, one has $\zeta_b^2/K \le e_b \le \zeta_b^2$. A uniform projector bound therefore controls the reported span error. Pathwise use requires the uniform margin in Proposition 8. The observed margins are well separated from the frozen cutoffs. Across all 20 accepted K_0 candidates, the largest field and bracket residuals are 1.924×10^{-4} and 2.85×10^{-4}, the smallest pointwise-rank fraction is 0.96875, and every capacity fraction is 1. All 20 null candidates pass Rank but have bracket residual in $[0.0640, 0.1248]$, above the 0.05 cutoff. The six above-boundary candidates that pass Bracket have capacity fractions in $[0.3229, 0.6458]$, below 0.80. These are full-cohort extrema reported after the frozen decisions and serve as descriptive margins. Recomputing the intersection of the pointwise-rank and capacity indicators from all accepted positive tensors gives a minimum joint fraction of 0.96875; every capacity fraction is one. This retrospective check confirms overlap on the recorded positives. The formal decision retains the original two marginal 80% screens.

The supplement contains the frozen protocol, immutable aggregate, and 211-entry integrity manifest. Deterministic replay verifies every recorded entry and reproduces the reported decisions.

G.2 Learned-law preservation gate

The v15 candidate barrier contains every v14 analytic-boundary linear and ReLU model. Before optimization, a second barrier stores three fresh utility tasks for each model, giving 60 tasks. Each task has 64 examples, an initialization on the radial shell $[0.8, 1.0]$, and a teacher generated by 50 RK4 steps of size 0.01 along a vector field tangent to the analytic level sets. A task is accepted only when its maximum analytic-law error is at most 10^{-9}. All five arms use float64 and the same Adam hyperparameters: learning rate 0.005, default moment coefficients $(0.9, 0.999)$, numerical

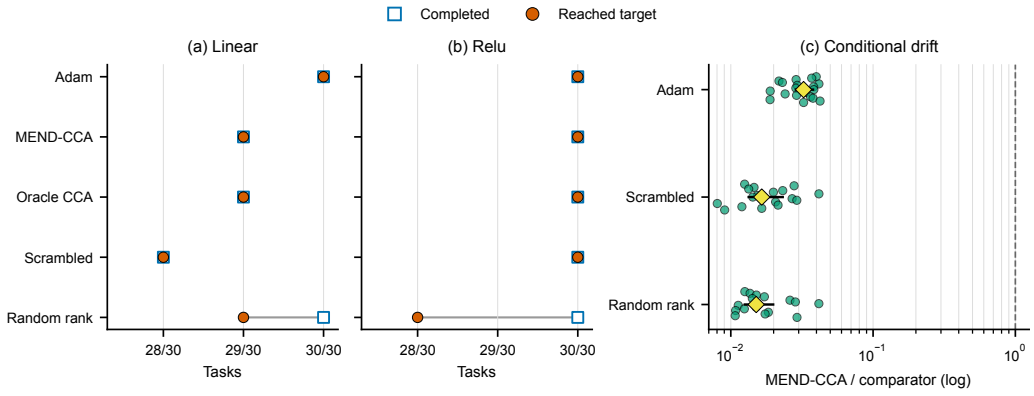


Figure 5: Pathwise events determining the learned-law preservation gate. Panels (a)–(b) compare completed arms (open squares) with arms reaching the matched loss target (filled circles) among 30 frozen tasks per family. Coincident markers indicate equal counts. Panel (c) shows every stored complete-cluster MEND-CCA drift ratio, together with the recorded median and bootstrap interval; the dashed line is equal drift. The complete-cluster panel covers 19, 17, and 16 of the 20 clusters because an arm aborted or missed its target. The prespecified all-cluster criterion therefore has formal status FAIL; the conditional ratios remain descriptive.

constant 10^{-8}, 500 updates, and the matched target $0.2L(\boldsymbol{\theta}_0)$. The four constraint arms use zero-ridge orthogonal projection, refresh their Jacobian every step, and apply the relative-rank tolerance 0.05; plain Adam supplies the unprojected reference. Every arm is subject to the operational support limit $\|\boldsymbol{\theta}\|_2 \leq 2$.

All 60 task files, 20 candidate hashes, and 300 arm records are retained. Four arms abort under the declared conditioning safeguard: MEND-CCA and oracle CCA on the same linear task $(22229, 15102528)$, plus scrambled-MEND CCA on $(22409, 15104326)$ and $(22511, 15105528)$. The shared oracle/MEND task begins with analytic relative minimum singular value 0.0723, only slightly above the frozen 0.05 threshold, and crosses the threshold during optimization. Three otherwise finite random-rank arms miss the matched target. The record retains every scheduled task and arm.

Table 11: Conditional v15 drift summaries. Effects use clusters with all three paired task endpoints; the prespecified criterion requires 20/20 complete clusters. Ratios below one favor MEND-CCA. The all-cluster criterion has formal status FAIL; the complete-cluster statistics are descriptive.

Comparator	Complete clusters	Favorable	Median drift ratio	Gate
Adam	19/20	19/19	0.0325	FAIL
Scrambled MEND-CCA	17/20	17/17	0.0165	FAIL
Random-rank CCA	16/20	16/16	0.0151	FAIL

The supplement contains the frozen protocol, the complete 60-task aggregate, the candidate and task barriers, and a 134-entry integrity manifest. Replay verifies the recorded status and reconstructs the conditional summaries.

H REPRODUCIBILITY AND FROZEN DECISIONS

The anonymized supplementary material contains the implementation, executable run scripts, dependency specification, and seed-resolved artifacts used for the reported controls. The v14 protocol fixes the candidate family, training schedule, seeds, support splits, bracket construction, numerical thresholds, and cohort rule before outer evaluation. Its candidate freeze precedes every outer tensor. The v15 protocol then consumes exactly the 20 v14 analytic-boundary positive-control candidates, freezes a global candidate barrier, and freezes all 60 utility tasks before executing any of the five op-

Algorithm 1 Discover → Certify → Preserve decision protocol

Require: Training and development fields, protocol-declared candidate dimensions $\mathcal{K}$, certificate thresholds, and outer cohort rules

for $K \in \mathcal{K}$ **do**
    Fit $C_\phi^{(K)}$ on training environments by Eq. (7)
    Select the checkpoint by development objective and freeze the declared candidate
end for
Serialize and hash every declared candidate, the protocol, and all source fingerprints before constructing outer tensors
Materialize held-out fields and first Lie brackets
for each frozen candidate **do**
    Evaluate Field, Rank, Bracket, and Full certificates, where Full adds Eq. (14)
end for
Apply the frozen architecture and cohort rules without retuning
Record the analytic row-space audit separately from data-only decisions
if the certificate-transfer gate passes **then**
    Freeze the learned-candidate cohort and every utility task
    Execute all matched optimizer arms and apply the separate utility gate
    if the utility gate status is FAIL **then**
        Retain every abstention and report the formal status
    end if
else
    Retain the certificate-transfer status and keep utility tasks locked
end if

timizer arms. Both formal runs retain source fingerprints, protocol hashes, freeze records, safeguard events, and complete manifests. Historical v3 and v5 runs retain separate provenance from the v14 and v15 decisions.

The certificate and preservation bundles retain the frozen protocols, candidate barriers, source fingerprints, seed-level records, and complete manifests. The preservation aggregate contains 60 task records and 300 arm records, including every safeguard activation and target miss. The exact-law attention control retains its three seed records and original loop indexing. The reproduction notebook verifies these records and recomputes Table 2, using the sample standard deviation for the last recorded loss and seed means for final-parameter drift.

The analytic-audit separation in Algorithm 1 describes information flow. In the consumed executable, the inherited evaluator computes analytic row-space diagnostics before the capacity stage attaches the final FULL flag. Those diagnostics are reserved for the aggregate positive-control criterion, which subsequently includes analytic span recovery. This is the implemented information flow for the frozen cohort.

H.1 SINGLE-ENVIRONMENT DIAGNOSTIC

A distinct frozen diagnostic uses one shared GELU environment for train/development and 12 fresh outer environments. It is reported as a separate diagnostic with its own objective and budget. Its purpose is to show the gap between a single-field tangency fit and out-of-environment confirmation.

Algorithm 2 CCA algebraic update and orthogonal-mode conditioning safeguard

Require: Loss L, constraint-Jacobian routine $\widetilde{\mathbf{J}}(\cdot)$, step sizes $\{\eta_t\}$, mode $m \in \{\textsc{Orthogonal}, \textsc{Ridge}\}$, ridge $\delta = 0$ in orthogonal mode and $\delta > 0$ in ridge mode, refresh period T_J, and orthogonal-mode tolerance $\tau_{\mathrm{rank}} > 0$

Initialize $\boldsymbol{\theta}_0$, $\mathbf{m}_0 \leftarrow \mathbf{0}_d$, and $\mathbf{v}_0 \leftarrow \mathbf{0}_d$
for $t = 0, 1, 2, \ldots$ **do**
  $\mathbf{g}_t \leftarrow \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$
  Update $(\mathbf{m}_{t+1}, \mathbf{v}_{t+1}, \mathbf{u}_t)$ by Eq. (31)
  if $\mathbf{u}_t$ is nonfinite **then**
    return abstention with the recorded reason
  end if
  if $t = 0$ **or** $t \bmod T_J = 0$ **then**
    $\widetilde{\mathbf{J}}_t \leftarrow \widetilde{\mathbf{J}}(\boldsymbol{\theta}_t)$
  else
    $\widetilde{\mathbf{J}}_t \leftarrow$ cached Jacobian from the last refresh
  end if
  if $\widetilde{\mathbf{J}}_t$ is nonfinite **then**
    return abstention with the recorded reason
  end if
  if $m = \textsc{Orthogonal}$ **then**
    if $K > d$ **then**
      return abstention with the recorded reason
    end if
    Compute $\sigma_1 \geq \cdots \geq \sigma_K$ of $\widetilde{\mathbf{J}}_t$
    if $\sigma_1 = 0$ or $\sigma_K \leq \tau_{\mathrm{rank}} \sigma_1$ **then**
      return abstention with the recorded reason
    end if
    Compute a thin QR factorization $\widetilde{\mathbf{J}}_t^\top = \mathbf{Q}_t \mathbf{R}_t$
    $\mathbf{d}_t \leftarrow \mathbf{u}_t - \mathbf{Q}_t(\mathbf{Q}_t^\top \mathbf{u}_t)$
  else
    Factor $\mathbf{G}_t \leftarrow \widetilde{\mathbf{J}}_t \widetilde{\mathbf{J}}_t^\top + \delta \mathbf{I}_K$
    Solve $\mathbf{G}_t \mathbf{a}_t = \widetilde{\mathbf{J}}_t \mathbf{u}_t$ and set $\mathbf{d}_t \leftarrow \mathbf{u}_t - \widetilde{\mathbf{J}}_t^\top \mathbf{a}_t$
  end if
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t \mathbf{d}_t$
end for