CONVERGENT EVOLUTION OF NEURAL NETWORKS: ZIPPERING AND WEAK-STRONG EQUIVALENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Why and when should independently trained neural networks develop the same internal mechanisms? We develop a theory of this convergent evolution through *identifiability*: what a network's end-to-end function determines about its internal computation. Our *zippering* theorems establish when agreement at the end forces corresponding representations throughout a nonlinear network, up to linear transformations. Our *weak-strong equivalence* theorems then show when this correspondence forces individual channels and gates to recur. Tasks that require a greater fraction of these components force stronger convergence. We prove exact and approximate versions of both results. Most strikingly, when our theory is applied to modern transformer architectures, it predicts *privileged attention heads*: individual head computations that recur across models. We find privileged heads and gates in both vision and language, ordered alignment across unequal depths, and stronger recurrence of more-used components. Task constraints can therefore force a shared computational hierarchy and reproducible mechanisms across models trained on the same tasks. These results thus give mechanistic interpretability stable structures to explain and AI safety research recurring targets for intervention in models spanning different tasks, modalities, and architectures.

1 INTRODUCTION

Over the past fifteen years, studies of deep neural networks have repeatedly found a form of *convergent evolution*: networks trained from different random seeds learn similar representations, arrange them in the same order across depth, and sometimes recover individually similar features (Li et al., 2015; Raghu et al., 2017; Kornblith et al., 2019; Khosla & Williams, 2023; Kapoor et al., 2025). Therefore, a natural question from these results is: *Why* should independent learning produce the same internal solutions, and can we characterize under what condition(s) does such internal correspondence arise in modern architectures and tasks?

This central question of representational convergence is important to several distinct fields: identifiability theory, mechanistic interpretability, and NeuroAI. In identifiability, we are concerned with what function the network computes end-to-end to determine about what is present internally, and when that function determines the units or parameters, up to computation-preserving transformations (Fefferman & Markel, 1993; Phuong & Lampert, 2020; Vlačić & Bölcskei, 2021; Bona-Pellissier et al., 2023). Prior results are specific to particular activation families and, in some cases, restricted architectures: they include complete identifiability for arbitrary-depth tanh-type networks (Vlačić & Bölcskei, 2021), open sets of identifiable parameters in broad classes of arbitrary-depth ReLU networks (Grillo & Montúfar, 2026), and complete symmetry classifications for three-layer bottleneck ReLU networks (Gegenfurtner et al., 2026). We connect identifiability to task-driven convergence, deriving exact and approximate bounds on how task agreement between networks propagates through depth and how the fraction of nonlinear components required by a task constrains their recurrence, with broad applicability to modern architectures *across* nonlinearities and without fixing every parameter ahead of time. We then empirically test the predictions of our theory in trained convolutional networks and transformers on vision and language tasks.

Mechanistic interpretability seeks to explain behavior through the computations of particular components, such as attention heads (Elhage et al., 2021). But does independent optimization reliably produce that mechanism at all? Otherwise, an explanation of one trained network may tell us little

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.

about another. For control and AI safety, identifying computations makes it possible to target them: interventions can change factual predictions and refusal behavior (Meng et al., 2022; Arditi et al., 2024). Finally, NeuroAI asks whether the same demands on artificial and biological systems can force shared internal solutions (Yamins & DiCarlo, 2016; Zador et al., 2023). The recurrence of individual features makes this convergent-evolution hypothesis especially striking (Khosla et al., 2024).

Our strategy has a global part and a local part. Globally, we ask why networks of the same architecture, differing in initialization and training order, should match throughout their hierarchies when they agree at the end. Here we allow an affine map to predict one representation from linear combinations of units in the other, which we call *weak alignment*. Locally, we ask how much finer-grained structure this agreement forces. *Strong alignment* requires individual channels or nonlinear gates to match one-to-one, up to elementary architectural symmetries. These reproducible coordinates are *privileged axes*: their identity matters, not just the information their mixtures carry. We prove two sets of results:

1. **Zippering.** For layerwise-minimal networks whose nonlinear expansions identify their inputs, agreement at a terminal representation forces agreement at the preceding layer. Repeating this argument *zippers* alignment backward through the hierarchy. End-to-end agreement can therefore determine the sequence of internal computations.
2. **Weak–strong equivalence.** Weak alignment on both sides of a coordinatewise nonlinearity forces every task-used axis to match. A used ReLU gate leaves a kink trace, and a Softplus gate leaves a characteristic curvature ridge. Mixing its coordinate moves that signature and breaks downstream task agreement. The fewest axes any accurate solution requires therefore *lower-bounds* the strongly aligned fraction.

We give exact and approximate versions of both results. Together, they formalize the prior informal (and philosophical) *contravariance principle*: harder tasks admit fewer adequate solutions, making those solutions more alike (Cao & Yamins, 2024). We formalize this through task-required nonlinear computation and the redundant or canceling computations that layerwise minimality excludes.

Most strikingly, our theory predicts *privileged attention heads* in transformers (Vaswani et al., 2017), not just privileged MLP gates. We find these privileged heads in transformers trained on either vision or language: learned gates and heads recur more strongly than random mixtures carrying the same information or total attention update. Attention head computations matter, rather than the query/key or value/output coordinates used to express them. Across AlexNet, ResNets, vision transformers (ViTs) (Dosovitskiy et al., 2020), and Pythia language models (Biderman et al., 2023), we also find our theory's predicted alternating alignment, ordered layer correspondences across unequal depths, and stronger recurrence of more task-used components. We explain how tasks force recurring mechanisms within artificial networks and potentially between models and brains.

2 Related work

Identifiability from the input–output function. Fefferman & Markel (1993) studied recovery of a network from its output. Phuong & Lampert (2020) construct ReLU networks identifiable up to permutation and positive rescaling, while Bona-Pellissier et al. (2023) give sufficient conditions for parameter recovery from the function on a domain. Vlačić & Bölcskei (2021) relate identifiability to activation symmetries and the absence of nontrivial zero-output networks. For finite samples, Bona-Pellissier et al. (2022) characterize local identification of deep ReLU parameters through geometric and Jacobian-rank conditions. We connect identifiability to *task-optimization demands*, namely: hierarchical convergence, task-used component recurrence, and transformer heads. Finally, while Nayebi (2026, Corollary 5) showed that two minimal systems solving the same tasks optimally must admit an invertible mapping, our results relate more directly to modern DNN architectures, and identify the *complexity* of this mapping as linear for ReLU and Softplus DNNs.

Interpretable mechanisms and control. Circuit analysis explains attention-head computations (Elhage et al., 2021), and causal interventions enable factual edits (Meng et al., 2022) and alter refusal (Arditi et al., 2024). We ask when the task and architecture require the components we interpret or intervene on to recur across trained models.

Representational convergence and NeuroAI. Linear prediction, canonical correlation, CKA, and soft matching reveal shared representations (Li et al., 2015; Raghu et al., 2017; Kornblith et al.,

2019; Khosla & Williams, 2023; Raghu et al., 2021), while privileged-axis comparisons ask whether individual features recur (Khosla et al., 2024). Goal-driven NeuroAI asks these questions about models and brains (Yamins & DiCarlo, 2016), including functional organization in language models (Rathi et al., 2025). Our theory explains when task demands force this convergence and why metric disagreement need not imply different computations (Soni et al., 2024; Bo et al., 2024). We show that canonicalized RSA and CKA recover shared task geometry after removing task-irrelevant scaling and redundancy, and establish quantitative links to ridge and Procrustes (Supp. Sec. E). Huang et al. (2025) empirically find that harder RNN tasks produce more similar dynamics even as weights diverge, as our contravariance theory predicts. Somewhat relatedly, the *Platonic Representation Hypothesis* (PRH) instead concerns convergence across *different* tasks or modalities in a common world (Huh et al., 2024), whereas contravariance concerns the *same task*. Our results therefore identify what a theory of PRH must explain: the constraints different tasks inherit from the world.

3 Zippering and weak–strong equivalence

Let $\Omega \subseteq \mathbb{R}^{d_0}$ be the task-input set, where d_0 is the input dimension, and let μ be a sampling probability distribution on Ω. For a neural network $N \in \{A, B\}$, layer ℓ has width d_ℓ^N, preactivation $z_\ell^N : \Omega \to \mathbb{R}^{d_\ell^N}$, and activation

$$z_\ell^N(x) = W_\ell^N h_{\ell-1}^N(x) + b_\ell^N \in \mathbb{R}^{d_\ell^N}, \qquad h_\ell^N(x) = \varphi(z_\ell^N(x)), \tag{1}$$

where $W_\ell^N \in \mathbb{R}^{d_\ell^N \times d_{\ell-1}^N}$, $b_\ell^N \in \mathbb{R}^{d_\ell^N}$, and φ acts coordinatewise. We prove the exact results for ReLU $\rho(t) = \max\{t, 0\}$ and Softplus $\sigma(t) = \log(1 + e^t)$. Supp. Sec. A.2 develops their adjacent-layer theory, and Supp. Sec. D.2 discusses extensions to other activations.

Let $M_\ell^N = z_\ell^N(\Omega) \subseteq \mathbb{R}^{d_\ell^N}$ be the realized layer-ℓ representation. A *task patch* $C \subseteq M_\ell^A$ has nonempty relative interior in its affine hull, and $\Omega_C := \{x \in \Omega : z_\ell^A(x) \in C\}$. It is a locally full-dimensional piece of the realized representation on which the task can expose nonlinear boundaries or curvature. Write $[d] = \{1, \dots, d\}$ and let $J_\ell^N \subseteq [d_\ell^N]$ index coordinate axes after counting exact duplicates once; d_ℓ^N remains the full width. The two notions of alignment are:

Definition 1 (Weak and strong alignment)**.** A layer-ℓ comparison is an injective affine map

$$E_\ell : \mathbb{R}^{d_\ell^A} \to \mathbb{R}^{d_\ell^B}, \qquad E_\ell(z) = T_\ell z + a_\ell,$$

where $T_\ell \in \mathbb{R}^{d_\ell^B \times d_\ell^A}$ has rank d_ℓ^A and $a_\ell \in \mathbb{R}^{d_\ell^B}$, so $d_\ell^A \le d_\ell^B$. It gives *weak alignment* on C when $z_\ell^B(x) = E_\ell(z_\ell^A(x))$ for all $x \in \Omega_C$. Let e_i be the ith standard basis vector of $\mathbb{R}^{d_\ell^B}$. When matching a unit in network A to a unit in network B, we compare the latter's preactivation with α times the former's. Let $\mathcal{H}_\varphi$ be the set of allowed factors α: $\mathcal{H}_\rho = \mathbb{R} \setminus \{0\}$ for ReLU and $\mathcal{H}_\sigma = \{1\}$ for Softplus. Axes $j \in J_\ell^A$ and $i \in J_\ell^B$ *strongly align* on C when $e_i^\top E_\ell(z) = \alpha z_j$ for every $z \in C$ and some $\alpha \in \mathcal{H}_\varphi$. Quantitatively,

$$\mathrm{AxisAlign}_\ell(A, B) := \frac{1}{d_\ell^A} \max_{\pi : J_\ell^A \hookrightarrow J_\ell^B} \#\Big\{ j \in J_\ell^A : \exists \alpha_j \in \mathcal{H}_\varphi,\ e_{\pi(j)}^\top E_\ell(z) = \alpha_j z_j \ \forall z \in C \Big\}.$$

Weak alignment allows mixtures of units in one network to match the other. Strong alignment requires the units themselves to match one-to-one, up to the activation symmetries defined above. $\mathrm{AxisAlign}_\ell(A, B)$ measures the fraction of layer-ℓ units in network A that can be matched one-to-one to units in network B.

3.1 Zippering: terminal agreement propagates upstream

If two networks agree at the end, how differently can they compute in the middle? We compare networks whose terminal representations match up to an injective affine map on every task input. Redundant or canceling computations can change internal representations without changing outputs, so we require layerwise "minimality" (defined below). If each layer's function identifies its preceding nonlinear expansion, terminal agreement forces alignment backward through the hierarchy, which we term "zippering" (visualized in Fig. 1a).

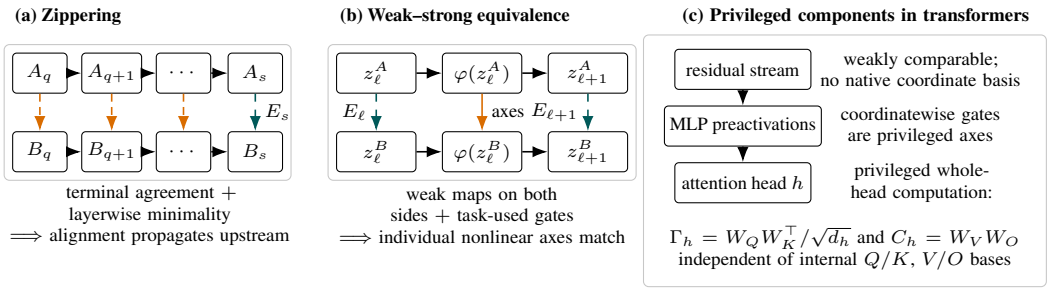


Fig. 1: **Terminal agreement forces a shared hierarchy and recurring individual components.** (a) Identifiability of minimal nonlinear expansions propagates terminal alignment backward. (b) Weak alignment on both sides of task-used gates forces their axes to match. (c) Our transformer theory predicts privileged MLP gates and whole attention heads.

Write a one-step expansion as

$$F(x) = b + \sum_{j=1}^{k} v_j \varphi(g_j(x)). \tag{2}$$

Here $x \in \Omega$, $b, v_j \in \mathbb{R}^m$, $g_j : \Omega \to \mathbb{R}$, and k counts nonlinear terms. The expansion is *minimal* when no zero coefficient, constant argument, duplicate, opposite, or affine cancellation can remove a term. Distinct minimal expansions that realize the same F form a *null pair*. Under our minimality and regularity assumptions, a layer's outputs across all task inputs determine the individual nonlinear units feeding into it, up to unit order and, for ReLU, positive rescaling (Supp. Thms. 5 and 4). Our proofs identify these units from the complex singularities introduced by Softplus and the kink boundaries at a given ReLU layer. Under the regularity conditions below, nontrivial null pairs occupy a lower-dimensional, measure-zero parameter set.

Theorem 1 (Exact zippering (informal))**.** *Let network A and network B use ReLU or Softplus, and suppose an injective affine map weakly aligns their terminal representations at layer s. If the two one-step expansions compared at every backward step from $s-1$ to q are minimal and their nonlinear signatures uniquely identify their scalar arguments, then the representations strongly align at every $r = q, \ldots, s-1$: up to permutation for Softplus, and permutation plus positive rescaling for ReLU.*

Our proof repeats one-step identification backward through the hierarchy (Supp. Thm. 6):

For our exact-zippering result, we treat unit permutations and positive ReLU rescalings compensated in the outgoing weights as equivalent descriptions of the same computation. The *one-step realization map* sends expansion parameters to the function they compute. *Jacobian regularity of this map* means full column rank after removing these function-preserving freedoms: every remaining infinitesimal parameter change alters the function. We also exclude full-dimensional families of non-identifiable, functionally identical expansion pairs; *full-dimensional* means as many degrees of freedom as the expansion space. Then failures form a measure-zero set. Supp. Defs. 21 and 22 formalize both conditions, and Supp. Cor. 5 proves our result.

To compare *approximately* aligned networks, define the weak error

$$\omega_\ell(E_\ell; A, B) := \left(\mathbb{E}_{x\sim\mu}\|z_\ell^B(x) - E_\ell z_\ell^A(x)\|_2^2\right)^{1/2}.$$

Let $\|\cdot\|_{L^2(C)}$ denote the task-distribution norm restricted to Ω_C. Writing $z = z_\ell^A(x)$, define the score comparing axis j of network A with the predicted response of axis i of network B by

$$s_{ij}^{\varphi}(E_\ell) := \sup_{\alpha\in\mathcal{H}_\varphi}\left[1 - \frac{\|(E_\ell z)_i - \alpha z_j\|_{L^2(C)}}{\|(E_\ell z)_i\|_{L^2(C)} + |\alpha|\|z_j\|_{L^2(C)}}\right].$$

For $\theta \in (0,1)$, let $\mathrm{AxisAlign}_\ell^\theta(A, B; E_\ell)$ be the largest fraction of layer-ℓ axes in A matched to distinct target axes with $s_{ij}^{\varphi}(E_\ell) \geq \theta$. For zippering, $\mathrm{AxisAlign}_\ell^\theta(A, B)$ instead compares actual responses: replace $(E_\ell z)_i$, network A's prediction of unit i's preactivation in network B through the affine map E_ℓ, with the actual preactivation $z_{\ell,i}^B$ throughout the score and restrict ReLU scales to $\alpha > 0$. Both scores use normalized tolerance $1-\theta$, while ω_ℓ measures whole-layer prediction error.

For a numerical error bound, we impose Jacobian regularity on network B at every backward step. It prevents a small downstream discrepancy from hiding an arbitrarily large upstream change and gives constants L_ℓ such that $\|z_\ell^B - E_\ell z_\ell^A\| \le L_\ell \|z_{\ell+1}^B - E_{\ell+1} z_{\ell+1}^A\|$.

Theorem 2 (Soft zippering (informal))**.** *For compact families with quantitative-minimality margin $c > 0$ and a full-rank Jacobian for network B's one-step realization map at every backward step, sufficiently small terminal mismatch $e_s := \|z_s^B - E_s z_s^A\|_{L^2(\mu)}$ implies*

$$\mathrm{AxisAlign}_r^\theta(A, B) \ge 1 - \frac{e_s^2 \prod_{\ell=r}^{s-1} L_\ell^2}{(1-\theta)^2 c^2 d_r^A}, \tag{3}$$

with $c = 1$ for normalized ReLU.

Small terminal error forces a large fraction of upstream axes to align. The bound weakens when errors amplify more strongly at the intervening layers, as measured by the L_ℓ (Supp. Thm. 8). The same induction permits different widths (Supp. Rem. 2). For unequal depths, a *depth warp* compares a layer in one network with a window of layers in the other (Supp. Def. 23 and Thm. 10).

3.2 From global zippering to layerwise weak-to-strong alignment: nonlinearities force task-used units to align strongly

Zippering addresses the global question of when terminal agreement forces strong alignment throughout the hierarchy. Locally, what can we infer from weak alignment at just two adjacent layers? We show that the fraction of nonlinear units required by the task lower-bounds the fraction that must match individually.

A dense T_ℓ can mix every coordinate at one layer, but alignment after the next nonlinear layer removes that freedom. If an affine map aligns two networks before a used ReLU gate but mixes their axes, the networks place their kinks on different hyperplanes. After ReLU, no affine map can put them back together. Softplus has no kink, but its affine-ridge curvature plays the same role. The locations of these nonlinear signatures identify the native axes.

Only a nonlinearity that affects the task can identify an axis. We call a ReLU axis j *used* on C when the task crosses its zero trace $\Gamma_j(C) = \{z \in C : z_j = 0\}$ in relative interior and its outgoing column is nonzero. We call a Softplus argument used when it varies, has a nonzero outgoing coefficient, and remains after we remove exact duplicates, opposite arguments, and the affine cancellation $\sigma(t) - \sigma(-t) = t$. A coordinate counts as used only when the task makes its nonlinear signature visible and the next affine map reads it out.

We measure layerwise task difficulty by the fewest nonlinear axes an accurate solution can use. Let $L(f)$ be task loss, ε a target loss, and $\mathcal{F}_{\ell,\le k}$ the solutions using at most k layer-ℓ axes. Define

$$m_\ell(\varepsilon) = \min\{k : \exists f \in \mathcal{F}_{\ell,\le k},\ L(f) \le \varepsilon\}. \tag{4}$$

The minimum is $+\infty$ if the set is empty. The task is harder at layer ℓ when every ε-accurate solution must expose more nonlinear coordinates there.

Theorem 3 (Exact weak–strong equivalence (informal))**.** *If injective affine maps weakly align network A with network B at adjacent layers ℓ and $\ell+1$, every task-used axis of network A at layer ℓ matches a distinct axis of network B. Hence, for loss at most ε,*

$$\mathrm{AxisAlign}_\ell(A, B) \ge \frac{m_\ell(\varepsilon)}{d_\ell^A}.$$

Adjacent weak alignment equates applying network A's layer followed by $E_{\ell+1}$ with applying E_ℓ followed by network B's layer on the task patch. Used kink traces or curvature ridges must therefore match, and distinct signatures make this match injective (Supp. Thms. 1 and 2).

Theorem 4 (Soft weak–strong equivalence (informal))**.** *For a compact family $\mathcal{K}$ of comparison problems, there is a separation margin $\kappa_{\mathcal{K}}(\theta) > 0$ such that*

$$\mathrm{AxisAlign}_\ell^\theta(A, B; E_\ell) \ge \frac{m_\ell(\varepsilon)}{d_\ell^A} - \frac{(\|W_{\ell+1}^B\|_{\mathrm{op}} \omega_\ell + \omega_{\ell+1})^2}{\kappa_{\mathcal{K}}(\theta)^2 d_\ell^A}, \tag{5}$$

Here $\|\cdot\|_{\mathrm{op}}$ is the operator norm. When weak errors vanish, the bound recovers the exact conclusion. Poor axis matches require adjacent weak error. Supp. Thm. 3 gives the formal statement, separation margin $\kappa_{\mathcal{K}}(\theta)$, and proof.

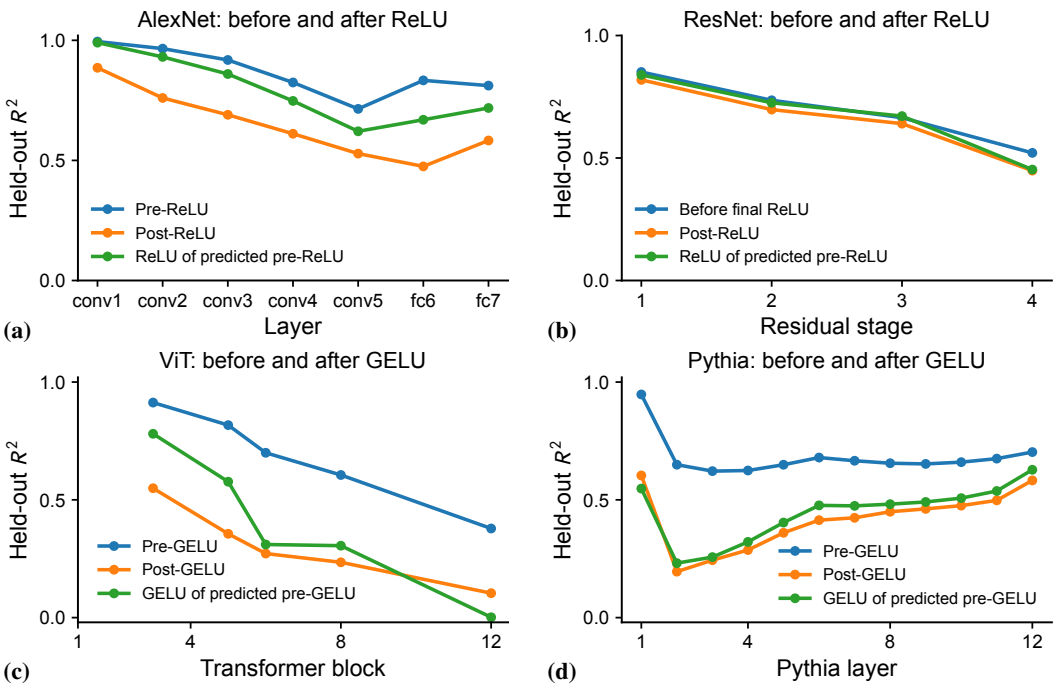


Fig. 2: **Affine alignment drops after nonlinear gates in all four model families.** Curves average independently trained pairs (Pythia-160M in d). Blue compares preactivations, orange post-activations, and green applies the target ReLU/GELU to its affinely predicted preactivation. In (b), "before final ReLU" is the residual-plus-skip sum. Supp. Sec. F gives the protocol.

3.3 Transformers have privileged gates and attention heads

Our theory predicts privileged gates and attention heads in transformers. A head's mechanism is its routing and output computation, not its particular parameters. For a head $h \in [H_\ell^N]$ in model width d, let $W_{Q,h}, W_{K,h} \in \mathbb{R}^{d\times d_h}$, $W_{V,h} \in \mathbb{R}^{d\times d_v}$, and $W_{O,h} \in \mathbb{R}^{d_v\times d}$. Suppressing h, the exact transformations, for any $R \in \mathrm{GL}(d_h)$ and $S \in \mathrm{GL}(d_v)$,

$$W'_Q = W_Q R, \quad W'_K = W_K R^{-\top}, \qquad W'_V = W_V S, \quad W'_O = S^{-1} W_O$$

change individual $Q/K/V/O$ coordinates but preserve the head computation and the quantities

$$\Gamma_h = \frac{W_{Q,h} W_{K,h}^\top}{\sqrt{d_h}}, \qquad C_h = W_{V,h} W_{O,h}. \tag{6}$$

For a token matrix $X \in \mathbb{R}^{n\times d}$, these determine the routing $P_h(X) = \mathrm{softmax}(X\Gamma_h X^\top)$ and whole-head contribution $O_h(X) = P_h(X) X C_h$. A *learned head* means this whole contribution from one trained head, not an individual $Q/K/V/O$ coordinate. We call these transformations *gauges*. *Quotienting* compares quantities they leave fixed (Supp. Prop. 9 and Cor. 6). Residual-stream coordinates also admit changes of basis, so we can compare them only weakly. Our weak–strong theory instead predicts recurring MLP preactivations: the scalar gates of a coordinatewise *nonlinear* expansion (Supp. Thm. 11).

We call a head "task-used" when (P_h, O_h) varies on the task patch and O_h is not identically zero. Let $m_{\mathrm{attn},\ell}(\varepsilon)$ be the fewest used layer-ℓ heads in any ε-accurate solution, and let $\mathrm{HeadAlign}_\ell$ be the largest fraction of A-side heads whose attention patterns and outputs match distinct B-side heads (Supp. Def. 25). A *no-splitting* condition requires one target head, rather than several target heads acting together, to reproduce a source head's distinguishing attention pattern and output (Supp. Def. 26, item A3). Under head usedness and minimality, one-head identification, and no splitting, attention weak–strong equivalence gives $\mathrm{HeadAlign}_\ell \geq m_{\mathrm{attn},\ell}(\varepsilon)/H_\ell^A$ (Supp. Thm. 11). Residual peeling propagates the weak maps through the attention and MLP branches (Supp. Lemma 17 and Thm. 12). Thus, our theory predicts recurring gates and whole heads, with the aligned fraction lower-bounded by the fraction the task requires.

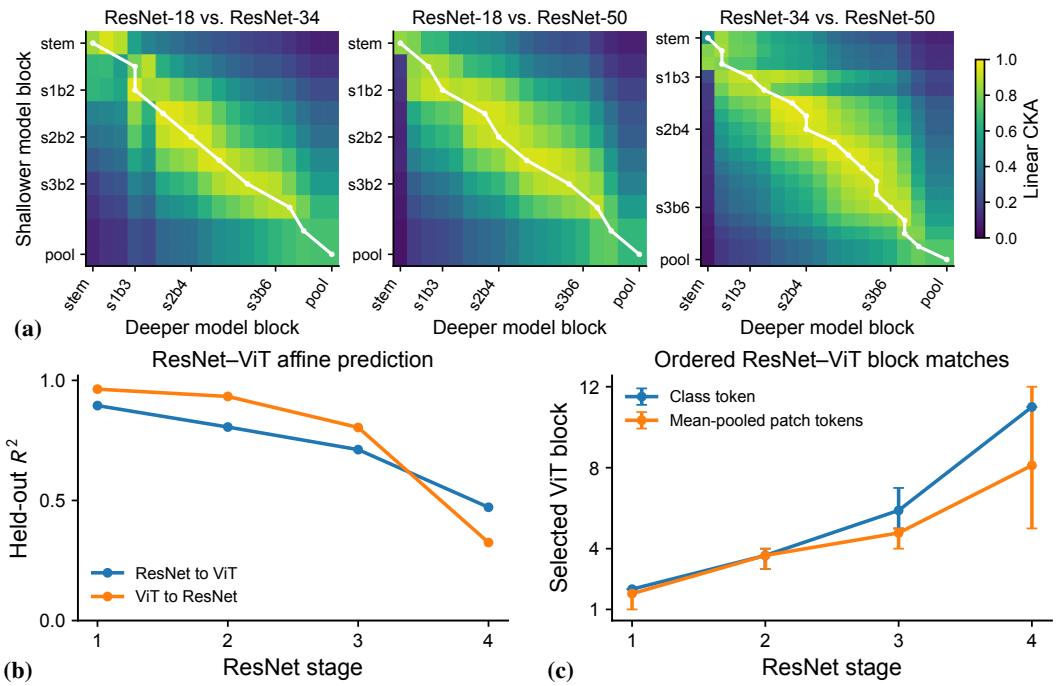


Fig. 3: **ResNets and ViTs align in order through depth.** (a) White lines trace the highest-CKA monotone paths through ResNet-18/34/50 (mean CKA 0.890–0.901); s2b4 means stage 2, residual block 4. (b) Bidirectional held-out affine R^2 compares ResNet stages with mean-pooled ViT patch tokens. (c) Class-token and mean-pooled matches both remain ordered. Supp. Sec. F gives the path-fitting protocol.

Our GELU experiments test the same before-versus-after-gate prediction, though our exact theorems cover ReLU and Softplus. Extending zippering to GELU requires uniquely identifying hidden arguments from their nonlinear expansions (Supp. Sec. D.2).

4 From Theorems to Measurements

We fit ridge regressions and select one-to-one component matches on training inputs, then evaluate both on held-out inputs. Ridge R^2 measures affine alignment before and after each nonlinear gate. For ResNets we use CKA, invariant to rotations and uniform scaling (Kornblith et al., 2019). We fit monotone paths through unequal depths. Random coordinate rotations preserve the linear span but change individual features, providing our control. For heads, we preserve the summed attention update (Supp. Eq. 138). We also follow head matches through three models, $A \to B \to C \to A$, and measure how often they return to the starting head.

Our theory predicts that the most-used components recur most strongly. On held-out inputs, we score channels and MLP gates by $u_j = \mathrm{Var}[\varphi(z_j)]\|W_{2,:,j}\|_2^2$, and heads by $u_h = \mathbb{E}\|O_h(X)\|_2^2$, with $W_{2,:,j}$ the outgoing/down-projection column. These scores rank the use of components by the task within a trained network. We rank the source alone, then test whether its selected component recurs or has the same intervention effect in another network. Supp. Sec. F gives the protocol.

5 Empirical Results

We first test global zippering: alignment throughout the hierarchy, including ordered correspondences across unequal depths. We then test local weak–strong equivalence: recurrence of individual channels, MLP gates, and attention heads, especially those the task uses most strongly.

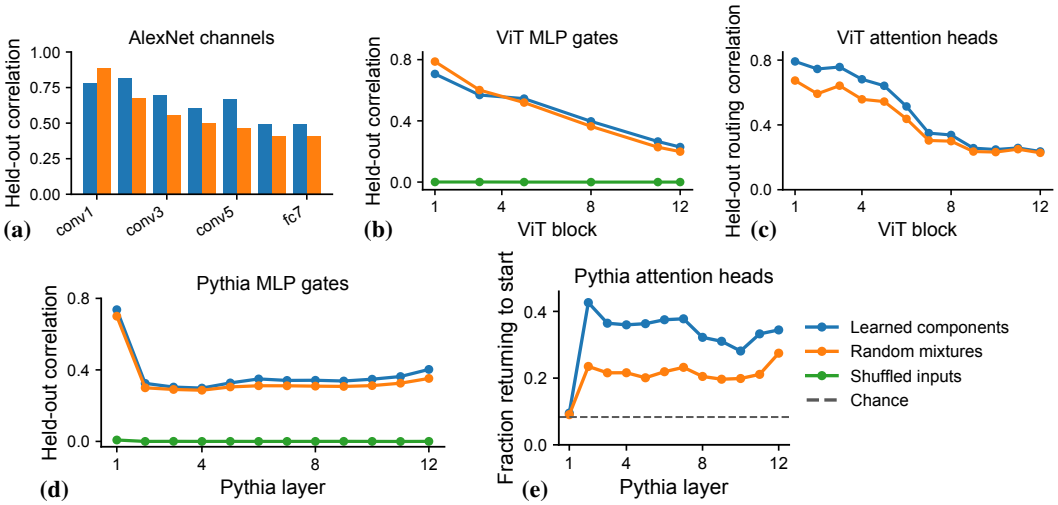


Fig. 4: **Learned channels, MLP gates, and whole attention heads recur across independently trained models.** Learned components match more closely than random mixtures preserving their linear span (a,b,d), and whole heads outperform mixtures preserving the summed update (c; Supp. Eq. 138). In (e), following a head through matches from $A \to B \to C \to A$ returns to the starting head above $1/12$ chance.

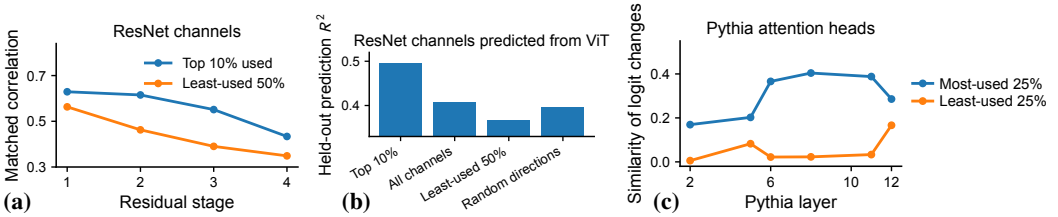


Fig. 5: **Components used most strongly by the source model recur most strongly in another model.** Top-used ResNet channels match better across seeds (a) and are predicted better from ViTs (b; $R^2 = .495$ versus .395 for random directions, $18/18$ pairs). Mean-ablating top-used Pythia heads produces more similar next-token-logit changes across seeds than ablating bottom-used heads (c). We rank components using only the source model.

5.1 Weak alignment spans the network hierarchy

Our zippering theorems predict that terminal agreement extends to earlier representations (Thms. 1 and 2; Supp. Thm. 12 for transformers). Across AlexNet, ResNets, ViTs, and Pythia, we find affine alignment throughout the hierarchy (Fig. 2). Our weak–strong theorems also predict weaker post-nonlinearity alignment when used axes do not match (Thms. 3 and 4). We observe these before-after drops in *every* model family. The ReLU comparisons test our theorem's consequence directly, while the GELU comparisons test the same predicted pattern at MLP gates.

5.2 Alignment follows the computational hierarchy

Our zippering theory predicts more than similar layers: it predicts their *order*. Our exact and soft zippering results predict ordered layerwise alignment, while our unequal-depth result allows corresponding units to occur within depth windows (Thms. 1 and 2; Supp. Thm. 10). ResNet-18/34/50 recover such paths with mean CKA 0.890–0.901 (Fig. 3a). For ViTs, we compare the class token with the mean of all image-patch tokens, excluding the class token. The harder ResNet-ViT test, beyond our proved setting, also advances in order and reaches held-out $R^2 \approx 0.70$, or 81.6% of prediction between independently trained models of the target architecture (Fig. 3b,c). Different blocks and depths still express the shared computation in the same sequence.

5.3 PRIVILEGED CHANNELS, GATES, AND HEADS RECUR

Our weak–strong equivalence theorems predict that task-used nonlinear axes must themselves match (Thms. 3 and 4). We compare learned components with random mixtures that preserve their linear span or summed attention update. If only the shared information mattered, then mixing would make no statistical difference. Instead, AlexNet channels and ViT/Pythia gates match more closely than random orthogonal mixtures, while whole heads outperform sum-preserving mixtures and return to themselves in three model families above chance (Fig. 4). These are privileged attention heads in both vision and language, as our transformer weak–strong theorem predicts (Supp. Thm. 11). We compare routing patterns and whole-head contributions, which are unchanged by equivalent Q/K and V/O parameterizations (Eq. 6). We find in Supp. Fig. 18(a,b) that a second target head improves learned-head matches much less than random mixtures in ViT and Pythia, supporting our transformer weak–strong equivalence theorem's premise that a source head's computation has an individual counterpart rather than requiring several target heads acting together (Supp. Thm. 11). Furthermore, Pythia heads matched by activity produce more similar logit changes than random matches in 248 of 270 model-pair/layer comparisons (Supp. Fig. 18c). Head recurrence is sparse, consistent with our theorem's restriction to task-used heads. We next test whether usedness predicts *which* heads have these *reproducible* effects on model predictions.

5.4 USEDNESS PREDICTS COMPONENT RECURRENCE

Our weak–strong equivalence theorems relate the aligned fraction to how many nonlinear axes the task requires (Thms. 3 and 4; Supp. Thm. 11 for heads). We therefore ask whether components used most strongly in one network recur most strongly in another. Top-used ResNet channels match more closely across seeds and are predicted from ViTs at $R^2 = .495$ rather than .395 for random directions in all 18 pairs. Top-used Pythia heads produce more similar logit changes (Fig. 5).

6 DISCUSSION AND LIMITATIONS

Hard tasks force convergent computation. Over the past fifteen years, study after study has found that independent networks agree under linear maps, line up across depth, and recover some of the same channels, gates, and heads (Li et al., 2015; Raghu et al., 2017; Kornblith et al., 2019; Khosla & Williams, 2023; Kapoor et al., 2025; Khosla et al., 2024). Here we develop a theory that addresses *why* and *when* this happens. *Why?* The more nonlinear computation a task requires, the less freedom two minimal solutions have to differ internally. *When?* Terminal agreement propagates backward when each layer is minimal and its nonlinear expansion identifies its inputs, with regularity controlling the propagation of approximate agreement.

We predict recurring channels, gates, and whole attention heads, and find them in vision and language.

Contravariance identifies what a theory of PRH must explain. Although contravariance concerns the same task (Cao & Yamins, 2024), our theory sets an "agreement floor" which shows that explaining PRH's proposed agreement across *different* tasks or modalities (Huh et al., 2024) requires identifying the common tasks agents need to survive in their world, their shared constraints, and the computations these force. Our exact results cover ReLU and Softplus, and our GELU experiments test their predicted signature.

Representational convergence in NeuroAI is mathematically inevitable. Goal-driven modeling asks whether behavior alone can make a brain-like internal solution emerge (Yamins & DiCarlo, 2016). Modulo the biological validity of our minimality and learning-dynamics assumptions, the major NeuroAI results of the past fifteen years are not accidental: for minimal artificial and biological solutions to sufficiently hard common tasks, convergent evolution is mathematically inevitable.

Prior work has found that ridge regression, RSA, and CKA can disagree about which representations match brain data (Soni et al., 2024; Bo et al., 2024). Our theory explains why: RSA and CKA retain *task-irrelevant* feature scaling, duplication, and measurement bias from neuron or task-data sampling. We show that canonicalization removes these scaling and duplication distortions and recovers shared task geometry (Supp. Secs. E.1 and E.2; sampling analyses in Secs. E.1.7 and E.2.4). NeuroAI should therefore distinguish metric disagreement from computational difference and ask which hard, natural tasks make convergence unavoidable, and which computations models and brains must share.

REPRODUCIBILITY STATEMENT

We define the empirical statistics in Section 4, including how our measured usedness scores differ from the task-level minimum in our theorem. Supplement Section F gives the model families, numbers of seeds and examples, fitting and evaluation protocol, null constructions, interventions, sample counts, and qualifications. Our complete theorem statements, assumptions, and proofs appear in Supplement Sections A-E.

AI USE STATEMENT

The authors have for some years known the statement and proof outline for the exact weak–strong theorem. We have also suspected that a zippering theorem had to be correct, both from intuition and empirical results. We developed this theory through an iterative research process with assistance from ChatGPT-5.5-Pro. Before these interactions, we had established the statement and proof outline of exact weak–strong equivalence and conjectured zippering from theoretical intuition and empirical observations. Our exchanges with ChatGPT helped develop the zippering arguments, including the formulation of Softplus minimality, a proof of Softplus one-step rigidity, and the use of Jacobian regularity for quantitative identifiability. We verified all final proofs and take full responsibility for the mathematical claims and the paper's content.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv preprint arXiv:2411.14633*, 2024. Version 2 posted September 14, 2025.

Joachim Bona-Pellissier, François Malgouyres, and François Bachoc. Local identifiability of deep ReLU neural networks: the theory. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b0ae046e198a5e43141519868a959c74-Paper-Conference.pdf.

Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward relu neural network. *Machine learning*, 112(11):4431–4493, 2023.

Lukas Braun, Erin Grant, and Andrew M Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *ICML*, 2025.

Martin D Buhmann and Allan Pinkus. Identifying linear combinations of ridge functions. *Advances in Applied Mathematics*, 22(1):103–118, 1999.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85:101200, 2024.

Trinity Chung, Yuchen Shen, Nathan Kong, and Aran Nayebi. Task-optimized convolutional recurrent networks align with tactile processing in the rodent brain. *Advances in Neural Information Processing Systems*, 38:68925–68951, 2026.

Nathan Cloos, Moufan Li, Markus Siegel, Scott L. Brincat, Earl K. Miller, Guangyu Robert Yang, and Christopher J. Cueva. Differentiable optimization of similarity scores between models and brains. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9edf04e9b128ed5085189205fc94e21f-Abstract-Conference.html.

David A. Cox, John B. Little, and Henry K. Schenck. *Toric Varieties*, volume 124 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011. ISBN 978-0-8218-4819-7. doi: 10.1090/gsm/124.

Jörn Diedrichsen, Serge Provost, and Hossein Zareamoghaddam. On the distribution of cross-validated mahalanobis distances. *arXiv preprint arXiv:1607.01371*, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

David Eisenbud and Bernd Sturmfels. Binomial ideals. *Duke Mathematical Journal*, 84(1):1–45, 1996.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 907–940. PMLR, 2016.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLoS computational biology*, 20(1):e1011792, 2024.

Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test, 2025.

Charles Fefferman and Scott Markel. Recovering a feed-forward net from its output. In *Advances in Neural Information Processing Systems 6*, pp. 335–342. Morgan Kaufmann, 1993.

Otto Forster. *Lectures on Riemann Surfaces*, volume 81 of *Graduate Texts in Mathematics*. Springer, New York, 1981.

Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16 (7):974–981, 2013.

William Fulton. *Introduction to toric varieties*. Number 131. Princeton university press, 1993.

Johanna Marie Gegenfurtner, Moritz Grillo, and Guido Montúfar. The symmetries of three-layer ReLU networks. *arXiv preprint arXiv:2605.18319*, 2026. URL https://arxiv.org/abs/2605.18319.

Phillip Griffiths and Joseph Harris. *Principles of Algebraic Geometry*. Pure and Applied Mathematics. John Wiley & Sons, New York, 1978.

Moritz Grillo and Guido Montúfar. Most ReLU networks admit identifiable parameters. *arXiv preprint arXiv:2605.03601*, 2026. URL https://arxiv.org/abs/2605.03601.

Sarah E. Harvey, Brett W. Larsen, and Alex H. Williams. Duality of bures and shape distances with implications for comparing neural representations. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pp. 11–26. PMLR, 2024a. URL https://proceedings.mlr.press/v243/harvey24a.html.

Sarah E. Harvey, David Lipshutz, and Alex H. Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 140–151. PMLR, 2024b. URL https://proceedings.mlr.press/v285/harvey24a.html.

Melika Honarmand, Ayati Sharma, Badr AlKhamissi, Johannes Mehrer, and Martin Schrimpf. Inducing dyslexia in vision language models. *arXiv preprint arXiv:2509.24597*, 2025.

Ann Huang, Satpreet H Singh, Flavio Martinelli, and Kanaka Rajan. Measuring and controlling solution degeneracy across task-trained recurrent neural networks. *ArXiv*, pp. arXiv–2410, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024.

Chaitanya Kapoor, Sudhanshu Srivastava, and Meenakshi Khosla. Bridging critical gaps in convergent learning: How representational alignment evolves across layers, training, and distribution shifts. *arXiv preprint arXiv:2502.18710*, 2025.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Reece Keller, Alyn Kirsch, Felix Pei, Xaq Pitkow, Leo Kozachkov, and Aran Nayebi. Intrinsic goals for autonomous agents: Model-based exploration in virtual zebrafish predicts ethological behavior and whole-brain dynamics. *Advances in Neural Information Processing Systems*, 38:35872–35903, 2026.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Meenakshi Khosla and Alex H. Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *arXiv preprint arXiv:2311.09466*, 2023.

Meenakshi Khosla, Alex H. Williams, Josh McDermott, and Nancy Kanwisher. Privileged representational axes in biological and artificial neural networks, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Nikolaus Kriegeskorte, Rhodri Cusack, and Peter Bandettini. How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? *NeuroImage*, 49(3): 1965–1976, 2010. doi: 10.1016/j.neuroimage.2009.09.059.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

J. M. Landsberg. *Tensors: Geometry and Applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212. PMLR, 2015.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L. K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7, 2024. doi: 10.1016/j.neuron.2024.04.018.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 33–44. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.4.

Jonathan A Michaels, Stefan Schaffelhofer, Andres Agudelo-Toro, and Hansjörg Scherberger. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proceedings of the national academy of sciences*, 117(50):32124–32135, 2020.

Sabine Muzellec and Kohitij Kar. Reverse predictivity: Going beyond one-way mapping to compare artificial neural network models and brains. *bioRxiv*, 2025. doi: 10.1101/2025.08.08.669382.

Aran Nayebi. What capable agents must know: Selection theorems for robust decision-making under uncertainty. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 4773–4795. PMLR, 2026. URL https://proceedings.mlr.press/v337/nayebi26a.html.

Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, et al. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, 34:12167–12179, 2021.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30, 2017.

Jason Phang, Haokun Liu, and Samuel R. Bowman. Fine-tuned transformers show clusters of similar representations across layers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 529–538. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.blackboxnlp-1.42.

Mary Phuong and Christoph H Lampert. Functional vs. parametric equivalence of relu networks. In *International conference on learning representations*, 2020.

Allan Pinkus. *Ridge Functions*. Cambridge University Press, Cambridge, 2015.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

Neil Rathi, Johannes Mehrer, Badr AlKhamissi, Taha Binhuraib, Nicholas M. Blauch, and Martin Schrimpf. TopoLM: Brain-like spatio-functional organization in a topographic language model. In *International Conference on Learning Representations*, 2025.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Ansh Soni, Sudhanshu Srivastava, Marvin Maechler, Konrad Kording, and Meenakshi Khosla. Conclusions drawn from neural network to brain alignment depend strongly on the chosen similarity measure. *bioRxiv*, 2024. doi: 10.1101/2024.08.07.607035. Version posted July 8, 2026.

Elias M. Stein and Rami Shakarchi. *Fourier Analysis: An Introduction*, volume 1 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, NJ, 2003.

David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7): 1025–1033, 2015.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Diverse perceptual representations across visual pathways emerge from a single objective. *bioRxiv*, pp. 2025–07, 2025.

Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539. PMLR, 2016.

Marvin Theiss, Lukas Braun, Andrew M Saxe, and Erin Grant. Parameter symmetries determine representational geometry in overparameterized nonlinear networks. In *ICML 2026 Workshop on Weight-Space Symmetries: from Foundations to Practical Applications*, 2026.

Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob Prince, Rosa Cao, and Daniel Yamins. Model-brain comparison using inter-animal transforms, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137: 188–200, 2016. doi: 10.1016/j.neuroimage.2015.12.012.

Nikolaus Weiskopf, Chloe Hutton, Oliver Josephs, Robert Turner, and Ralf Deichmann. Optimized EPI for fMRI studies of the orbitofrontal cortex: compensation of susceptibility-induced gradients in the readout direction. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 20(1): 39–49, 2007. doi: 10.1007/s10334-006-0067-6.

Alex H. Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 10–23. PMLR, 2024. URL `https://proceedings.mlr.press/v285/williams24a.html`.

Christopher Wolfram and Aaron Schein. Layers at similar depths generate similar activations across LLM architectures. *arXiv preprint arXiv:2504.08775*, 2025.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Similarity analysis of contextual word representation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4638–4655. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.422.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. doi: 10.1038/nn.4244.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

APPENDIX TABLE OF CONTENTS

A Formal foundations of weak–strong equivalence **17**

A.1 Common setup and alignment notions 17

A.2 Exact Weak–strong equivalence 18

A.3 Asymptotic weak–strong equivalence 25

B From one-step identifiability to representational zippering **27**

B.1 Minimality, regularity, and exact zippering 27

B.2 Quantitative minimality, Jacobian regularity, and soft zippering 40

B.3 Genericity 47

C Unequal depth, residual networks, and transformers **56**

C.1 Depth-warped weak–strong equivalence 56

C.2 Residual and transformer circuits 60

C.3 Regularity and depth faithfulness 74

D Failure modes and architectural scope **76**

D.1 A null-network atlas: redundant solutions and how to exclude them 76

D.2 Other architecture types and activation functions 82

E Representation metrics beyond ridge **85**

E.1 RSA and weak–strong equivalence 85

E.2 CKA and weak–strong equivalence 105

F Empirical protocols, controls, and additional results **113**

A FORMAL FOUNDATIONS OF WEAK–STRONG EQUIVALENCE

This section gives the common setup and complete proofs of the exact and asymptotic weak–strong equivalence theorems. Section B turns one-step identifiability into exact, asymptotic, and soft zippering; Section C treats unequal depth, residual networks, and transformers; Section D isolates the failure modes and architectural scope; and Section E develops the corresponding theory for RSA and CKA. Together these sections contain the full formal development behind the main paper.

A.1 COMMON SETUP AND ALIGNMENT NOTIONS

We primarily consider ReLU and Softplus activation functions

$$\varphi \in \{\sigma, \rho\}, \qquad \sigma(t) = \log(1 + e^t), \qquad \rho(t) = \max\{t, 0\}.$$

We define affine–nonlinear networks $N \in \{A, B\}$ by alternating an affine step with a pointwise nonlinear step:

$$z_\ell^N(x) = W_\ell^N h_{\ell-1}^N(x) + b_\ell^N, \qquad h_\ell^N(x) = \varphi(z_\ell^N(x)),$$

with φ applied coordinatewise. The full-layer operation is denoted by

$$\Phi_{\ell,\varphi}^N(z) = W_{\ell+1}^N \varphi(z) + b_{\ell+1}^N.$$

The layer-ℓ image of the input set is $M_\ell^N = z_\ell^N(\Omega)$. It will often be convenient to work on a *task patch* $C \subseteq M_\ell^A$ with nonempty relative interior in its affine hull.

Capturing the intuitive definitions of weak alignment as affine mappability, we define:

Definition 1 (Weak alignment)**.** A layer-ℓ comparison map is an injective affine map

$$E_\ell(z) = T_\ell z + a_\ell, \qquad T_\ell : \mathbb{R}^{d_\ell^A} \to \mathbb{R}^{d_\ell^B}.$$

The networks are *weakly equivalent* at layer ℓ on task patch C if

$$E_\ell z = z_\ell^B(x) \quad \text{whenever } z = z_\ell^A(x) \in C.$$

The following basic commutation lemma is useful below.

Lemma 1 (Adjacent commutation)**.** *If exact weak alignment holds at layers ℓ and $\ell + 1$, then the adjacent square in Figure 1(b) commutes, i.e.*

$$\Phi_{\ell,\varphi}^B(E_\ell z) = E_{\ell+1} \Phi_{\ell,\varphi}^A(z) \qquad \forall z \in C.$$

Proof. For any $z \in C \subseteq M_\ell^A$, choose $x \in \Omega$ such that $z = z_\ell^A(x)$. Then

$$\begin{aligned}
\Phi_\ell^B(E_\ell z) &= \Phi_\ell^B(E_\ell z_\ell^A(x)) \\
&= \Phi_\ell^B(z_\ell^B(x)) \\
&= z_{\ell+1}^B(x) \\
&= E_{\ell+1} z_{\ell+1}^A(x) \\
&= E_{\ell+1} \Phi_\ell^A(z_\ell^A(x)) \\
&= E_{\ell+1} \Phi_\ell^A(z).
\end{aligned}$$

□

It will sometimes be necessary to account for scalar symmetries when comparing axes. The allowed scalar symmetry depends on the activation: Softplus has no nontrivial real scaling symmetry, while ReLU is homogeneous and therefore permits nonzero real rescalings in adjacent weak–strong comparisons. This motivates the following scalar symmetry sets:

$$\mathcal{H}_\sigma = \{1\}, \qquad \mathcal{H}_\rho = \mathbb{R} \setminus \{0\}.$$

We can now define strong alignment precisely:

Definition 2 (Strong alignment)**.** Given φ, E_ℓ, an A-axis j and pulled-back B-axis i are *strongly aligned* on C if

$$(E_\ell z)_i = \alpha z_j \qquad \forall z \in C$$

for some $\alpha \in \mathcal{H}_\varphi$. Let $J_\ell^{N,\varphi}(C)$ denote the set of axes of network N on task patch C, and define

$$\mathrm{AxisAlign}_{\ell,\varphi}(A, B; E_\ell, C) = \frac{1}{d_\ell^A} \max_\pi \#\left\{ j \in J_\ell^{A,\varphi}(C) : (E_\ell z)_{\pi(j)} = \alpha_j z_j,\ \alpha_j \in \mathcal{H}_\varphi \right\},$$

where π ranges over injective matchings into the counted B-axes.

It is useful to introduce the concept of a one-layer softplus expansion: the input-output formula for any given softplus layer r, viewed as a function of a set of scalar-valued functions g_j realizing the network operations in layers below r. This notation separates the lower part of the network, which produces scalar hidden axes, from the next layer, which applies softplus to those axes and linearly recombines them:

Definition 3 (One-step expansion)**.** A one-step expansion is

$$G = \big(b, \{(v_j, g_j)\}_{j=1}^k\big), \qquad Y_G^\varphi(x) = b + \sum_{j=1}^k v_j \varphi(g_j(x)).$$

In an actual network layer, this notation is simply the normal layer update rewritten with the "dictionary"

$$g_j = z_{r,j}^N, \qquad v_j = W_{r+1}^N[:, j], \qquad b = b_{r+1}^N.$$

If $E(y) = Ty + a$ is affine, define

$$EG = \big(Tb + a, \{(Tv_j, g_j)\}_{j=1}^k\big), \qquad Y_{EG}^\varphi = E(Y_G^\varphi).$$

When T is injective, many properties introduced below (such as usedness and affine sign-noncancellation) are preserved.

A.2 Exact Weak–strong equivalence

The common strategy of the exact weak-strong equivalence is to define "usedness" for a nonlinear element of the network, and then prove that an affine transformation cannot undo a used nonlinearity. But, the concrete definition of usedness is a little bit different depending on the activation function. For ReLU, we observe "kinks" where the nonlinearity acts; for softplus, we use conditions on the non-redundancy of the expansion coefficients.

A.2.1 ReLU: usedness as visible kink traces

The key object for ReLU usedness is the trace associated with an axis (sometimes called a "gate"):

Definition 4 (Used ReLU gates.)**.** For a ReLU axis j, define its trace on a task patch C by

$$\Gamma_j^N(C) = \{z \in C : z_j = 0\}.$$

Given a comparison map E, the pulled-back B-trace is

$$\Gamma_i^B(E; C) = \{z \in C : (Ez)_i = 0\}.$$

Let V be the affine hull of C. A trace $\Gamma(C) = C \cap \{a(z) = 0\}$ is *regular at point* $p \in \Gamma(C)$ if $a|_V$ is not the zero affine functional, so near p inside V the trace is a codimension-one affine slice. The patch C *crosses* this trace at p if, for every $\eta > 0$, there are $z^+, z^- \in C \cap B_\eta(p)$ with $a(z^+) > 0$ and $a(z^-) < 0$. A set of traces is *regular and separated on* C if every trace in the set contains a relatively open patch of regular crossed points p, and no two distinct counted traces agree on a relatively open subset of such a patch (in other terms, the trace has a "private" regular crossed piece). A ReLU gate j is *used on* C if (i) its trace $\Gamma_j^N(C)$ contains a regular crossed patch, (ii) and the outgoing column is nonzero: $W_{\ell+1}^N[:, j] \neq 0$.

The word "used" is deliberately broader than the technical downstream condition $W_{\ell+1}[:, j] \neq 0$. It says both that the task crosses the nonlinear boundary and that the next affine map reads out the resulting kink. This is the notion that belongs in the nonlinear gate budget.

The next lemma explains why used gates are the right budgeted object for measuring task difficulty: a gate that is not used may still be physically present in the network, but it is not needed as a nonlinear boundary on the task patch.

Lemma 2 (Shrinkability of non-used ReLU gates)**.** *For the adjacent block*

$$\Phi_{\ell,\rho}(z) = W_{\ell+1}\rho(z) + b_{\ell+1}$$

on a task patch C, a gate that is not used contributes no indispensable nonlinear boundary on C. More precisely: if $W_{\ell+1}[:, j] = 0$, the gate can be deleted; if C does not cross $\{z_j = 0\}$, then $\rho(z_j)$ is either 0 or affine on each component of C and its contribution can be absorbed into the affine part of the block; if two counted affine traces coincide on a regular patch, their ReLU contributions can be merged, up to an affine residual coming from $\rho(t) - \rho(-t) = t$.

Proof. The first case is immediate. In the second case, if $z_j < 0$ on the patch then $\rho(z_j) = 0$, while if $z_j > 0$ then $\rho(z_j) = z_j$ and the contribution $W_{\ell+1}[:, j]z_j$ is affine in the layer coordinate. If two affine traces coincide on $V = \mathrm{aff}(C)$, the defining affine functions are proportional: $h = \alpha g$ or $h = -\alpha g$ with $\alpha > 0$. In the positive case, $v\rho(g) + \widetilde{v}\rho(\alpha g) = (v + \alpha\widetilde{v})\rho(g)$. In the negative case, use $\rho(g) - \rho(-g) = g$ to rewrite the pair as one ReLU term plus an affine function. Hence at most one nonlinear boundary remains. □

Accordingly, the gate budget below counts task-visible nonlinear slope changes, not physically present but redundant coordinates.

Definition 5 (Required used ReLU-gate budget)**.** Let $\mathcal{F}^{\rho}_{\ell,\le k}$ be the fixed-macroarchitecture networks solving the task with at most k used layer-ℓ ReLU gates. Define

$$m^{\rho}_{\ell}(\varepsilon) = \min\{k : \exists f \in \mathcal{F}^{\rho}_{\ell,\le k},\ L(f) \le \varepsilon\}.$$

We are now in a position to state and prove the exact ReLU version of weak-strong alignment equivalence.

Theorem 1 (Exact ReLU weak–strong equivalence)**.** *Fix a task patch $C \subseteq z^A_\ell(\Omega)$ and let $\Omega_C = \{x : z^A_\ell(x) \in C\}$. Suppose exact weak equivalence holds at layers ℓ and $\ell+1$ on Ω_C. Assume the counted A-side and pulled-back B-side ReLU trace families are used on C in the sense of Definition 4. Then there is an injective map*

$$\pi : J^{A,\rho}_{\ell}(C) \to J^{B,\rho}_{\ell}(E_\ell; C)$$

with

$$(E_\ell z)_{\pi(j)} = \alpha_j z_j, \qquad \alpha_j \neq 0, \qquad z \in C.$$

If A reaches loss at most ε, then

$$\mathrm{AxisAlign}_{\ell,\rho}(A, B; E_\ell, C) \ge \frac{m^{\rho}_{\ell}(\varepsilon)}{d^A_\ell}.$$

Proof. Write $V = \mathrm{aff}(C)$. On each relatively open region of C on which the signs of all coordinates are fixed, ρ is affine and therefore both $\Phi^A_{\ell,\rho}$ and $\Phi^B_{\ell,\rho} \circ E_\ell$ are affine. Fix $j \in J^{A,\rho}_{\ell}(C)$ and choose a private regular point

$$p \in \Gamma^A_j(C)$$

which lies on no other counted trace. In a neighborhood of p, every term except the jth is affine through the trace. If $u \in V - V$ is a crossing direction with $u_j \neq 0$, then the one-sided directional derivatives of the jth contribution differ by

$$W^A_{\ell+1}[:, j]u_j.$$

Equivalently, the full derivative matrices on the two sides differ by

$$W^A_{\ell+1}[:, j]e_j^{\top}\big|_{V-V}.$$

This jump is nonzero because the task patch crosses the trace and $W^A_{\ell+1}[:, j] \neq 0$. Hence a relatively open part of every used trace belongs to the nondifferentiability set of the restricted block. Conversely, away from the union of the coordinate traces the sign pattern is locally constant, so the block is affine. Thus, up to lower-dimensional intersections of traces,

$$\mathrm{Sing}(\Phi^A_{\ell,\rho}|_C) = \bigcup_{j \in J^{A,\rho}_\ell(C)} \Gamma^A_j(C).$$

The analogous statement holds for the pulled-back B-block. Its used trace family is

$$\Gamma^B_i(E_\ell; C) = \{z \in C : (E_\ell z)_i = 0\},$$

and, again up to lower-dimensional trace intersections,

$$\mathrm{Sing}((\Phi^B_{\ell,\rho} \circ E_\ell)|_C) = \bigcup_{i \in J^{B,\rho}_\ell(E_\ell;C)} \Gamma^B_i(E_\ell; C).$$

By Lemma 1,

$$\Phi^B_{\ell,\rho} \circ E_\ell = E_{\ell+1} \circ \Phi^A_{\ell,\rho} \qquad \text{on } C.$$

Writing $E_{\ell+1}(y) = T_{\ell+1}y + a_{\ell+1}$, a derivative jump J of the A-block is sent to $T_{\ell+1}J$. Since $T_{\ell+1}$ is injective, $J \neq 0$ implies $T_{\ell+1}J \neq 0$. Thus postcomposition cannot erase an A-side kink. The equality above also prevents the left side from having a kink where the right side is differentiable. Therefore the two displayed singular sets coincide on the task patch.

Now fix a connected private open patch $U \subset \Gamma^A_j(C)$. It is contained in the finite union of the pulled-back B-traces. If no one of those traces contained a relatively open subset of U, then each intersection with U would have codimension at least one inside U and finitely many such intersections could not cover U. Hence some B-trace agrees with $\Gamma^A_j(C)$ on a relatively open patch. The trace-separation part of usedness makes the match unique, and two distinct A-traces cannot be assigned to the same separated B-trace. We obtain an injective matching $i = \pi(j)$.

For such a matched pair, the nonzero affine functions

$$z \longmapsto z_j, \qquad z \longmapsto (E_\ell z)_i$$

have the same regular codimension-one zero set on an open subset of V. Two nonzero affine functionals on V with the same zero hyperplane are proportional, so there is $\alpha_j \neq 0$ with

$$(E_\ell z)_{\pi(j)} = \alpha_j z_j \qquad \forall z \in C.$$

This proves the axis-matching statement.

Finally, a task solution of loss at most ε must contain at least $m^\rho_\ell(\varepsilon)$ used layer-ℓ gates. Every such gate is matched by the construction above, so at least that many A-axes contribute to the numerator of $\mathrm{AxisAlign}_{\ell,\rho}$. Dividing by d^A_ℓ yields the stated lower bound. □

A.2.2 SOFTPLUS: USEDNESS VIA AFFINE RIDGE CURVATURE

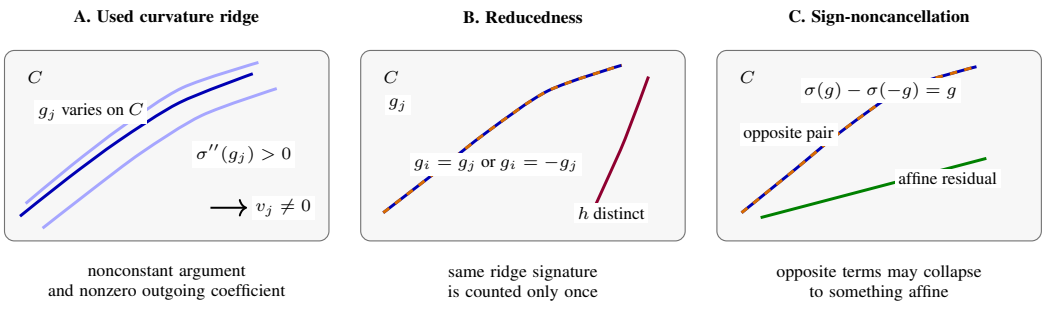


Figure 6: Softplus usedness as a curvature-based analogue of ReLU usedness. Panel A shows the basic requirement: a Softplus term must have a nonconstant argument and a nonzero outgoing coefficient, so that its smooth curvature is visible on the task patch. Panel B illustrates reducedness: duplicate or opposite arguments generate the same curvature signature and should not be counted as distinct nonlinear axes. Panel C illustrates sign-noncancellation: because $\sigma(g) - \sigma(-g) = g$, a collection of opposite-sign terms can leave only an affine residual unless this possibility is explicitly ruled out.

In the Softplus case, there is no exact kink. Instead, every nonconstant Softplus argument contributes smooth curvature through $\sigma''(g)$. Figure 6 illustrates why this is the analogue of ReLU usedness. A ReLU gate is used when the task crosses a kink trace and the next affine map reads out the resulting slope jump. A Softplus term is used when the argument genuinely varies, its outgoing coefficient is nonzero, and the resulting curvature ridge is not being counted redundantly through duplicates, opposites, or affine sign-cancellations. Intuitively, in proving the weak–strong equivalence in this setting, our job will be to show that adjacent affine mappability forces all relevant slopes within an expansion to be matched.

We posit that the conditions analogous to ReLU usedness are:

Definition 6 (Used Softplus expansion)**.** After deleting unused terms and absorbing constant arguments into the bias, a Softplus expansion is *used* if:

(i) every $v_j \neq 0$ and every g_j is nonconstant;

(ii) it is reduced: $g_i = \pm g_j$ implies $i = j$;

(iii) it is sign-noncancelling: for every nonempty S,

$$\inf_{a \in \mathbb{R}^m} \left\| \sum_{j \in S} v_j g_j - a \right\|_{L^2} > 0.$$

To understand what motivated these conditions as replacement for the ReLU usedness condition, we show that these conditions enable a softplus version of shrinkability. For ReLU, a gate can be shrinkable because the patch may not cross its zero trace; then $\rho(z_j)$ is either zero or affine on that patch. For Softplus, there is no "inactive side": if g is nonconstant and the outgoing coefficient is nonzero, then $\sigma(g)$ contributes genuine smooth curvature on any open patch, because $\sigma''(t) > 0$ everywhere. The Softplus shrinkability result therefore needs to be about algebraic redundancies: zero coefficients, constant arguments, duplicate arguments, opposite arguments, and the sign ambiguity coming from the softplus identity $\sigma(t) - \sigma(-t) = t$. Specifically:

Lemma 3 (Shrinkability of Softplus degeneracies)**.** *Let C be a task patch with nonempty relative interior in $V = \mathrm{aff}(C)$. Consider a vector-valued Softplus expansion on C*

$$F(z) = a(z) + \sum_{j=1}^{k} v_j \sigma(g_j(z)), \qquad z \in C,$$

where $a : V \to \mathbb{R}^m$ is affine, each $v_j \in \mathbb{R}^m$, and each $g_j : V \to \mathbb{R}$ is affine. Then the following operations preserve F and do not increase the number of nonlinear Softplus terms:

1. *If $v_j = 0$, the j-th term may be deleted.*
2. *If $g_j \equiv \gamma$ on V, the j-th term may be deleted and the affine part replaced by*
$$a(z) \longmapsto a(z) + v_j \sigma(\gamma).$$
3. *If $g_i = g_j$ on V, then the two terms may be replaced by one term:*
$$v_i \sigma(g_i) + v_j \sigma(g_j) = (v_i + v_j)\sigma(g_j).$$
If $v_i + v_j = 0$, both nonlinear terms disappear.
4. *If $g_i = -g_j$ on V, then the two terms may be replaced by one Softplus term plus an affine residual:*
$$v_i \sigma(g_i) + v_j \sigma(g_j) = (v_i + v_j)\sigma(g_j) - v_i g_j.$$
Since g_j is affine, the residual $-v_i g_j$ can be absorbed into the affine part a. If $v_i + v_j = 0$, no nonlinear Softplus term remains from this pair.
5. *More generally, if for a nonempty set S*
$$\sum_{j \in S} v_j g_j$$
is constant on C, then simultaneously replacing g_j by $-g_j$ for all $j \in S$ changes the expansion only by a bias term.

Consequently, modulo affine residuals, every finite Softplus expansion can be reduced to one in which all remaining nonlinear terms have nonzero coefficient, nonconstant argument, and no two remaining arguments are equal up to sign. The additional affine sign-noncancellation condition rules out the remaining orientation ambiguity caused by $\sigma(t) - \sigma(-t) = t$.

Proof. The zero-coefficient and constant-argument cases are immediate. If $v_j = 0$, the term contributes nothing. If $g_j \equiv \gamma$, then $v_j\sigma(g_j) = v_j\sigma(\gamma)$, which is constant and can be folded into the affine part.

If $g_i = g_j$, then

$$v_i\sigma(g_i) + v_j\sigma(g_j) = (v_i + v_j)\sigma(g_j),$$

so duplicate arguments merge.

If $g_i = -g_j$, use the Softplus identity

$$\sigma(t) - \sigma(-t) = t, \qquad \text{equivalently} \qquad \sigma(-t) = \sigma(t) - t.$$

Writing $g_i = -g_j$, we get

$$v_i\sigma(g_i) + v_j\sigma(g_j) = v_i\sigma(-g_j) + v_j\sigma(g_j) = v_i(\sigma(g_j) - g_j) + v_j\sigma(g_j),$$

hence

$$v_i\sigma(g_i) + v_j\sigma(g_j) = (v_i + v_j)\sigma(g_j) - v_ig_j.$$

The last term is affine in z, because g_j is affine on V, so it can be absorbed into a.

For the final claim, suppose

$$\sum_{j\in S} v_jg_j \equiv c$$

is constant on C. Then

$$\sum_{j\in S} v_j\sigma(-g_j) = \sum_{j\in S} v_j(\sigma(g_j) - g_j) = \sum_{j\in S} v_j\sigma(g_j) - c.$$

Thus flipping all signs in S changes the represented function only by the constant $-c$, which can be absorbed into the bias. This is exactly the orientation ambiguity excluded by affine sign-noncancellation. □

Thus, we can let $m^\sigma_\ell(\varepsilon)$ denote the budget count for such softplus used axes, exactly in analogy the ReLU case.[1] We can now also formulate and prove the softplus version of weak-strong equivalence:

Theorem 2 (Exact Softplus weak–strong equivalence)**.** *Fix $C \subseteq z^A_\ell(\Omega)$ and suppose exact weak equivalence holds at layers ℓ and $\ell + 1$. Assume the transformed A-side and pulled-back B-side adjacent expansions are used in the sense of Definition 6. Then there is an injective matching*

$$\pi : J^{A,\sigma}_\ell(C) \to J^{B,\sigma}_\ell(E_\ell; C)$$

with

$$(E_\ell z)_{\pi(j)} = z_j \qquad \forall z \in C.$$

If A reaches loss at most ε, then

$$\mathrm{AxisAlign}_{\ell,\sigma}(A, B; E_\ell, C) \geq \frac{m^\sigma_\ell(\varepsilon)}{d^A_\ell}.$$

Proof. As above, write $E_r(y) = T_ry + a_r$. As with the ReLU case, by Lemma 1, exact weak equivalence at layers ℓ and $\ell + 1$ gives

$$\Phi^B_{\ell,\sigma}(E_\ell z) = E_{\ell+1}\Phi^A_{\ell,\sigma}(z), \qquad z \in C.$$

[1] The intricacy of the above argument is why, for the weak-strong equivalence-type of result, the ReLU case is easier to think about that the softplus case. Interestingly, the opposite will be true for the zippering-type results, where monodromy makes our lives easier. Ideally we could find a unification of all the arguments that reduce all this complexity.

After deleting unused terms and absorbing constant arguments into the bias, this equality can be written as

$$\widehat{b} + \sum_{j=1}^{k} \widehat{v}_j \sigma(g_j(z)) = \widetilde{b} + \sum_{i=1}^{\widetilde{k}} \widetilde{v}_i \sigma(\widetilde{g}_i(z)), \qquad z \in C, \tag{7}$$

where the notations are described above, e.g. $g_j(z) = z_j$, $\widetilde{g}_i(z) = (E_\ell z)_i$, $\widehat{b} = T_{\ell+1} b^A_{\ell+1} + a_{\ell+1}$, $\widehat{v}_j = T_{\ell+1} W^A_{\ell+1}[:, j]$, $\widetilde{b} = b^B_{\ell+1}$, and $\widetilde{v}_i = W^B_{\ell+1}[:, i]$. Since $T_{\ell+1}$ is injective, it preserves nonzero outgoing coefficients: $W^A_{\ell+1}[:, j] \neq 0$ implies $\widehat{v}_j \neq 0$. It also preserves sign-noncancellation. Indeed, if for some nonempty set S the transformed sum

$$\sum_{j \in S} \widehat{v}_j g_j = T_{\ell+1} \left(\sum_{j \in S} W^A_{\ell+1}[:, j] g_j \right)$$

were constant on C, then taking differences between any two points of C and using injectivity of $T_{\ell+1}$ would imply that $\sum_{j \in S} W^A_{\ell+1}[:, j] g_j$ was already constant on C, contrary to the sign-noncancellation assumption. Thus the transformed A-side expansion remains used, reduced, nonconstant, and sign-noncancelling; the pulled-back B-side expansion has these properties by hypothesis.

Now, let V be the affine hull of the task patch C and let L be its tangent space. We choose z_0 in the relative interior of C and choose $\xi \in L$ "generically": we exclude all pairs (z_0, ξ) for which some nonconstant argument has zero slope on the line ξ. We also exclude accidental sign-coincidences on the chosen line, i.e. for every pair q, q' and every $\delta \in \{+1, -1\}$ such that $q \neq \delta q'$ as affine functions on V, we exclude the set

$$\big\{(z_0, \xi) : q(z_0) - \delta q'(z_0) = 0,\ (\mathrm{lin}(q) - \delta\, \mathrm{lin}(q'))(\xi) = 0\big\}$$

where $\mathrm{lin}(q)$ is the linear part of the operation q. This is a finite union of proper affine-algebraic subsets of $\mathrm{relint}(C) \times L$, and so has measure 0. Thus we can generically choose (z_0, ξ) outside this "bad" exceptional set.

Then every restricted argument has nonzero slope, and two restricted affine functions agree up to sign only if the original affine functions already agree up to that sign on all of V. For $|t|$ small, the line $z(t) = z_0 + t\xi$ lies in C. Along this line write

$$g_j(z_0 + t\xi) = s_j t + r_j, \qquad \widetilde{g}_i(z_0 + t\xi) = \widetilde{s}_i t + \widetilde{r}_i,$$

with $s_j \neq 0$ and $\widetilde{s}_i \neq 0$ for every i, j. Restricting (7) to the line gives equality of real-analytic functions on a nonempty interval, hence on all of $\mathbb{R}$:

$$\widehat{b} + \sum_{j=1}^{k} \widehat{v}_j \sigma(s_j t + r_j) = \widetilde{b} + \sum_{i=1}^{\widetilde{k}} \widetilde{v}_i \sigma(\widetilde{s}_i t + \widetilde{r}_i). \tag{8}$$

Now put

$$\eta(t) = \sigma''(t) = \frac{1}{4 \cosh^2(t/2)}.$$

Differentiating (8) twice yields:

$$\sum_{j=1}^{k} \widehat{v}_j s_j^2 \eta(s_j t + r_j) = \sum_{i=1}^{\widetilde{k}} \widetilde{v}_i \widetilde{s}_i^2 \eta(\widetilde{s}_i t + \widetilde{r}_i). \tag{9}$$

This is a so-called "ridge identity": an equality between finite sums of translated and rescaled curvature bumps obtained by slicing the original affine ridges along the generic line. The curvature term $z \mapsto \eta(q(z))$, is a "ridge function": it is constant along the affine slices $q(z) = \mathrm{constant}$ and varies only in the normal direction. Restricting to a generic line converts this geometric ridge into a translated and rescaled one-dimensional bump $t \mapsto \eta(st + r)$, whose slope and offset can then be identified by Fourier decay and phase. This line-restriction method is standard in the ridge-function literature; see Pinkus (2015).

Following this same line of literature, we now take Fourier transforms to separate the one-D bumps for analysis. For Softplus curvature,

$$\widehat{\eta}(\omega) = \frac{\pi\omega}{\sinh(\pi\omega)}.$$

For every $s \neq 0$,

$$\mathcal{F}\{s^2\eta(st + r)\}(\omega) = |s|e^{i\omega r/s}\widehat{\eta}(\omega/s).$$

Taking Fourier transforms in (9) gives

$$\sum_j \widehat{v}_j|s_j|e^{i\omega r_j/s_j}\widehat{\eta}(\omega/s_j) = \sum_i \widetilde{v}_i|\widetilde{s}_i|e^{i\omega\widetilde{r}_i/\widetilde{s}_i}\widehat{\eta}(\omega/\widetilde{s}_i). \tag{10}$$

As $\omega \to +\infty$,

$$|s|\,\widehat{\eta}(\omega/s) = 2\pi\omega e^{-\pi\omega/|s|}(1 + o(1)). \tag{11}$$

Thus the terms with largest absolute slope have the slowest exponential decay and can be isolated from all smaller-slope terms. To that end, let

$$S = \max\Big(\{|s_j| : 1 \le j \le k\} \cup \{|\widetilde{s}_i| : 1 \le i \le \widetilde{k}\}\Big).$$

Multiply (10) by $e^{\pi\omega/S}/(2\pi\omega)$ and use (11). Every term with absolute slope strictly smaller than S tends to zero. The maximal-slope terms satisfy

$$P_S(\omega) := \sum_{|s_j|=S} \widehat{v}_j e^{i\omega r_j/s_j} - \sum_{|\widetilde{s}_i|=S} \widetilde{v}_i e^{i\omega\widetilde{r}_i/\widetilde{s}_i} = o(1). \tag{12}$$

For a maximal-slope term, define its phase frequency by $\tau = r/s$. Among affine functions with the same absolute slope S, equality of τ is exactly equality up to sign as one-dimensional affine functions. Indeed, if $s' = \pm s$ and $r'/s' = r/s$, then $r' = \pm r$ with the same sign, so $s't + r' = \pm(st + r)$.

Group the terms in (12) by this common phase frequency. For each frequency τ, define the grouped vector coefficient

$$c_\tau = \sum_{\substack{j:\ |s_j|=S \\ r_j/s_j=\tau}} \widehat{v}_j - \sum_{\substack{i:\ |\widetilde{s}_i|=S \\ \widetilde{r}_i/\widetilde{s}_i=\tau}} \widetilde{v}_i. \tag{13}$$

With this notation, $P_S(\omega) = \sum_\tau c_\tau e^{i\omega\tau}$. We claim that all of these coefficients c_τ are zero. Since $P_S(\omega) = o(1)$, we have

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \|P_S(\omega)\|^2\, d\omega = 0.$$

On the other hand,

$$\begin{aligned}\frac{1}{T}\int_0^T \|P_S(\omega)\|^2\, d\omega &= \frac{1}{T}\int_0^T \left\|\sum_\tau c_\tau e^{i\omega\tau}\right\|^2 d\omega \\ &= \sum_{\tau,\tau'}\langle c_\tau, c_{\tau'}\rangle \left(\frac{1}{T}\int_0^T e^{i\omega(\tau-\tau')}\, d\omega\right).\end{aligned} \tag{14}$$

If $\tau = \tau'$, the parenthesized factor is 1. If $\tau \neq \tau'$, then

$$\frac{1}{T}\int_0^T e^{i\omega(\tau-\tau')}\, d\omega = \frac{e^{iT(\tau-\tau')} - 1}{iT(\tau - \tau')} \longrightarrow 0.$$

Letting $T \to \infty$ in (14) gives

$$0 = \sum_\tau \|c_\tau\|^2.$$

Therefore $c_\tau = 0$ for every maximal-slope frequency τ. By reducedness and by the generic choice of the line, each frequency group contains at most one A-side term and at most one B-side term. Since all coefficient vectors are nonzero, $c_\tau = 0$ implies that every maximal-slope A-term is paired with a unique maximal-slope B-term, and conversely. For such a pair (j, i),

$$\widetilde{g}_i(z_0 + t\xi) = \varepsilon_j g_j(z_0 + t\xi) \qquad \text{for all } t,$$

for some $\varepsilon_j \in \{+1, -1\}$, and $\widetilde{v}_i = \widehat{v}_j$. Because η is an even function and the slopes differ only by sign, the paired terms contribute identically to the ridge identity (9). Subtract all such matched maximal-slope pairs from both sides of (9). If any terms remain, repeat the same largest-slope Fourier isolation with the next-largest absolute slope. Since the number of terms is finite, this iteration terminates.

Thus all restricted affine arguments have been matched. We obtain a bijection

$$\pi : \{1, \dots, k\} \to \{1, \dots, \widetilde{k}\}$$

and signs $\varepsilon_j \in \{+1, -1\}$ such that

$$\widetilde{g}_{\pi(j)}(z_0 + t\xi) = \varepsilon_j g_j(z_0 + t\xi) \qquad \text{for all } t, \tag{15}$$

and

$$\widetilde{v}_{\pi(j)} = \widehat{v}_j. \tag{16}$$

The generic-line choice promotes this linewise equality to equality on all of V. Indeed, if $\widetilde{g}_{\pi(j)} - \varepsilon_j g_j$ were not identically zero on V, then (15) would mean that the chosen pair (z_0, ξ) lies in one of the excluded accidental sign-coincidence sets. Hence

$$\widetilde{g}_{\pi(j)}(z) = \varepsilon_j g_j(z) \qquad \forall z \in V. \tag{17}$$

It remains to remove the negative orientations. Substitute (17) and (16) into the original expansion identity (7). After reindexing by π,

$$\widehat{b} + \sum_j \widehat{v}_j \sigma(g_j) = \widetilde{b} + \sum_j \widehat{v}_j \sigma(\varepsilon_j g_j).$$

Let $S_- = \{j : \varepsilon_j = -1\}$. Cancelling the positively oriented matched terms gives

$$\widehat{b} - \widetilde{b} + \sum_{j \in S_-} \widehat{v}_j \big(\sigma(g_j) - \sigma(-g_j)\big) = 0.$$

Using the Softplus identity $\sigma(t) - \sigma(-t) = t$, we obtain

$$\widehat{b} - \widetilde{b} + \sum_{j \in S_-} \widehat{v}_j g_j = 0. \tag{18}$$

If S_- were nonempty, (18) would say that a nonempty downstream-weighted sum of arguments is constant on C, contradicting sign-noncancellation. Therefore $S_- = \varnothing$. All signs are positive, and so

$$\widetilde{g}_{\pi(j)}(z) = g_j(z) \quad \forall z \in V.$$

Since $g_j(z) = z_j$ and $\widetilde{g}_i(z) = (E_\ell z)_i$, this is exactly

$$(E_\ell z)_{\pi(j)} = z_j \qquad \forall z \in C.$$

This proves the strong Softplus axis-matching statement.

Finally, if A reaches loss at most ε, then by definition of the Softplus used-axis budget it uses at least $m_\ell^\sigma(\varepsilon)$ layer-ℓ Softplus axes. The argument above matches every such used A-side axis to a distinct pulled-back B-side axis. Therefore at least $m_\ell^\sigma(\varepsilon)$ axes contribute to the numerator of $\mathrm{AxisAlign}_{\ell,\sigma}$, and

$$\mathrm{AxisAlign}_{\ell,\sigma}(A, B; E_\ell, C) \geq \frac{m_\ell^\sigma(\varepsilon)}{d_\ell^A}.$$

□

A.3 Asymptotic weak–strong equivalence

We now develop a soft versions of weak-strong equivalence, by measuring quantitative defects in alignment and then relating them. The structure of the definitions and theorem proof is identical for both ReLU and softplus activations.

Definition 7 (Weak alignment defect)**.** Define

$$\omega_\ell(E_\ell; A, B) = \left(\mathbb{E}_\Omega \| z_\ell^B(x) - E_\ell z_\ell^A(x) \|_2^2\right)^{1/2}.$$

By definition, $\omega_\ell = 0$ implies exact weak alignment.

Definition 8 (Soft axis score)**.** For $\varphi \in \{\sigma, \rho\}$, define

$$s_{ij}^{\varphi}(E_\ell; C) = \sup_{\alpha \in \mathcal{H}_\varphi} \left[1 - \frac{\|(E_\ell z)_i - \alpha z_j\|_{L^2(C)}}{\|(E_\ell z)_i\|_{L^2(C)} + |\alpha| \|z_j\|_{L^2(C)}}\right].$$

Let $\mathrm{AxisAlign}_{\ell,\varphi}^{\theta}$ be the maximal fraction of counted A-axes that can be injectively matched with score at least θ.

Theorem 3 (Asymptotic weak–strong equivalence)**.** *Fix $\varphi \in \{\sigma, \rho\}$. Let $\mathcal{K}$ be a compact family of adjacent comparison tuples $p = (A, B, E_\ell, E_{\ell+1})$. For every $\theta \in (0, 1)$ there is $\kappa_{\mathcal{K},\varphi}(\theta) > 0$ such that*

$$\mathrm{AxisAlign}_{\ell,\varphi}^{\theta}(A, B; E_\ell, C) \geq \frac{m_\ell^{\varphi}(\varepsilon)}{d_\ell^A} - \frac{\left(\|W_{\ell+1}^B\|\, \omega_\ell + \omega_{\ell+1}\right)^2}{\kappa_{\mathcal{K},\varphi}(\theta)^2 d_\ell^A}.$$

Moreover, if $\{p_n = (A_n, B_n, E_{\ell,n}, E_{\ell+1,n})\}$ is a sequence of comparison tuples in $\mathcal{K}$ for which the defect sequences $\omega_{\ell,n} := \omega_\ell(E_{\ell,n}; A_n, B_n)$ and $\omega_{\ell+1,n} := \omega_{\ell+1}(E_{\ell+1,n}; A_n, B_n)$ both converge to zero, then there exists N such that, for all $n \geq N$,

$$\mathrm{AxisAlign}_{\ell,\varphi}^{\theta}(A_n, B_n; E_{\ell,n}, C_n) \geq \frac{m_\ell^{\varphi}(\varepsilon)}{d_{\ell,n}^A}.$$

In words, every axis forced by the task budget is eventually θ-aligned.

Proof. For a comparison tuple $p = (A, B, E_\ell, E_{\ell+1})$, let $B^\theta(p)$ be the minimum number of counted task-required A-axes that fail threshold θ under the best injective matching into the counted B-axes. Define the *commutation defect*

$$\Delta_{\ell,\varphi}(p) = \|\Phi_{\ell,\varphi}^B(E_\ell z_\ell^A) - E_{\ell+1}\Phi_{\ell,\varphi}^A(z_\ell^A)\|_{L^2(\Omega)}.$$

Now let

$$\kappa_{\mathcal{K},\varphi}(\theta) = \inf_{\substack{p \in \mathcal{K} \\ B^\theta(p) > 0}} \frac{\Delta_{\ell,\varphi}(p)}{\sqrt{B^\theta(p)}}.$$

If the indexing set is empty, every tuple in $\mathcal{K}$ already has all counted axes θ-aligned and the theorem is immediate. Otherwise we first show that the infimum is positive.

Assume for contradiction that it is zero. Then there is a sequence $p_n \in \mathcal{K}$ with $B^\theta(p_n) > 0$ and

$$\frac{\Delta_{\ell,\varphi}(p_n)^2}{B^\theta(p_n)} \longrightarrow 0.$$

The number of counted axes is uniformly bounded by d_ℓ^A, so $\Delta_{\ell,\varphi}(p_n) \to 0$. Compactness gives a convergent subsequence $p_n \to p_* \in \mathcal{K}$. The network operations, affine comparison maps, and L^2 norms vary continuously on the bounded family, hence

$$\Delta_{\ell,\varphi}(p_*) = 0.$$

The zero-defect member satisfies the hypotheses of the appropriate exact weak–strong theorem. It therefore admits an injective matching π_* for which every counted task-required axis has score exactly one. There are only finitely many matched pairs, and each pairwise score is continuous in the tuple parameters on the quantitatively nondegenerate family. Since $\theta < 1$, the same fixed matching has every score greater than θ for all sufficiently large n. This implies $B^\theta(p_n) = 0$, contradicting the choice of the sequence. Thus

$$\kappa_{\mathcal{K},\varphi}(\theta) > 0.$$

By the definition of the infimum,

$$B^\theta(p) \leq \frac{\Delta_{\ell,\varphi}(p)^2}{\kappa_{\mathcal{K},\varphi}(\theta)^2}.$$

It remains to relate the commutation defect to the two weak-alignment errors. Put $Z = z_\ell^A(x)$. Since $\Phi_{\ell,\varphi}^B(z_\ell^B(x)) = z_{\ell+1}^B(x)$,

$$\Phi_{\ell,\varphi}^B(E_\ell Z) - E_{\ell+1}\Phi_{\ell,\varphi}^A(Z)$$

$$= \big[\Phi^B_{\ell,\varphi}(E_\ell z^A_\ell(x)) - \Phi^B_{\ell,\varphi}(z^B_\ell(x))\big] + \big[z^B_{\ell+1}(x) - E_{\ell+1} z^A_{\ell+1}(x)\big].$$

Both ReLU and Softplus are one-Lipschitz, so

$$\|\Phi^B_{\ell,\varphi}(u) - \Phi^B_{\ell,\varphi}(v)\|_2 \le \|W^B_{\ell+1}\|\,\|u - v\|_2.$$

Taking $L^2(\Omega)$ norms in this identity and applying the triangle inequality yields

$$\Delta_{\ell,\varphi}(p) \le \|W^B_{\ell+1}\|\,\omega_\ell(p) + \omega_{\ell+1}(p).$$

At least $m^\varphi_\ell(\varepsilon)$ used axes are task-required, and at most $B^\theta(p)$ of those fail threshold θ. Therefore

$$\mathrm{AxisAlign}^\theta_{\ell,\varphi} \ge \frac{m^\varphi_\ell(\varepsilon) - B^\theta(p)}{d^A_\ell}.$$

Combining the bad-axis bound with the defect bound proves the displayed inequality.

Finally, suppose $p_n \in \mathcal{K}$ and both weak errors tend to zero. Compactness uniformly bounds $\|W^{B_n}_{\ell+1}\|$, so the defect bound gives $\Delta_{\ell,\varphi}(p_n) \to 0$. The bad-axis bound then gives $B^\theta(p_n) \to 0$. Since $B^\theta(p_n)$ is a nonnegative integer, it equals zero for all sufficiently large n. Thus every task-required counted axis is eventually θ-aligned. □

B From one-step identifiability to representational zippering

B.1 Minimality, regularity, and exact zippering

Weak–strong alignment views the nonlinear block as a function of the layer coordinate z. Its arguments are affine: z_j on one side and $(E_\ell z)_i$ on the other. Zippering starts with only the next representation aligned and asks whether the unknown earlier scalar arguments $z^N_{r,j}(x)$ can be recovered as functions of the original input. The obstruction is a *collision*: two networks with the same deep input–output relationship but different internal structures.

Our high-level strategy for doing this is to:

1. First strengthen the notion of usedness from single axes or gates to *minimality* of entire layerwise expansions (§B.1.2). Minimality is an intrinsic condition on each expansion: every counted unit leaves a private task-visible nonlinear signature. If two one-step expansions are equal, each private signature on one side must be reproduced on the other, giving a unique cross-match.
2. We then seek to understand what happens when two minimal layers collide. We show that the property of *identifiability* – that is, the recovery of equal internal constituent operations – can be ensured by a set of very simple conditions on *regularity* (§B.1.3 and §B.1.4).
3. We then show how identifiability, based on minimality and regularity, leads to exact zippering (§B.1) under perfect terminal equivalence of the last-layer representations.

With an additional Jacobian-regularity condition, the same strategy applies when terminal equivalence is only approximate (§B.2). We then prove that the regularity conditions are generic (§B.3).

B.1.1 Null pairs and identifiability

We start with some basic definitions:

Definition 9 (Realization map, null pairs, collisions, and identifiability)**.** Fix $\varphi \in \{\sigma, \rho\}$. Let Θ^φ_r be the manifold of layer-r one-step expansions. Define the *realization map* by

$$F^\varphi_r : \Theta^\varphi_r \longrightarrow L^2(\Omega; \mathbb{R}^{d_{r+1}}), \qquad F^\varphi_r : \vartheta \longmapsto Y^\varphi_{G_\vartheta},$$

so called because F^φ_r is the function that sends one-step parameters to the function they realize. Two parameter values $\vartheta, \widetilde{\vartheta}$ form a *null pair*, or *collision*, if

$$F^\varphi_r(\vartheta) = F^\varphi_r(\widetilde{\vartheta})$$

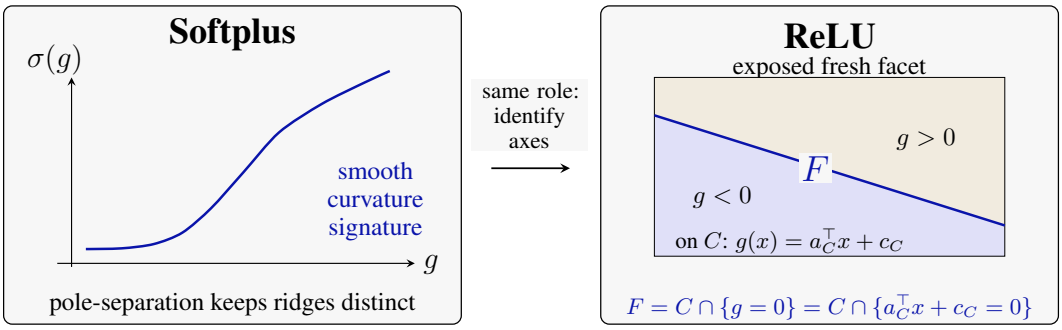


Figure 7: Softplus and ReLU use different nonlinear signatures to identify hidden axes. For softplus, the identifying signature is a smooth curvature pattern, with pole-separation ensuring that distinct ridges are not confused. For ReLU, the analogous signature is an exposed fresh facet: inside a cell C where the inherited lower computation is affine, $g(x) = a_C^\top x + c_C$, and the zero set $F = C \cap \{g = 0\} = C \cap \{a_C^\top x + c_C = 0\}$ separates the regions $g < 0$ and $g > 0$.

e.g. they represent the same function on Ω. The corresponding *collision set* is

$$\mathcal{C}_r^\varphi = \{(\vartheta, \widetilde{\vartheta}) : F_r^\varphi(\vartheta) = F_r^\varphi(\widetilde{\vartheta})\}.$$

Each activation function has some unavoidable hidden-unit symmetries–permutation for Softplus, and permutation plus positive rescaling for ReLU–the "obvious self-collisions" form the *diagonal*

$$\Delta_r^\varphi = \{(\vartheta, \vartheta) : \vartheta \in \Theta_r^\varphi\}.$$

A collision is *nontrivial* if it lies away from this diagonal after the allowed hidden-unit matching has been fixed. When $\mathcal{C}_r^\varphi$ is decomposed into smooth pieces, a *collision stratum* means one such smooth piece. An *off-diagonal collision stratum* is a smooth piece contained in $\mathcal{C}_r^\varphi \setminus \Delta_r^\varphi$.

The opposite of collision is *identifiability*. In the Softplus case, identifiability means that, after reindexing,

$$\widetilde{g}_j = g_j, \qquad \widetilde{v}_j = v_j, \qquad \widetilde{b} = b.$$

In the ReLU case, because $\rho(\alpha_j g_j) = \alpha_j \rho(g_j)$, identifiability means that, after reindexing, there are $\alpha_j > 0$ such that

$$\widetilde{g}_j = \alpha_j g_j, \qquad \alpha_j \widetilde{v}_j = v_j, \qquad \widetilde{b} = b.$$

The exact zippering proof is a common backward induction once this one-step identifiability target has been verified. We now verify it by two different signature theories.

B.1.2 Intrinsic minimality and private cross-matching

Intrinsic minimality is the part of the argument that does not yet try to reconstruct an entire hidden scalar function. Its job is simpler: it ensures that each counted unit has a private nonlinear signature, and therefore that equality of two one-step expansions forces a unique cross-side match for that signature. For Softplus the private signature is an "outer pole patch" (Fig. 7, left). For ReLU it is an "exposed fresh facet" (Fig. 7, right).

Softplus minimality and private pole cross-matching. The adjacent Softplus theorem used only affine-ridge usedness. In a deep one-step expansion the arguments $g_j = z_{r,j}(x)$ are no longer affine functions of the compared layer coordinate; they inherit all lower-layer nonlinear structure. Softplus minimality adds the internal pole-separation condition saying that each sign-class leaves a private pole signature that cannot be confused with another sign-class in the same expansion:

Definition 10 (Softplus pole terminology and minimality)**.** Let g be a real-valued argument, and let $g^{\mathbb{C}}$ denote a chosen analytic continuation on a common complex continuation domain. Write

$$Q_g = 1 + e^{g^{\mathbb{C}}}.$$

The zero set $\{Q_g = 0\}$ is the *outer Softplus pole divisor* of g. (The word "outer" is used to signify that this divisor comes from the current Softplus factor $\sigma(g^{\mathbb{C}}) = \log(1 + e^{g^{\mathbb{C}}})$, not from singularities

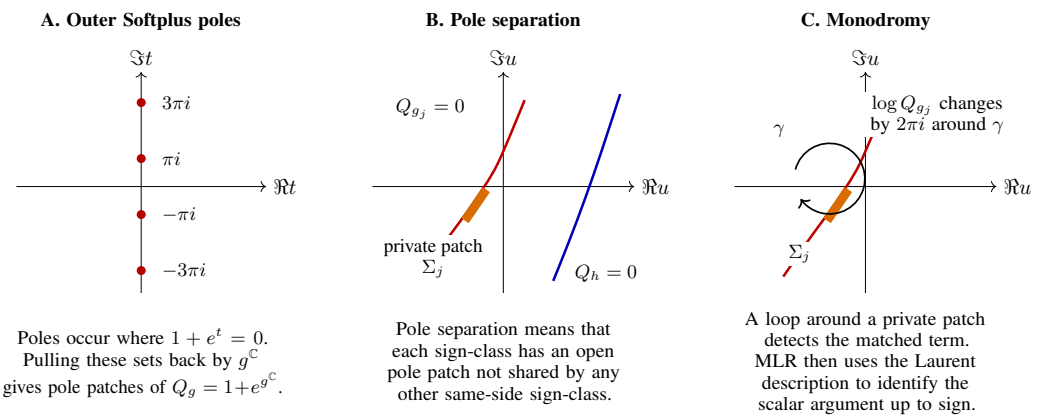


Figure 8: Softplus pole signatures used in the zippering proof. Panel A shows the complex pole locations that define the outer factors $Q_g = 1 + e^{g^{\mathbb{C}}}$. Panel B illustrates pole separation: a sign-class must have a private open pole patch not shared by other sign-classes in the same expansion. Panel C illustrates the monodromy test used for cross-matching and regularity. A small loop around a private patch changes $\log Q_{g_j} = \sigma(g_j^{\mathbb{C}})$ by $2\pi i$; equality of two one-step expansions forces the same monodromy on the other side, and MLR is the condition that turns the resulting matched pole data into identification of the underlying scalar argument up to sign.

already inherited inside the lower-depth argument $g^{\mathbb{C}}$. The term *divisor* is used in the analytic geometry sense of being the zero set of a holomorphic factor Q_g.) An *outer Softplus pole patch* of g is a connected smooth codimension-one patch $\Sigma \subset \{Q_g = 0\}$ on which $dQ_g \neq 0$, or, equivalently, a small loop around Σ changes the branch of $\log Q_g = \sigma(g^{\mathbb{C}})$ by $2\pi i$. Since

$$Q_{-g} = 1 + e^{-g^{\mathbb{C}}} = e^{-g^{\mathbb{C}}} Q_g,$$

the arguments g and $-g$ have the same outer Softplus pole divisor. Thus pole patches naturally belong to the sign-class $[g] = \{g, -g\}$. A pole patch for $[g]$ is *private* to an expansion if no different sign-class in the same expansion shares a nonempty open piece of that same smooth pole surface. The expansion is *pole-separated modulo sign* if every sign-class has such a private outer Softplus pole patch. For two sign-classes $[g]$ and $[h]$, we say that they are *divisor-matched* if an outer Softplus pole patch of $[g]$ and an outer Softplus pole patch of $[h]$ share a nonempty open piece of the same smooth pole surface, or, equivalently, $[g]$ and $[h]$ have the same local Softplus pole signature on that patch. When the two sign-classes lie on opposite sides of a null-pair equality, we also call this a *cross-match*. A Softplus one-step expansion is *minimal* if it is used in the sense of Definition 6 and is pole-separated modulo sign. Thus Softplus minimality is adjacent usedness plus the internal pole-separation condition needed to make each sign-class visible.

Figure 8 sketches the pole-separation and monodromy ideas implicit in this definition. The key result of the softplus minimality definition is the following result:

Lemma 4 (A private Softplus pole patch has a unique cross match)**.** *Suppose*

$$b + \sum_{j=1}^{k} v_j \sigma(g_j) = \widetilde{b} + \sum_{i=1}^{\widetilde{k}} \widetilde{v}_i \sigma(\widetilde{g}_i)$$

on Ω. If both expansions are used and individually pole-separated modulo sign, every private pole patch on one side is shared by a unique sign-class on the other side.

Proof. We prove the statement for a private pole patch on the left; the argument with the two sides interchanged is identical. Choose a private pole patch Σ for the sign-class $[g_j]$. Since

$$Q_{-g_j} = 1 + e^{-g_j^{\mathbb{C}}} = e^{-g_j^{\mathbb{C}}}(1 + e^{g_j^{\mathbb{C}}}) = e^{-g_j^{\mathbb{C}}} Q_{g_j},$$

g_j and $-g_j$ have the same outer pole divisor. Thus, after choosing the representative if necessary, we may write

$$\Sigma \subset \{Q_j = 0\}, \qquad Q_j := Q_{g_j} = 1 + e^{g_j^{\mathbb{C}}}, \qquad dQ_j \neq 0 \text{ on } \Sigma.$$

Pick a point $p \in \Sigma$ generically: avoid singular points of all pole divisors and avoid all lower-dimensional intersections of Σ with other pole divisors. More explicitly, if a same-side sign-class $[g_a] \neq [g_j]$ does not share a relatively open subset of Σ, then choose p outside

$$\Sigma \cap \{Q_{g_a} = 0\}.$$

Do the same for the right-side pole divisors that do not contain a relatively open subset of Σ. Since there are only finitely many divisors, this removes only a finite union of lower-dimensional analytic subsets of Σ.

Let U be a small complex neighborhood of p such that $U \cap \Sigma$ is smooth and connected. Because $dQ_j(p) \neq 0$, Q_j is a local defining equation for Σ. Choose a small loop γ in $U \setminus \Sigma$ linking Σ once positively. Equivalently, in local coordinates one may take

$$Q_j(\gamma(\theta)) = \varepsilon e^{i\theta}, \qquad 0 \leq \theta \leq 2\pi,$$

with the remaining coordinates fixed. For any locally nonzero holomorphic factor Q, define its monodromy around γ by

$$\Delta_\gamma \log Q := \text{analytic continuation of } \log Q \text{ along } \gamma - \log Q = \int_\gamma \frac{dQ}{Q}.$$

Then

$$\Delta_\gamma \log Q_j = 2\pi i.$$

By the choice of p and U, every other left-side sign-class has no pole in U; hence, for $a \neq j$,

$$\Delta_\gamma \log Q_{g_a} = 0.$$

Suppose, for contradiction, that no right-side sign-class shares a relatively open subset of Σ. Then our generic choice of p and smallness of U ensure that every right-side factor $Q_{\widetilde{g}_i}$ is also nonzero on U, so

$$\Delta_\gamma \log Q_{\widetilde{g}_i} = 0 \qquad \text{for all } i.$$

Analytically continue the null-pair identity

$$b + \sum_a v_a \log Q_{g_a} = \widetilde{b} + \sum_i \widetilde{v}_i \log Q_{\widetilde{g}_i}$$

around γ and subtract the original branch. The constants have zero monodromy, so the left side changes by

$$\sum_a v_a \Delta_\gamma \log Q_{g_a} = 2\pi i\, v_j,$$

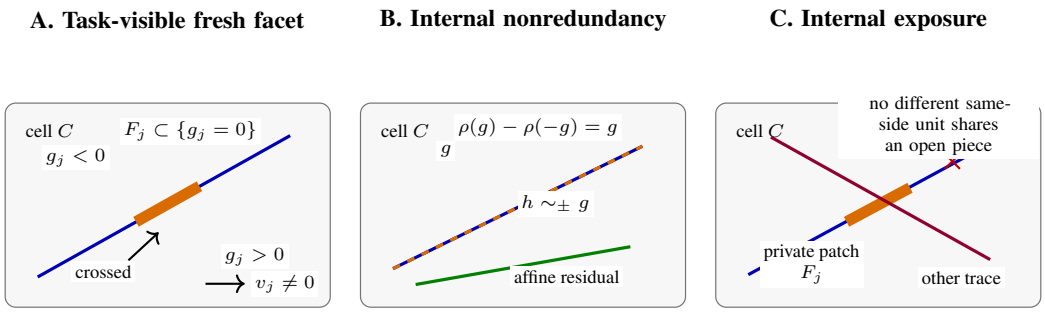
while the right side changes by

$$\sum_i \widetilde{v}_i \Delta_\gamma \log Q_{\widetilde{g}_i} = 0.$$

Thus equality would force $2\pi i\, v_j = 0$, contradicting usedness, which gives $v_j \neq 0$.

Therefore some right-side sign-class $[\widetilde{g}_i]$ must share a relatively open subset of Σ; this is the desired cross-match. It is unique: if two different right-side sign-classes shared a relatively open subset of the same patch, then they would share an open pole patch with each other, contradicting right-side pole separation modulo sign. Reducedness of the used expansion then prevents two distinct terms from representing the same sign-class. Hence the private left patch has a unique right-side match. Repeating the same argument with left and right interchanged gives the converse statement for private right-side patches. □

ReLU minimality and private facet cross-matching. We now develop a ReLU analog to the minimality concept and results of the previous subsection. In a deep one-step expansion the scalar arguments are continuous and piecewise affine functions $g_j = z_{r,j}(x)$, so the crossed trace must be strengthened by a crossed *fresh facet* inside a lower-layer affine cell (Fig 9). The exposed-facet part plays the same role as Softplus pole separation: it makes the nonlinear signature of one unit private rather than confused with another unit's signature.



A counted ReLU unit has a fresh zero facet inside an inherited affine cell, the task crosses it, and the next layer reads out the kink.

Sign-scale copies are not distinct axes. Opposite ReLU pairs can also collapse to an affine term, so affine sign-cancellations are excluded.

Each marked fresh facet must contain an open patch not shared by any genuinely different same-expansion argument.

Figure 9: EFF-minimality for ReLU one-step expansions. Panel A illustrates task-visible nonlinear use: inside an inherited affine cell, the argument has a fresh zero facet, the task crosses it, and the outgoing coefficient is nonzero. Panel B illustrates internal nonredundancy: sign-scale duplicates are not counted as separate axes, and opposite pairs can leave only an affine residual through $\rho(g) - \rho(-g) = g$. Panel C illustrates internal exposure: a marked fresh facet must have a private open patch not shared by another genuinely different same-side argument.

Definition 11 (ReLU facet terminology)**.** A scalar function is *CPWA* if it is continuous and piecewise affine. For a finite family of scalar CPWA functions, a *common inherited affine complex* is a finite polyhedral subdivision $\mathscr{P}$ of Ω on whose full-dimensional cells every function is affine. This complex records the kink structure inherited from lower layers. For scalar functions g, h, write $h \sim_+ g$ if $h = \alpha g$ for some $\alpha > 0$, and $h \sim_{\pm} g$ if $h = \varepsilon \alpha g$ for $\varepsilon \in \{\pm 1\}$ and $\alpha > 0$. If $g|_C(x) = a_C^\top x + c_C$ on a full-dimensional inherited cell C, a *fresh zero facet* is a nonempty relatively open codimension-one patch

$$F \subset \operatorname{relint}(C) \cap \{g = 0\}$$

with $a_C \neq 0$ and both signs of g present in C. The facet is fresh because the kink of $\rho(g)$ across F is created by the current outer ReLU, not inherited from ∂C. A finite list of such facets for g_j is a *marked fresh-facet atlas*. In analogy with the softplus private pole, a marked facet is *private*, or *internally exposed*, if no other same-expansion argument vanishes on an open subset of it unless that argument is globally sign-scale equivalent to g_j.

We can now define ReLU minimality, again by adding to the usedness condition a signature separation condition, here a played by fresh facets:

Definition 12 (Minimal ReLU one-step expansion)**.** A ReLU one-step expansion is *Exposed-Fresh-Face (EFF)-minimal*, if it satisfies the following three groups of conditions.

(A) **Task-visible nonlinear use.** Every counted term has $v_j \neq 0$, and every argument g_j has at least one marked fresh zero facet. This is the deep version of adjacent ReLU usedness: the task crosses a newly created ReLU boundary and the next layer reads out the resulting kink.

(B) **Internal nonredundancy.** Distinct same-side arguments are not sign-scale equivalent, and the expansion is affine sign-noncancelling:

$$\inf_{a \in \mathbb{R}^m} \left\| \sum_{j \in S} v_j g_j - a \right\|_{L^2} > 0$$

for every nonempty S. This excludes duplicate axes and the affine identity $\rho(g) - \rho(-g) = g$.

(C) **Internal signature separation.** The marked fresh-facet atlases are internally exposed: no genuinely different same-expansion argument shares an open piece of a marked fresh facet. This is the ReLU analogue of Softplus pole separation.

As with the softplus case, the purpose of EFF-minimality is expressed by the following cross-match lemma.

Lemma 5 (A private marked facet has a unique cross match)**.** *Suppose*

$$b + \sum_j v_j \rho(g_j) = \widetilde{b} + \sum_i \widetilde{v}_i \rho(\widetilde{g}_i).$$

If both expansions are EFF-minimal, every private marked facet has a unique cross-side gate carrying the same open fresh zero patch.

Proof. We prove the claim for a private marked facet on the left; the converse follows by exchanging the two expansions. Refine the inherited affine complex, if necessary, so that all left- and right-side arguments are affine on each full-dimensional cell of the refinement. Let F be a private marked facet of the left-side argument g_j, and let C be a full-dimensional inherited cell containing F in its relative interior. On C write

$$g_j(x) = a_j^\top x + c_j, \qquad a_j \neq 0, \qquad F \subset \operatorname{relint}(C) \cap \{a_j^\top x + c_j = 0\}.$$

Because F is fresh, both signs of g_j occur in C. Choose a normal vector n with

$$a_j^\top n > 0.$$

For $p \in F$ generic and $t > 0$ small, the points $p + tn$ and $p - tn$ lie in the two sides of F inside C.

Choose p away from lower-dimensional intersections with all other marked facets and with any right-side zero facet that does not contain a relatively open subset of F. Near such a p, all left-side terms except the jth are affine through F. Define the one-sided derivative matrices of the represented left function

$$Y_L(x) := b + \sum_a v_a \rho(g_a(x))$$

by

$$D^+ Y_L(p) := \lim_{t \downarrow 0} DY_L(p + tn), \qquad D^- Y_L(p) := \lim_{t \downarrow 0} DY_L(p - tn).$$

For the jth term,

$$D^+\big(v_j \rho(g_j)\big)(p) - D^-\big(v_j \rho(g_j)\big)(p) = v_j a_j^\top,$$

up to the harmless sign determined by the choice of n. All other left terms have the same derivative on the two sides of F, so

$$D^+ Y_L(p) - D^- Y_L(p) = v_j a_j^\top.$$

This rank-one matrix is nonzero because $v_j \neq 0$ by task-visible nonlinear use and $a_j \neq 0$ by freshness.

The two one-step expansions represent the same function, so for

$$Y_R(x) := \widetilde{b} + \sum_i \widetilde{v}_i \rho(\widetilde{g}_i(x))$$

we must also have

$$D^+ Y_R(p) - D^- Y_R(p) = v_j a_j^\top \neq 0.$$

Suppose, for contradiction, that no right-side argument vanishes on a relatively open subset of F. By the generic choice of p, no right-side zero set passes through p, and after the common refinement all right-side arguments are affine on the cell containing the two small one-sided neighborhoods. Hence each sign $\operatorname{sgn}(\widetilde{g}_i)$ is locally constant across F near p, so each term $\widetilde{v}_i \rho(\widetilde{g}_i)$ is affine with the same derivative on the two sides. Therefore

$$D^+ Y_R(p) - D^- Y_R(p) = 0,$$

contradicting the nonzero jump above.

Thus at least one right-side argument $\widetilde{g}_i$ has a zero facet containing a relatively open subset of F. This gives a cross-side gate carrying the same open fresh zero patch. The match is unique: if two different right-side arguments vanished on an open subset of that patch, internal exposure on the right would force them to be globally sign-scale equivalent, and the internal nonredundancy condition rules out two distinct same-side terms in the same sign-scale class. Reversing the roles of the two expansions proves the same statement for private right-side facets. □

B.1.3 SOFTPLUS REGULARITY AND IDENTIFIABILITY

Softplus minimality gives unique cross-matches of private pole patches. The remaining question is whether a shared pole patch determines the underlying scalar argument – that is identiability. Ensuring this happens is where regularity enters. Specifically, we introduce a relationship between monodromy and an algebraic object called the Laurent ring. The point of this formalism is to build a bridge from complex-analytic information like monodromy to elementary algebra in a Laurent polynomial ring. The basic intuition for doing this is that the bad objects we need to rule out (nontrivial collisions) can be characterized by algebraic conditions on the ring of polynomials containing (analytic continuations of) softplus expansions. Essentially, we find that collisions are non-primitive polynomials in the Laurent polynomial ring, while good objects (non-collisions) are the primitive monomials in this ring. Analytically, a collision is a pair of different expansions whose pole data is not enough to distinguish them. Algebraically, the same phenomenon appears as torsion, composite exponents, or unexpected common factors among Laurent binomials. The condition that such things not happen is regularity – in this case, for softplus, *Monodromy-Laurent Regularity* (MLR). Having introduced the MLR concept, we show that it leads to identifiability as desired. For background on the idea of a complex-analytic/algebraic geometry interface and on Laurent monomial coordinates, see (Griffiths & Harris, 1978; Cox et al., 2011).

From monodromy to Laurent coordinates. A Softplus layer contains outer factors

$$Q_g = 1 + e^{g^{\mathbb{C}}}.$$

When we analytically continue around a lower-layer pole loop, logarithmic pieces inside $g^{\mathbb{C}}$ may change by multiples of $2\pi i$. Consequently, $e^{g^{\mathbb{C}}}$ may come back multiplied by a nonzero complex scalar. We record these scalar multipliers by the concept of a *character* in a *monodromy coordinate chart*.

Definition 13 (Monodromy coordinates and characters)**.** Let $\widehat{\mathcal{U}}$ be a domain on which the exponentials e^{g_α} are single-valued and meromorphic. Number the poles of softplus arguments $1, \dots, p$. Then let $\tau_1, \dots, \tau_p$ denote the analytic continuations associated with each pole. For $n = (n_1, \dots, n_p) \in \mathbb{Z}^p$, write the expression

$$\tau^n := \tau_1^{n_1} \cdots \tau_p^{n_p}.$$

This means a path going around the first selected loop n_1 times, the second n_2 times, and so on. We call this fixed choice of continuation domain and selected loops a *monodromy coordinate chart*: the local coordinate system in which we compare the monodromy of all relevant arguments. With the chart fixed, an argument g_α has *monodromy character* $\chi_\alpha : \mathbb{Z}^p \to \mathbb{C}^*$ if $\tau^n(e^{g_\alpha}) = \chi_\alpha(n) e^{g_\alpha}$ for all $n \in \mathbb{Z}^p$. Equivalently, if $\tau_a(e^{g_\alpha}) = \lambda_{\alpha,a} e^{g_\alpha}$ for $\lambda_{\alpha,a} \in \mathbb{C}^*$ then

$$\chi_\alpha(n) = \prod_{a=1}^{p} \lambda_{\alpha,a}^{n_a}.$$

Characters are multiplicative:

$$\chi_\alpha(n + n') = \chi_\alpha(n)\chi_\alpha(n').$$

For an affine-Softplus layer, we can compute the characters explicitly (the so-called "character rule"). Suppose

$$g = \beta + \sum_a w_a H_a, \qquad H_a = \log Q_a.$$

If continuation by τ^n sends

$$H_a \longmapsto H_a + 2\pi i\, \nu_a(n),$$

then

$$e^g \longmapsto \exp\left(2\pi i \sum_a w_a \nu_a(n)\right) e^g.$$

Thus

$$\chi_g(n) = \exp\left(2\pi i \sum_a w_a \nu_a(n)\right).$$

Affine weights are therefore monodromy exponents. Exact integer or rational relations among these exponents are the source of the "bad" character resonances causing collision that we exclude below with the regularity concept. To get there, first we set up some basic properties of characters.

Lemma 6 (Independence of distinct characters)**.** *Let $f_1, \dots, f_m$ be nonzero meromorphic functions in a fixed monodromy setup. Suppose each f_a has character χ_a, meaning*

$$\tau^n f_a = \chi_a(n) f_a \qquad \forall n \in \mathbb{Z}^p.$$

If the characters $\chi_1, \dots, \chi_m$ are pairwise distinct, then $f_1, \dots, f_m$ are linearly independent over K.

Proof. Suppose

$$\sum_{a=1}^{m} c_a f_a = 0, \qquad c_a \in K,$$

is a nontrivial relation with the smallest possible number of nonzero terms. After dividing by one nonzero coefficient, assume $c_1 = 1$. Since the characters are distinct, choose $n \in \mathbb{Z}^p$ with

$$\chi_1(n) \neq \chi_2(n).$$

Applying τ^n to the relation gives

$$\sum_a c_a \chi_a(n) f_a = 0,$$

because every coefficient $c_a \in K$ is unchanged by continuation. Subtracting $\chi_1(n)$ times the original relation gives

$$\sum_a c_a \big(\chi_a(n) - \chi_1(n)\big) f_a = 0.$$

The $a = 1$ term is eliminated, while the $a = 2$ coefficient remains nonzero. This is a shorter nontrivial relation, contradicting minimality. Hence no such relation exists. □

The previous lemma says that different characters behave like independent monomials. We now make that statement algebraic. Let $\Gamma = \langle \chi_\alpha \rangle$ be the finitely generated group of characters produced by the arguments under comparison. When Γ has no torsion, choose basic characters $\gamma_1, \dots, \gamma_s$ freely generating Γ, and choose meromorphic functions $T_1, \dots, T_s$ whose characters are $\gamma_1, \dots, \gamma_s$. These T_a's may be taken locally as products and quotients of the e^{g_α}'s corresponding to a chosen character basis. Then every e^{g_α} can be written uniquely as

$$e^{g_\alpha} = c_\alpha T^{m_\alpha}, \qquad c_\alpha \in K^*, \qquad m_\alpha \in \mathbb{Z}^s,$$

where $T^{m_\alpha} = T_1^{m_{\alpha,1}} \cdots T_s^{m_{\alpha,s}}$.

Thus, the exponentials e^{g_α} become invariant coefficients times Laurent monomials. Let K be the set of meromorphic functions that do not change under continuation loops:

$$K = \{f \in \mathcal{M}(\widehat{\mathcal{U}}) : \tau^n f = f \text{ for every } n \in \mathbb{Z}^p\}.$$

Then the algebra generated by these exponentials is identified the *Laurent polynomial ring*

$$K[T_1^{\pm 1}, \dots, T_s^{\pm 1}],$$

whose elements are finite K-linear combinations of Laurent monomials. The point of these coordinates is that each Softplus outer factor $1 + e^g$ becomes a Laurent with two terms (a *binomial*):

Definition 14 (Laurent binomial and primitive exponent)**.** A *Laurent binomial* is an expression

$$B_{c,m}(T) = 1 + cT^m, \qquad c \in K^*, \quad m \in \mathbb{Z}^s \setminus \{0\}.$$

In our setting this is exactly the outer Softplus factor $1 + e^g$ after writing $e^g = cT^m$.

The exponent m is *primitive* if it is not a nontrivial integer multiple of another lattice vector:

$$m \neq q m_0 \qquad \text{for every integer } |q| > 1.$$

Equivalently,

$$\gcd(|m_1|, \dots, |m_s|) = 1.$$

Nonprimitive exponents are what allow composite resonances such as

$$1 + Z^3 = (1 + Z)(1 - Z + Z^2).$$

The primitive-exponent condition rules out this kind of hidden factorization.

Lemma 7 (Primitive Laurent binomials are prime)**.** *If m is primitive, then $1 + cT^m$ is prime in $K[T_1^{\pm 1}, \dots, T_s^{\pm 1}]$.*

Proof. A primitive vector $m \in \mathbb{Z}^s$ can be completed to an integral basis

$$e'_1 = m, \qquad e'_2, \dots, e'_s$$

of $\mathbb{Z}^s$. Equivalently, there is a determinant ± 1 ("unimodular") matrix in $GL_s(\mathbb{Z})$ whose first basis vector is m. The associated monomial change of variables is a Laurent-ring isomorphism obtained by setting

$$S_1 = T^{e'_1} = T^m, \qquad S_\nu = T^{e'_\nu} \quad (\nu = 2, \dots, s).$$

Because m is unimodular, the Laurent monomial T^u is uniquely expressible in the S_ν coordinates, and conversely. Hence

$$K[T_1^{\pm 1}, \dots, T_s^{\pm 1}] \cong K[S_1^{\pm 1}, \dots, S_s^{\pm 1}],$$

and $1 + cT^m$ becomes $1 + cS_1$.

The quotient by $1+cS_1$ is obtained by setting $S_1 = -c^{-1}$, which is a unit because $c \in K^*$. Therefore

$$K[S_1^{\pm 1}, \dots, S_s^{\pm 1}]/(1 + cS_1) \cong K[S_2^{\pm 1}, \dots, S_s^{\pm 1}],$$

an integral domain. Hence the ideal generated by $1 + cT^m$ is prime, and the binomial is a prime element of the unique-factorization Laurent ring. □

Lemma 8 (Associates of two-term Laurent binomials)**.** *Two Laurent polynomials are* associates *if they differ by multiplication by a Laurent unit aT^k. If*

$$1 + dT^n = aT^k(1 + cT^m),$$

then either

$$n = m, \qquad d = c, \qquad k = 0, \qquad a = 1,$$

or

$$n = -m, \qquad d = c^{-1}, \qquad k = n, \qquad a = d.$$

For real-valued arguments this means $h = g$ or $h = -g$.

Proof. Expanding the right side gives

$$aT^k + acT^{k+m}.$$

Laurent monomials are linearly independent over K, so the two-element support sets must agree:

$$\{0, n\} = \{k, k + m\}.$$

If $k = 0$ and $k + m = n$, then $m = n$ and coefficient comparison gives $a = 1$ and $c = d$. If $k = n$ and $k + m = 0$, then $n = -m$ and coefficient comparison gives $a = d$ and $dc = 1$. These are the only two matchings of the two support sets. Returning to $e^g = cT^m$ and $e^h = dT^n$ gives $e^h = e^g$ in the first case and $e^h = e^{-g}$ in the second; real-valuedness on Ω gives $h = g$ or $h = -g$. □

The character and primitivity conditions rule out the algebraic resonances inside the Laurent ring itself: torsion characters, commensurate characters, and composite binomials such as $1 + Z^3 = (1 + Z)(1 - Z + Z^2)$. There remains one further way a divisor match could be misleading (e.g. be a nontrivial collision): two distinct Laurent polynomials might appear to share a pole patch only after being pulled back through a degenerate lower-feature map. Laurent divisor-faithfulness rules out this pullback accident: a shared analytic pole patch must come from a genuine common Laurent factor before pullback.

Definition 15 (Laurent divisor-faithfulness)**.** Suppose $e^g = cT^m$ and $e^h = dT^n$ are two coordinates. Let Σ be a shared outer pole patch, meaning a connected patch on which both $1 + e^g$ and $1 + e^h$ vanish. The pair is *Laurent divisor-faithful* at Σ if the two Laurent binomials $1 + cT^m$ and $1 + dT^n$ have a common nonunit factor in the Laurent ring

$$K[T_1^{\pm 1}, \dots, T_s^{\pm 1}].$$

In words, the shared pole patch must come from a real shared Laurent factor, not merely from a collapse caused by the lower-feature map.

The genericity section §B.3 gives a rank criterion for verifying the faithfulness condition is a real-world reasonable assumption (Lemma 13). For now, the exact zippering theorem below will simply assume Laurent divisor-faithfulness as part of regularity:

Definition 16 (Monodromy–Laurent regularity)**.** A collision pair (g, h), together with a shared outer pole patch Σ, is *monodromy–Laurent regular* (MLR) if:

(R1) the character group generated by the compared arguments has no torsion;

(R2) the Laurent exponent vectors are primitive;

(R3) the pair is Laurent divisor-faithful at Σ.

Thus MLR says that the monodromy data can be represented by true Laurent monomials, the relevant Softplus factors do not secretly factor through a composite exponent, and any shared pole patch reflects a genuine common Laurent divisor. The "shared pole patch" in (R3) is exactly the shared outer pole patch Σ on which $[g]$ and $[h]$ collide.

All of this setup culminates in the key identifiability result:

Theorem 4 (Softplus one-step identifiability under MLR)**.** *Suppose*

$$b + \sum_{j=1}^{k} v_j \sigma(g_j) = \widetilde{b} + \sum_{i=1}^{\widetilde{k}} \widetilde{v}_i \sigma(\widetilde{g}_i)$$

on Ω. Assume both one-step Softplus expansions are minimal, and assume that every cross-matched pair supplied by Lemma 4 is MLR on its selected shared pole patch. Then $k = \widetilde{k}$ and, after a permutation,

$$\widetilde{g}_j = g_j, \qquad \widetilde{v}_j = v_j, \qquad \widetilde{b} = b.$$

Proof. Lemma 4 gives a unique cross-match for every private sign-class on either side. Since every sign-class in a minimal expansion has a private pole patch, these cross-matches define a bijection between the left and right sign-classes. After reindexing, the matched pairs are $[g_j]$ and $[\widetilde{g}_j]$.

Fix one such matched pair and let Σ_j be the selected shared outer pole patch witnessing the match. By MLR, the monodromy setup is torsion-free, so after choosing Laurent coordinates we may write

$$e^{g_j} = c_j T^{m_j}, \qquad e^{\widetilde{g}_j} = d_j T^{n_j},$$

with $m_j, n_j \in \mathbb{Z}^s$ primitive. The associated outer Softplus factors are the Laurent binomials

$$P_j(T) = 1 + c_j T^{m_j}, \qquad R_j(T) = 1 + d_j T^{n_j}.$$

The shared pole patch means that the pullbacks of P_j and R_j vanish on the same regular analytic hypersurface patch. Laurent divisor-faithfulness says that this shared pulled-back divisor is not merely an accident of the lower-feature map: P_j and R_j have a common nonunit factor in the Laurent ring.

By primitivity, Lemma 7 says that both P_j and R_j are prime. Two prime Laurent binomials with a common nonunit factor must be associates, so

$$1 + d_j T^{n_j} = a T^k (1 + c_j T^{m_j})$$

for some Laurent unit aT^k. Lemma 8 leaves only two possibilities: either $n_j = m_j$ and $d_j = c_j$, or $n_j = -m_j$ and $d_j = c_j^{-1}$. In the first case $e^{\widetilde{g}_j} = e^{g_j}$; in the second $e^{\widetilde{g}_j} = e^{-g_j}$.

On the real task domain g_j and $\widetilde{g}_j$ are real-valued. Hence $e^{\widetilde{g}_j} = e^{g_j}$ implies $\widetilde{g}_j - g_j \in 2\pi i\mathbb{Z}$ and therefore $\widetilde{g}_j = g_j$, while $e^{\widetilde{g}_j} = e^{-g_j}$ implies $\widetilde{g}_j + g_j \in 2\pi i\mathbb{Z}$ and therefore $\widetilde{g}_j = -g_j$. Thus

$$\widetilde{g}_j = \varepsilon_j g_j, \qquad \varepsilon_j \in \{+1, -1\}.$$

The coefficient vectors also match. Write the two equal represented functions near the private pole patch as

$$F_L = b + \sum_k v_k \log(1 + e^{g_k}), \qquad F_R = \tilde{b} + \sum_i \tilde{v}_i \log(1 + e^{\tilde{g}_i}).$$

Choose a small loop γ around a point where $1 + e^{g_j} = 0$, and where no other left-side term has such a zero. Then every left-side term except j is single-valued around γ, while

$$\log(1 + e^{g_j}) \longmapsto \log(1 + e^{g_j}) + 2\pi i.$$

Thus, after analytic continuation once around γ,

$$F_L \longmapsto F_L + 2\pi i\, v_j.$$

Because this private pole patch is matched on the right side, the same loop winds around exactly one right-side term, namely $\tilde{g}_{\pi(j)}$, so

$$F_R \longmapsto F_R + 2\pi i\, \tilde{v}_{\pi(j)}.$$

Since $F_L = F_R$ as analytically continued functions, their changes around the same loop must be equal:

$$2\pi i\, v_j = 2\pi i\, \tilde{v}_{\pi(j)}.$$

Hence $\tilde{v}_{\pi(j)} = v_j$.

It remains only to eliminate negative orientations. Let $S = \{j : \varepsilon_j = -1\}$. After cancelling every positively oriented matched term and using the coefficient equality, the null relation becomes

$$b - \widetilde{b} + \sum_{j \in S} v_j \big[\sigma(g_j) - \sigma(-g_j)\big] = 0.$$

The Softplus identity $\sigma(t) - \sigma(-t) = t$ turns this into

$$b - \widetilde{b} + \sum_{j \in S} v_j g_j = 0.$$

If S were nonempty, this would say that a nonempty downstream-weighted sum of arguments is constant, contradicting sign-noncancellation. Hence $S = \varnothing$. All arguments and coefficients agree after the permutation, and the remaining equality of constants gives $b = \widetilde{b}$. □

B.1.4 ReLU regularity and identifiability

Here we introduce the ReLU analog of the regularity concept and prove the corresponding identifiability result. The reader will probably feel that the math in this section is easier (or at any rate, more elemantary) than in the previous section. The price we pay for that is that the assumptions are stronger. Because of the lack of analytic structure for ReLU, there is no way we know of to translate ReLU expansions into characters and apply any kind of algebraic reasoning. As a result, the arguments are in a sense simpler but require more blunt force (as expressed by needing *multi-facet signatures* and an explicit assumption of *coherence*, defined below).

Because of the nature of ReLU, the regularity arguments below (and the associated genericity arguments in §B.3) necessarily end up having to study sets that are not smooth everywhere (such as the ReLU collision sets). A *stratum* is one smooth piece of such a set. A *stratification* is a decomposition into smooth pieces, usually locally finite so that only finitely many pieces meet any small neighborhood. We use strata so that tangent spaces and dimension counts are well-defined piece by piece. Thus an "X-stratum" means a smooth piece of the set X.

We can now define Multi-Facet Regularity, the ReLU analog of MLR:

Definition 17 (Multi-facet signature and MFR)**.** Fix a ReLU one-step comparison and a common inherited affine complex $\mathscr{P}$. An *inherited cell* means a full-dimensional cell $C \in \mathscr{P}$, so that every argument is affine on C. Choose marked fresh-facet slots

$$F_\nu \subset C_\nu, \qquad \nu = 1, \dots, L.$$

For a scalar argument g, write its affine formula on the inherited cell C_ν as

$$g|_{C_\nu}(x) = a_\nu^\top x + c_\nu, \qquad q_\nu(g) := (a_\nu, c_\nu).$$

The zero facet is unchanged if $q_\nu(g)$ is multiplied by a nonzero scalar, so we record only the scalar-multiple class $[q_\nu(g)]$. Then the *multi-facet signature* of g is the finite list

$$\Sigma_r(g) = \big(C_\nu, \nu, [q_\nu(g)]\big)_{\nu=1}^{L}.$$

The entries remember which inherited cell and which marked-facet slot they came from. The signature *identifies* g if, for any other scalar argument h,

$$\Sigma_r(h) = \Sigma_r(g) \quad \Longrightarrow \quad h = \varepsilon\alpha g \quad \text{for some } \varepsilon \in \{\pm 1\},\ \alpha > 0.$$

In words, the marked facets determine the whole argument, up to the unavoidable ReLU sign and positive-scale ambiguity. In a null-pair equality, lemma 5 matches individual fresh facets across the two sides. The matching is *coherent* for a pair (g, h) if all marked facets of g are matched to marked facets of the same single competitor h, and conversely. A coherently matched pair (g, h) is *multi-facet regular* (MFR), if the coherently matched facet records have the same multi-facet signature and that signature identifies the argument:

$$\Sigma_r(h) = \Sigma_r(g) \quad \Longrightarrow \quad h = \varepsilon\alpha g, \qquad \varepsilon \in \{\pm 1\}, \quad \alpha > 0.$$

With this definition in mind, we can get the needed identifiability result:

Theorem 5 (ReLU identifiability under MFR)**.** *Suppose*

$$b + \sum_{j=1}^{k} v_j \rho(g_j) = \widetilde{b} + \sum_{i=1}^{\widetilde{k}} \widetilde{v}_i \rho(\widetilde{g}_i)$$

on Ω. Assume both expansions are EFF-minimal, the induced cross-facet matching is coherent, and every coherently matched scalar pair is MFR. Then $k = \widetilde{k}$ and, after a permutation, there are $\alpha_j > 0$ such that

$$\widetilde{g}_j = \alpha_j g_j, \qquad \alpha_j \widetilde{v}_j = v_j, \qquad \widetilde{b} = b.$$

Proof. By Lemma 5, private facets match across the equality. Coherence says that all marked facets of a given g_j match one competitor $\widetilde{g}_{\pi(j)}$. MFR then reconstructs the whole scalar argument from its labelled signature, giving

$$\widetilde{g}_{\pi(j)} = \varepsilon_j \alpha_j g_j, \qquad \varepsilon_j \in \{\pm 1\}, \quad \alpha_j > 0.$$

Reducedness makes this matching bijective.

Now compare the derivative jumps across a marked facet of g_j. If $\varepsilon_j = +1$, then $\rho(\widetilde{g}_{\pi(j)}) = \alpha_j \rho(g_j)$ near the facet. If $\varepsilon_j = -1$, then the active side is reversed, but the jump magnitude is still $\alpha_j \widetilde{v}_{\pi(j)} a_j^\top$. Equality of the represented functions therefore gives

$$v_j a_j^\top = \alpha_j \widetilde{v}_{\pi(j)} a_j^\top,$$

and since $a_j \neq 0$, $v_j = \alpha_j \widetilde{v}_{\pi(j)}$.

Let $S = \{j : \varepsilon_j = -1\}$. After substituting the matched coefficients and cancelling the positively oriented terms, the equality becomes

$$b - \widetilde{b} + \sum_{j \in S} v_j \big[\rho(g_j) - \rho(-g_j)\big] = 0.$$

Using $\rho(t) - \rho(-t) = t$, this reduces to

$$b - \widetilde{b} + \sum_{j \in S} v_j g_j = 0.$$

Affine sign-noncancellation forces $S = \varnothing$, and then the remaining constant terms give $b = \widetilde{b}$. □

Remark 1. To reiterate, while this math is shorter and more elementary than in the previous softplus section, the MFR definition is basically clobbering things with its multi-face and coherence conditions, whereas in the MLR case, we can prove many things without having to assume them. The saving grace is that coherent MFR can be shown to be generic (as seen in §B.3) and supports the same genericity reasoning as the softplus condition, leading to the powerful activation-independent null-net theorem 9.

B.1.5 MINIMAL NULL PAIRS AND EXACT ZIPPERING

We now return to the organizing strategy stated at the beginning of the section. The Softplus and ReLU arguments have supplied the activation-specific inputs: minimality gives unique private cross-matches, and regularity makes those matches identify the scalar arguments. The remaining step is activation-independent. At a backward zippering step, two one-step expansions represent the same next-layer function; that is, they form a null pair. The following corollary records exactly what the two activation-specific identifiability theorems have proved: such a null pair determines the previous-layer arguments after the only unavoidable hidden-unit matching, namely permutation for Softplus and permutation plus positive rescaling for ReLU.

Corollary 1 (Activation-specific one-step identifiability)**.** *Let a minimal null pair satisfy the relevant regularity hypothesis: MLR on every Softplus cross-match, or coherent MFR on every ReLU cross-match. Then the two expansions are the same after the activation's unavoidable hidden-unit matching. Concretely, after a permutation,*

$$\widetilde{g}_j = g_j, \qquad \widetilde{v}_j = v_j, \qquad \widetilde{b} = b$$

in the Softplus case, while in the ReLU case there are $\alpha_j > 0$ such that

$$\widetilde{g}_j = \alpha_j g_j, \qquad \alpha_j \widetilde{v}_j = v_j, \qquad \widetilde{b} = b.$$

Proof. This is just the combination of the two activation-specific identifiability theorems: Theorem 4 for Softplus and Theorem 5 for ReLU. □

Theorem 6 (Abstract exact zippering)**.** *Let A, B be same-depth, same-width affine-φ networks, and let $q < s$. Suppose terminal weak equivalence holds:*

$$z_s^B = E_s z_s^A$$

for an injective affine map E_s. Assume that at each backward step $r = s-1, \dots, q$, the two one-step expansions compared at layer r satisfy the appropriate hypotheses of Corollary 1: Softplus minimality plus MLR, or ReLU EFF-minimality plus coherent MFR. Then alignment zippers upstream. For every $r = q, \dots, s-1$, the layer-r arguments are exactly matched: in the Softplus case, after a permutation,

$$z^B_{r,\pi_r(j)} = z^A_{r,j},$$

and in the ReLU case, after a permutation and positive scales $\alpha_{r,j} > 0$,

$$z^B_{r,\pi_r(j)} = \alpha_{r,j} z^A_{r,j}.$$

Proof. At the top step, expand z_s^B and $E_s z_s^A$ through layer $s-1$:

$$z_s^B = b_s^B + \sum_i W_s^B[:,i]\varphi(z^B_{s-1,i}),$$

$$E_s z_s^A = T_s b_s^A + a_s + \sum_j T_s W_s^A[:,j]\varphi(z^A_{s-1,j}).$$

These are two equal one-step expansions. Corollary 1 applies, so the arguments at layer $s-1$ match: by a permutation for Softplus, and by a permutation together with positive coordinatewise scales for ReLU. Let E_{s-1} denote the corresponding concrete coordinate map; then

$$z^B_{s-1} = E_{s-1} z^A_{s-1}.$$

Now assume inductively that layer $r+1$ has already been matched, so

$$z^B_{r+1} = E_{r+1} z^A_{r+1}$$

for the coordinate map obtained in the previous step. Expanding both sides through layer r gives two equal one-step expansions. By the recursive hypothesis, they satisfy the relevant Softplus or ReLU assumptions, so Corollary 1 again matches their scalar arguments. Thus, layer r is aligned in the same activation-specific sense. Repeating this argument from $s-1$ down to q proves the theorem. □

Remark 2 (Unequal depths and counted widths.). We note that the same-depth, same-width hypotheses in Theorem 6, and the common layer indexing in Theorems 7–8, are mainly notational to keep proofs straightforward, but they hold without loss of generality for unequal depths and widths. Namely, more generally, the proofs can use only a paired sequence of one-step comparison problems

$$a_q < a_{q+1} < \cdots < a_s, \qquad b_q < b_{q+1} < \cdots < b_s,$$

with the same number of backward zippering steps. Replacing layer r by the paired layers (a_r, b_r) gives the same induction, with constants L_r attached to the paired step $(a_r, b_r) \to (a_{r+1}, b_{r+1})$.

For unequal raw widths, the statement should be read in terms of the counted minimal one-step expansions. If extra units are unused, constant, duplicate, or otherwise outside the counted minimal expansion, they may be deleted or absorbed as in the shrinkability reductions. On the remaining counted axes, the exact theorem matches the two minimal expansions after the unavoidable activation symmetry. Equivalently, in a directional A-to-B formulation, one may write an injective matching

$$\pi_r : J^{A,\varphi}_{a_r} \hookrightarrow J^{B,\varphi}_{b_r},$$

provided the unmatched B-side coordinates are not additional counted minimal terms. Otherwise those terms are genuine extra nonlinear signatures and must be matched, controlled, or the comparison direction/axis set must be changed.

With this notation, the soft bound has the same form, except that d_r should be read as the number of counted A-side axes at the paired layer,

$$d^A_{a_r} := |J^{A,\varphi}_{a_r}|,$$

and the product of inverse constants is taken over the paired zippering steps:

$$\mathrm{AxisAlign}^{\theta}_{a_r,b_r,\varphi}(A, B) \geq 1 - \frac{e_s^2 \prod_{t=r}^{s-1} L_t^2}{(1-\theta)^2 c^2 d^A_{a_r}}.$$

This reduces to our above theorems when the paired layers have a common indexing and common counted width. This same notational flexibility is distinct from, and weaker than, the depth-warped result of Appendix §C.1, where a preferred axis may correspond to a window of nearby layers rather than to one fixed layer.

Corollary 2 (Exact MLR Softplus zippering)**.** *If every backward comparison is individually minimal and every cross-matched scalar pair is MLR, then terminal weak equivalence zippers upstream into exact permutation alignment.*

Proof. Theorem 4 supplies Softplus one-step identifiability at each backward step, so Theorem 6 applies. □

Corollary 3 (Exact MFR ReLU zippering)**.** *If every backward comparison is EFF-minimal, coherently matched, and MFR, then terminal weak equivalence zippers upstream into alignment by permutation plus positive coordinatewise rescaling.*

Proof. Theorem 5 supplies ReLU one-step identifiability at each step. □

B.2 Quantitative minimality, Jacobian regularity, and soft zippering

B.2.1 Quantitative minimality

The exact theory in §B.1 uses qualitative minimality: each counted unit must leave a private nonlinear signature, and MLR or MFR must make the resulting cross-match determine the scalar argument. For asymptotic and soft zippering we need the same conditions with positive margins. A convergent sequence of comparison problems should not be able to lose a counted unit in the limit, for example by letting an outgoing coefficient vanish, making an argument constant, merging two arguments, creating an affine sign-cancellation, or losing the private pole or facet certificate. The definitions below are exactly these positive-margin versions of the minimality conditions in §B.1. Their role is to produce compact comparison families whose exact null limits still satisfy the qualitative hypotheses of §B.1. The same margins also keep the denominators in the soft axis scores from degenerating.

Definition 18 (Quantitative Softplus minimality)**.** A k-term Softplus expansion is (M, c)-minimal if its finite-dimensional realization parameters are bounded by M and:

(S1) $\|v_j\| \geq c$ and $\mathrm{Var}(g_j) \geq c^2$;

(S2) $\|g_i - g_j\|_{L^2}, \|g_i + g_j\|_{L^2} \geq c$ for $i \neq j$;

(S3) every nonempty signed subset has affine-cancellation distance at least c;

(S4) every sign-class has a regular pole certificate separated from the other same-expansion pole sets by at least c.

Definition 19 (Quantitative ReLU EFF-minimality)**.** A normalized k-term ReLU expansion is (M, c)-EFF-minimal if its parameters are bounded by M, every argument is normalized by $\|g_j\|_{L^2} = 1$, and:

(R1) $\|v_j\| \geq c$ and both $\|\rho(g_j)\|_{L^2}$ and $\|\rho(-g_j)\|_{L^2}$ are at least c;

(R2) distinct arguments are separated modulo sign and positive scale by at least c;

(R3) every nonempty signed subset has affine-cancellation distance at least c;

(R4) each marked facet atlas has quantitative lower bounds on slope, patch size, two-sided crossing, and separation from other same-side fresh traces.

Lemma 9 (Quantitative minimality gives compact comparison families)**.** *For a fixed finite-dimensional architecture chart, fixed term count, and fixed M, c, each quantitative family above is compact after fixing the finite hidden-unit order. Moreover, every limit point remains qualitatively minimal in the sense used in §B.1. In particular, the counted arguments have score denominators bounded away from zero uniformly in the family: in the Softplus case $\|g_j\|_{L^2} \geq c$, and in the normalized ReLU case $\|g_j\|_{L^2} = 1$.*

Proof. The parameter bound gives a compact ambient set. Every quantitative condition is a closed inequality on the chosen analytic chart or fixed ReLU activation-pattern stratum, meaning a smooth piece where the relevant ReLU cell and facet pattern is fixed. Imposing all of these inequalities selects a closed subset of that compact set. Choosing one ordering of the finitely many hidden units preserves compactness. For ReLU we also normalize each argument as in the preceding definition, so multiplying an argument by a positive constant is not treated as a distinct point.

The positive margins are precisely what keeps limits from falling out of the qualitative hypotheses of §B.1. If a coefficient vanished, an argument became constant, two arguments became duplicate or opposite, a signed affine cancellation appeared, or a private pole or facet certificate disappeared, then one of the displayed margins would have to go to zero. Since all margins remain at least c, every limit point remains qualitatively minimal. Finally, $\mathrm{Var}(g_j) \geq c^2$ implies $\|g_j\|_{L^2} \geq c$ for Softplus, while ReLU arguments are normalized to have L^2 norm one. □

B.2.2 Minimality, regularity, and asymptotic zippering

Throughout this subsection and the next, the phrase *the appropriate activation-specific quantitative minimality and regularity hypotheses* means the following. First, the comparison pair lies in one of the compact quantitative-minimal families supplied by Lemma 9. Second, any exact null pair that appears as a limit inside that family satisfies the corresponding regularity condition in §B.1: MLR on the exact cross-matched pole pairs for Softplus, and coherent MFR on the exact cross-matched facet pairs for ReLU. This is not a new assumption; it is just the way the exact one-step identifiability results of §B.1 are made stable under compact limits.

We will measure hidden-argument error with one notation for both activations.

Definition 20 (Matched-argument distance)**.** Let G and $\widetilde{G}$ be two one-step expansions, and write g and $\widetilde{g}$ for their scalar argument vectors. The *allowed matching matrices* are $E = P$ in the Softplus case, where P is a permutation matrix, and $E = DP$ in the ReLU case, where P is a permutation matrix and D is a positive diagonal matrix. For such an E, define

$$d_{\mathrm{arg},\varphi}(G, \widetilde{G}; E) = \|\widetilde{g} - Eg\|_{L^2(\Omega;\ell_2)}. \tag{19}$$

When the matching is not specified, set

$$d_{\arg,\varphi}(G,\widetilde{G}) = \inf_{\text{allowed } E} d_{\arg,\varphi}(G,\widetilde{G};E).$$

For axis-count statements in what follows, use the soft-axis score from Definition 8 applied to the layer-r zippering match. The only change is that in the ReLU zippering score the scale parameter is restricted to $\alpha > 0$, because exact ReLU zippering only allows positive rescaling.

The next lemma says that quantitative minimality leads output function distance to be uniformly continuous in configuration space distance:

Lemma 10 (Uniform inverse error bound from minimality and regularity)**.** *Fix $\varphi \in \{\sigma,\rho\}$ and let $\mathcal{K}_r^{\varphi}$ be a compact family of one-step comparison pairs $(G,\widetilde{G})$ satisfying the appropriate activation-specific quantitative minimality and regularity hypotheses. Then there is a nondecreasing error-bound function $\omega_r^{\varphi}(t) \to 0$ as $t \downarrow 0$ such that, for every $(G,\widetilde{G}) \in \mathcal{K}_r^{\varphi}$ with*

$$\|Y_G^{\varphi} - Y_{\widetilde{G}}^{\varphi}\|_{L^2} \le t,$$

there is a matching map of the allowed form E satisfying

$$d_{\arg,\varphi}(G,\widetilde{G};E) \le \omega_r^{\varphi}(t).$$

Proof. We first show that hidden-argument error must go to zero whenever the represented-function error goes to zero. Suppose not. Then there are pairs $(G_n,\widetilde{G}_n) \in \mathcal{K}_r^{\varphi}$ and a number $\eta > 0$ such that

$$\|Y_{G_n}^{\varphi} - Y_{\widetilde{G}_n}^{\varphi}\|_{L^2} \longrightarrow 0, \qquad d_{\arg,\varphi}(G_n,\widetilde{G}_n) \ge \eta.$$

By compactness, after passing to a subsequence,

$$(G_n,\widetilde{G}_n) \longrightarrow (G_*,\widetilde{G}_*) \in \mathcal{K}_r^{\varphi}.$$

Continuity of the realization map gives $Y_{G_*}^{\varphi} = Y_{\widetilde{G}_*}^{\varphi}$, so the limit is an exact null pair. The hypotheses of the family are exactly the activation-specific hypotheses needed to apply Corollary 1: Softplus minimality plus MLR, or ReLU EFF-minimality plus coherent MFR. Therefore the limiting null pair has the activation-specific one-step identifiability form. Concretely, there is a matching map of the allowed form E_* such that $\widetilde{g}_* = E_* g_*$, and hence

$$d_{\arg,\varphi}(G_*,\widetilde{G}_*;E_*) = 0.$$

Using the same nearby matching for n large, and in the ReLU case using the nearby positive rescalings supplied by the local coordinates, gives

$$d_{\arg,\varphi}(G_n,\widetilde{G}_n) \longrightarrow 0,$$

contradicting $d_{\arg,\varphi}(G_n,\widetilde{G}_n) \ge \eta$. Now define

$$\omega_r^{\varphi}(t) = \sup\big\{d_{\arg,\varphi}(G,\widetilde{G}) : (G,\widetilde{G}) \in \mathcal{K}_r^{\varphi}, \|Y_G^{\varphi} - Y_{\widetilde{G}}^{\varphi}\|_{L^2} \le t\big\},$$

with $\omega_r^{\varphi}(t) = 0$ if the set is empty. The preceding paragraph says exactly that $\omega_r^{\varphi}(t) \to 0$ as $t \downarrow 0$. Replacing it by $\sup_{0\le s\le t} \omega_r^{\varphi}(s)$ if necessary makes it nondecreasing. By the definition of $d_{\arg,\varphi}$, after increasing $\omega_r^{\varphi}(t)$ harmlessly on $t > 0$ if needed, a matching map of the allowed form E can be chosen with the displayed bound. □

The above result enables us to prove an asymptotic version of zippering:

Theorem 7 (Asymptotic zippering)**.** *Fix $\varphi \in \{\sigma,\rho\}$ and $q < s$. Let $(A_n,B_n,E_{s,n})$ be comparison problems with terminal error*

$$e_{s,n} = \|z_{s,n}^B - E_{s,n} z_{s,n}^A\|_{L^2} \to 0.$$

Assume that, at each backward step $r = s-1,\dots,q$, the one-step comparison pairs encountered in the zippering induction lie in a fixed compact family $\mathcal{K}_r^{\varphi}$ satisfying the appropriate activation-specific quantitative minimality and regularity hypotheses. Then, for every $r = q,\dots,s-1$ and every fixed $\theta < 1$,

$$\mathrm{AxisAlign}_{r,\varphi}^{\mathrm{zip},\theta}(A_n,B_n) \longrightarrow 1,$$

and in fact this quantity equals 1 for all sufficiently large n.

Proof. We prove the statement by moving one layer at a time from s back to q.

At the top step, compare the transformed A-side expansion of $E_{s,n}z^A_{s,n}$ with the B-side expansion of $z^B_{s,n}$. These are the represented functions of

$$E_{s,n}G^A_{s-1,n} \qquad \text{and} \qquad G^B_{s-1,n}.$$

Their represented-function distance is exactly the terminal mismatch:

$$\|Y^{\varphi}_{E_{s,n}G^A_{s-1,n}} - Y^{\varphi}_{G^B_{s-1,n}}\|_{L^2} = \|E_{s,n}z^A_{s,n} - z^B_{s,n}\|_{L^2} = e_{s,n}.$$

Lemma 10 applied at layer $s-1$ gives a matching map of the allowed form $E_{s-1,n}$ with

$$e_{s-1,n} := d_{\arg,\varphi}\big(E_{s,n}G^A_{s-1,n}, G^B_{s-1,n}; E_{s-1,n}\big) \le \omega^{\varphi}_{s-1}(e_{s,n}).$$

By the definition of $d_{\arg,\varphi}$, this is the same as

$$\|z^B_{s-1,n} - E_{s-1,n}z^A_{s-1,n}\|_{L^2} \le \omega^{\varphi}_{s-1}(e_{s,n}).$$

Since $e_{s,n} \to 0$ and $\omega^{\varphi}_{s-1}(t) \to 0$, we get $e_{s-1,n} \to 0$.

Now suppose the construction has reached layer $r+1$, so that a matching map of the allowed form $E_{r+1,n}$ has been found and

$$e_{r+1,n} := d_{\arg,\varphi}\big(E_{r+2,n}G^A_{r+1,n}, G^B_{r+1,n}; E_{r+1,n}\big) \to 0.$$

Equivalently,

$$\|z^B_{r+1,n} - E_{r+1,n}z^A_{r+1,n}\|_{L^2} = e_{r+1,n}.$$

The B-side one-step expansion through layer r represents $z^B_{r+1,n}$, while the transformed A-side expansion $E_{r+1,n}G^A_{r,n}$ represents $E_{r+1,n}z^A_{r+1,n}$. Hence

$$\|Y^{\varphi}_{E_{r+1,n}G^A_{r,n}} - Y^{\varphi}_{G^B_{r,n}}\|_{L^2} = e_{r+1,n}.$$

Applying Lemma 10 at layer r gives a matching map of the allowed form $E_{r,n}$ with

$$e_{r,n} := d_{\arg,\varphi}\big(E_{r+1,n}G^A_{r,n}, G^B_{r,n}; E_{r,n}\big) \le \omega^{\varphi}_{r}(e_{r+1,n}).$$

Because $e_{r+1,n} \to 0$, the right-hand side tends to zero. Iterating this estimate gives the displayed composition of the error-bound functions.

It remains to translate matched-argument closeness into the terms of the axis alignment score function. Fix a layer r. The preceding paragraph gives

$$d_{\arg,\varphi}\big(E_{r+1,n}G^A_{r,n}, G^B_{r,n}; E_{r,n}\big) \longrightarrow 0.$$

Thus every matched coordinate error in (19) tends to zero. In the Softplus case, this is ordinary coordinatewise convergence after the identified permutation. In the ReLU case, the same positive rescaling used in the matched-argument distance is allowed in the zippering score. Therefore the score of every recovered matched pair tends to 1. Since there are only finitely many coordinates, every score is eventually above any fixed $\theta < 1$. Hence $\mathrm{AxisAlign}^{\mathrm{zip},\theta}_{r,\varphi}(A_n, B_n) = 1$ for all sufficiently large n. □

B.2.3 Jacobian regularity and soft zippering

Ideally, we can go further than just having an asymptotic result and actually figure out how much error upstream is bounded by downstream error. The missing ingredient in doing this is controlling the ill-conditioning of the Jacobian of the realization map from configurations to outputs. Specifically:

Definition 21 (Expansion map and derivative Gram matrix)**.** Let Θ^{φ}_r be a local finite-dimensional chart of one-step expansions after fixing the hidden-unit order. In the ReLU case, each argument is normalized in the chart by imposing $\|g_j\|_{L^2} = 1$; this simply avoids listing the same ReLU unit many times after multiplying its argument by a positive constant and dividing its outgoing coefficient by the same constant. Define the realization map as

$$F^{\varphi}_r(\vartheta) = Y^{\varphi}_{G_\vartheta}$$

so that its Jacobian is

$$J^{\varphi}_r(\vartheta) = DF^{\varphi}_r(\vartheta)^* DF^{\varphi}_r(\vartheta).$$

Then the expansion is *Jacobian-regular* if

$$\lambda_{\min}(J^{\varphi}_r) > 0.$$

For Softplus, we can explicitly calculate:

$$\delta Y = \delta b + \sum_j \delta v_j \sigma(g_j) + \sum_j v_j \sigma'(g_j)\delta g_j.$$

For ReLU within an affine cell,

$$\delta Y = \delta b + \sum_j \delta v_j \rho(g_j) + \sum_j v_j \mathbf{1}_{\{g_j>0\}}\delta g_j.$$

(For ReLU this derivative is taken after the normalization just described; otherwise the exact rescaling identity $(v, g) \mapsto (\alpha^{-1}v, \alpha g)$ would appear as a zero direction even though it does not change the unit.)

With this in mind we can actually bound the output-configuration relationship in an explicit way:

Lemma 11 (Quantitative one-step error inversion)**.** *Let $\mathcal{K}_r^\varphi$ be a compact comparison family satisfying the appropriate activation-specific quantitative minimality and regularity hypotheses. Let G_r be a one-step expansion in this family, and suppose that G_r is Jacobian-regular in a local chart. Then there are $\varepsilon_r > 0$ and $L_r < \infty$ such that*

$$\|Y^\varphi_{G_r} - Y^\varphi_{\widetilde{G}_r}\|_{L^2} \le \varepsilon_r$$

implies a matching map of the allowed form E satisfying

$$d_{\mathrm{arg},\varphi}(G_r, \widetilde{G}_r; E) \le L_r \|Y^\varphi_{G_r} - Y^\varphi_{\widetilde{G}_r}\|_{L^2}.$$

Proof. The proof has two parts. First, compactness and the exact theory of §B.1 show that a sufficiently close near-collision must use the same local matching as G_r. Second, the Jacobian lower bound gives a linear estimate inside that local chart.

Step 1: localization in one local chart. Choose a local chart U around G_r small enough that its hidden-unit matching is fixed. For ReLU, keep the arguments normalized throughout U by dividing each argument by its L^2 norm and putting the scale into the outgoing coefficient. Let d_{full} be any metric inducing the topology of the compact comparison family and measuring all expansion data, including coefficients and bias. The compactness argument from Lemma 10 gives

$$\|Y^\varphi_G - Y^\varphi_{\widetilde{G}}\|_{L^2} \to 0 \quad \Longrightarrow \quad d_{\mathrm{full}}(G, \widetilde{G}) \to 0.$$

Indeed, otherwise a convergent subsequence would limit to an exact null pair whose hidden arguments are not identified by the allowed activation-specific matching, contradicting Corollary 1. Therefore there is $\varepsilon_r^{(0)} > 0$ such that

$$\|Y^\varphi_{G_r} - Y^\varphi_{\widetilde{G}_r}\|_{L^2} \le \varepsilon_r^{(0)}$$

implies that $\widetilde{G}_r$ lies inside U, after the unique nearby permutation and, in the ReLU case, after the same simple normalization of scalar arguments.

Step 2: a lower singular-value estimate in coordinates. Let ϑ_0 be the coordinate of G_r and write the coordinate of $\widetilde{G}_r$ as $\vartheta_0 + h$. Set

$$\sigma_r = \sqrt{\lambda_{\min}(J_r^\varphi(\vartheta_0))} > 0.$$

Since $J_r^\varphi = DF_r^{\varphi *} DF_r^\varphi$, this means

$$\|DF_r^\varphi(\vartheta_0)h\|_{L^2} \ge \sigma_r \|h\| \qquad \text{for every coordinate tangent } h.$$

Continuity of the derivative lets us shrink U so that

$$\|DF_r^\varphi(\vartheta) - DF_r^\varphi(\vartheta_0)\|_{\mathrm{op}} \le \frac{\sigma_r}{2} \qquad \forall \vartheta \in U.$$

Take $\varepsilon_r \le \varepsilon_r^{(0)}$ small enough that the segment $\vartheta_0 + th$, $0 \le t \le 1$, remains in U. The integral form of Taylor's theorem gives

$$F_r^\varphi(\vartheta_0 + h) - F_r^\varphi(\vartheta_0) = DF_r^\varphi(\vartheta_0)h$$

$$+\int_0^1 \big[DF_r^{\varphi}(\vartheta_0+th)-DF_r^{\varphi}(\vartheta_0)\big]h\,dt.$$

Using the lower singular-value estimate at ϑ_0 and the derivative-variation bound,

$$\|F_r^{\varphi}(\vartheta_0+h)-F_r^{\varphi}(\vartheta_0)\|_{L^2}\geq \sigma_r\|h\|-\int_0^1 \frac{\sigma_r}{2}\|h\|\,dt\geq \frac{\sigma_r}{2}\|h\|.$$

Thus

$$\|h\|\leq \frac{2}{\sigma_r}\|Y^{\varphi}_{G_r}-Y^{\varphi}_{\widetilde{G}_r}\|_{L^2}.$$

On the fixed local chart, the map from coordinates to the matched scalar arguments is locally Lipschitz. For ReLU, the positive scale allowed in E is just the scale recorded by the local coordinates before normalization. Therefore, for some $C_r^{\arg}<\infty$,

$$d_{\arg,\varphi}(G_r,\widetilde{G}_r;E)\leq C_r^{\arg}\|h\|$$

for the fixed local matching. Taking

$$L_r=\frac{2C_r^{\arg}}{\sigma_r}$$

gives the displayed estimate. □

With that explicit error bounding at each single step, we can state a quantitative soft zippering theorem:

Theorem 8 (Soft zippering)**.** *Fix $\varphi\in\{\sigma,\rho\}$ and $q<s$. Assume every backward comparison encountered in the zippering induction lies in compact families satisfying the appropriate activation-specific quantitative minimality and regularity hypotheses, and assume the realized B-side expansion at each step is Jacobian-regular, with constants L_r,ε_r. If the terminal error $e_s=\|z_s^B-E_sz_s^A\|_{L^2}$ satisfies*

$$e_s\prod_{t=r+1}^{s-1}L_t\leq\varepsilon_r\qquad(r=q,\dots,s-1)$$

for the Lemma 11 constants L_r,ε_r for every $r=q,\dots,s-1$, then

$$\mathrm{AxisAlign}^{\theta}_{r,\varphi}(A,B)\geq 1-\frac{e_s^2\prod_{t=r}^{s-1}L_t^2}{(1-\theta)^2c^2d_r}$$

where for softplus c is the quantitative bound from definition 18, and for ReLU $c=1$.

Proof. We carry out the backward induction explicitly. For a one-step expansion G_r^N, recall that

$$Y^{\varphi}_{G_r^N}=z^N_{r+1}.$$

If $E(y)=Ty+a$ is affine, then EG_r^A denotes the transformed expansion obtained by replacing the bias by $Tb+a$ and every outgoing column by Tv_j; its represented function is Ez^A_{r+1}. The scalar arguments of EG_r^A are still the layer-r arguments $z^A_{r,j}$.

Step 1: the terminal backward step. The two layer-$(s-1)$ expansions are

$$E_sz_s^A=T_sb_s^A+a_s+\sum_j T_sW_s^A[:,j]\varphi(z^A_{s-1,j}),$$

$$z_s^B=b_s^B+\sum_i W_s^B[:,i]\varphi(z^B_{s-1,i}).$$

Their represented-function distance is exactly

$$\|Y^{\varphi}_{E_sG^A_{s-1}}-Y^{\varphi}_{G^B_{s-1}}\|_{L^2}=\|E_sz_s^A-z_s^B\|_{L^2}=e_s.$$

The small-error hypothesis for $r=s-1$ says $e_s\leq\varepsilon_{s-1}$. Lemma 11 therefore gives a matching map of the allowed form E_{s-1} such that

$$e_{s-1}:=d_{\arg,\varphi}\big(E_sG^A_{s-1},G^B_{s-1};E_{s-1}\big)\leq L_{s-1}e_s.$$

By (19), this is the same as

$$\|z^B_{s-1} - E_{s-1} z^A_{s-1}\|_{L^2} \le L_{s-1} e_s.$$

This proves the desired bound at the top zippering step.

Step 2: the inductive step. Suppose that for some $r \le s-2$ a matching map of the allowed form E_{r+1} has already been constructed and

$$e_{r+1} := d_{\arg,\varphi}\big(E_{r+2} G^A_{r+1}, G^B_{r+1}; E_{r+1}\big) \le e_s \prod_{t=r+1}^{s-1} L_t.$$

Equivalently,

$$\|z^B_{r+1} - E_{r+1} z^A_{r+1}\|_{L^2} = e_{r+1}.$$

The B-side expansion of z^B_{r+1} is

$$z^B_{r+1} = b^B_{r+1} + \sum_i W^B_{r+1}[:, i]\varphi(z^B_{r,i}),$$

and the transformed A-side expansion is

$$E_{r+1} z^A_{r+1} = E_{r+1} b^A_{r+1} + \sum_j E_{r+1} W^A_{r+1}[:, j]\varphi(z^A_{r,j}),$$

where the affine translation part of E_{r+1} is included in the transformed bias. These are precisely the represented functions of G^B_r and $E_{r+1} G^A_r$, and their distance is e_{r+1}. The assumed smallness condition gives

$$e_{r+1} \le e_s \prod_{t=r+1}^{s-1} L_t \le \varepsilon_r.$$

Applying Lemma 11 at layer r produces a matching map of the allowed form E_r and

$$e_r := d_{\arg,\varphi}\big(E_{r+1} G^A_r, G^B_r; E_r\big) \le L_r e_{r+1} \le e_s \prod_{t=r}^{s-1} L_t.$$

Backward induction proves the $d_{\arg,\varphi}$ bound for every $r = q, \dots, s-1$.

Step 3: convert the matched argument error into an axis count. Fix a layer r and let B^θ_r be the number of coordinates that fail the zippering threshold under the matching supplied by E_r. The induction gives

$$d_{\arg,\varphi}\big(E_{r+1} G^A_r, G^B_r; E_r\big)^2 \le e_s^2 \prod_{t=r}^{s-1} L_t^2.$$

For Softplus, E_r is a permutation. If π_r is that permutation, the preceding display reads

$$\sum_{j=1}^{d_r} \|z^B_{r,\pi_r(j)} - z^A_{r,j}\|^2_{L^2} \le e_s^2 \prod_{t=r}^{s-1} L_t^2.$$

If coordinate j has score below θ, then

$$\begin{aligned} \|z^B_{r,\pi_r(j)} - z^A_{r,j}\|_{L^2} &> (1-\theta)\big(\|z^B_{r,\pi_r(j)}\|_{L^2} + \|z^A_{r,j}\|_{L^2}\big) \\ &\ge (1-\theta)c, \end{aligned}$$

because quantitative nonconstancy gives $\|z^A_{r,j}\|_{L^2} \ge c$. Every bad coordinate therefore contributes more than $(1-\theta)^2 c^2$ to the total matched squared argument error. Hence

$$B^\theta_r (1-\theta)^2 c^2 \le e_s^2 \prod_{t=r}^{s-1} L_t^2.$$

Since $\mathrm{AxisAlign}^{\mathrm{zip},\theta}_{r,\sigma} = 1 - B^\theta_r / d_r$, this gives the Softplus bound.

For ReLU, write the positive scale in the matched coordinate as $\alpha_j > 0$, so that the coordinate error controlled by $d_{\arg,\rho}$ is

$$\|z^B_{r,\pi_r(j)} - \alpha_j z^A_{r,j}\|_{L^2}.$$

If a matched pair has positive-scale score below θ, then this particular positive scale also fails the score test, so

$$\|z^B_{r,\pi_r(j)} - \alpha_j z^A_{r,j}\|_{L^2} > (1-\theta)\big(\|z^B_{r,\pi_r(j)}\|_{L^2} + \alpha_j \|z^A_{r,j}\|_{L^2}\big).$$

For normalized ReLU arguments, $\|z^B_{r,\pi_r(j)}\|_{L^2} = 1$ and $\|z^A_{r,j}\|_{L^2} = 1$, so the right-hand side is $(1-\theta)(1+\alpha_j) \geq (1-\theta)$. Thus every bad ReLU coordinate contributes more than $(1-\theta)^2$ to the total matched squared argument error, and

$$(1-\theta)^2 B^\theta_r \leq e_s^2 \prod_{t=r}^{s-1} L_t^2.$$

Dividing by d_r gives the ReLU bound. The stronger factor $4(1-\theta)^2$ would follow if the controlled error were the normalized comparison at the fixed scale $\alpha = 1$. Here, however, $d_{\arg,\rho}$ controls the error at the recovered positive scale α_j, which is the scale needed for the zippering map itself. □

B.3 Genericity

B.3.1 The overall genericity strategy

This section asks why the minimality, regularity, and Jacobian assumptions used above should hold for typical parameterized networks. Propositions 1 and 2 show that qualitative minimality fails only on lower-dimensional bad sets; Proposition 3 then upgrades this to positive c-margins with high probability on compact parameter families. Those are the compact quantitative-minimal families used in Lemma 10.

We will establish that:

1. Intrinsic minimality fails only on analytic or algebraic degeneracy sets, conditional on task-usedness and task exposure;
2. The regularity mechanisms fail only on resonance, incidence, or rank-defect sets, provided the required condition holds at one point on each relevant stratum;
3. Jacobian regularity, after fixing the standard hidden-unit order and ReLU normalization, fails on the zero set of an analytic Gram determinant, provided the full-rank condition holds at least at one point.

The first two give generic identifiability. The third controls the generic geometry of exact null-pair collisions. The generic null-net theorem below combines them and converts pairwise genericity into generic uniqueness of the set of parameters realizing the same function as an individual network. Proposition 3 explains how the qualitative full-measure statements feed back into the quantitative hypotheses of §B.2.

B.3.2 Genericity of minimality

In this section, we show that qualitative softplus and ReLU minimality are generic, and that the quantitative versions are high probability conditional on qualitative minimality holding.

Proposition 1 (Generic qualitative Softplus minimality)**.** *Fix a set of Softplus expansions for which at least one private regular pole witness exists for each counted sign-class. Then failure of qualitative Softplus minimality is contained in a finite union of proper real-analytic subsets and has empty interior and measure zero.*

Proof. Work on one connected analytic parameter chart and fix the finitely many marked pole neighborhoods used by the definition. The elementary failure modes can be represented by nonnegative analytic defect functions. For example, set

$$U_j(\vartheta) = \|v_j(\vartheta)\|^2,$$

$$N_j(\vartheta) = \|g_j(\vartheta) - \mathbb{E}_\Omega g_j(\vartheta)\|_{L^2}^2,$$
$$D_{ij}^{\pm}(\vartheta) = \|g_i(\vartheta) \pm g_j(\vartheta)\|_{L^2}^2.$$

For a nonempty subset S, write

$$f_{S,\vartheta} = \sum_{j\in S} v_j(\vartheta) g_j(\vartheta), \qquad C_S(\vartheta) = \|f_{S,\vartheta} - \mathbb{E}_\Omega f_{S,\vartheta}\|_{L^2}^2.$$

Then $U_j = 0$ is unusedness, $N_j = 0$ is constancy, $D_{ij}^{\pm} = 0$ is duplication or opposition, and $C_S = 0$ is affine sign-cancellation. Bounded analytic dependence on the parameters and compactness of Ω permit differentiation under the integral, so these functions are real analytic. The existence of one minimal witness ensure that none of the relevant defects vanishes identically. Hence each corresponding bad set is a proper analytic zero set.

For the pole clause, let $Q_j = 1 + e^{g_j^{\mathbb{C}}}$ on the marked continuation chart. Failure of regularity of the selected pole patch is the vanishing of the appropriate differential minors of Q_j. Accidental sharing of a relatively open pole component by two distinct sign-classes is, after fixing the local incidence type, an analytic component-incidence condition. The private-pole witness makes each such incidence set proper. There are finitely many counted indices, pairs, subsets, and marked patches. Their union therefore has empty interior and measure zero. □

Proposition 2 (Generic qualitative ReLU EFF-minimality)**.** *Fix a finite-dimensional ReLU cell/facet chart of one-step expansions that contains at least one EFF-minimal expansion. Then the set of parameters in the chart for which the expansion is not EFF-minimal is contained in a finite union of proper real-algebraic subsets, and thus has empty interior and measure zero in the chart.*

Proof. On this fixed cell/facet chart, the lower-layer ReLU pattern does not change. Hence there is a fixed finite list of inherited cells and marked fresh-facet slots, and on each cell every counted argument has the affine form

$$g_j^{\vartheta}|_C(x) = a_{j,C}(\vartheta)^\top x + c_{j,C}(\vartheta).$$

The coefficient vectors $a_{j,C}(\vartheta)$ and constants $c_{j,C}(\vartheta)$ are polynomial in the raw parameters, and real-analytic after the normalization used to remove positive ReLU scale. Therefore all algebraic relations among the cellwise affine pieces are described by finitely many polynomial, or after normalization real-analytic, equations.

We now record the possible failures of EFF-minimality in these coordinates. First, task-visible fresh-facet use can fail only if one of the selected fresh-facet certificates degenerates: for example, the relevant affine normal vanishes, the marked zero facet ceases to be fresh inside the inherited cell, or the fixed two-sided crossing condition collapses. In the chosen chart these are detected by finitely many polynomial or analytic certificate functions.

Second, internal exposure can fail if another counted argument shares an open piece of a marked fresh facet without being globally sign-scale equivalent. On the cell containing that facet, sharing an open piece of the zero set forces proportionality of the two augmented affine coefficient vectors,

$$(a_{j,C}(\vartheta), c_{j,C}(\vartheta)) \parallel (a_{i,C}(\vartheta), c_{i,C}(\vartheta)).$$

This proportionality is exactly the vanishing of the corresponding 2×2 minors. Global sign-scale duplication is also described by finitely many such proportionality equations, imposed compatibly across all inherited cells.

Third, affine sign-cancellation can fail. For a fixed nonempty subset S of counted terms, the cellwise expression

$$\sum_{j\in S} v_j(\vartheta) g_j^{\vartheta}$$

is constant on the task set only if, on every full-dimensional inherited cell C,

$$\sum_{j\in S} v_j(\vartheta) a_{j,C}(\vartheta)^\top = 0,$$

and the corresponding cellwise constants agree on cells lying in the same connected component. These are again finitely many polynomial, or after normalization real-analytic, equations.

There are only finitely many counted gates, inherited cells, marked facets, pairs of arguments, and nonempty subsets S. Thus the total EFF-minimality failure set is contained in a finite union of algebraic or analytic zero sets.

It remains only to see that these zero sets are proper. By assumption, the chart contains at least one EFF-minimal expansion. At that parameter value none of the above EFF-failure conditions holds. Therefore none of the corresponding certificate functions vanishes identically on the chart. Each bad set is consequently a proper real-algebraic subset, or after normalization a proper real-analytic subset. A finite union of such proper subsets has empty interior and measure zero. □

Having established that the qualitative forms of minimality are generic, we now establish that the quantitative margins version is high probability:

Proposition 3 (Positive margins for quantitative minimality are generic)**.** *Fix a layer and work inside one compact subset K of a finite-dimensional parameter or solution chart, with the relevant Softplus pole-patch data or ReLU cell/facet data fixed. Let μ be a probability law on K. Suppose the qualitative minimality failure set,*

$$\mathscr{B}_{\mathrm{qual}} = \{\vartheta \in K : \textit{the layer is not qualitatively minimal at } \vartheta\},$$

has μ-measure zero. Then for every $\delta > 0$, there exist $M < \infty$ and $c > 0$ such that, with μ-probability at least $1 - \delta$, the layer satisfies the corresponding quantitative minimality conditions with parameter bound M and margin c.

Proof. On the fixed chart, the qualitative minimality conditions are tested by finitely many continuous nonnegative certificate functions $d_1, \dots, d_N$. These include the relevant coefficient norms, variation or crossing quantities, duplicate/opposite separation distances, affine sign-cancellation distances, and private pole/facet margins. Qualitative minimality holds exactly when all certificates are positive. Therefore, with

$$d(\vartheta) = \min_{1 \le a \le N} d_a(\vartheta),$$

we have $\mathscr{B}_{\mathrm{qual}} = \{d = 0\}$. By assumption, $\mu(\{d = 0\}) = 0$. Since the sets $\{d < \eta\}$ decrease to $\{d = 0\}$ as $\eta \downarrow 0$, continuity of probability from above gives $\mu(\{d < \eta\}) \to 0$. Choose $c > 0$ such that $\mu(\{d < c\}) \le \delta$. Then, outside a set of probability at most δ, every certificate satisfies $d_a \ge c$.

Finally, K is compact, so the parameter norm is bounded on K; choose M larger than this bound. The constants M and c are exactly the boundedness and positive-margin inputs in the corresponding quantitative minimality definition. If finitely many layers are considered at once, apply the same argument to each layer and use a finite union bound. □

B.3.3 Genericity of Jacobian regularity

We now show that Jacobian regularity is generi.

Proposition 4 (Generic Jacobian regularity)**.** *Fix a finite-dimensional chart for one-step expansions after the activation symmetries have been removed. If the Jacobian regularity condition holds at least at one point, then the Jacobian-singular set*

$$\{\vartheta : \det J_r^{\varphi}(\vartheta) = 0\}$$

is a proper real-analytic subset of the chart and therefore has empty interior and measure zero.

Proof. Let $d = \dim \Theta_r^{\varphi}$. In the chosen coordinates, J_r^{φ} is the $d \times d$ Gram matrix

$$(J_r^{\varphi})_{ab}(\vartheta) = \langle \partial_a F_r^{\varphi}(\vartheta), \partial_b F_r^{\varphi}(\vartheta) \rangle_{L^2}.$$

The expansion is full-rank at ϑ exactly when this Gram matrix is nonsingular, or equivalently when $\det J_r^{\varphi}(\vartheta) \neq 0$.

For Softplus, the first-variation features are analytic functions of the parameters and of the input. On compact Ω, the integrals defining the Gram entries may be differentiated under the integral sign. Hence every entry of J_r^{σ} is real analytic in ϑ.

For ReLU, the reason for fixing the activation pattern is simple: if the pattern does not change, then the same finite cell decomposition of Ω is used throughout the chart. On each cell, each ReLU gate is either active or inactive, so the indicator $\mathbf{1}_{\{g_j>0\}}$ appearing in the derivative formula is constant on that cell. The derivative features are therefore affine in x on each cell, with coefficients depending polynomially or analytically on the parameters. If the cell boundaries move with the parameters, subdivide the cells into finitely many simplices and pull each integral back to a fixed reference simplex. The simplex vertices and the integrand coefficients vary analytically, so each cell integral, and hence each Gram entry, is real analytic on the fixed-pattern chart.

Thus $\det J_r^{\varphi}$ is a real-analytic function. The full-rank condition at one point says this determinant is nonzero at least at one point, so it is not identically zero. Its zero set is therefore a proper real-analytic subset of the chart. Such a subset has empty interior and measure zero. This is the claimed generic Jacobian regularity. □

B.3.4 Genericity of the Softplus MLR regularity

The MLR hypotheses remove the concrete ways a Softplus pole match can fail to determine an argument. There are two main exceptional mechanisms:

- First, the monodromy characters can satisfy exact integer relations: this produces torsion, cross-character commensurability, or nonprimitive exponents such as $1+Z^3=(1+Z)(1-Z+Z^2)$. These are composite resonances.
- Second, even when the Laurent binomials are irreducible, the lower representation map may collapse the relevant Laurent coordinates, so two distinct Laurent hypersurfaces appear to share a pulled-back divisor. This is a failure of the rank condition for the lower-feature map.

The next two results show that both mechanisms are described by affine integer-resonance equations or by vanishing analytic minors, and thus form lower-dimensional exceptional sets.

Lemma 12 (Character and composite resonances are lower-dimensional)**.** *After fixing a monodromy setup, relations of the form*

$$\chi_i^q = \chi_j^{q'}$$

for nonzero integers q, q', torsion relations, and nonprimitive exponent relations are exact integer-affine constraints on the layer weights. Unless forced by tying, their union is a countable union of proper affine subsets and has measure zero.

Proof. Choose a basis of the finitely generated character group in this setup. The character of a current argument is determined by an affine-linear function of its incoming weights. A relation $\chi_i^q = \chi_j^{q'}$, or a torsion relation, therefore has the form

$$q\,w_i - q'\,w_j - \nu = 0, \qquad \nu \in \mathbb{Z}^p,$$

with the analogous finite integer-linear expression for more characters. For each fixed integer tuple this is an affine subspace of the untied weight parameter space. If the relation is not forced, the subspace is proper. Nonprimitivity of an exponent is likewise an exact divisibility relation $m = qm_0$, $|q|>1$, in the character lattice. There are only countably many integer tuples and lattice elements, so the union of all such proper affine sets has Lebesgue measure zero and empty interior. □

The exact Softplus theorem in §B.1 assumes Laurent divisor-faithfulness as part of MLR. The next lemma is a "verification tool" for that clause: it says that a shared pulled-back divisor is genuine whenever the lower-feature map has large enough rank.

Lemma 13 (Rank criterion for Laurent divisor-faithfulness)**.** *Suppose*

$$e^g = cT^m, \qquad e^h = dT^n$$

in a fixed monodromy chart, and let Σ be a shared outer pole patch. If near a generic point of Σ there is a complex submanifold L on which c, d are constant and

$$\operatorname{rank}_{\mathbb{C}} D(T|_L) = s,$$

then the pair is Laurent divisor-faithful at Σ.

Proof. Write

$$P(T) = 1 + cT^m, \qquad R(T) = 1 + dT^n.$$

On L, the pulled-back functions are $P(T(x)) = 1 + e^{g(x)}$ and $R(T(x)) = 1 + e^{h(x)}$. The shared patch $\Sigma \cap L$ is a regular piece of the zero set of $P(T(x))$, and by assumption $R(T(x))$ also vanishes on it.

Since $D(T|_L)$ has rank s at a generic point of the patch, $T|_L$ is locally a submersion there. Therefore the image under T of a relatively open piece of $\Sigma \cap L$ contains a relatively open piece of the Laurent hypersurface $\{P = 0\}$ in $(\mathbb{C}^*)^s$. The second Laurent binomial R vanishes on that image, because its pullback $R(T(x))$ vanishes on the shared patch.

Thus R vanishes on a nonempty analytic open subset of an irreducible component of $\{P = 0\}$. If m is primitive, P is prime by Lemma 7; more generally, on the selected irreducible component this means the corresponding irreducible Laurent factor of P divides R. Hence the common pulled-back divisor comes from a genuine common Laurent factor. This is exactly Laurent divisor-faithfulness. □

Once a particular local Softplus pole match has been fixed, the next proposition shows that the rank condition in Lemma 13 fails only on a lower-dimensional exceptional subset; hence Laurent divisor-faithfulness is generic within that fixed pole-matching family.

Proposition 5 (Generic Laurent divisor-faithfulness)**.** *Let S be a family of Softplus comparison parameters and g_θ, h_θ a pair of divisor-matched arguments in S. Suppose the rank condition of Lemma 13 holds at one point of S. Then Laurent divisor-faithfulness holds for an open dense full-measure subset of S.*

Proof. Let S denote the connected divisor-incidence stratum in the proposition. Thus, for every parameter value in S, the same pair of Softplus pole divisors is being required to share the same kind of local pole patch. We work in a local analytic coordinate chart on S; a locally finite collection of such charts gives the same conclusion globally.

For a parameter value $\theta \in S$, write the two Laurent forms as

$$e^{g_\theta} = c_\theta T_\theta^m, \qquad e^{h_\theta} = d_\theta T_\theta^n,$$

where

$$T_\theta = (T_{\theta,1}, \ldots, T_{\theta,s})$$

is the Laurent-coordinate map into $(\mathbb{C}^*)^s$. In other words, the functions $T_{\theta,a}$ are the local Laurent coordinates in which the two Softplus factors become

$$P_\theta(T) = 1 + c_\theta T^m, \qquad R_\theta(T) = 1 + d_\theta T^n.$$

The rank criterion in Lemma 13 is checked on a local complex submanifold L_θ through the shared pole patch, chosen so that the invariant coefficients c_θ, d_θ are constant on L_θ. Choose a reference point on the shared pole patch. The local submanifolds L_θ, the reference points x_θ, and the restrictions

$$T_\theta|_{L_\theta} : L_\theta \to (\mathbb{C}^*)^s$$

vary analytically with θ. Now let

$$J_\theta := D\big(T_\theta|_{L_\theta}\big)_{x_\theta}.$$

The condition required by Lemma 13 is $\operatorname{rank}_{\mathbb{C}} J_\theta = s$. This is equivalent to saying that at least one $s \times s$ minor of J_θ is nonzero. Define

$$\Delta(\theta) = \sum_M \left|\det M(J_\theta)\right|^2,$$

where the sum is over all relevant $s \times s$ minors. This is a real-analytic nonnegative function on the coordinate chart in S, and

$$\Delta(\theta) > 0 \quad \Longleftrightarrow \quad \operatorname{rank}_{\mathbb{C}} J_\theta = s.$$

By hypothesis, the rank condition of Lemma 13 holds at at least one point of the connected stratum. Hence Δ is not identically zero. Therefore its zero set

$$Z = \{\theta \in S : \Delta(\theta) = 0\}$$

is a proper real-analytic subset of S. Such a subset has empty interior and measure zero. Consequently, $S \setminus Z$ is open, dense, and full-measure in S. For every $\theta \in S \setminus Z$, the restricted Laurent-coordinate map $T_\theta|_{L_\theta}$ has full rank s at the chosen point. Thus Lemma 13 applies. It says that the shared pulled-back pole patch comes from a genuine common Laurent factor of

$$P_\theta(T) = 1 + c_\theta T^m \qquad \text{and} \qquad R_\theta(T) = 1 + d_\theta T^n.$$

That is exactly Laurent divisor-faithfulness at the selected shared pole patch. Hence Laurent divisor-faithfulness holds on an open dense full-measure subset of S. □

B.3.5 Genericity of ReLU MFR regularity

The ReLU exceptional mechanisms parallel the Softplus ones, but with kink-facet signatures replacing pole divisors. Once the task has exposed enough fresh facets, there are, as with the softplus case, two main ways MFR can fail:

- First, two distinct normalized piecewise affine arguments may have the same labelled multi-facet signature; this is a failure of the signature map to be locally injective. The first proposition below rules out persistent off-diagonal equal signatures when the secant derivative has the required rank at one point.
- Second, the facets of one source argument may be split across several competitor arguments, so that no single competitor carries the whole signature. The second proposition rules out persistent split matching when the analogous split-incidence derivative has the required rank at one point.

Together, these results justify the ReLU part of the general-position assumption on collision strata to be used in the generic null-net theorem below in Section B.3.

Let $\Sigma_r : \Xi_r \to \mathbb{R}^{N_r}$ be the signature map from Definition 17, with $p_r = \dim \Xi_r$. An off-diagonal signature-collision stratum is a smooth piece of the set of distinct normalized arguments with the same labelled multi-facet signature. The rank test below uses the derivative of the difference map

$$H(\xi, \widetilde{\xi}) = \Sigma_r(\xi) - \Sigma_r(\widetilde{\xi}).$$

Following the standard secant-variety terminology, we call this derivative the *secant derivative* (Landsberg, 2012):

$$\mathcal{T}_{\xi,\widetilde{\xi}}(h, \widetilde{h}) = D\Sigma_r(\xi)h - D\Sigma_r(\widetilde{\xi})\widetilde{h}.$$

Proposition 6 (Generic multi-facet recovery)**.** *Suppose every connected off-diagonal signature-collision stratum contains one point with*

$$\operatorname{rank} \mathcal{T}_{\xi,\widetilde{\xi}} \geq p_r + 1.$$

Then every off-diagonal signature-collision stratum has dimension at most $p_r - 1$. Hence ambiguous equal-signature values occur only on a lower-dimensional subset of the normalized argument chart.

Proof. Let S be a connected smooth off-diagonal stratum of

$$\{(\xi, \widetilde{\xi}) : \Sigma_r(\xi) = \Sigma_r(\widetilde{\xi}),\ \xi \neq \widetilde{\xi}\}.$$

At the assumed point $p = (\xi, \widetilde{\xi})$, the tangent space satisfies

$$T_p S \subseteq \ker \mathcal{T}_{\xi,\widetilde{\xi}},$$

because differentiating the signature equality along S gives zero. The domain of $\mathcal{T}$ has dimension $2p_r$, and its rank at that point is at least $p_r + 1$. Therefore

$$\dim S = \dim T_p S \leq 2p_r - (p_r + 1) = p_r - 1.$$

The relevant rank condition is the nonvanishing of finitely many $(p_r + 1) \times (p_r + 1)$ minors. Since one such minor is nonzero at that point, the rank inequality holds on an open dense, full-measure subset of the connected stratum. Finally, projection to the first argument chart is locally Lipschitz and cannot increase Hausdorff dimension. Thus normalized arguments with a distinct equal-signature partner form a set of dimension at most $p_r - 1$. □

This proposition is the ReLU analogue of generic Laurent divisor-faithfulness for Softplus: it says that ambiguous same-signature pairs are lower-dimensional. It handles the case where one source argument is matched to one competitor argument. The remaining issue is split matching, where different facets of one source argument are explained by different competitors.

For split matching, write the signature componentwise as

$$\Sigma_r = (\Sigma_{r,1}, \dots, \Sigma_{r,L}).$$

Fix $q \geq 2$ and a surjective assignment $\tau : \{1, \dots, L\} \to \{1, \dots, q\}$, meaning that facet ν of a source argument is matched by competitor $\eta_{\tau(\nu)}$. Define

$$H_\tau(\xi, \eta_1, \dots, \eta_q) = \big(\Sigma_{r,\nu}(\xi) - \Sigma_{r,\nu}(\eta_{\tau(\nu)})\big)_{\nu=1}^{L}.$$

A split-incidence stratum is a smooth piece of the set $H_\tau^{-1}(0)$, where this fixed split assignment is imposed.

Proposition 7 (Split matching is generically lower-dimensional)**.** *Suppose that, for every nontrivial split assignment τ, every connected smooth stratum of the split-incidence set $H_\tau^{-1}(0)$ contains a point at which*

$$\operatorname{rank} DH_\tau \geq qp_r + 1.$$

Then the projection of the split-matching set to the source argument chart has Hausdorff dimension at most $p_r - 1$.

Proof. The domain of H_τ has dimension $(q+1)p_r$. Let S_τ be a smooth incidence stratum, and choose a point of S_τ where the displayed rank condition holds. At that point,

$$TS_\tau \subseteq \ker DH_\tau,$$

so

$$\dim S_\tau \leq (q+1)p_r - (qp_r + 1) = p_r - 1.$$

As above, the rank inequality is the nonvanishing of analytic minors and therefore holds generically on the stratum once it holds at one point. Projection to the source coordinate cannot increase Hausdorff dimension. For a fixed finite facet atlas and finite width bound, only finitely many nontrivial assignments τ occur, so the union of their projected sets still has dimension at most $p_r - 1$. □

Taken together, propositions 6 and 7 explain why the ReLU signature-recovery hypotheses hold generically on any full-dimensional collision stratum, after discarding lower-dimensional signature and split-incidence exceptions.

B.3.6 The generic null-net theorem

The preceding subsections gave activation-specific genericity inputs: Softplus MLR failures and ReLU MFR failures occur only on lower-dimensional exceptional sets, provided the relevant rank conditions hold somewhere on the stratum being considered. The purpose of this subsection is to combine those activation-specific inputs with the geometry of the realization map. The conclusion is not a complete classification of all null networks. Rather, it says that nontrivial one-step null partners cannot occur on an open, full-dimensional family.

Recall the definitions of realization map F_r^φ, collision set $\mathcal{C}_r^\varphi$, and off-diagonal collision strata from Definition 9. Throughout this subsection we work inside a fixed normalized finite-dimensional realization chart Θ_r^φ, where "normalized" means that the hidden-unit order has been fixed, and in the ReLU case the positive-rescaling freedom has also been fixed.

Lemma 14 (Dimension bound for collision strata)**.** *Let S be a smooth off-diagonal collision stratum, and suppose that F_r^φ is an immersion at both members of every collision pair in S:*

$$\operatorname{rank} DF_r^\varphi(\vartheta) = \operatorname{rank} DF_r^\varphi(\widetilde{\vartheta}) = d_r^\Theta, \qquad (\vartheta, \widetilde{\vartheta}) \in S.$$

Then $\dim S \leq d_r^\Theta$.

Proof. Fix $p = (\vartheta, \widetilde{\vartheta}) \in S$. Define the secant derivative

$$\mathcal{S}_p(h, \widetilde{h}) = DF_r^{\varphi}(\vartheta)h - DF_r^{\varphi}(\widetilde{\vartheta})\widetilde{h}.$$

If $(h, \widetilde{h}) \in T_pS$, then differentiating the collision identity $F_r^{\varphi}(\vartheta(t)) = F_r^{\varphi}(\widetilde{\vartheta}(t))$ along a curve in S gives

$$\mathcal{S}_p(h, \widetilde{h}) = 0.$$

Hence $T_pS \subseteq \ker \mathcal{S}_p$. The restriction

$$h \longmapsto \mathcal{S}_p(h, 0) = DF_r^{\varphi}(\vartheta)h$$

is injective by the immersion hypothesis, and has rank d_r^{Θ}. Therefore $\operatorname{rank} \mathcal{S}_p \geq d_r^{\Theta}$. By rank-nullity,

$$\dim S = \dim T_pS \leq \dim \ker \mathcal{S}_p \leq 2d_r^{\Theta} - d_r^{\Theta} = d_r^{\Theta}.$$

□

Remark 3 (Full-dimensional collision strata). The dimension equality case in the lemma above has useful geometric meaning. If $\dim S = d_r^{\Theta}$, then the first-coordinate projection

$$\pi_1 : S \to \Theta_r^{\varphi}, \qquad \pi_1(\vartheta, \widetilde{\vartheta}) = \vartheta,$$

is locally a diffeomorphism. Indeed, if

$$(0, \widetilde{h}) \in \ker D\pi_1(p),$$

then $(0, \widetilde{h}) \in T_pS$, so

$$DF_r^{\varphi}(\widetilde{\vartheta})\widetilde{h} = 0,$$

and the immersion hypothesis at $\widetilde{\vartheta}$ forces $\widetilde{h} = 0$. Thus $D\pi_1(p)$ is injective, hence an isomorphism between spaces of dimension d_r^{Θ}. Therefore, locally,

$$S = \{(\vartheta, \Psi(\vartheta))\}, \qquad F_r^{\varphi}(\Psi(\vartheta)) = F_r^{\varphi}(\vartheta).$$

Similarly, the second-coordinate projection is locally a diffeomorphism. Thus a full-dimensional off-diagonal collision stratum is locally a nontrivial realization-preserving reparameterization. This explains why full-dimensional collision strata are the dangerous case.

Definition 22 (General position on full-dimensional collision strata)**.** A normalized realization chart satisfies *general position on full-dimensional collision strata* if every d_r^{Θ}-dimensional off-diagonal collision stratum contains at least one point where the activation-specific one-step identifiability result applies. Concretely, at that point the null pair is intrinsically minimal and MLR in the Softplus case; in the ReLU case it is EFF-minimal, coherently matched, and MFR.

The genericity results above explain how this condition is meant to be checked: the Softplus rank criterion controls the lower-feature-rank failure in Laurent divisor-faithfulness, and the ReLU signature and split-matching results control the same-signature and split-facet failures. Thus the condition says that a full-dimensional collision stratum is not trapped entirely inside the relevant exceptional sets.

Theorem 9 (Generic activation-independent null-net theorem)**.** *Work in the normalized finite-dimensional chart and collision stratification described above. Assume:*

- *(i) F_r^{φ} is Jacobian-regular on the chart, i.e. DF_r^{φ} has full column rank d_r^{Θ} at every point of the chart;*
- *(ii) the general-position condition on full-dimensional collision strata holds.*

Then every off-diagonal collision stratum has dimension at most $d_r^{\Theta} - 1$. Consequently the set of parameters with a nontrivial collision,

$$\mathcal{E}_r^{\varphi} = \{\vartheta : \exists \widetilde{\vartheta} \neq \vartheta,\ F_r^{\varphi}(\widetilde{\vartheta}) = F_r^{\varphi}(\vartheta)\},$$

has Hausdorff dimension at most $d_r^{\Theta} - 1$, hence empty interior and measure zero in the chart.

Proof. Let S be any smooth off-diagonal collision stratum. By Jacobian regularity, F_r^{φ} is an immersion at every point of the chart. Hence the hypothesis of Lemma 14 applies at both coordinates of every collision pair in S. Therefore $\dim S \le d_r^{\Theta}$. We next rule out the full-dimensional case. Suppose, for contradiction, that $\dim S = d_r^{\Theta}$. By general position, S contains a point $p_0 = (\vartheta_0, \widetilde{\vartheta}_0)$ where the activation-specific one-step identifiability theorem applies. Since p_0 is a collision, the two represented one-step functions are equal:

$$F_r^{\varphi}(\vartheta_0) = F_r^{\varphi}(\widetilde{\vartheta}_0).$$

The activation-specific identifiability result then says that the two one-step expansions differ only by the unavoidable hidden-unit matching: permutation for Softplus, and permutation plus positive coordinatewise scaling for ReLU. But the chart is normalized precisely so that this hidden-unit matching has already been fixed. Therefore the two parameter values are equal in the chart: $\vartheta_0 = \widetilde{\vartheta}_0$. Equivalently, $p_0 \in \Delta_r^{\varphi}$, where

$$\Delta_r^{\varphi} = \{(\vartheta, \vartheta) : \vartheta \in \Theta_r^{\varphi}\}$$

is the diagonal of trivial self-collisions. This contradicts the fact that S is off diagonal, i.e. $S \subset \mathcal{C}_r^{\varphi} \setminus \Delta_r^{\varphi}$. Hence no off-diagonal collision stratum has dimension d_r^{Θ}, and so every off-diagonal collision stratum satisfies

$$\dim S \le d_r^{\Theta} - 1.$$

It remains to pass from collision pairs to individual parameters. The set $\mathcal{E}_r^{\varphi}$ is not itself a set of pairs; it is the set of first coordinates of off-diagonal collision pairs. In other words,

$$\mathcal{E}_r^{\varphi} = \pi_1\big(\mathcal{C}_r^{\varphi} \setminus \Delta_r^{\varphi}\big), \qquad \pi_1(\vartheta, \widetilde{\vartheta}) = \vartheta.$$

Thus the first-coordinate projection simply forgets the nontrivial partner and remembers the parameter value that has such a partner.

On each local coordinate chart, π_1 is Lipschitz, so Hausdorff dimension cannot increase:

$$\dim_H \pi_1(S) \le \dim_H S \le d_r^{\Theta} - 1$$

for every off-diagonal stratum S. Since the off-diagonal collision set is a countable locally finite union of such smooth strata, and since

$$\mathcal{E}_r^{\varphi} = \bigcup_S \pi_1(S),$$

countable stability of Hausdorff dimension gives

$$\dim_H \mathcal{E}_r^{\varphi} \le d_r^{\Theta} - 1.$$

A subset of a d_r^{Θ}-dimensional smooth chart with Hausdorff dimension at most $d_r^{\Theta} - 1$ has zero chart-Lebesgue measure and cannot contain a nonempty open set. This proves the measure-zero and empty-interior conclusions. □

Corollary 4 (Generic Softplus and ReLU null-pair uniqueness)**.** *In any normalized chart satisfying the ambient setup above, Jacobian regularity plus the appropriate general-position condition implies generic null-pair uniqueness. For Softplus, this means that non-permutation null pairs form a measure-zero set. For ReLU, this means that nontrivial null pairs modulo permutation and positive coordinatewise scaling form a measure-zero set.*

Proof. Apply Theorem 9 with $\varphi = \sigma$ and use the Softplus specialization of the activation-specific one-step identifiability result; then apply it with $\varphi = \rho$ and use the ReLU specialization. The diagonal of obvious self-collisions is ordinary equality after choosing the Softplus permutation representative, and ordinary equality after choosing the ReLU permutation and positive-scale normalization. □

Corollary 5 (Generic exact zippering from generic null triviality)**.** *Suppose that, at each of finitely many compared layers, the relevant normalized one-step chart satisfies the ambient setup of this subsection and the two hypotheses of Theorem 9. Then, outside a finite union of layerwise measure-zero sets of parameters with nontrivial collision partners, exact terminal weak equivalence zippers upstream. The recovered maps are permutations for Softplus and permutations followed by positive coordinatewise scalings for ReLU.*

Proof. Begin at layer s. Exact terminal weak equivalence identifies the two layer-$(s-1)$ one-step represented functions: the B-side expansion of z_s^B and the transformed A-side expansion of $E_s z_s^A$ form a null pair. If the realized B-side expansion is outside the projected set of parameters with a nontrivial collision partner at layer $s-1$, Theorem 9 shows that this null pair is matched by the allowed hidden-unit map. Thus it produces the preceding comparison map E_{s-1}: a permutation for Softplus, or a permutation followed by positive coordinatewise scaling for ReLU.

Assume inductively that E_{r+1} has been recovered and

$$z_{r+1}^B = E_{r+1} z_{r+1}^A.$$

Expanding both sides through layer r again creates a one-step null pair. Outside the layer-r set of parameters with a nontrivial collision partner, generic null triviality produces E_r. Backward induction reaches every layer $r = q, \dots, s-1$. There are only finitely many compared layers, so the union of the layerwise exceptional sets remains measure zero. □

The preceding MLR and MFR genericity results verify the activation-specific identifiability input in the general-position condition, while Jacobian regularity supplies the geometric input used in Lemma 14. Neither ingredient alone is enough. Pairwise MLR/MFR genericity cannot exclude a realization-preserving graph that is lower-dimensional in the product but projects onto an open set of individual networks, while Jacobian regularity alone cannot exclude two distinct immersed parameter families with the same realized functions. General position on full-dimensional collision strata combines the two.

Remark 4 (Relation to complete null-network classifications). The null-network program of Vlačić and Bölcskei seeks an exhaustive activation-specific description of all regular zero-realizing networks and thereby obtains global identifiability modulo exact activation symmetries. The theorem above deliberately asks for less: it rules out full-dimensional families of nontrivial collisions without classifying every exceptional resonant null network. For Softplus, this avoids the hardest (and unsolved) missing step in a literal complete classification. One need not prove a factorization-completeness theorem for every divisor-free unit or nested resonance. Composite integer-weight, nested-identity constructions or lower-feature-rank-defect null networks may survive on lower-dimensional exceptional sets; they do not obstruct generic zippering unless an entire open family of realizations is trapped there. The same logic avoids a complete ReLU classification of every construction based on changes on ReLU-inactive regions or split-facet matching, which would probably be even harder than the Softplus case. The generic null-net theorem is therefore weaker than universal parameter identifiability, but exactly strong enough for the alignment and zippering conclusions pursued here.

Remark 5 (Optimizer nonconcentration and generic asymptotic/soft zippering). The role of genericity is only to justify that in the space of parameters, the conditions needed for zippering are full-measure. However, they do not guarantee that the exceptional set is avoided by learning algorithms. We are implicitly assuming, in interpreting the zippering theorems as applying to optimized networks, that the distribution produced by training, restricted to the relevant solution or comparison stratum, does not concentrate on the intrinsic-minimality, signature-recovery, nontrivial collision, or Jacobian-singular exceptional sets. Proving that this assumption is correct is an important topic for future work.

C Unequal depth, residual networks, and transformers

C.1 Depth-warped weak–strong equivalence

The weak–strong theorem compares adjacent layers. Empirically, however, networks of different depths often appear to have corresponding single axes at nearby but not identical depths. The right comparison is then not layer a of A against one prescribed layer of B, but layer a of A against a window of B-layers. We call this *depth warping*.

The results in the previous sections perform, in their essence, a kind of "signature matching" (where signatures are ReLU kink traces or Softplus poles): representations align because we can show that linear alignment forces alignment of signatures. In thinking about depth warping, we work with the same *kinds* of signatures as above, but we can no longer assume that the signatures live in exactly corresponding layers. The new issue is that, when B is deeper than A, the counterpart of an A-axis may appear anywhere in a short window of B-layers. Thus a whole layer of A-axes can be distributed

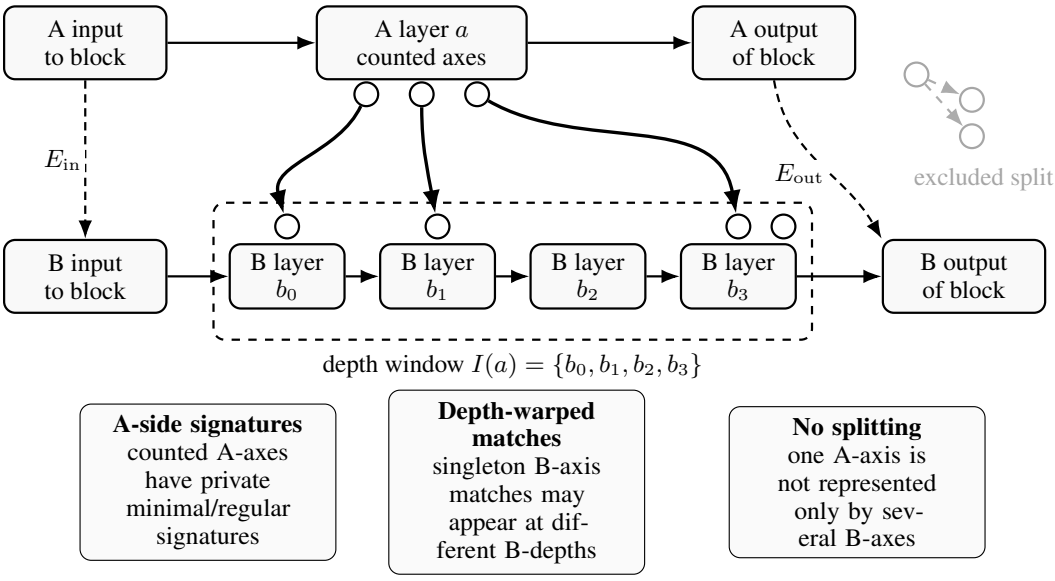


Figure 10: Depth-warped singleton alignment. A layer a in a shallower network is compared to a window $I(a)$ of layers in a deeper network. The private signatures are the same minimal/regular signatures used earlier: ReLU facets or Softplus pole/monodromy patches. The new feature is that the matched B-side axes may be spread over several layers in the window. The singleton theorem rules out the split alternative, where one A-axis is represented only by a packet of several B-axes.

across several nearby layers of B. This is the formal version of the intuition that a deeper model, such as a large ResNet, may smear the preferred axes of a shallower AlexNet-like or ResNet-18-like model across a longer block.

Definition 23 (Depth warp and depth-warped axis alignment)**.** Let $\Omega_0 \subseteq \Omega$ denote the task-input subset on which we compare the two network blocks. Let A have a block of layers $a_0, \dots, a_1$, and let B have a block of layers $b_0, \dots, b_1$. A *depth warp* assigns to each compared A-layer a a nonempty window

$$I(a) \subseteq \{b_0, \dots, b_1\}.$$

For each such a, let $J_a^A \subseteq \{1, \dots, d_a^A\}$ be the counted A-side axes, and let

$$J_{I(a)}^B = \{(b, i) : b \in I(a),\ 1 \le i \le d_b^B\}$$

be the B-side axes available in the corresponding depth window. The *exact depth-warped axis-alignment score* is

$$\mathrm{DWAxisAlign}_{a,I(a),\varphi}(A, B) = \frac{1}{d_a^A} \max_{\pi} \#\left\{ j \in J_a^A : z_{\pi(j)}^B = \alpha_j z_{a,j}^A \text{ on } \Omega_0 \text{ for some } \alpha_j \in \mathcal{H}_\varphi \right\},$$

where π ranges over injective maps $J_a^A \hookrightarrow J_{I(a)}^B$, and if $\pi(j) = (b, i)$ we write $z_{\pi(j)}^B = z_{b,i}^B$.

A depth window is useful only if it is the right window. We will call $I(a)$ *signature-complete* for the counted A-axes if the B-side signatures corresponding to those A-axes occur inside $J_{I(a)}^B$. If $I(a)$ is the whole B-block this is automatic; for a smaller window, it is the assertion that the proposed depth warp has selected the correct local range of B-depths.

The split obstruction. The only genuinely new obstruction in this subsection is *splitting*. The terminology is the same as in the ReLU MFR genericity argument. There, a split match occurs when different facets of one source argument are explained by different competitor arguments, so no single competitor recovers the whole multi-facet signature. In the depth-warp setting, the same obstruction can occur across depth: different pieces of one source axis's block signature may be explained by different B-axes in the window $I(a)$. If that happens, the correct conclusion is not singleton axis alignment but packet alignment. The exact theorem below assumes this packet alternative is absent; the genericity proposition then gives a rank/transversality condition under which the split obstruction is lower-dimensional.

Definition 24 (Depth-warped signature pieces and split assignments)**.** Fix a depth warp $a \mapsto I(a)$. For each counted A-axis $j \in J_a^A$, choose a finite "signature record"

$$\Sigma_j^A = (\Sigma_{j,1}^A, \dots, \Sigma_{j,L_j}^A),$$

i.e. just a list of the ReLU fresh-facet pieces, or Softplus pole/monodromy pieces, defining the j axis. For a B-axis $q = (b, i) \in J_{I(a)}^B$, say that *$q$ matches the piece* $\Sigma_{j,\nu}^A$ if the corresponding B-side piece agrees with $\Sigma_{j,\nu}^A$ in the fixed ReLU or Softplus signature coordinates. A *split assignment* for j is a map

$$\tau : \{1, \dots, L_j\} \to J_{I(a)}^B$$

whose image contains at least two B-axes. The split assignment is *realized* if, for every ν, the B-axis $\tau(\nu)$ matches the piece $\Sigma_{j,\nu}^A$.

This definition is deliberately parallel to the ReLU split-matching definition in Proposition 7. In that proposition, the source signature is written as

$$\Sigma_r = (\Sigma_{r,1}, \dots, \Sigma_{r,L}),$$

and a split assignment sends different components $\Sigma_{r,\nu}$ to different competitors. The depth-warped definition is the same construction, except that the competitors may live at different depths inside $I(a)$.

Now we can state the depth-warped analog of the weak-strong alignment theorem.

Theorem 10 (Exact depth-warped weak–strong alignment)**.** *Let A and B be affine-φ networks, with $\varphi \in \{\sigma, \rho\}$. Let*

$$\Psi_A = \Phi_{a_1-1}^A \circ \cdots \circ \Phi_{a_0}^A, \qquad \Psi_B = \Phi_{b_1-1}^B \circ \cdots \circ \Phi_{b_0}^B$$

be blocks of possibly different depths. Suppose there are injective affine maps

$$E_{\text{in}} : \mathbb{R}^{d_{a_0}^A} \to \mathbb{R}^{d_{b_0}^B}, \qquad E_{\text{out}} : \mathbb{R}^{d_{a_1}^A} \to \mathbb{R}^{d_{b_1}^B},$$

such that the block endpoints are exactly weakly aligned:

$$\Psi_B(E_{\text{in}} z) = E_{\text{out}} \Psi_A(z)$$

for all z in the input task patch of the A-block. Assume (i) that the usual minimality-regularity conditions from earlier (non-depth-warped) identifiability results, and (ii) no split assignment from Definition 24 is realized. Then there is an injective matching $\pi : J_a^A \hookrightarrow J_{I(a)}^B$ such that, for every $j \in J_a^A$, if $\pi(j) = (b, i)$, then $z_{b,i}^B = \alpha_j z_{a,j}^A$ for some $\alpha_j \in \mathcal{H}_\varphi$. Consequently,

$$\text{DWAxisAlign}_{a,I(a),\varphi}(A, B) \geq \frac{|J_a^A|}{d_a^A}.$$

Proof. Fix $j \in J_a^A$. By assumption (i), the A-axis j has a private visible nonlinear signature record: a Softplus pole/monodromy record or a ReLU fresh-facet/multi-facet record. Endpoint weak equality says that the two block maps represent the same function after applying E_{in} and E_{out}. Since E_{out} is injective, it cannot erase a nonzero kink jump, monodromy vector, or marked readout. Hence every marked piece of the A-signature must be matched somewhere on the B-side. Since the chosen depth window is signature-complete, those matched pieces lie inside $J_{I(a)}^B$.

Choose, for each marked piece $\Sigma_{j,\nu}^A$, one B-axis in the window that matches it. This gives a map

$$\tau_j : \{1, \dots, L_j\} \to J_{I(a)}^B.$$

If the image of τ_j contained two or more B-axes, then τ_j would be a realized split assignment. This is excluded by assumption (ii). Therefore all marked pieces of the private signature record are matched by one and the same B-axis. Call this axis

$$\pi(j) = (b, i).$$

Now apply the activation-specific regularity assumption (i) to this singleton match. In the Softplus case, MLR identifies the matched scalar argument up to the Softplus sign ambiguity, with sign-noncancellation selecting the admissible orientation. In the ReLU case, MFR identifies the matched CPWA argument up to the ReLU scalar symmetry. Thus

$$z_{\pi(j)}^B = \alpha_j z_{a,j}^A \quad \text{on } \Omega_0, \qquad \alpha_j \in \mathcal{H}_\varphi.$$

It remains only to check that $j \mapsto \pi(j)$ is injective; this is not a separate no-splitting assumption, but a consequence of same-side separation. Suppose, toward a contradiction, that two distinct counted A-axes $j_1 \neq j_2$ are assigned to the same B-axis q. Then q has the full private signature record of both j_1 and j_2. By the singleton identifiability just used,

$$z_q^B = \alpha_1 z_{a,j_1}^A \quad \text{and} \quad z_q^B = \alpha_2 z_{a,j_2}^A \qquad \text{on } \Omega_0,$$

with $\alpha_1, \alpha_2 \in \mathcal{H}_\varphi$. Hence z_{a,j_1}^A and z_{a,j_2}^A have the same full signature record up to the allowed scalar symmetry, contradicting the same-side separation part of assumption (i). Therefore π is injective.

Every counted A-axis contributes to the numerator of $\mathrm{DWAxisAlign}_{a,I(a),\varphi}$, and the displayed lower bound follows after dividing by d_a^A. □

Proposition 8 (Split matching is generically lower-dimensional in a depth warp)**.** *Suppose that every possible nontrivial split assignment has a split-rank witness: after allowing the B-side packet variables to vary, the equations expressing that split impose at least one independent constraint on the source A-axis signature. Then the set of source A-axis parameters that admit a split match in $I(a)$ is lower-dimensional and measure zero in the source chart.*

Proof. We now spell out the rank condition in coordinates. Let Θ_A be the local chart for a source A-axis signature, and let $p_A = \dim \Theta_A$. Fix a nontrivial split assignment

$$\tau : \{1, \dots, L_j\} \to J_{I(a)}^B$$

whose image contains $q \geq 2$ distinct B-axes. Let Θ_B^τ be the local chart for those q competitor B-axes, and put $p_B^\tau = \dim \Theta_B^\tau$. The split assignment imposes a finite system of incidence equations

$$H_\tau(\xi, \eta) = 0, \qquad \xi \in \Theta_A, \quad \eta \in \Theta_B^\tau,$$

saying that each marked piece of the A-signature is matched by the B-axis assigned to it by τ. The variables η are nuisance variables: they describe how the competitor B-axes are allowed to move while trying to fit the pieces of the fixed A-signature.

The split-rank witness assumption means that every connected smooth stratum of $H_\tau^{-1}(0)$ contains a point where

$$\operatorname{rank} DH_\tau \geq p_B^\tau + 1.$$

The domain of H_τ has dimension $p_A + p_B^\tau$. Let S_τ be a connected smooth stratum of $H_\tau^{-1}(0)$. At a rank-witness point,

$$TS_\tau \subseteq \ker DH_\tau,$$

so rank-nullity gives

$$\dim S_\tau \leq (p_A + p_B^\tau) - (p_B^\tau + 1) = p_A - 1.$$

Because S_τ is a smooth stratum, its dimension is constant; hence the whole stratum has dimension at most $p_A - 1$.

Project the split-incidence stratum to the source A-axis chart:

$$\pi_A : S_\tau \to \Theta_A.$$

This projection simply forgets the competitor B-axes and records which A-axis parameters admit that split explanation. Since π_A is locally Lipschitz, it cannot increase Hausdorff dimension:

$$\dim_H \pi_A(S_\tau) \leq \dim S_\tau \leq p_A - 1.$$

There are only finitely many counted A-axes, marked signature pieces, B-axes in the finite window, and split assignments τ. Taking the finite union over all of them preserves the dimension bound. This proves that split matching is lower-dimensional and measure zero.

The inequality $\operatorname{rank} DH_\tau \geq p_B^\tau + 1$ has the same meaning as the split-incidence rank condition in the ReLU MFR proof. Rank p_B^τ could be used merely to solve for the nuisance B-competitors. The extra $+1$ says that the proposed split imposes at least one genuine constraint on the source A-axis, so it cannot occur on a full-dimensional set of source axes. □

If $I(a)$ consists of a single B-layer and the block has one nonlinear step, the depth-warped statement reduces to the ordinary adjacent weak–strong theorem. The activation-specific minimality/regularity condition is exactly the old private-signature condition. The new content is that, for unequal-depth blocks, the axis matches are allowed to land at different depths inside $I(a)$. The theorem says that the preferred axes of a shallower layer may be distributed across a window of a deeper network: different A-axes from the same layer can have their singleton matches in different B-layers. This is the sense in which a deeper model can smear a shallower model's axis set across depth.

This is different from a "packet match". A packet match would mean that one single A-axis is represented only by a collection of several B-axes, with no single B-axis having the whole private signature. The exact singleton theorem assumes that this split packet alternative is absent, and Proposition 8 gives a concrete rank-witness condition under which the split alternative is lower-dimensional. If that rank condition fails, the right conclusion would be a block-packet alignment theorem rather than a singleton axis-alignment theorem.

A useful way to understand the split-rank witness assumption required above is that it says packet explanations are not free parameters of the architecture. A single axis is a "coherent marked object": its local signature pieces all come from one scalar argument and share the same downstream mark. A split packet would have to imitate that coherence using several independently parameterized B-side axes. Critically, in an ordinary MLP, CNN, or ResNet block, different channels have separate filters and separate outgoing readout columns; convolution shares one channel across spatial positions, but it does not force several distinct channels to coordinate as one marked unit. Residual depth creates more locations where a feature can appear, but it also does not impose equations requiring several units at different depths to share the same mark and jointly reproduce one source signature. Thus a split-packet explanation requires extra coincidences: matching the right marked pieces, matching compatible readout directions, and doing so while the packet variables are allowed to move. The rank-witness assumption is the formal expression of this non-forcing claim. It says that, after the B-packet variables have been adjusted as freely as possible, the proposed split still imposes at least one independent constraint on the A-side source signature. That is why the split set becomes lower-dimensional rather than a generic alternative. The assumption would actually be somewhat suspect in architectures that *do* structurally force packet behavior, such as explicitly tied multi-branch modules, grouped factorizations, hard bottlenecks, or other mechanisms that make several units share one effective mark by design. So analysis of such networks would have to done on a case-by-case basis. But at least in "ordinary" independently parameterized architectures, however, packet coherence is an extra incidence constraint, not a built-in symmetry.

C.2 Residual and transformer circuits

Here we present a version of the Weak–Strong Equivalence and Zippering theorems for transformer architectures. Because transformers are residual circuits, along the way we will end up also resolving the weak–strong equivalence relationship for residual blocks more generally.

This section proceeds in four steps. First, we identify the "strong objects" in the MLP stream (the usual privileged axes) and attention stream ("gauge-quotiented heads"). We then explain why the native coordinates of the residual stream do not obey an analogous weak–strong theorem. Next, we introduce a technique for handling general residual blocks ("peeling") and apply it twice within each Transformer block to obtain weak-alignment zippering of the residual stream. Finally, we show that the additional failures specific to residual peeling occur only on lower-dimensional exceptional sets under the same kind of witness assumptions used earlier.

C.2.1 Strong objects in Transformer branches

Transformer block notation. For a Transformer network $N \in \{A, B\}$, write $x_\ell^N(\omega) \in \mathbb{R}^{d_\ell^N}$ for the residual stream at block ℓ on a task input $\omega \in \Omega$. In a pre-normalization block, write

$$u_{\ell,\mathrm{attn}}^N = \mathrm{LN}_{\ell,\mathrm{attn}}^N(x_\ell^N), \qquad y_\ell^N = x_\ell^N + \mathcal{A}_\ell^N(u_{\ell,\mathrm{attn}}^N),$$

and

$$u_{\ell,\mathrm{mlp}}^N = \mathrm{LN}_{\ell,\mathrm{mlp}}^N(y_\ell^N), \qquad x_{\ell+1}^N = y_\ell^N + \mathcal{M}_\ell^N(u_{\ell,\mathrm{mlp}}^N).$$

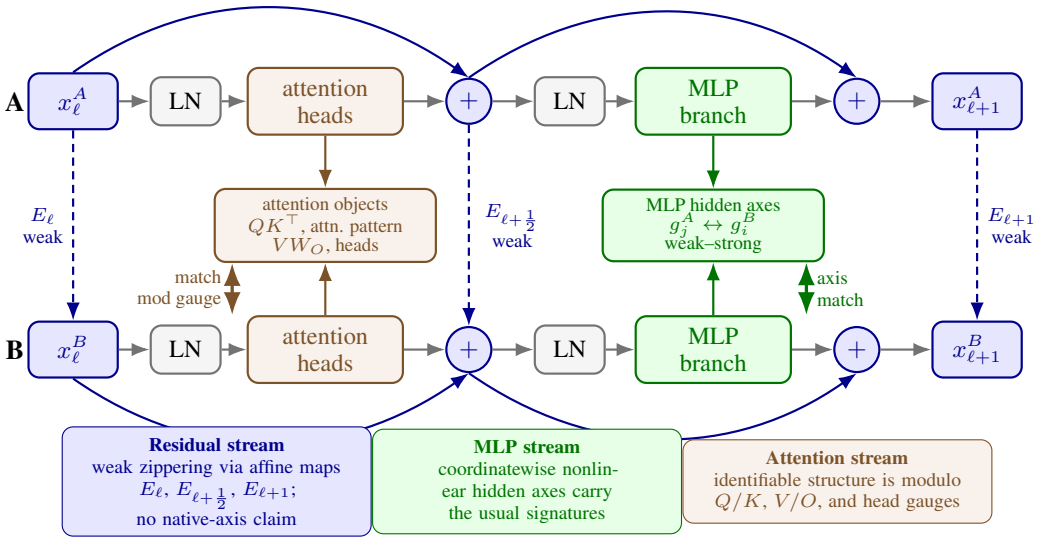


Figure 11: Transformer weak–strong structure under residual peeling. Residual states zipper weakly across depth via affine maps, but this does not privilege the native residual coordinates. Inside each block, the MLP branch has ordinary coordinatewise-nonlinear hidden axes, while attention has task-visible heads and head signatures modulo its exact internal gauges. Peeling the MLP and attention sublayers recovers the intermediate weak maps.

The MLP branch has the one-step form

$$\mathcal{M}_\ell^N(u) = b_{\ell,\mathrm{out}}^N + \sum_j v_{\ell,j}^N \psi(g_{\ell,j}^N(u)), \qquad g_{\ell,j}^N(u) = \langle w_{\ell,j}^N, u\rangle + b_{\ell,j}^N,$$

where $\psi \in \{\rho, \sigma\}$ is applied coordinatewise.

A single attention head maps a token matrix X through

$$Q_h = XW_{Q,h}, \qquad K_h = XW_{K,h}, \qquad V_h = XW_{V,h},$$

and then produces

$$L_h(X) = \frac{Q_h K_h^\top}{\sqrt{d_h}}, \qquad P_h(X) = \mathrm{softmax}(L_h(X)), \qquad O_h(X) = P_h(X)V_h W_{O,h}.$$

We summarize the behavior of head h on an input X by the pair

$$\Sigma_h(X) := (P_h(X), O_h(X)),$$

consisting of its attention pattern and its output. One may also keep the logit matrix L_h, with two logit matrices treated as equivalent when they differ by adding a constant to every entry of a row, since such rowwise shifts do not change softmax. For the statements below, however, the attention pattern and output are the quantities that matter. Biases or affine operations on the branch input can be included by appending a constant coordinate to each token vector, so no generality is lost by using the simpler linear notation.

Attention gauges. Attention has exact internal symmetries, which prevent its raw internal coordinates from being strong objects. We therefore quotient these symmetries out and compare heads only up to the function-preserving changes of coordinates below. Let $R, S \in GL_{d_h}(\mathbb{R})$, and define

$$W_Q' = W_Q R, \qquad W_K' = W_K R^{-\top}, \qquad W_V' = W_V S, \qquad W_O' = S^{-1} W_O.$$

Proposition 9 (Attention coordinate gauges)**.** *The primed and unprimed parameter choices compute exactly the same attention head for every input X.*

Proof. The transformed query and key matrices are

$$Q' = XW_QR = QR, \qquad K' = XW_KR^{-\top} = KR^{-\top}.$$

Thus

$$Q'(K')^\top = QRR^{-1}K^\top = QK^\top,$$

so the logits and attention weights are unchanged. Also

$$V'W'_O = XW_VSS^{-1}W_O = VW_O.$$

Hence the head output is unchanged for every X. □

Two attention heads are regarded as the same head, for purposes of strong comparison, if they are related by these Q/K and V/O gauges, together with the usual permutation of head labels. Thus the raw coordinates of Q, K, V are not identifiable. What can be identified is the head's attention pattern and output, modulo gauge.

Corollary 6 (No raw-coordinate weak–strong theorem for attention)**.** *Weak alignment before and after an attention head cannot, by itself, identify raw query, key, or value coordinate axes. At most it can identify quantities unchanged by the gauges: the attention pattern, the logit matrix up to rowwise softmax shifts, the value–output map on the task inputs, and head identity up to permutation.*

Proof. Proposition 9 gives a continuum of different raw Q, K, V coordinate systems with exactly the same represented attention function. A generic dense choice of R or S mixes the raw coordinates, so the transformed coordinates are not permutations or scalar rescalings of the original coordinates. Thus raw attention-coordinate alignment is not identifiable from the represented function. Any strong theorem for attention must be stated modulo the gauges. □

Definition 25 (Usedness, budgets, and alignment scores)**.** The branch-specific components that can be strongly aligned are:

(i) *MLP hidden axes $g_{\ell,j}$.* *Usedness* here is exactly in the sense of the earlier adjacent ReLU/Softplus weak–strong theorem: ReLU-type axes have crossed kink traces that are visible in the branch output and a nonzero readout, while Softplus-type axes have nonconstant, nonredundant curvature ridges and a nonzero readout. Let $m^{\psi}_{\mathrm{mlp},\ell}(\epsilon)$ be the minimum number of used MLP hidden axes at branch ℓ needed, within the fixed Transformer macroarchitecture, to solve the task to loss at most ϵ. The corresponding denominator is the A-side MLP hidden width $d^A_{\mathrm{mlp},\ell}$.

(ii) *Attention heads h.* A head is *used* if the pair $\Sigma_h = (P_h, O_h)$ varies on the task set and its output contribution is not identically zero after the output readout. A used head is minimal if there is some task region on which its attention pattern and output are not shared by any other counted head, and if the head cannot be deleted, combined with the branch's linear map and bias, or merged with another counted head without changing the branch function on the task set. Let $m_{\mathrm{attn},\ell}(\epsilon)$ be the minimum number of used attention heads at block ℓ needed, within the fixed Transformer macroarchitecture, to solve the task to loss at most ϵ. The corresponding denominator is the A-side number of heads H^A_ℓ.

For MLP branches, $\mathrm{AxisAlign}_{\mathrm{mlp},\ell,\psi}(A, B)$ is the usual axis-alignment score applied at the hidden MLP preactivation axes. For attention branches, define

$$\mathrm{HeadAlign}_{\mathrm{attn},\ell}(A, B) = \frac{1}{H^A_\ell} \max_{\pi} \#\{h : \Sigma^A_h \text{ matches } \Sigma^B_{\pi(h)} \text{ on } \Omega, \text{ modulo gauge}\},$$

where π ranges over injective maps from A-side heads to B-side heads. For attention, set

$$\mathrm{AxisAlign}_{\mathrm{attn},\ell}(A, B) := \mathrm{HeadAlign}_{\mathrm{attn},\ell}(A, B).$$

A term that is affine in the branch input X can be combined with the branch's existing linear map and bias, so it need not be counted as a separate nonlinear component. Whether that same term can also be combined with the residual skip is a different question, taken up in §C.2.3.

With these definitions in mind, we can state the usual shrinkability result needed to justify the usedness definitions:

Lemma 15 (Shrinkability of non-used Transformer axes and heads)**.** *The following reductions do not increase the number of counted MLP axes or attention heads.*

(i) Non-used MLP hidden axes may be deleted, merged, or reduced by the earlier ReLU/Softplus shrinkability lemmas. Any affine remainder produced by those reductions can be combined with the branch's existing linear map and bias.

(ii) An attention head with $O_h \equiv 0$ on the task set may be deleted.

(iii) If P_h is constant on the task set, then the head output is linear in the branch input. It may therefore be combined with the branch's existing linear map and bias.

(iv) Two heads with the same attention pattern may be merged when the sum of their value-output products can be represented by one allowed head.

Thus the branch budgets count MLP axes and attention heads whose nonlinear behavior is observable on the task inputs, rather than merely counting physically present coordinates or heads.

Proof. For MLP axes, the first statement is exactly the earlier shrinkability calculation. Zero coefficients delete. Constant arguments contribute constants. Duplicate arguments merge. For either $\psi = \rho$ or $\psi = \sigma$, an opposite pair obeys

$$v_+\psi(g) + v_-\psi(-g) = (v_+ + v_-)\psi(g) - v_-g,$$

because $\psi(g) - \psi(-g) = g$. In the adjacent branch theorem, g is affine in the common branch input, so the last term is part of the branch's linear map and bias.

If $O_h \equiv 0$, the head contributes nothing and can be deleted. If $P_h = P$ is constant, then

$$O_h(X) = PXW_{V,h}W_{O,h},$$

which is linear in X. Finally, if two heads have the same pattern P, their combined contribution is

$$PXW_{V,1}W_{O,1} + PXW_{V,2}W_{O,2} = PX\big(W_{V,1}W_{O,1} + W_{V,2}W_{O,2}\big).$$

This can be represented by one head with the same pattern whenever the matrix in parentheses has rank no larger than the head's allowed internal dimension. These operations prove all four claims. □

Definition 26 (Branch-specific weak–strong hypotheses)**.** For an MLP branch, the weak–strong hypotheses are exactly the activation-specific adjacent usedness hypotheses from the earlier ReLU/Softplus weak–strong theorem. For an attention branch, the weak–strong hypotheses are:

(A1) *head usedness/minimality:* every counted A-side head, and every B-side head after its input has been composed with the weak input map, is used and minimal; in particular, each has some attention-pattern-and-output behavior on the task set that is not shared by another counted head on the same side;

(A2) *one-head identification:* once one A-side head and one B-side head have been separated from their respective multihead sums and shown to have the same behavior on the task inputs, they are the same head modulo the gauges of Proposition 9 and head permutation;

(A3) *no splitting:* the distinguishing behavior of each source head is reproduced by one head on the other side, not only by several heads acting together.

The no-splitting condition is needed because a head has two linked pieces of observable behavior—its attention pattern and its output—and several heads could in principle reproduce those pieces jointly. This is not a new assumption for MLP branches: the earlier adjacent theorem already handles the corresponding issue through usedness, reducedness, and the identifying nonlinear behavior of each axis.

We next give a concrete and easily checked sufficient condition for (A2). Define the two product matrices of a head by

$$\Gamma_h := \frac{W_{Q,h}W_{K,h}^\top}{\sqrt{d_h}}, \qquad C_h := W_{V,h}W_{O,h}.$$

Then

$$P_h(X) = \mathrm{softmax}(X\Gamma_h X^\top), \qquad O_h(X) = P_h(X)XC_h.$$

Choose finitely many task inputs $\mathcal{X} = \{X^1, \dots, X^R\}$, and write x_i^r for row i of X^r. For a matrix D, set

$$\mathcal{Q}_{\mathcal{X}}(D) := \big((x_i^r)^\top D(x_j^r - x_k^r)\big)_{r,i,j,k},$$

and, for fixed Γ,

$$\mathcal{V}_{\Gamma,\mathcal{X}}(D) := \big(P_\Gamma(X^r)X^r D\big)_{r=1}^R.$$

The first map records every logit difference that rowwise softmax can detect; the second records how changes in the value–output product affect the head output. A component of Γ or C is retained here exactly when changing it can change the attention pattern or output on at least one of the selected task inputs. We say that the selected task inputs contain enough variation to determine the head when the two maps have full column rank on these retained components.

Lemma 16 (A concrete sufficient condition for identifying one head)**.** *Suppose two heads have already been separated from their respective multihead sums using (A1) and (A3), and suppose their attention patterns and outputs agree on $\mathcal{X}$. Then*

$$\Gamma - \widetilde{\Gamma} \in \ker \mathcal{Q}_{\mathcal{X}}, \qquad C - \widetilde{C} \in \ker \mathcal{V}_{\Gamma,\mathcal{X}}.$$

Consequently, if the selected task inputs contain enough variation in the sense just defined, the two heads have the same components of Γ and C that affect their behavior on those inputs. If, in addition, no unused internal head coordinates remain—equivalently, after restriction to these task-relevant products, the query–key and value–output factorizations use latent dimensions equal to the ranks of the corresponding products—then the individual matrices W_Q, W_K, W_V, W_O are related exactly by the Q/K and V/O gauges of Proposition 9.

Proof. We prove the query–key claim, the value–output claim, and then the claim about the individual matrices.

For two row vectors a, b, equality $\mathrm{softmax}(a) = \mathrm{softmax}(b)$ implies

$$e^{a_j - a_k} = \frac{e^{a_j}}{e^{a_k}} = \frac{e^{b_j}}{e^{b_k}} = e^{b_j - b_k}$$

for every j, k. Hence $a_j - a_k = b_j - b_k$, so $a - b$ is constant along the row. Apply this row by row to

$$P_\Gamma(X^r) = P_{\widetilde{\Gamma}}(X^r).$$

Writing $D = \Gamma - \widetilde{\Gamma}$, we obtain

$$(x_i^r)^\top D(x_j^r - x_k^r) = 0$$

for all r, i, j, k. This is exactly $D \in \ker \mathcal{Q}_{\mathcal{X}}$.

The attention patterns now agree on every selected input. Equality of the two head outputs therefore gives

$$P_\Gamma(X^r)X^r(C - \widetilde{C}) = 0 \qquad (r = 1, \dots, R),$$

which is precisely $C - \widetilde{C} \in \ker \mathcal{V}_{\Gamma,\mathcal{X}}$. Full column rank on the retained components forces both differences to vanish there.

It remains to identify the individual matrices once unused internal coordinates have been removed. Suppose

$$W_Q W_K^\top = \widetilde{W}_Q \widetilde{W}_K^\top$$

and both product representations use a latent dimension equal to the rank of this common product. Then W_Q and $\widetilde{W}_Q$ have full column rank and the same column space, namely the column space of the product. Hence $\widetilde{W}_Q = W_Q R$ for some invertible R. Let L_Q be a left inverse of W_Q. Substitution gives

$$W_Q W_K^\top = W_Q R \widetilde{W}_K^\top, \qquad W_K^\top = R\widetilde{W}_K^\top,$$

and therefore $\widetilde{W}_K = W_K R^{-\top}$. The same argument applied to $W_V W_O = \widetilde{W}_V \widetilde{W}_O$ gives

$$\widetilde{W}_V = W_V S, \qquad \widetilde{W}_O = S^{-1} W_O$$

for an invertible S. These are exactly the two attention gauges. □

These results now allow us to state the core branch-level Weak–Strong Equivalence result for transformers:

Theorem 11 (Transformer branch weak–strong equivalence)**.** *Consider either an MLP branch or an attention branch of a Transformer block. Suppose exact weak alignment holds at the branch input and branch output. For an MLP branch, assume the MLP one-step expansions satisfy the activation-specific adjacent usedness hypotheses. For an attention branch, assume (A1)–(A3) of Definition 26. Then every used A-side MLP hidden axis or attention head has a distinct B-side match: MLP hidden axes match as scalar axes, and attention heads match modulo attention gauge. Consequently,*

$$\mathrm{AxisAlign}_{\mathrm{mlp},\ell,\psi}(A,B) \geq \frac{m^{\psi}_{\mathrm{mlp},\ell}(\epsilon)}{d^{A}_{\mathrm{mlp},\ell}}$$

for MLP branches, and

$$\mathrm{AxisAlign}_{\mathrm{attn},\ell}(A,B) = \mathrm{HeadAlign}_{\mathrm{attn},\ell}(A,B) \geq \frac{m_{\mathrm{attn},\ell}(\epsilon)}{H^{A}_{\ell}}$$

for attention branches.

Proof. For an MLP branch, write the adjacent weak-alignment equations as

$$u^B = E_{\mathrm{in}} u^A, \qquad \mathcal{M}^B(u^B) = E_{\mathrm{out}} \mathcal{M}^A(u^A).$$

Substitute the first equality into the second. The B-side branch becomes

$$b^B_{\mathrm{out}} + \sum_i v^B_i \psi(g^B_i(E_{\mathrm{in}} u^A)),$$

while the transformed A-side branch is

$$E_{\mathrm{out}} b^A_{\mathrm{out}} + \sum_j (E_{\mathrm{out}} v^A_j) \psi(g^A_j(u^A)).$$

These are two one-step affine–coordinatewise-nonlinear expansions of the same common input variable u^A. The earlier adjacent weak–strong theorem therefore gives an injective match of every used A-side MLP hidden axis to a pulled-back B-side axis, with the activation's allowed scalar symmetry. Counting the task-required used axes proves the MLP bound.

For attention, first compose the B-side branch input with the weak input map so that both branches are functions of the same task variable, and apply the weak output map to the A-side branch output so that the two outputs are written in the same coordinates. We then have equality of two multihead branch functions. By (A1), each counted A-side head has some attention-pattern-and-output behavior not shared by another counted A-side head, so that behavior must also appear on the B side. By (A3), one B-side head reproduces the complete behavior, rather than several heads reproducing different parts jointly. Condition (A2) then identifies that head modulo the exact attention gauges and head permutation. Lemma 16 gives a concrete sufficient rank test for this step.

The assignment is injective. If two distinct A-side heads were assigned to the same B-side head, condition (A2) would identify both A-side heads with that same B-side head. Their attention patterns and outputs would therefore agree, contrary to the distinguishing behavior required by (A1). Thus every task-required A-side head has a distinct B-side match. Counting those heads gives the displayed attention bound. □

As elsewhere in the appendix, we next ask whether (A1)–(A3) are generic. The corresponding MLP conditions are already generic by §B.3. For attention, we use the same witness principle used there. Here a *witness* is one parameter choice at which the relevant nonvanishing or rank condition holds. For (A3), the split-rank witness additionally says that, even after the competitor heads are allowed to vary, a proposed split still imposes at least one independent condition on the source head.

Proposition 10 (Genericity of the attention hypotheses)**.** *Fix a finite-dimensional real-analytic family of attention-branch comparisons, together with the finite task measurements used to test (A1)–(A3). If the corresponding witnesses exist, including the split-rank witness of Proposition 8 for every proposed split, then (A1)–(A3) hold generically: the set on which any of them fails is lower-dimensional, and hence has empty interior and measure zero.*

Proof. For (A1), fix the finite task regions or evaluations used in the comparison. A zero head output, loss of variation, or equality between two heads that were supposed to differ is again expressed by finitely many analytic equations. The assumed witness shows that these equations do not vanish identically.

For (A2), failure of either map $\mathcal{Q}_{\mathcal{X}}$ or $\mathcal{V}_{\Gamma,\mathcal{X}}$ to have full rank is detected by the vanishing of all of its largest relevant minors, meaning the square subdeterminants that certify the desired rank. Failure of the reduced query–key or value–output factors to have full rank is likewise detected by vanishing minors. Whenever the required rank occurs at one point, these determinant equations define a proper lower-dimensional subset.

It remains to handle splitting (A3). Fix one proposed way, denoted by τ, for several B-side heads to reproduce one source head. Let ξ be local coordinates for the source head, of dimension p, and let η collect the parameters of the B-side heads used in the proposed split, of dimension n_τ. The requirement that those heads reproduce the selected attention-pattern-and-output behavior of the source head can be written as finitely many equations

$$H_\tau(\xi, \eta) = 0.$$

Consider any connected smooth piece S_τ of the solution set of these equations. By the split-rank witness, that piece contains a point where

$$\operatorname{rank} DH_\tau \geq n_\tau + 1.$$

At such a point its tangent space lies in $\ker DH_\tau$, so rank–nullity gives

$$\dim S_\tau \leq (p + n_\tau) - (n_\tau + 1) = p - 1.$$

Forgetting the B-side parameters and retaining only ξ cannot increase Hausdorff dimension. At fixed width there are only finitely many proposed splits, and the solution sets can be decomposed into smooth pieces with only finitely many meeting any small neighborhood. Their union therefore remains lower-dimensional and measure zero. Combining the three kinds of failure proves the proposition. □

This genericity result does not settle the separate question of where learning dynamics place trained networks, but it shows that the branch conditions fail only for specially constrained parameter choices.

C.2.2 Why residual coordinates need not be privileged

The branch theorem identifies privileged units and heads within branches. Exact changes of residual basis can mix native residual coordinates while preserving the represented network.

Proposition 11 (Residual weak alignment does not imply residual-coordinate alignment)**.** *Consider an idealized class of residual-stream networks closed under invertible residual-basis changes. Thus, for every invertible $R \in GL_d(\mathbb{R})$ and every network A with residual states x_ℓ^A, there is an equivalent network B with residual states*

$$x_\ell^B = R x_\ell^A$$

at every layer, obtained by conjugating all residual-stream linear maps by R. Suppose the coordinate functions of x_ℓ^A are linearly independent on the task set. For a generic dense R, the two networks are exactly weakly aligned at layer ℓ, but have no native residual-coordinate strong alignment at that layer.

Proof. The equation $x_\ell^B = R x_\ell^A$ is exact weak alignment with comparison map $E_\ell = R$. If a B-side residual coordinate were strongly aligned with an A-side residual coordinate, then for some row $r_i^\top$ of R, some coordinate j, and some nonzero scalar α,

$$r_i^\top x_\ell^A(\omega) = \alpha x_{\ell,j}^A(\omega) \qquad (\omega \in \Omega).$$

Linear independence of the residual-coordinate functions forces

$$r_i^\top = \alpha e_j^\top.$$

A generic dense invertible matrix has no row proportional to a coordinate vector. Hence exact weak alignment can hold with no native residual-coordinate strong alignment. □

Thus the appropriate residual-stream conclusion is weak affine alignment, not native-coordinate alignment. We now explain how such a weak map can nevertheless be recovered backward through a residual block.

C.2.3 RESIDUAL BLOCKS AND THE PEELING PRINCIPLE

Turning our attention to zippering, we now ask how weak alignment can be propagated one step backward through a residual sublayer. The earlier branch results identify the nonlinear units or heads inside a branch. A residual sublayer adds an affine skip path, and that skip provides the route for recovering the sublayer input once the branch contribution has been removed. We call this procedure *residual peeling*. This subsection proves the result for one abstract residual sublayer; the next subsection applies it first to the MLP sublayer and then to the attention sublayer of each Transformer block.

For one network $N \in \{A, B\}$, write a residual sublayer as

$$\mathcal{F}^N(z) = S^N z + c^N + F^N(z), \tag{20}$$

where $S^N z + c^N$ is the affine skip and F^N is the nonlinear branch. Suppose an injective affine map $E_+(r) = T_+ r + a_+$ already aligns the two sublayer outputs:

$$\mathcal{F}^B(z^B(\omega)) = E_+ \mathcal{F}^A(z^A(\omega)). \tag{21}$$

The goal of one backward zippering step is to construct an injective affine map E_- satisfying $z^B = E_- z^A$ on the task set. The skip suggests how to do this: identify which branch terms match, subtract those terms from the output equality, and then solve the remaining skip equation for z^B.

The point requiring care is the subtraction step. The branch-identification results may match an MLP argument with the same or the opposite orientation. Same-orientation terms cancel directly. An opposite-orientation match need not: both ReLU and Softplus satisfy $\psi(t) - \psi(-t) = t$, so subtracting the two matched nonlinear terms can leave a term proportional to the branch argument. In an ordinary residual block, where the argument is affine in the residual input, $g(z) = a^\top z + b$, this causes no new nonlinear difficulty:

$$Sz + c + v\psi(g(z)) = \big(S + va^\top\big)z + (c + vb) + v\psi(-g(z)). \tag{22}$$

The remaining term $vg(z)$ is affine in z, so it merely changes the affine skip. After that change, one can still solve an affine equation for the input, provided the resulting map is injective.

A pre-LayerNorm branch is different. Its scalar argument has the form $g(\mathrm{LN}(z))$, which is generally not affine in z. An opposite-orientation match can therefore leave a function of $\mathrm{LN}(z)$ that cannot be included in an affine map of the residual state. The following example shows that this can genuinely prevent the aligned outputs from determining affinely related inputs.

Example (LayerNorm can prevent direct recovery of the residual input). Let $\psi \in \{\rho, \sigma\}$, so that $\psi(t) - \psi(-t) = t$. Use two residual features and standard LayerNorm with $\epsilon > 0$, unit gain, and zero learned bias. On the task interval $-1 < t < 1$, let

$$y(t) = (t, -t), \qquad \mathrm{LN}(y(t)) = \frac{(t, -t)}{\sqrt{t^2 + \epsilon}}.$$

Define

$$g(y) := e_1^\top \mathrm{LN}(y), \qquad f(t) := g(y(t)) = \frac{t}{\sqrt{t^2 + \epsilon}},$$

and choose a nonzero readout vector in the all-ones direction, which LayerNorm does not detect:

$$v = c(1, 1), \qquad c \neq 0.$$

The A-side residual sublayer is

$$F_A(y) = y + v\psi(g(y)).$$

Now change the B-side unnormalized residual input by the input-dependent shift

$$\widetilde{y}(t) := y(t) + vg(y(t)) = y(t) + vf(t).$$

Because LayerNorm is unchanged by adding a scalar multiple of the all-ones vector, $\mathrm{LN}(\widetilde{y}(t)) = \mathrm{LN}(y(t))$. Give the B-side unit the opposite argument

$$\widetilde{g}(z) := -e_1^\top \mathrm{LN}(z).$$

Then $\widetilde{g}(\widetilde{y}(t)) = -g(y(t))$, and

$$\begin{aligned} F_B(\widetilde{y}(t)) &= \widetilde{y}(t) + v\psi(\widetilde{g}(\widetilde{y}(t))) \\ &= y(t) + vg(y(t)) + v\psi(-g(y(t))) \\ &= y(t) + v\psi(g(y(t))) \\ &= F_A(y(t)). \end{aligned}$$

Nevertheless, no affine map E satisfies $\widetilde{y}(t) = E(y(t))$ on the interval. Every affine function of $y(t) = t(1, -1)$ is affine in t, whereas the all-ones component of $\widetilde{y}(t)$ is $cf(t)$ and

$$f''(t) = -\frac{3\epsilon t}{(t^2 + \epsilon)^{5/2}}$$

is not identically zero. Thus the two residual-sublayer outputs agree exactly even though their unnormalized inputs are not affinely related.

The example shows exactly what must be checked in addition to branch identification. The earlier results can correctly pair the two nonlinear units, but after their contributions are compared, the term $v(g \circ \mathrm{LN})$ remains. A change in the residual input that LayerNorm does not detect can cancel this term through the skip. Consequently, output alignment determines an affine relation between the inputs only when every term remaining after the branch comparison is itself affine in the unnormalized residual state.

We use the following terminology for that requirement. Relative to a residual-state family $z : \Omega \to \mathbb{R}^d$, a term r is *affine in the unnormalized residual state* if

$$r(\omega) = Az(\omega) + c \qquad (\omega \in \Omega) \tag{23}$$

for a fixed matrix A and vector c. Notice that being affine in the branch input $X = \mathrm{LN}(z)$ is not enough: the weak map we seek must be affine in z itself.

For the peeling calculation, we now write the same residual sublayer in a form that keeps this point explicit. Apply the shrinkability reductions of Lemma 15 to delete or merge non-used branch terms. If a term produced by those reductions has the form (23), combine it with the skip. If it is affine only in $\mathrm{LN}(z)$, leave it in the branch. This is merely a way of organizing the same represented function; it ensures that the affine part contains only terms that really are affine in the residual state. Thus write

$$\mathcal{F}^N(z^N) = S^N z^N + c^N + \mathcal{B}^N(z^N), \tag{24}$$

where $S^N z^N + c^N$ contains the skip together with any branch terms already shown to be affine in z^N, and $\mathcal{B}^N$ contains all remaining branch contributions. For an untouched identity skip, $S^N = I$.

We can now state precisely what remains after the branch-identification step. Let the relevant MLP or attention result provide a matching between the A-side and B-side branch objects. Reindex the B-side objects according to that matching. For attention, Proposition 9 also allows us to replace a matched head by a gauge-equivalent parameterization without changing the function it computes. These notational choices let us compare each matched pair directly. Subtract the matched branch contributions from (21), placing any difference that does not cancel into a single term $R(\omega)$. The remaining equality is

$$S^B z^B + c^B = T_+ S^A z^A + T_+ c^A + a_+ + R(\omega). \tag{25}$$

If R is affine in z^A, this is an affine equation for z^B. If R is not affine in z^A, the LayerNorm example shows that no affine input comparison need follow.

There is one further requirement. Solving the affine equation must produce an *injective* map, because weak alignment is defined using an injective affine comparison. This gives the following two conditions.

Definition 27 (Conditions for peeling one residual sublayer)**.** Assume the current branch in (21) has been identified by the relevant earlier MLP or attention result, and let R be the term in (25). The residual sublayer can be peeled when:

(i) R is affine in the unnormalized A-side residual state,

$$R(\omega) = Lz^A(\omega) + d \tag{26}$$

for fixed L, d; and

(ii) S^B is invertible and $T_+S^A + L$ is injective.

The first condition ensures that branch subtraction leaves an affine equation in the residual states. The second ensures that solving this equation gives a valid injective weak comparison map.

Lemma 17 (Peeling one residual sublayer)**.** *Suppose the current branch satisfies the relevant earlier identification hypotheses—Softplus minimality plus MLR, the corresponding ReLU private-zero-set conditions, or attention conditions (A1)–(A3)—and suppose the two conditions of Definition 27 hold. Then there is an injective affine map E_- such that*

$$z^B = E_- z^A$$

on the task set. Every counted MLP axis or attention head identified in the branch comparison is matched as well.

Proof. We first identify the branch terms inside the full residual-sublayer equality. For Softplus, the private-pole and MLR argument applies to the composed scalar functions. For attention, (A1)–(A3) separate each head from the multihead sum and identify it modulo gauge. For ReLU, let g_j be one counted scalar function after LayerNorm and choose a point p on a distinctive smooth piece of $\{g_j = 0\}$, away from the other counted zero sets, with $dg_j(p) \neq 0$. Across this piece,

$$D_+\big(v_j\rho(g_j)\big) - D_-\big(v_j\rho(g_j)\big) = v_j\, dg_j(p) \neq 0.$$

The skip and the other current terms are smooth there, so an equal realization on the other side must have a matching zero-set piece. The same coherence and regularity conditions used earlier collect the matched pieces onto one scalar function and determine it up to sign and positive scale; comparing the derivative jumps determines the corresponding readout coefficient. Applying an injective affine map to the output cannot erase the nonzero jump.

Let π denote the resulting matching, and reindex the B-side terms by π. In the attention case, choose gauge-equivalent representatives for the matched heads; Proposition 9 shows that this does not change their outputs. Matched attention-head outputs then cancel directly.

For an MLP branch, positively oriented matches also cancel directly. To compute what remains from negative orientations, write the output comparison as

$$S^B z^B + c^B + \sum_i v_i^B \psi(h_i) = T_+S^A z^A + T_+c^A + a_+ + \sum_j T_+ v_j^A \psi(g_j). \tag{27}$$

In the Softplus case, $h_{\pi(j)} = \varepsilon_j g_j$ and $v^B_{\pi(j)} = T_+ v_j^A$; in the ReLU case, $h_{\pi(j)} = \varepsilon_j \alpha_j g_j$ and $\alpha_j v^B_{\pi(j)} = T_+ v_j^A$, with $\alpha_j > 0$. Let $S_- := \{j : \varepsilon_j = -1\}$. For $j \in S_-$, the identity $\psi(-g) - \psi(g) = -g$ gives

$$v^B_{\pi(j)}\psi(h_{\pi(j)}) - T_+ v_j^A \psi(g_j) = -T_+ v_j^A g_j,$$

where ReLU positive homogeneity absorbs α_j. Summing over the matched pairs gives

$$S^B z^B + c^B = T_+S^A z^A + T_+c^A + a_+ + R_{S_-}, \qquad R_{S_-} := \sum_{j\in S_-} T_+ v_j^A g_j. \tag{28}$$

Thus in the MLP case the term R from (25) is explicitly R_{S_-}; in the attention case it is whatever remains after the matched head outputs have been subtracted.

By condition (i), this remaining term has the form $R = Lz^A + d$. Substitution into (25) gives

$$S^B z^B + c^B = T_+S^A z^A + T_+c^A + a_+ + Lz^A + d.$$

Since S^B is invertible, solving for z^B yields

$$E_-(z) = (S^B)^{-1}(T_+S^A + L)z + (S^B)^{-1}(T_+c^A + a_+ + d - c^B). \tag{29}$$

Condition (ii) makes the linear part of E_- injective, so E_- is a valid weak comparison map. The branch-object matching was obtained before solving this affine equation and therefore remains in force. □

C.2.4 Peeling and Zippering for Transformers

If we are comparing corresponding blocks in two transformers A and B, every compared block has the two-step form

$$\begin{aligned} y_\ell^N &= x_\ell^N + \mathcal{A}_\ell^N\big(\mathrm{LN}_{\ell,\mathrm{attn}}^N(x_\ell^N)\big)\,, \\ x_{\ell+1}^N &= y_\ell^N + \mathcal{M}_\ell^N\big(\mathrm{LN}_{\ell,\mathrm{mlp}}^N(y_\ell^N)\big)\,, \qquad N \in \{A, B\}. \end{aligned} \tag{30}$$

The forward computation applies attention first and the MLP second: thus, the two-step backward peeling argument we need to do proceeds in the reverse order. Starting from the weak map $E_{\ell+1}$ at the right side of Fig. 11, we first peel the MLP sublayer and obtain the intermediate map $E_{\ell+\frac{1}{2}}$ at y_ℓ; we then peel the attention sublayer and obtain E_ℓ at x_ℓ. These are the three dashed residual-stream maps shown in the figure. Formally:

Lemma 18 (Two-step backward peeling of a Transformer block)**.** *Suppose*

$$x_{\ell+1}^B = E_{\ell+1} x_{\ell+1}^A$$

on the task set. Assume that the MLP sublayer satisfies its activation-specific branch-identification hypotheses and the two peeling conditions of Definition 27, and that the attention sublayer satisfies (A1)–(A3) and the same two peeling conditions. Then there are injective affine maps $E_{\ell+\frac{1}{2}}$ and E_ℓ such that

$$y_\ell^B = E_{\ell+\frac{1}{2}} y_\ell^A, \qquad x_\ell^B = E_\ell x_\ell^A.$$

The first step also matches all counted MLP hidden axes, and the second matches all counted attention heads modulo gauge.

Proof. The proof consists of the two backward peeling steps displayed in Fig. 11.

First step: peel the MLP sublayer. The second line of (30) gives

$$x_{\ell+1}^N = y_\ell^N + \mathcal{M}_\ell^N\big(\mathrm{LN}_{\ell,\mathrm{mlp}}^N(y_\ell^N)\big)\,.$$

Substituting this formula for $N = A, B$ into the assumed comparison at $x_{\ell+1}$ yields

$$y_\ell^B + \mathcal{M}_\ell^B\big(\mathrm{LN}_{\ell,\mathrm{mlp}}^B(y_\ell^B)\big) = E_{\ell+1}\big[y_\ell^A + \mathcal{M}_\ell^A\big(\mathrm{LN}_{\ell,\mathrm{mlp}}^A(y_\ell^A)\big)\big]\,.$$

This is exactly a comparison of two residual sublayers with downstream map $E_{\ell+1}$ and raw residual inputs y_ℓ^A, y_ℓ^B. Lemma 17 first identifies the counted MLP axes inside this equality, then subtracts their matched contributions, and finally solves the remaining affine skip equation. It therefore produces an injective affine map $E_{\ell+\frac{1}{2}}$ satisfying

$$y_\ell^B = E_{\ell+\frac{1}{2}} y_\ell^A.$$

The half-index emphasizes that this map compares the intermediate residual states after attention and before the MLP: it is the middle dashed arrow in Fig. 11.

Second step: peel the attention sublayer. The first line of (30) gives

$$y_\ell^N = x_\ell^N + \mathcal{A}_\ell^N\big(\mathrm{LN}_{\ell,\mathrm{attn}}^N(x_\ell^N)\big)\,.$$

Substituting this formula into the newly obtained comparison at y_ℓ gives

$$x_\ell^B + \mathcal{A}_\ell^B\big(\mathrm{LN}_{\ell,\mathrm{attn}}^B(x_\ell^B)\big) = E_{\ell+\frac{1}{2}}\big[x_\ell^A + \mathcal{A}_\ell^A\big(\mathrm{LN}_{\ell,\mathrm{attn}}^A(x_\ell^A)\big)\big]\,.$$

This is a second residual-sublayer comparison, now with downstream map $E_{\ell+\frac{1}{2}}$ and raw inputs x_ℓ^A, x_ℓ^B. Applying Lemma 17 again identifies and subtracts the attention branch, matches the counted heads modulo gauge, and produces an injective affine map E_ℓ satisfying

$$x_\ell^B = E_\ell x_\ell^A.$$

This is the left dashed arrow in Fig. 11. Lemma 16 gives the more concrete statement that the parts of each matched head product that affect the task inputs agree and, after unused internal coordinates have been removed, the individual head matrices agree modulo gauge.

Notice that neither step tries to commute an affine comparison map through LayerNorm. The branch objects are identified inside the full residual-sublayer equality, and the unnormalized residual input is recovered only after that branch has been removed. □

We can now just apply this two-step backward peeling argument iteratively to achieve weak zippering:

Theorem 12 (Weak zippering for Transformer residual streams)**.** *Let A and B be Transformer networks whose compared blocks have the pre-normalization form* (30)*; thus LayerNorms may occur throughout the networks, but each one is applied before its local attention or MLP branch. Suppose exact weak alignment holds at a terminal residual layer s:*

$$x_s^B = E_s x_s^A$$

for an injective affine map E_s. Assume that, at every MLP residual sublayer from block $s-1$ down to q, the existing activation-specific hypotheses identify the current MLP axes; at every attention residual sublayer, (A1)–(A3) hold; and each residual sublayer satisfies the two peeling conditions of Definition 27. Then residual weak alignment zippers upstream. For every $\ell = q, \dots, s-1$, there are injective affine maps $E_{\ell+\frac{1}{2}}$ and E_ℓ such that

$$y_\ell^B = E_{\ell+\frac{1}{2}} y_\ell^A, \qquad x_\ell^B = E_\ell x_\ell^A \qquad \text{on } \Omega.$$

Proof. We apply Lemma 18 repeatedly, moving from the terminal residual state toward the input. The induction invariant is that, before processing block ℓ, an injective affine map $E_{\ell+1}$ has already been obtained at the block's output:

$$x_{\ell+1}^B = E_{\ell+1} x_{\ell+1}^A.$$

For the first step, take $\ell = s-1$. The required output comparison is the assumed terminal relation $x_s^B = E_s x_s^A$. Apply the block lemma to block $s-1$. Its first peeling step removes the MLP branch and gives

$$y_{s-1}^B = E_{s-\frac{1}{2}} y_{s-1}^A;$$

its second peeling step removes the attention branch and gives

$$x_{s-1}^B = E_{s-1} x_{s-1}^A.$$

Thus one backward block step follows the three dashed maps in Fig. 11 from right to left:

$$E_s \implies E_{s-\frac{1}{2}} \implies E_{s-1}.$$

Now suppose that the induction invariant holds for some block ℓ, with $q \le \ell \le s-2$. The hypotheses of the theorem ensure that both residual sublayers of block ℓ satisfy the assumptions of the block lemma. Applying it first to the MLP sublayer gives the intermediate comparison

$$y_\ell^B = E_{\ell+\frac{1}{2}} y_\ell^A,$$

and applying it next to the attention sublayer gives

$$x_\ell^B = E_\ell x_\ell^A.$$

The latter relation is exactly the output comparison needed to begin the same argument for block $\ell - 1$. Each newly produced map is injective because the one-sublayer peeling lemma constructs it from the two conditions in Definition 27.

Repeating this right-to-middle-to-left step for $\ell = s-1, s-2, \dots, q$ produces every map $E_{\ell+\frac{1}{2}}$ and E_ℓ in the statement. At the same time, the MLP peel at each block matches the counted MLP axes, and the attention peel matches the counted attention heads modulo gauge. This completes the backward zippering induction. □

Combining all our key results about transformers together, we get:

Corollary 7 (Strong objects throughout Transformer branches)**.** *Under Theorem 12, every peeling step also matches the task-used strong objects in the branch being removed. Hence, at every compared block,*

$$\mathrm{AxisAlign}_{\mathrm{mlp},\ell,\psi}(A,B) \ge \frac{m^{\psi}_{\mathrm{mlp},\ell}(\epsilon)}{d^A_{\mathrm{mlp},\ell}}, \qquad \mathrm{HeadAlign}_{\mathrm{attn},\ell}(A,B) \ge \frac{m_{\mathrm{attn},\ell}(\epsilon)}{H^A_\ell}.$$

Thus, whenever the exact hypotheses of the zippering theorem hold, Transformers have strong objects throughout their non-residual branches: every task-required MLP hidden axis matches as a scalar function on the task set, and every task-required attention head matches through its attention pattern and output modulo gauge. If every physical MLP axis or attention head is task-required, the corresponding score is one.

Proof. At each block, the first step in the block lemma identifies every counted MLP axis before subtracting the MLP branch, while the second step identifies every counted attention head before subtracting the attention branch. By definition of the two task budgets, any solution of loss at most ϵ contains at least the displayed number of task-used axes or heads. Dividing those matched counts by the A-side physical widths gives the two inequalities. If the budgets equal the physical counts, every counted axis or head is matched. □

Remark 6 (How this explains recent findings from Kapoor–Srivastava–Khosla). Theorem 12, Proposition 11, and Corollary 7 allow us to make sense of some recent results about transformer models. Kapoor, Srivastava, and Khosla compare vision models using linear regression, orthogonal Procrustes, permutation, and soft-matching metrics (Kapoor et al., 2025). For ViTs, they find strong hierarchical representational alignment under linear and Procrustes-style metrics, but weak evidence for privileged native axes under permutation-style tests. Linear and Procrustes metrics test weak alignment of the residual or CLS-token representation, which the zippering theorem predicts under its stated conditions. Permutation and soft-matching tests on residual coordinates instead ask whether the observed residual basis itself is privileged; Proposition 11 shows that this need not be true. The positive strong-alignment prediction is the branch-specific one in Corollary 7: task-used MLP hidden axes should align, and task-used attention heads should align modulo gauge. Thus a negative result in the residual basis is not evidence against Transformer weak–strong equivalence; it indicates that the test was applied to the stream in which the theory makes no native-axis claim.

C.2.5 Why the peeling conditions are generic

The exact zippering theorem assumes the two peeling conditions in Definition 27. We now ask how restrictive those conditions are. They can fail in two different ways. First, after the nonlinear branch has been removed, the affine map obtained from the skip equation might fail to be injective. This is an ordinary matrix-rank failure: it means that the candidate upstream map collapses some residual-stream direction and therefore is not a valid weak comparison map. Second, the remainder after branch subtraction might still be nonlinear in the unnormalized residual state, as in the LayerNorm example. This is the genuinely new residual-block failure. The next lemma treats the first issue; the proposition that follows treats the second.

Lemma 19 (The affine injectivity condition)**.** *Let*

$$M_- := T_+ S^A + L.$$

The injectivity requirement in Definition 27 has the following properties.

(a) If S^A and S^B are invertible and $L = 0$, then it follows from injectivity of T_+. In particular, it is automatic for untouched identity skips.

(b) If S^B is invertible, $T_+ S^A$ is injective, and

$$\|L\|_{\mathrm{op}} < \sigma_{\min}(T_+ S^A),$$

then M_- is injective. Thus injectivity is preserved under sufficiently small affine changes.

(c) In a finite-dimensional real-analytic local coordinate system, or in a fixed ReLU region with an unchanged activation pattern, failure of injectivity is lower-dimensional whenever injectivity holds at one point of each connected coordinate region.

Proof. For (a), an invertible S^A followed by an injective T_+ is injective, and multiplication by the invertible matrix $(S^B)^{-1}$ does not change injectivity. For (b),

$$\sigma_{\min}(T_+ S^A + L) \geq \sigma_{\min}(T_+ S^A) - \|L\|_{\mathrm{op}} > 0.$$

For (c), S^B is invertible exactly when $\det S^B \neq 0$, and M_- is injective exactly when $\det(M_-^\top M_-) \neq 0$. These determinants are analytic functions of the local coordinates (and polynomial on a fixed ReLU region). A point where the rank condition holds is a witness that the relevant determinant is not identically zero, so its zero set is a proper lower-dimensional subset. □

Thus the affine injectivity condition is automatic in the basic identity-skip case with no affine remainder, stable under small changes, and generic whenever it holds somewhere in the parameter

family. It is not another nonlinear-identification assumption; it only ensures that the affine relation left after peeling qualifies as weak alignment.

The remaining question is whether a nonlinear LayerNorm cancellation can persist over an open family of residual sublayers. A *witness* below means one parameter choice at which the stated rank is actually attained, showing that the corresponding equations are not identically dependent. We say that the architecture *forces* a cancellation only when an exact symmetry or tied mechanism produces it for every nearby parameter choice, rather than merely at specially tuned values.

Proposition 12 (Genericity of residual peeling)**.** *Fix finite-dimensional analytic local coordinate systems for a source residual sublayer, a B-side sublayer, and the affine output comparison map; for ReLU, restrict when needed to a fixed smooth region with an unchanged zero-trace pattern. Assume the usual first-order regularity after the exact attention gauges, hidden-unit order, and positive ReLU scales have been fixed. Also assume that every proposed nonlinear cancellation has a rank witness showing that, even after all B-side and comparison parameters are allowed to vary, the cancellation still imposes at least one independent condition on the source sublayer, and that no additional cancellation is forced by the architecture over an open parameter family. Then failure of residual peeling is lower-dimensional, and hence has empty interior and measure zero, in the source local coordinate system.*

Proof. First fix one proposed cancellation mechanism τ. Let the source coordinates be $\xi \in \Theta_A$, with $\dim \Theta_A = p$, and let $\eta \in \mathbb{R}^{n_\tau}$ collect all B-side, residual-input, and comparison parameters allowed to vary. Write the finitely many task-evaluation, derivative, pole-or-zero-trace, and attention-output equations required by the cancellation as

$$H_\tau(\xi, \eta) = 0.$$

At a rank witness,

$$\operatorname{rank} DH_\tau \geq n_\tau + 1.$$

Hence every connected smooth piece S_τ of the solution set satisfies

$$\dim S_\tau \leq (p + n_\tau) - (n_\tau + 1) = p - 1.$$

Forgetting η and retaining only ξ cannot increase Hausdorff dimension, so the source parameters admitting this cancellation have dimension at most $p - 1$. A finite union over proposed mechanisms has the same bound. The extra $+1$ is important: rank n_τ could merely determine the B-side and comparison parameters, while rank $n_\tau + 1$ says that the cancellation also constrains the source itself.

For a possible cancellation not listed in advance, consider a smooth family S of bad exact pairs on the region where the branch-identification and affine-rank conditions already hold. First-order regularity says that, after exact symmetries are fixed, varying only the B-side and comparison parameters changes the represented output unless those parameters do not change. Therefore the map from S to the source coordinates has injective derivative, so $\dim S \leq p$. If equality held, the inverse function theorem would make the B-side and comparison parameters smooth functions of the source over an open set. By the assumption that the architecture does not force the cancellation, this family contains a point where the leftover after branch subtraction is affine in the unnormalized source residual state. Lemma 17 then produces an injective affine input map, contradicting that S consists of peeling failures. Thus these unlisted failures also have dimension at most $p - 1$ after projection to the source coordinates.

Finally, failure of branch identification is lower-dimensional by Proposition 10 and the earlier MLP genericity results, while the affine rank failure is lower-dimensional by Lemma 19. Adding these exceptional sets preserves the conclusion. □

This result of course does not say that the LayerNorm example is impossible (since it obviously exists!); rather, it says that such a cancellation cannot persist throughout an open parameter family unless the architecture itself supplies an exact symmetry or tied mechanism that produces it for every nearby parameter choice. As elsewhere in the paper, applying the generic conclusion to learned networks also requires the optimizer not to concentrate on the exceptional set, which we have not shown.

C.3 Regularity and depth faithfulness

From here on out, the formal proof of our main results is done. However, the regularity assumptions also have useful interpretive consequences, which are covered in this and the following sections.

The purpose of this section is to explain why regularity makes network depth a meaningful measure of complexity. For ordinary affine–ReLU and affine–Softplus layers, once a private current nonlinear signature has been matched, regularity says that the unit cannot be recreated by functions already available at lower depth, except through the activation's unavoidable symmetries. The new layer therefore contributes genuinely new nonlinear structure rather than merely rewriting a shallower computation. This is a local version of the usual depth-separation idea in neural-network theory (Eldan & Shamir, 2016; Telgarsky, 2016).

The residual-block and Transformer argument in §C.2 adds one distinction. A term may be affine in the normalized branch input $\mathrm{LN}(x)$ while remaining nonlinear in the raw residual state x. Such a term cannot simply be moved into the residual skip; only a term of the form $Lx + c$ can. The LayerNorm example in §C.2.3 gives an exact special-case counterexample in which LayerNorm hides a nonlinear branch term. Residual peeling handles it in the same way as the earlier null-pair examples: first recover the current branch, then show that the remaining nonaffine branch–skip cancellations occur only on a lower-dimensional exceptional set.

C.3.1 Softplus units cannot be pushed to lower layers

At a backward zippering step r, the scalar arguments $g_j = z_{r,j}$ have already been produced by the lower part of the network. The current layer applies σ to these arguments and then linearly recombines them. To say that a Softplus unit at this stage is genuinely new, we compare it to a fixed space of functions that were already available before applying this current outer Softplus.

Definition 28 (Softplus lower-depth space)**.** Fix a backward step r and a chosen lower-depth model class, such as the layers below r with specified widths, depth, and parameter bounds. A *Softplus lower-depth space* is a closed linear subspace $\mathscr{L}_r^\sigma \subset L^2(\Omega)$ representing scalar functions already available before the current outer Softplus stage. By construction, we assume that it contains constants, the current preactivations,

$$1, \qquad z_{r,1}, \dots, z_{r,d_r},$$

and any other scalar functions representable by the chosen lower-depth model class. Thus "lower-depth" means: available from the input and earlier layers, without using another copy of the current Softplus stage.

Now suppose two matched Softplus arguments g and h share the same local outer pole patch. Writing

$$Q_g = 1 + e^{g^{\mathbb{C}}}, \qquad Q_h = 1 + e^{h^{\mathbb{C}}},$$

this means locally that the common pole divisor can be factored out:

$$Q_h = uQ_g,$$

where u is holomorphic and nonzero along the shared patch. On the real task domain, with a compatible logarithm branch, logarithms are well defined and we have:

$$\log u = \sigma(h) - \sigma(g).$$

So the difference $\sigma(h) - \sigma(g)$ is the leftover part after the shared Softplus pole signature has been removed.

Definition 29 (Softplus depth-faithfulness)**.** Let $\mathscr{L}_r^\sigma$ be a Softplus lower-depth space. A matched pair (g, h) is *Softplus depth-faithful over* $\mathscr{L}_r^\sigma$ if a trivial remainder can only happen if there is equality up to sign:

$$\sigma(h) - \sigma(g) \in \mathscr{L}_r^\sigma \quad \Longrightarrow \quad h = g \quad \text{or} \quad h = -g.$$

In the above definition, the negative case is unavoidable for a single pair because

$$\sigma(-g) - \sigma(g) = -g$$

and by definition g itself is already a lower-depth preactivation. The depth-faithfulness definition basically says: depth is a meaningful notion of computational complexity.

Now, the key point is that MLR is sharp condition for forcing depth faithfullness, which explains why regularity is the "right" notion for ruling out null pairs in the first place:

Proposition 13 (MLR implies matched-pair faithfulness)**.** *If the matched pair (g, h) is MLR, then $h = \pm g$. In a full minimal expansion, the actual match $h = g$ obtains.*

Proof. The matched-pair argument inside Theorem 4 shows that an MLR divisor match has only two possibilities:

$$h = g \quad \text{or} \quad h = -g.$$

In the first case there is no leftover term:

$$\sigma(h) - \sigma(g) = 0 \in \mathscr{L}_r^\sigma.$$

In the second case,

$$\sigma(h) - \sigma(g) = \sigma(-g) - \sigma(g) = -g \in \mathscr{L}_r^\sigma,$$

because g is one of the lower-depth preactivations. Thus MLR gives exactly the Softplus depth-faithfulness dichotomy for this pair. Finally, if a negatively oriented match occurred inside an equality of complete expansions, the terms with negative orientation would contribute a lower-depth affine combination

$$\sum_{j \in S} v_j g_j,$$

which is forbidden by sign-noncancellation unless S is empty. Hence the complete minimal expansion keeps only the positive matches. □

C.3.2 ReLU exposed facets cannot be pushed to lower layers

We now illustrate that the exat analogous depth faithfulness arises for ReLU from the MFR condition (this is actually in a way how MFR was inspired in the first place). For ReLU, the lower part of the network already determines a polyhedral cell decomposition $\mathscr{P}$. On each cell of $\mathscr{P}$, all lower-depth scalar functions are affine. A current ReLU unit is genuinely new only if its fresh facet creates a kink that cannot be represented by a function affine on those inherited cells.

Definition 30 (ReLU lower-depth space and depth-faithfulness test)**.** Let $\mathscr{P}$ be the common inherited affine complex for the ReLU scalar arguments being compared. The *ReLU lower-depth space* associated with $\mathscr{P}$ is

$$\mathscr{L}(\mathscr{P}) = \{f \in L^2(\Omega) : f|_C \text{ is affine on every full-dimensional cell } C \in \mathscr{P}\}.$$

Thus $\mathscr{L}(\mathscr{P})$ is the space of scalar functions that do not introduce a new ReLU kink across the marked fresh facets. For scalar arguments g, h and a scale $\alpha > 0$, define the matched residual

$$R_\alpha(h, g) = \rho(h) - \alpha\rho(g).$$

The pair (g, h) is *lower-depth* if $R_\alpha(h, g) \in \mathscr{L}(\mathscr{P})$ for some $\alpha > 0$, and *ReLU depth-faithful over* $\mathscr{L}(\mathscr{P})$ if this can happen only in the two sign-scale cases

$$\rho(h) - \alpha\rho(g) \in \mathscr{L}(\mathscr{P}) \quad \Longrightarrow \quad h = \alpha g \quad \text{or} \quad h = -\alpha g.$$

Proposition 14 (MFR implies ReLU matched-pair depth faithfulness)**.** *If (g, h) is coherently MFR, then there are $\alpha > 0$ and $\varepsilon \in \{+1, -1\}$ such that $h = \varepsilon\alpha g$. For this α,*

$$\rho(h) - \alpha\rho(g) = \begin{cases} 0, & \varepsilon = +1, \\ -\alpha g, & \varepsilon = -1, \end{cases} \in \mathscr{L}(\mathscr{P}).$$

Thus MFR supplies the ReLU depth-faithfulness dichotomy for the matched pair. In a full minimal expansion, only the positive scale remains valid.

Proof. By Definition 17, coherent MFR gives

$$h = \varepsilon\alpha g$$

for some $\alpha > 0$ and $\varepsilon \in \{+1, -1\}$. If $\varepsilon = +1$, positive homogeneity gives

$$\rho(h) = \rho(\alpha g) = \alpha\rho(g),$$

so the residual is zero. If $\varepsilon = -1$, then

$$\rho(h) - \alpha\rho(g) = \alpha\big[\rho(-g) - \rho(g)\big] = -\alpha g.$$

Because g is affine on every inherited cell, $-\alpha g \in \mathscr{L}(\mathscr{P})$. Thus the only lower-depth residuals allowed by MFR are the two sign-scale possibilities. At the expansion level, the negative case would create exactly the affine cancellation excluded by EFF-minimality, while the positive rescaling is the standard ReLU symmetry. □

Taking the last two propositions together, we see that MLR and MFR do more than just identify matched units, but also explain *why those units belong at their current depth*. After the private Softplus pole or ReLU fresh facet has been matched, any remaining difference is lower-depth only in the trivial sign or rescaling cases. This is the sense in which the regularity assumptions make depth a faithful measure of the network's nonlinear complexity.

C.3.3 Residual blocks: what may be moved into the skip

The previous two subsections asked whether a current Softplus or ReLU unit can be replaced by functions that were already available at lower depth. In a residual block, there is one additional possibility: part of the nonlinear branch might instead be moved into the skip connection.

For a residual representation $x : \Omega \to \mathbb{R}^d$, the skip can carry functions of the form

$$\mathcal{A}_x = \{\omega \mapsto Lx(\omega) + c : L \in \mathbb{R}^{m\times d},\ c \in \mathbb{R}^m\}.$$

These are exactly the functions that are affine in the representation already carried by the skip. A function can be linear in $\mathrm{LN}(x)$ without being affine in x, so such a function cannot in general be moved into the skip.

Proposition 15 (Only affine branch remainders can be moved into the skip)**.** *Suppose the current MLP axes or attention heads can be identified using the branch results of §C.2. Outside the lower-dimensional exceptional set of Proposition 12, a counted branch object cannot be removed by compensating for it with a nonlinear change in the residual representation. Any remainder that can be moved into the skip must belong to $\mathcal{A}_x$.*

Proof. The results of §C.2 first identify the counted objects in the two branches.

For an MLP branch, matched units with the same orientation cancel exactly. A match with the opposite orientation can leave a term of the form

$$R_{S_-} = \sum_{j\in S_-} T_+ v_j g_j.$$

If this remainder is affine in x, then it belongs to $\mathcal{A}_x$ and can simply be combined with the affine skip. If it is not affine in x, removing it would require an input-dependent change in the residual representation that LayerNorm does not detect. This is exactly the exceptional cancellation analyzed in §C.2; outside the lower-dimensional exceptional set, it cannot occur.

For attention, conditions (A1)–(A3) identify each counted head modulo attention gauge. After the matched head outputs have been subtracted, a counted head could disappear only if its remaining effect were canceled by the same kind of nonlinear change in the residual representation. The generic peeling result excludes this possibility as well. Thus, for either kind of branch, only a remainder of the form $Lx + c$ can be moved into the skip. □

This is the residual-block version of depth faithfulness. A counted MLP axis or attention head contributes genuinely new nonlinear structure unless its remaining effect is already affine in the representation carried by the skip. The affine-rank condition in Lemma 19 has a separate role: after the nonlinear branch has been removed, it checks only that the resulting affine relation is an injective weak comparison map.

D Failure modes and architectural scope

D.1 A null-network atlas: redundant solutions and how to exclude them

Null networks are needlessly complex ways of realizing simpler input–output relationships. There are three essential ways they are ruled out in the alignment and zippering theory:

1. First, intrinsic minimality removes representations that are already redundant inside one expansion.
2. Second, MLR or MFR rules out a nontrivial equal-function pair at a point where the relevant nonlinear signatures are identifiable.

Mechanism	Softplus diagnosis	ReLU diagnosis
Unused or constant feature	The term is absorbed into the bias or removed; usedness and nonconstancy fail.	The branch produces no visible task kink; usedness or two-sided crossing fails.
Duplicate or opposite features	Coefficients can be redistributed, or the sign identity leaves an affine residual; reducedness, pole separation, and sign-noncancellation intervene.	Positive-scale duplicates are redundant under positive rescaling; opposite pairs can collapse to an affine term; sign-scale reduction and sign-noncancellation intervene.
Continuous feature symmetry	A postactivation shift can move in a downstream nullspace; the realization Jacobian has a kernel.	Positive scaling is removed by normalization; any additional continuous output-preserving motion violates Jacobian regularity after fixing the standard matching convention.
Shifted or nested identity	An affine output cannot retain private curvature divisors; deeper nested identities require exact factorization or integer-weight resonance.	An affine output has no kink jumps; deeper variants require coincident facets, invisibility on ReLU-inactive regions, or split matching.
LayerNorm branch–skip transfer	An opposite MLP match can leave $v(g \circ \mathrm{LN})$. A skip can hide this term only on a special parameter set; only terms affine in the raw residual state may be moved into the skip.	The same sign identity gives the ReLU version. Private smooth traces recover the unit, and the rank test makes the remaining nonaffine cancellations lower-dimensional.
Discrete cross-family nontrivial collision	Possible on special non-MLR resonance sets, including exact integer weights and linked biases.	Possible on special fixed activation-pattern and incidence cases with proportional cellwise hyperplanes or incomplete task exposure.

Table 1: A guide to the assumptions. Pointwise signature recovery excludes nontrivial collisions on the recovery set. The generic theorem permits exceptional nontrivial collisions, but makes their projection into individual-network space lower-dimensional.

3. Third, by not actually ruling them out but showing that they are rare. The generic null-net theorem does *not* claim that every exceptional identity is impossible: it allows resonant nontrivial collisions on lower-dimensional subsets, but prevents them from filling a full-dimensional family of regular minimal networks.

As an expository utility in this section, we show examples of each of these things. In what follows a *resonance set* means a lower-dimensional set of parameters where an exact algebraic relation, such as an integer-weight relation or a fixed ReLU activation-pattern coincidence, makes one of these exceptional identities possible. Quantitative assumptions add positive distance from these bad sets and are used for stable, soft statements. Table 1 summarizes the types of silly networks that are then explained in greater detail in each of the following subsections.

D.1.1 Unused, constant, duplicate, and opposite features

The easiest null relations are internal redundancies. If $v_j = 0$, the term $v_j\varphi(g_j)$ is absent from the represented function. If g_j is constant, then $v_j\varphi(g_j)$ can be folded into the bias. If two arguments coincide, then

$$v_1\varphi(g) + v_2\varphi(g) = (v_1 + v_2)\varphi(g),$$

so the division of coefficient mass between the two units is unidentifiable.

The sign-pair identities are especially important:

$$\sigma(t) - \sigma(-t) = t, \qquad \rho(t) - \rho(-t) = t. \tag{31}$$

For Softplus, g and $-g$ have the same curvature-pole set. For ReLU, they have the same zero facet with opposite active sides. Thus a pair of opposite units can erase its nonlinear signature and leave only an affine function. Same-side reducedness prevents duplicate or opposite sign-classes from being counted as separate minimal units, while affine sign-noncancellation rules out a collection of such residual affine terms being absorbed into the bias.

The quantitative margins serve a different purpose. Bounds such as $\|v_j\| \geq c$, lower variance or crossing margins, and positive separation from duplicates prevent a convergent sequence of counted

networks from approaching one of these degeneracies. Parameter boundedness by M does not by itself rule out a null relation; it supplies compactness, so that a sequence of near-counterexamples has a limiting counterexample to which the exact theory can be applied.

D.1.2 SHIFTED IDENTITY BLOCKS

A tempting construction inserts an exact identity block before a module G in one network and after G in another. The terminal functions agree, while the internal axes appear at different depths.

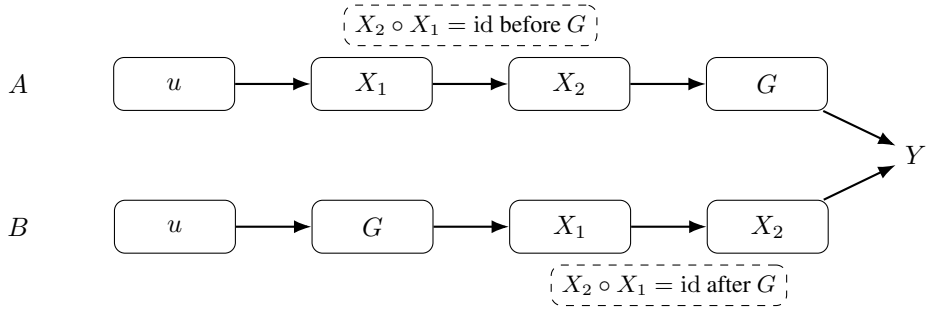


Figure 12: A shifted identity block. The correct place to test the construction is the first backward step at which the two one-step decompositions differ. There it creates a nontrivial null pair, so either pointwise signature recovery identifies the axes or the construction lies on an exceptional resonance set.

For a Softplus identity block, the last layer would give a one-step representation of an affine map,

$$u = b + \sum_j v_j \sigma(g_j(u)). \tag{32}$$

The left side has zero Hessian. On a private regular pole patch of a minimal sign-class,

$$D^2[\sigma(g_j)] = \sigma''(g_j)\, dg_j \otimes dg_j + \sigma'(g_j) D^2 g_j$$

contains a singular contribution that cannot be canceled by terms analytic on that patch. Hence an affine output is possible only if the term is unused or constant, if another term shares the same divisor, or if sign-opposite terms collapse through (31). These are exactly the roles of usedness, nonconstancy, reducedness, pole separation, and sign-noncancellation. Where MLR holds, the same conclusion is obtained more structurally: the outer Laurent binomial factors can match only through g versus $\pm g$; the negative case is then removed by sign-noncancellation.

For ReLU, the affine identity has no derivative jump. A used fresh facet of g_j contributes the rank-one jump

$$v_j \nabla g_j^\top.$$

It can disappear only if another unit has a coincident facet and an exactly cancelling jump. Same-side sign-scale reduction and sign-noncancellation rule out the basic opposite-pair construction. Deep ReLU blocks also permit ReLU-inactive-region and split-facet constructions, which is why one local facet is not sufficient and MFR uses a coherent atlas of facets across inherited cells.

The first-mismatch formulation is the most general. At the first backward zippering step where the two networks differ, the already aligned next representation has two equal one-step expansions. Pointwise MLR or MFR makes that null pair trivial. If the inserted block is genuinely nontrivial, it must therefore leave the recovery set. The generic null-net theorem allows this on a proper lower-dimensional exceptional set, but excludes a full-dimensional family of such shifted nontrivial collisions. Quantitative margins keep the realized networks away from the corresponding degeneracy or resonance sets.

D.1.3 LAYERNORM CAN TRADE A BRANCH TERM AGAINST THE SKIP

A pre-LayerNorm residual block has one null-pair mechanism that is absent from an ordinary affine–nonlinear expansion. LayerNorm is unchanged by adding a tokenwise shift in the all-ones feature

direction, while both activations used in this paper satisfy

$$\varphi(t) - \varphi(-t) = t.$$

A sign reversal of a current MLP argument can therefore leave a term $v(g \circ \mathrm{LN})$, and a sample-dependent shift of the residual stream can sometimes absorb it. The LayerNorm example in §C.2.3 gives an exact one-unit example: the residual-sublayer outputs agree, but the two raw residual-input families are not affinely related.

The example shows why ordinary branch minimality is not quite enough for raw residual zippering. Minimality can identify the current sign-class, yet the opposite orientation may leave a term that is linear in normalized coordinates and nonlinear in the raw residual state. This is also why Lemma 15 permits a term to enter the skip only after it has been shown to be raw-residual-affine. This nonaffine transfer is the new null-pair mechanism. The later check that the remaining affine map is injective is different: it is automatic for an untouched identity skip with no affine transfer, stable under small affine transfers, and generically fails only on a rank-defect set by Lemma 19.

The construction is nevertheless special rather than a free architectural symmetry. Its readout must point in a LayerNorm-invisible direction, the competitor must obey an exact sign relation, and the residual input must move in just the corresponding invisible direction. Proposition 12 writes these requirements as finite equations. If, after all competitor variables are allowed to move, those equations still impose at least one condition on the source network, then the networks admitting the transfer are lower-dimensional. Proposition 12 handles any additional same-output transfer not listed in advance. As elsewhere in the atlas, exact exceptional identities are allowed, but they cannot fill an open family of regular minimal networks.

D.1.4 SOFTPLUS POSTACTIVATION-SHIFT SYMMETRIES

A more subtle Softplus identity shifts an *activated* feature rather than canceling a sign pair. For $c > 0$, define

$$H_c(t) = \sigma^{-1}(\sigma(t) + c) = \log\big(e^c(1 + e^t) - 1\big). \tag{33}$$

Then

$$\sigma(H_c(t)) = \sigma(t) + c. \tag{34}$$

Let $V = [v_1 \ \cdots \ v_k]$ be the downstream coefficient matrix and choose a nonzero vector q with $Vq = 0$. For base arguments u_j and constants c_j, set

$$g_j^{(\tau)} = H_{c_j + \tau q_j}(u_j).$$

Then

$$Y_\tau = b + \sum_j v_j \sigma(g_j^{(\tau)}) \tag{35}$$

$$= b + \sum_j v_j\big(\sigma(u_j) + c_j + \tau q_j\big) = Y_0. \tag{36}$$

Thus τ is a continuous output-preserving coordinate even when every individual v_j is nonzero and every argument varies.

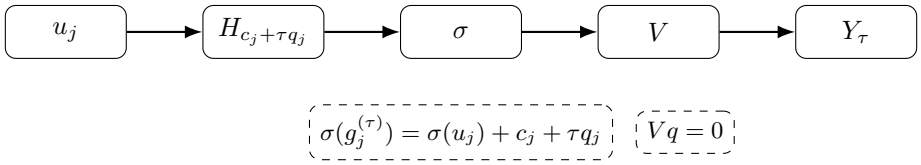


Figure 13: A postactivation-shift symmetry. The hidden activated coordinates move in a downstream nullspace while the represented function remains fixed.

Differentiating (36) gives

$$DF(\vartheta)\,\partial_\tau = 0.$$

Hence qualitative Jacobian regularity fails. This example is useful because it shows that intrinsic minimality alone does not remove every continuous redundancy. It does not threaten the regular generic null-net theorem: the entire continuous symmetry is removed when one passes to the regular part of the realization chart.

D.1.5 NESTED SOFTPLUS IDENTITIES

Softplus also admits identities that insert a nominal nonlinear stage while changing an activated feature only by a constant. A basic example is

$$\sigma(a + \sigma(t - a + \sigma(a))) = \sigma(a) + \sigma(t). \tag{37}$$

Equivalently, if

$$\widehat{t} = a + \sigma(t - a + \sigma(a)),$$

then the extra stage satisfies $\sigma(\widehat{t}) - \sigma(t) = \sigma(a)$, which the next bias can absorb. Such an identity is not generated by a generic perturbation of all weights and biases: it uses exact unit coefficients and linked offsets. In multiplicative form,

$$1 + e^{\widehat{t}} = e^{\sigma(a)}(1 + e^t),$$

so the two outer divisors coincide and their ratio is a constant unit. This is precisely a non-MLR factorization resonance.

The identity can be nested or shifted across several layers. It may also coexist with another branch that makes the whole network output genuinely depth-essential, so global output-depth minimality alone is not enough to remove every nested subbranch. What matters for the generic theorem is that the exact linked-weight equations define a lower-dimensional resonance set. Where MLR holds, the factorization is impossible except for the trivial $g \leftrightarrow \pm g$ cases, and general position on collision strata prevents the nested-identity set from supporting a full-dimensional collision stratum.

D.1.6 A REGULAR DISCRETE SOFTPLUS RESONANCE

Not every exceptional nontrivial collision is Jacobian-singular or globally depth-reducible. Let

$$\kappa = \sigma(\eta).$$

Then

$$\sigma(\eta - \sigma(t)) + \sigma(\eta - \sigma(\kappa - t)) = \kappa. \tag{38}$$

To verify this, write

$$Q_1 = 1 + e^{\eta - \sigma(t)} = \frac{1 + e^t + e^\eta}{1 + e^t},$$

and, since $e^\kappa = 1 + e^\eta$,

$$Q_2 = 1 + e^{\eta - \sigma(\kappa - t)} = \frac{(1 + e^\eta)(1 + e^t)}{1 + e^t + e^\eta}.$$

Thus $Q_1 Q_2 = 1 + e^\eta = e^\kappa$, and taking the real logarithm gives (38).

This identity yields two ordinary dense same-depth networks with equal outputs and different intermediate axes. For input (x, y), choose $n, m \neq 0$, generic η, γ, β, and an irrational $0 < \alpha < 1$. Let

$$z_1^A = (x, y), \tag{39}$$
$$z_2^A = \big(\eta - \sigma(x),\ \gamma + \alpha\sigma(y)\big), \tag{40}$$
$$z_3^A = \beta + n\sigma(z_{2,1}^A) + m\sigma(z_{2,2}^A), \tag{41}$$

and

$$z_1^B = (\kappa - x, y), \tag{42}$$
$$z_2^B = \big(\eta - \sigma(\kappa - x),\ \gamma + \alpha\sigma(y)\big), \tag{43}$$
$$z_3^B = \beta + n\kappa - n\sigma(z_{2,1}^B) + m\sigma(z_{2,2}^B). \tag{44}$$

Equation (38) gives

$$z_3^A(x, y) = z_3^B(x, y),$$

while generically

$$z_{2,1}^A \neq z_{2,1}^B, \qquad z_{2,1}^A \neq -z_{2,1}^B.$$

The common y-branch can be chosen to carry genuinely top-depth analytic structure; the irrational exponent α is a convenient way to prevent the entire output from being explained by the elementary x-branch identity alone. Moreover, the individual realization Jacobians can remain full rank: this is a discrete cross-family nontrivial collision, not a continuous symmetry.

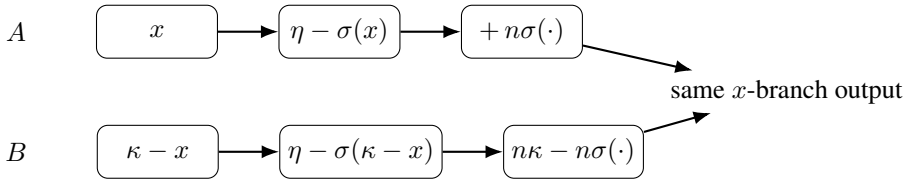


Figure 14: The regular discrete Softplus resonance. The identity is exact, but it requires a special reflected input, exact -1 coefficients, zero cross-couplings, and linked biases.

Why does this not contradict the generic null-net theorem? A fully dense 2–2–2–1 network has 15 parameters. The displayed nontrivial-collision family has 12 free parameters: six for two independent first-layer affine forms, three for (η, γ, α), and three for (n, m, β). It imposes, in particular, the exact second-layer conditions

$$W_{2,11} = -1, \qquad W_{2,12} = 0, \qquad W_{2,21} = 0.$$

Thus its projection into an individual dense-network chart is codimension at least three. The example is regular and potentially depth-essential, but it lies on an integer-weight and sparsity resonance set. The generic theorem is designed to permit exactly this sort of exceptional nontrivial collision while ruling out an open family of them. Actually this type of example (which ChatGPT-5.5-Pro spent several hours finding!) is the motivation behind the whole genericity approach we take.

D.1.7 ReLU positive homogeneity and changes on ReLU-inactive regions

The exact positive-rescaling symmetry

$$(v, g) \mapsto (\alpha^{-1} v, \alpha g), \qquad \alpha > 0, \tag{45}$$

is unavoidable because $\rho(\alpha g) = \alpha \rho(g)$. It is removed by the ReLU normalization convention before Jacobian regularity or null-pair uniqueness is stated.

A genuinely different obstruction is that ReLU erases everything an argument does while it remains negative. For $\lambda > 0$, let

$$g(x) = x, \qquad h_\lambda(x) = x - \lambda \rho(-x - 1) = \begin{cases} (1 + \lambda)x + \lambda, & x < -1, \\ x, & x \geq -1. \end{cases} \tag{46}$$

For $x < -1$, $h_\lambda(x) < 0$; for $x \geq -1$, $h_\lambda(x) = x$. Hence

$$\rho(h_\lambda) = \rho(g)$$

although $h_\lambda \not\sim_\pm g$. The two arguments share the exposed zero facet at $x = 0$ but differ on an inactive inherited cell.

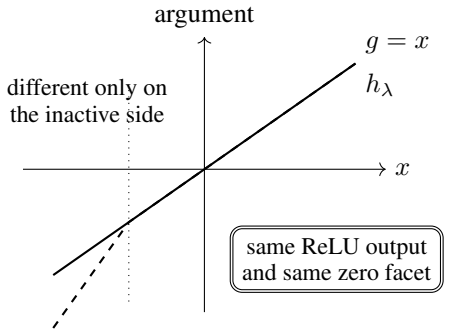


Figure 15: Why one exposed facet is not enough for ReLU. MFR records enough labelled facets across inherited cells to recover the complete normalized CPWA argument rather than only its visible active boundary.

If λ is a free coordinate in one realization chart, (46) is a continuous fiber and Jacobian regularity after fixing the standard matching convention fails. A discrete version may instead connect two separate regular families. In that case the task exposes too little of the argument for one-facet reasoning, while the MFR hypothesis explicitly requires a task-complete multi-cell signature. Under the signature secant-derivative and general-position assumptions on collision strata, ambiguous ReLU-inactive-region nontrivial collisions can survive only on lower-dimensional strata.

D.1.8 SPLIT-FACET MATCHING AND SPECIAL ACTIVATION-PATTERN CASES

A final ReLU-specific failure is *split matching*. Suppose one deep argument g has marked facets F_1, F_2, F_3 in three inherited cells. Equality of represented kink sets might match F_1 to a competitor h_1, F_2 to h_2, and F_3 to h_3, even though no single competitor is globally sign-scale equivalent to g.

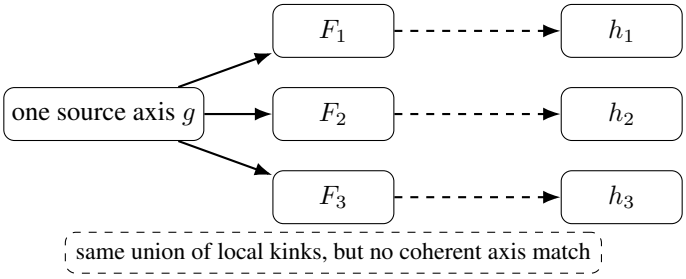


Figure 16: Split-facet matching. Coherent MFR requires all marked facets of one source argument to be explained by one competitor argument.

The coherence condition excludes this possibility. Generically, a prescribed split assignment imposes simultaneous proportionality equations for several cellwise affine germs, together with activation-pattern incidence equations. On a fixed ReLU activation-pattern stratum these are polynomial or piecewise-analytic constraints. The split-signature result (Proposition 7) shows that, when the required rank condition holds at one point, the projected split-matching set has codimension at least one. The ReLU general-position condition on collision strata then prevents a union of split assignments from filling a full-dimensional collision stratum.

Remark 7. That this split-facet matching situation exists is totally reasonable, but the fact it has to be ruled out by the clobberingly strong coherence condition, rather than the elegant resonant Laurent-nonprimitive resonance characterization that works for softplus, makes us think we do not fully understand how best to handle the ReLU case. Now, this type of case is non-generic, and the dimensionality analysis of exceptional sets is very clean and activation-independent, so "all ends well". But it would be nice if coherence could get cleaned up in future work – perhaps some form of discrete algebraic geometry will enable this.

D.2 OTHER ARCHITECTURE TYPES AND ACTIVATION FUNCTIONS

The arguments above were written for affine–ReLU and affine–Softplus networks because those models make the nonlinear signatures especially transparent. The reusable part of the theory is more general. For a new activation or architecture, one needs four ingredients: the natural symmetry convention, a task-visible nonlinear signature for adjacent weak–strong alignment, a pointwise one-step identifiability theorem, and a genericity argument showing that ambiguous full-dimensional collision strata are lower-dimensional. Once those ingredients are available, the compactness, inverse-continuity, and Jacobian-regularity arguments give the asymptotic and soft versions with little change.

Setting	Natural signature/symmetry	Likely exact theory	Main additional difficulty
CNNs	ReLU/Softplus signatures at channel-position axes; convolutional weight sharing	Affine–nonlinear proofs apply to convolutional preactivation maps	Pooling, subsampling, and the choice between channel-level and channel-position axes
RNNs	Unrolled kink/ridge signatures with parameters tied across time	Feedforward theorem applies to each fixed unrolled time and layer	Weak maps and genericity must respect time-tying and long-horizon stability
Tanh	Meromorphic poles; signed permutations	Close to an existing full null-net theory	Quantitative saturation and conditioning
GELU	Gaussian ridge or quadratic-exponential phase; sign identity	Adjacent theorem should be direct; deep theorem needs a new recovery module	Nonmonotone inverse branches and deep phase matching
ELU	Transition jump jet plus negative-branch exponential character	Hybrid ReLU/smooth proof	Negative-only translation symmetry and saturation

Table 2: High-level extension map. The common collision geometry and zippering induction remain; the nonlinear signature mechanism changes.

D.2.1 CNNs

Convolutional layers are affine maps with weight sharing. Therefore, before pooling or subsampling, a convolutional ReLU or Softplus block is still an affine–coordinatewise-nonlinear–affine block. The adjacent weak–strong argument applies to the channel-position preactivation axes: a used ReLU channel-position creates a crossed kink trace, and a used Softplus channel-position creates a curvature ridge. Weight sharing does not remove these signatures; it ties the parameters generating spatially translated copies of them.

The main bookkeeping issue is the level at which axes are counted. Before spatial pooling, the natural axes are channel-position coordinates. After global or local averaging, several positions of the same channel may be pooled into a coarser channel-level feature, so the theorem should either be applied before pooling or be restated for the pooled channel representation. Max-pooling is piecewise affine and introduces its own switching surfaces; it can be included by refining the ReLU cell decomposition, but it is cleaner to place the comparison immediately before or after the pooling operation.

D.2.2 RNNs

Unrolling an RNN over time turns it into a feedforward network with tied parameters. For a ReLU or Softplus recurrent update, the weak–strong theorem applies at a fixed time and layer of the unrolled computation exactly as it does for a feedforward affine–nonlinear block. The hidden state coordinates at that time have the usual kink or ridge signatures, and a pair of adjacent weak maps forces the used coordinates to align.

The additional issue is compatibility across time. The same recurrent weights are reused at every unrolled step, so a comparison map chosen at one time cannot be arbitrary if one wants a single time-consistent equivalence over a long trajectory. For zippering, the corresponding compactness and genericity statements must be made inside the tied-weight parameter chart. This restriction does not change the local nonlinear signature mechanism, but it can make the allowed family of weak maps and null pairs smaller and more structured.

D.2.3 Tanh networks

For $\tau(t) = \tanh t$, the unavoidable unit symmetry is

$$(v, g) \sim (-v, -g),$$

so the allowed hidden-unit matching uses permutations and coordinatewise sign flips rather than positive scales. Tanh is meromorphic with a regular complex pole lattice. Affine-ridge weak–strong alignment can therefore be proved by matching pole signatures, and deep one-step null relations can

be attacked by an activation-specific null-network classification. In fact, tanh-type activations are substantially closer than Softplus to settings where a complete regular null-net theorem is available. A tanh version of the present draft would likely replace MLR/MFR by a signed-pole recovery theorem and then reuse the common generic null-net and zippering sections almost verbatim.

The main difficulty is quantitative rather than exact. Since

$$\tau'(t) = \operatorname{sech}^2(t),$$

large preactivations are strongly saturated. Exact identifiability may survive, while the smallest realization Jacobian singular value becomes very small. Soft zippering therefore needs an explicit nonsaturation or Jacobian conditioning margin.

D.2.4 GELU NETWORKS

For exact GELU,

$$\gamma(t) = t\Phi(t),$$

where Φ is the Gaussian distribution function. It satisfies

$$\gamma(t) - \gamma(-t) = t,$$

so an affine sign-noncancellation condition again has a natural role. For adjacent affine arguments, derivatives of GELU produce Gaussian ridge functions. Restriction to generic lines and Fourier decay should identify affine arguments up to sign, giving a weak–strong theorem by a mechanism analogous to the Softplus affine-ridge proof.

Deep zippering is less immediate. The natural analytic signature is a quadratic-exponential phase of the form

$$\exp\bigl(-g^2/2\bigr)\,,$$

and one would need a quadratic-exponential analogue of MLR showing that persistent equal phases force $h = \pm g$ outside a resonance set. Exact GELU is also not globally injective, so part of its range has two inverse branches. A complete theorem must either restrict counted arguments to a common monotone branch, prove that branch-swapping compositions cannot occur at the same network complexity, or treat persistent branch nontrivial collisions by general position on collision strata. The common generic null-net framework is well suited to the last option, but the activation-specific recovery module is genuinely new. The widely used tanh approximation to GELU has different complex singularities and should be analyzed as a separate activation rather than silently identified with exact GELU.

D.2.5 ELU NETWORKS

For parameter $\alpha > 0$,

$$\eta_\alpha(t) = \begin{cases} t, & t > 0, \\ \alpha(e^t - 1), & t \le 0. \end{cases}$$

ELU combines a ReLU-like transition hypersurface with an exponential negative branch. If $\alpha \neq 1$, the first derivative jumps at zero; for $\alpha = 1$, the first derivative matches but the second derivative jumps. An adjacent weak–strong theorem can therefore use the first nonvanishing transition jet. Matching consecutive jet coefficients should fix both orientation and scale for a crossed affine argument.

Crossing is important because a unit confined to the negative branch has a translation–coefficient symmetry. If g and $g + c$ remain negative on the task set, then an appropriate rescaling of the downstream coefficient and a bias correction preserve the represented function. A zero crossing anchors the offset and removes this symmetry. A deep ELU recovery module would probably combine coherent multi-transition-jet signatures with exponential character recovery on negative cells. The remaining generic collision geometry is the same as here. Soft bounds must also account for negative-branch saturation, since $\eta'_\alpha(t) = \alpha e^t$ can be arbitrarily small.

E REPRESENTATION METRICS BEYOND RIDGE

E.1 RSA AND WEAK–STRONG EQUIVALENCE

In this work so far, we have focused on relating two basic metrics for comparing networks: linear regression (weak alignment) and privileged axes (strong alignment). We now turn to Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). RSA is a broad framework: practitioners may compare representational similarity matrices (RSMs) or representational dissimilarity matrices (RDMs), and different choices of centering and matrix comparison produce different statistics. This section fixes one common RSM-RSA convention: inner-product RSMs are reduced to their strict upper triangles and compared by Pearson correlation. It is important not to conflate this convention with double-centered squared-Euclidean RDM-RSA, which is exactly equivalent to linear CKA; kernel-induced nonlinear variants admit the same correspondence (Williams, 2024). We derive that equivalence explicitly in Appendix §E.2.1. RSA is one of the most popular frameworks for comparing neural networks to brain data, in part because it does not require fitting comparison parameters (unlike linear regression).

In this section, we ask: how does RSA fit into the weak–strong equivalence picture? In doing so, we rely heavily on the program laid out in (Theiss et al., 2026), synthesizing their results into the framework of this paper and slightly extending their reach. In what follows, we summarize and reformulate in our notation the main result of Theiss et al; strengthen the Theiss results in a few ways to support canonicalization; prove weak–strong equivalence for canonicalized RSA; develop a soft weak–strong alignment formula relating ridge error to canonical RSA distance; and show how sample bias can interact with all these issues.

E.1.1 RSM-RSA AND THE THEISS DECOMPOSITION

Fix an evaluation input set $\Omega_P = \{x^1, \dots, x^P\} \subseteq \Omega$. For a representation with scalar feature vectors $f_1, \dots, f_d \in \mathbb{R}^P$ on these stimuli, write

$$H = \begin{bmatrix} f_1^\top \\ \vdots \\ f_d^\top \end{bmatrix}, \qquad R(H) := H^\top H = \sum_{j=1}^{d} f_j f_j^\top. \tag{47}$$

The $P \times P$ matrix $R(H)$ is the representational similarity matrix (RSM): its (μ, ν) entry is the inner product between the population responses to stimuli x^μ and x^ν.

Let Π_{H} set the diagonal of a symmetric matrix to zero and subtract the mean of its off-diagonal entries. When $\Pi_{\mathsf{H}} M \neq 0$ and $\Pi_{\mathsf{H}} N \neq 0$, define the Pearson RSM-RSA score

$$\rho_{\mathrm{RSA}}(M, N) := \frac{\langle \Pi_{\mathsf{H}} M, \Pi_{\mathsf{H}} N \rangle_F}{\|\Pi_{\mathsf{H}} M\|_F \, \|\Pi_{\mathsf{H}} N\|_F}. \tag{48}$$

This is Pearson correlation between the strict upper-triangular RSM entries, a standard form of RSA (Kriegeskorte et al., 2008); Lemma B.3 of Theiss et al. (2026) gives the displayed cosine representation. Thus ρ_{RSA} is a similarity score, not an RSA distance.

Summary of Theiss *et al.* Theiss et al. study representations connected by a "catalog" of exact, function-preserving parameter symmetries. Although those parameter transformations have various forms, their action on hidden feature vectors reduces to three elementary operations: *addition* of redundant features, *duplication* of existing features, and featurewise *scaling*. Their Proposition 4.2 shows that every representation captured in the orbit of one such symmetry can be written in a single add–duplicate–scale form relative to a base member of that orbit. Their Proposition B.1 then separates the RSM into a *task-linked* part, whose generators come from the computationally relevant irreducible features of the base member, and a symmetry-induced part contributed by redundant added features. Scaling and duplication can still change the positive weights on the task-linked generators. Their Proposition C.3 ultimately shows that a minimum representation-norm or minimum weight-norm rule selects one RSM within a fixed orbit. The RSM decomposition itself originates in the linear-network analysis of Braun et al. (2025); Theiss et al. (2026) give the nonlinear symmetry-orbit form used here. The next lemma restates that decomposition in our notation, and we reproduce its short proof for completeness.

Lemma 20 (Theiss RSM decomposition)**.** *Let $H^\star \in \mathbb{R}^{m\times P}$ have rows $f_1^\top, \dots, f_m^\top$. Suppose H belongs to a Theiss symmetry orbit of $H^\star$, so that*

$$H = \operatorname{diag}(\alpha) D_\nu \begin{bmatrix} H^\star \\ U \end{bmatrix}, \tag{49}$$

where the rows $u_1^\top, \dots, u_K^\top$ of U are symmetry-added features, D_ν duplicates rows, and α contains the scales of the resulting physical coordinates. Then

$$R(H) = \sum_{j=1}^{m} \gamma_j f_j f_j^\top + \sum_{k=1}^{K} \eta_k u_k u_k^\top, \qquad \gamma_j, \eta_k > 0. \tag{50}$$

After applying Π_{H},

$$\begin{aligned} g := \Pi_{\mathsf{H}} R(H) = \underbrace{\sum_{j=1}^{m} \gamma_j q_j}_{\substack{\textit{task-linked} \\ \textit{essential-core part}}} + \underbrace{a}_{\substack{\textit{symmetry-induced} \\ \textit{added-feature part}}}, \\ q_j := \Pi_{\mathsf{H}}(f_j f_j^\top), \qquad a := \sum_{k=1}^{K} \eta_k \Pi_{\mathsf{H}}(u_k u_k^\top). \end{aligned} \tag{51}$$

Thus scaling and duplication reweight the original RSM generators q_j, whereas feature addition can introduce new generators.

Proof. Put

$$M := \begin{bmatrix} H^\star \\ U \end{bmatrix}.$$

Equation (49) gives

$$\begin{aligned} R(H) &= H^\top H \\ &= M^\top D_\nu^\top \operatorname{diag}(\alpha^{\odot 2}) D_\nu M. \end{aligned}$$

Every row of a duplication matrix contains exactly one entry equal to one. Consequently, two distinct columns of D_ν never have a nonzero entry in the same row. The weighted Gram matrix of D_ν is therefore diagonal:

$$D_\nu^\top \operatorname{diag}(\alpha^{\odot 2}) D_\nu = \operatorname{diag}(D_\nu^\top \alpha^{\odot 2}).$$

Let $\zeta := D_\nu^\top \alpha^{\odot 2}$. Every entry of ζ is positive, because it is the sum of the squared nonzero scales of all physical copies of one row of M. If $m_r^\top$ is row r of M, then

$$R(H) = M^\top \operatorname{diag}(\zeta) M = \sum_{r=1}^{m+K} \zeta_r m_r m_r^\top.$$

The first m rows of M are the $f_j^\top$ and the remaining K rows are the $u_k^\top$. Renaming the corresponding entries of ζ as γ_j and η_k proves (50). Applying the linear map Π_{H} term by term proves (51). □

Following Theiss et al. (2026), we call the first sum in (51) the *task-linked* (or task-relevant) part: its generators come from the irreducible feature directions of the base implementation, whereas a comes from features introduced only by function-preserving symmetries and is redundant to the represented layer function. The adjective "task-linked" does not mean that the weights γ_j are fixed by the task—scaling and duplication leave those weights free. We call the underlying task-linked directions the *essential core*. At this point, however, this is only the candidate essential core of one cataloged orbit. The rows of $H^\star$ are chosen from one irreducible representative, and Theiss et al. do not claim that every same-function realization lies in its orbit; a given function might in principle have two different irreducible representatives.

The next two subsections turn this orbit-relative decomposition into a function-level statement. The first proves an exact uniqueness result within the symmetries cataloged by Theiss et al: once the paper's existing minimality conditions are imposed, addition, duplication, and scaling do not produce a second minimal axis dictionary inside the same Theiss orbit. The second asks whether the same function can nevertheless have another minimal realization outside that orbit, and shows that such alternatives are generically exceptional. After these two steps, the first term of (51) be interpreted as a weighting of *the* essential core of the represented function.

E.1.2 Uniqueness within a Theiss symmetry orbit

Theiss et al. begin with one irreducible representative and describe all representations obtainable from it through their catalog of exact parameter symmetries. Their RSM decomposition therefore already suggests a distinguished set of original feature directions, but only relative to that chosen orbit. We first show that the existing minimality notions of this paper make that orbit-relative core exact and unambiguous.

Definition 31 (Minimal realization and minimal axis dictionary)**.** A *minimal realization* is a one-step expansion

$$G = \big(b, \{(v_j, g_j)\}_{j=1}^{k}\big), \qquad Y_G^{\varphi} = b + \sum_{j=1}^{k} v_j \varphi(g_j),$$

that satisfies the earlier activation-specific minimality condition: Softplus minimality from Definition 10, or ReLU EFF-minimality from Definition 12.

The *minimal axis dictionary* of G is the unordered set

$$\mathcal{J}_{\varphi}(G) := \{[(v_j, g_j)]_{\varphi} : 1 \le j \le k\}, \tag{52}$$

where Softplus identifies terms only up to permutation, while ReLU also identifies $(v, g) \sim (\alpha^{-1}v, \alpha g)$ for $\alpha > 0$.

Thus G is the complete one-step formula, including its bias and readout vectors, whereas $\mathcal{J}_{\varphi}(G)$ contains only the nonlinear axis/readout terms that remain in a minimal formula. Write $\mathcal{O}_{\mathrm{cat}}(G^{\star})$ for the set of realizations obtained from $G^{\star}$ by finite compositions of the exact symmetries cataloged by Theiss et al. The next proposition says that minimality gives one axis dictionary per such orbit.

Proposition 16 (Orbitwise uniqueness under the Theiss symmetries)**.** *Let $G^{\star}$ be a minimal ReLU or Softplus one-step realization. If $\widetilde{G} \in \mathcal{O}_{\mathrm{cat}}(G^{\star})$ is also minimal, then*

$$\mathcal{J}_{\varphi}(\widetilde{G}) = \mathcal{J}_{\varphi}(G^{\star}). \tag{53}$$

Hence every cataloged Theiss orbit has one minimal axis dictionary, up to permutation and the positive ReLU scale gauge.

Proof. Every realization in the cataloged orbit is obtained by composing feature permutation, duplication, addition, and scaling. It is therefore enough to check that none of these operations can create a second minimal dictionary.

Permutation. Reordering the summands changes neither the represented function nor the unordered set in (52).

Duplication and zero groups. Suppose q physical units share the same scalar argument g. Their total contribution is

$$\sum_{r=1}^{q} v_r \varphi(g) = \left(\sum_{r=1}^{q} v_r\right) \varphi(g).$$

Thus the copies represent one nonlinear term with aggregate readout $\sum_r v_r$. A minimal formula contains only that aggregate term; if the aggregate readout is zero, it contains no term at all. Duplication and zero groups therefore do not change the minimal dictionary.

Constant additions. If $g(x) \equiv c$ on the task set, then $v\varphi(g(x)) = v\varphi(c)$ is independent of x. It is absorbed into the bias, so it contributes no nonlinear axis to a minimal formula.

Aligned and opposite groups. Both activations considered here satisfy $\varphi(t) - \varphi(-t) = t$. Hence

$$\begin{aligned} v_+\varphi(g) + v_-\varphi(-g) &= v_+\varphi(g) + v_-\big(\varphi(g) - g\big) \\ &= (v_+ + v_-)\varphi(g) - v_- g. \end{aligned}$$

The first term is one nonlinear sign-class; the second is affine in the lower-layer argument and belongs to the affine part of the expansion. If $v_+ + v_- = 0$, the nonlinear term disappears. The linear and constant groups in the Theiss catalog are built from these same aligned/opposite and constant identities, and likewise leave at most one nonlinear class after minimality is imposed.

Positive ReLU scaling. For $\alpha > 0$, positive homogeneity gives

$$v\rho(g) = (\alpha^{-1}v)\rho(\alpha g).$$

These are two representatives of the same ReLU axis/readout class, which is why the equivalence relation in (52) identifies them.

Each cataloged primitive therefore preserves the minimal dictionary, and so does every finite composition of them. This proves (53). □

Proposition 16 says that the core suggested by the first sum in (51) is uniquely determined *within a fixed Theiss orbit*. Raw members of that orbit may still have very different RSMs—duplication changes multiplicity, scaling changes the weights of rank-one terms, and a canceling added feature contributes geometry even though it does not change the function—but all of them return to the same minimal axis dictionary.

The remaining question is larger in scope. The same represented function might, in principle, have another minimal realization that does not belong to the cataloged orbit of $G^\star$. The next subsection shows that such a second minimal core is generically absent.

E.1.3 Generic uniqueness beyond the Theiss catalog

Fix hidden-unit order and, for ReLU, the positive-scale normalization used in Appendix §B.2. For a represented function f, let

$$\mathcal{M}_\varphi(f) := \left\{ G_\vartheta : G_\vartheta \text{ is normalized and minimal, and } Y^\varphi_{G_\vartheta} = f \right\}. \tag{54}$$

This is simply the set of all normalized minimal one-step formulas for f; in geometric terminology it is a fiber over f, but no special use of that terminology is needed below. Proposition 16 says that each cataloged Theiss orbit contributes at most one normalized dictionary to this set. We now ask whether $\mathcal{M}_\varphi(f)$ can contain a second point outside that orbit.

There are two ways such nonuniqueness could occur. A family of alternatives might vary continuously with the original realization, in which case it is another continuous symmetry. Or the alternative might be a separate solution, possibly with a different width or parameterization type. Jacobian regularity handles the first case.

Proposition 17 (Continuous same-function families are Jacobian-singular)**.** *Suppose a nonconstant differentiable curve $t \mapsto \vartheta(t)$ of normalized minimal realizations represents the same function for every t:*

$$\mathcal{F}(\vartheta(t)) = \mathcal{F}(\vartheta(0)).$$

At every point where $\dot\vartheta(t) \neq 0$, the realization derivative has a nonzero kernel vector. Thus a continuous same-function symmetry must either be included in the gauge quotient or lie outside the Jacobian-regular locus.

Proof. Differentiate the identity $\mathcal{F}(\vartheta(t)) = \mathcal{F}(\vartheta(0))$ at a point, relabeled $t = 0$, at which $\dot\vartheta(0) \neq 0$. The chain rule gives

$$D\mathcal{F}(\vartheta(0))\dot\vartheta(0) = 0.$$

Thus the nonzero tangent vector $\dot\vartheta(0)$ lies in the kernel of the realization derivative. If the curve is a persistent structural symmetry, it belongs in the gauge quotient, just like hidden-unit permutation and positive ReLU scaling. If it has not been quotiented, the curve is contained in the Jacobian-singular locus. □

Consequently, after the ordinary gauges have been fixed, a Jacobian-regular realization cannot lie on an unaccounted continuous same-function path. The only remaining possibility is a separate minimal solution. Because such a solution may have another width, we allow a finite collection of normalized minimal-realization families $\{\Theta^\varphi_\tau\}_{\tau\in\mathfrak{T}}$, with realization maps

$$\mathcal{F}^\varphi_\tau(\vartheta) := Y^\varphi_{G_\vartheta}.$$

Here a chart is simply one smooth normalized parameter family of minimal one-step formulas.

Definition 32 (General position across minimal-realization families)**.** The collection $\{\Theta_\tau^\varphi\}_{\tau\in\mathfrak{T}}$ is in *general position across minimal-realization families* if the following holds. Take any two families and any smooth family of pairs $(\vartheta, \widetilde{\vartheta})$ satisfying

$$\mathcal{F}_\tau^\varphi(\vartheta) = \mathcal{F}_{\widetilde{\tau}}^\varphi(\widetilde{\vartheta}).$$

If the source parameters ϑ fill an open subset of the source family, then the pair family contains at least one point at which the existing activation-specific one-step identifiability theorem applies: Softplus MLR, or coherent ReLU MFR.

This is the cross-family version of Definition 22. The earlier condition says that a full-dimensional family of same-function pairs inside one chart cannot be trapped entirely in the exceptional set where the one-step signatures fail to identify the terms. Definition 32 asks for the same thing when the two minimal formulas may have different widths or belong to different parameter families.

The condition is natural because an open family of cross-family coincidences would be a robust reparameterization rule, not an accidental equality. One would expect such a rule to come from one of four sources:

(i) a genuine gauge that should be quotiented;

(ii) a systematic nonminimal construction, such as duplication or null-feature insertion;

(iii) an equivalence forced by the architecture, such as hard parameter tying or a built-in factorization;

(iv) a task patch too small to expose the nonlinear signatures needed for one-step identifiability.

The first possibility is removed by normalization and Proposition 17; Proposition 16 handles the second for the systematic Theiss constructions. Once those are removed, an ordinary independently parameterized architecture has no evident mechanism that would make a genuinely different minimal solution track every perturbation of the source network exactly. The local Softplus and ReLU genericity arguments support the same expectation because their failures are described by resonance equations or vanishing rank minors rather than by open sets.

This remains an assumption, not a theorem for every architecture. It can fail when hard parameter tying, grouped or multi-branch factorizations, fixed bottlenecks, or additional residual or attention gauges force alternative descriptions. With this definition in mind, however, we can establish generic uniqueness.

Theorem 13 (Generic uniqueness of the core)**.** *Assume the normalized minimal-realization families are Jacobian-regular and satisfy Definition 32. Fix a source family Θ_τ^φ of dimension d_τ, and write $f_\vartheta := Y_{G_\vartheta}^\varphi$. Then, outside a subset of Hausdorff dimension at most $d_\tau - 1$,*

$$G_{\widetilde{\vartheta}} \in \mathcal{M}_\varphi(f_\vartheta) \quad \Longrightarrow \quad \widetilde{\vartheta} = \vartheta. \tag{55}$$

Hence the exceptional set has empty interior and measure zero. Before fixing permutation and positive ReLU scale, the minimal realizations of a generic source function form one activation-gauge orbit.

Proof. Fix a source family Θ_τ^φ and a possible target family $\Theta_{\widetilde{\tau}}^\varphi$. The corresponding same-function collision set is

$$\mathcal{C}_{\tau,\widetilde{\tau}}^\varphi := \{(\vartheta, \widetilde{\vartheta}) : \mathcal{F}_\tau^\varphi(\vartheta) = \mathcal{F}_{\widetilde{\tau}}^\varphi(\widetilde{\vartheta})\}.$$

Call a pair *off-diagonal* when the two normalized realizations are different. By the stratified setup of Appendix §B.3, the off-diagonal collision set is a countable locally finite union of smooth strata. Let S be one such stratum and let

$$\pi_1 : S \to \Theta_\tau^\varphi, \qquad \pi_1(\vartheta, \widetilde{\vartheta}) = \vartheta,$$

be the source projection.

Step 1: the source projection is locally injective on tangent spaces. Take $(0, \dot{\widetilde{\vartheta}}) \in T_{(\vartheta,\widetilde{\vartheta})}S$. Choose a differentiable curve $t \mapsto (\vartheta(t), \widetilde{\vartheta}(t))$ in S with that tangent at $t = 0$. Along the curve,

$$\mathcal{F}_\tau^\varphi(\vartheta(t)) = \mathcal{F}_{\widetilde{\tau}}^\varphi(\widetilde{\vartheta}(t)).$$

Differentiating at $t = 0$ gives

$$D\mathcal{F}_{\tau}^{\varphi}(\vartheta)\dot{\vartheta} = D\mathcal{F}_{\widetilde{\tau}}^{\varphi}(\widetilde{\vartheta})\dot{\widetilde{\vartheta}}.$$

The source tangent is zero, so the left side vanishes. Jacobian regularity makes $D\mathcal{F}_{\widetilde{\tau}}^{\varphi}(\widetilde{\vartheta})$ injective; hence $\dot{\widetilde{\vartheta}} = 0$. Therefore

$$\ker(D\pi_1|_{TS}) = \{0\}.$$

The tangent map $D\pi_1|_{TS}$ is injective into the d_τ-dimensional source tangent space, and consequently

$$\dim S \le d_\tau. \tag{56}$$

Step 2: equality in (56) *is impossible.* Suppose $\dim S = d_\tau$. Then $D\pi_1|_{TS}$ is an isomorphism at every regular point. The inverse function theorem implies that, locally, S is the graph of a smooth same-function partner over an open subset of the source family. Thus there are an open set $V \subseteq \Theta_{\tau}^{\varphi}$ and a smooth map $\Psi : V \to \Theta_{\widetilde{\tau}}^{\varphi}$ such that

$$\mathcal{F}_{\widetilde{\tau}}^{\varphi}(\Psi(\vartheta)) = \mathcal{F}_{\tau}^{\varphi}(\vartheta) \qquad \text{for every } \vartheta \in V.$$

The source members of this collision family fill an open set. Definition 32 supplies one point at which the appropriate one-step identifiability theorem applies. At that point, equality of the represented functions forces the two minimal realizations to agree after permutation, and after positive coordinatewise scaling in the ReLU case. Those gauges have already been fixed, so the two normalized parameter values are equal. This contradicts that S is off-diagonal. Hence every off-diagonal stratum satisfies

$$\dim S \le d_\tau - 1.$$

Step 3: project to the exceptional source set. The smooth map π_1 is locally Lipschitz, so Hausdorff dimension cannot increase under projection:

$$\dim_{\mathrm{H}} \pi_1(S) \le \dim S \le d_\tau - 1.$$

Taking the countable locally finite union over all off-diagonal strata and all target families preserves the same Hausdorff-dimension bound. The resulting exceptional subset of the d_τ-dimensional source family has empty interior and Lebesgue measure zero.

Outside that set, G_ϑ has no different normalized minimal realization of the same function, which proves (55). Restoring permutation and positive ReLU scale turns the single normalized point into one activation-gauge orbit. □

Corollary 8 (Generic unique-core principle)**.** *For a generic source realization covered by Theorem 13, every minimal realization of its function f has the same minimal axis dictionary. We denote this common dictionary by $\mathcal{J}_\varphi(f)$ and call it the* essential core of f*. More explicitly, any two minimal realizations of f have the same number of terms and, after permutation,*

$$\widetilde{g}_j = g_j, \qquad \widetilde{v}_j = v_j \quad \textit{for Softplus},$$

whereas for ReLU there are $\alpha_j > 0$ such that

$$\widetilde{g}_j = \alpha_j g_j, \qquad \alpha_j \widetilde{v}_j = v_j. \tag{57}$$

On the evaluation set Ω_P, the nonzero projected rank-one terms of these axes determine the function-intrinsic set of privileged-axis RSM directions

$$\mathcal{D}_\varphi(f) := \{\mathbb{R}_{>0}\, \Pi_{\mathsf{H}}(f_j f_j^\top) : [(v_j, g_j)]_\varphi \in \mathcal{J}_\varphi(f)\}. \tag{58}$$

Thus the first sum in (51) *is, generically, a positive weighting of a uniquely determined set of RSM directions.*

Proof. Take two minimal realizations of the same generic function f and normalize their hidden-unit order and, for ReLU, their positive scales. If the normalized realizations were different, either would give an off-diagonal same-function partner forbidden by Theorem 13. Hence the normalized realizations are equal.

Undoing normalization leaves only the activation's unavoidable gauge. Softplus has no nontrivial positive-scale gauge. ReLU has $(v, g) \sim (\alpha^{-1}v, \alpha g)$, which gives (57). Positive rescaling multiplies $\Pi_{\mathsf{H}}(f_j f_j^\top)$ by a positive scalar without changing its ray, yielding (58). Lemma 20 then identifies the first sum in (51) as a positive weighting of representatives of these directions. □

Why the symmetry catalog need not be complete. Theorem 13 ranges over all normalized minimal-realization families, not only over realizations generated by the Theiss catalog. The following corollary is therefore a scope statement, not a second identifiability theorem: it shows explicitly that an off-catalog minimal partner is already included in the theorem's exceptional set.

Corollary 9 (Off-catalog collisions are part of the generic exceptional set)**.** *For source and target families, define*

$$\mathcal{C}^{\varphi}_{\mathrm{miss},\tau,\widetilde{\tau}} := \left\{ (\vartheta,\widetilde{\vartheta}) : \begin{array}{l} Y^{\varphi}_{G_\vartheta} = Y^{\varphi}_{G_{\widetilde{\vartheta}}}, \\ G_\vartheta, G_{\widetilde{\vartheta}} \text{ are normalized and minimal}, \\ G_{\widetilde{\vartheta}} \notin \mathcal{O}_{\mathrm{cat}}(G_\vartheta) \end{array} \right\}. \tag{59}$$

Under Theorem 13, if the source family has dimension d_τ, then

$$\dim_{\mathrm{H}} \pi_1 \left(\bigcup_{\widetilde{\tau}} \mathcal{C}^{\varphi}_{\mathrm{miss},\tau,\widetilde{\tau}} \right) \le d_\tau - 1. \tag{60}$$

Thus generic uniqueness of the essential core does not require a complete symmetry catalog.

Proof. Every pair in (59) is an off-diagonal normalized minimal same-function pair. It is therefore contained in the collision set analyzed in Theorem 13, and its source projection is contained in the theorem's exceptional source set. The dimension bound follows because a subset cannot have larger Hausdorff dimension than the set containing it. □

Remark 8 (Relation to the Vlačić–Bölcskei program). An exhaustive null-network classification would enumerate every exact realization-preserving identity and prove global identifiability after all such symmetries have been quotiented. The generic strategy used here asks for less. Known Theiss redundancies collapse to one minimal dictionary by Proposition 16; any missing continuous symmetry is another gauge or is Jacobian-singular by Proposition 17; and any remaining separate, off-catalog minimal partner is confined to the lower-dimensional exceptional set of Corollary 9. This is the same trade made by the generic null-net theorem earlier in the paper: generic uniqueness replaces exhaustive enumeration when the intended conclusion only requires a well-defined generic core.

E.1.4 The essential-core cone and raw RSA

Corollary 8 makes the positive part of the Theiss decomposition truly well-defined. Let f be a generic represented function, choose one nonzero sampled feature vector f_j from each class in $\mathcal{J}_\varphi(f)$, and put

$$q_j := \Pi_{\mathsf{H}}(f_j f_j^\top).$$

A different positive representative of the same ReLU class only rescales q_j. It therefore leaves unchanged the set of all nonnegative combinations of these directions. Define

$$Q_f : \mathbb{R}^m \to \mathsf{H}, \qquad Q_f\gamma := \sum_{j=1}^{m} \gamma_j q_j, \tag{61}$$

and the *essential-core cone*

$$\mathcal{C}_f := \mathrm{cone}\{q_1,\dots,q_m\} = \{Q_f\gamma : \gamma_j \ge 0\}. \tag{62}$$

Here a cone simply means the set of all nonnegative linear combinations of its generators. It contains all possible positive weightings of the function's privileged-axis RSM directions.

Proposition 18 (The function-intrinsic core cone)**.** *For a generic function f, the cone $\mathcal{C}_f$ is independent of the chosen minimal realization. Every raw realization obtained from a minimal realization of f through the cataloged Theiss operations has projected RSM*

$$g := \Pi_{\mathsf{H}} R(H) = Q_f\gamma + a, \qquad \gamma_j > 0, \tag{63}$$

where $Q_f\gamma \in \mathcal{C}_f$ is the weighted essential-core geometry and a is the projected contribution of redundant added features. Scaling and duplication move the core geometry within $\mathcal{C}_f$; an added feature enlarges the cone exactly when its projected rank-one term lies outside $\mathcal{C}_f$.

Moreover, if exact weak alignment holds on both sides of a nonlinear step and the one-step minimality and identifiability hypotheses apply, the two compared networks determine the same privileged-axis directions and hence the same cone.

Proof. Corollary 8 makes the essential classes, and therefore their positive RSM rays, independent of the minimal realization. Choosing different representatives merely rescales the generators and leaves their cone unchanged. Lemma 20 gives (63); Proposition 16 ensures that the cataloged duplicate, zero, constant, and aligned/opposite constructions return to the same minimal classes.

Changing scale or duplication count changes only γ, so $Q_f\gamma$ remains in $\mathcal{C}_f$. If an added feature has projected contribution $r_u := \Pi_{\mathsf{H}}(uu^\top)$, the enlarged cone is $\mathrm{cone}(\mathcal{C}_f \cup \{r_u\})$. This equals $\mathcal{C}_f$ when $r_u \in \mathcal{C}_f$ and is strictly larger otherwise.

Finally, exact weak alignment at the two adjacent layers gives equality of the transported one-step functions. The exact weak–strong and one-step identifiability results match the task-used minimal axes up to the activation gauge. Their privileged-axis RSM directions, and hence their cones, are therefore equal. □

Proposition 19 (RSA inside the essential-core cone)**.** *Suppose two representations have no added-feature geometry, so $g_A = Q_f\gamma_A$ and $g_B = Q_f\gamma_B$. Then if $K_{ij} := \langle q_i, q_j\rangle_F$ is the feature Gram matrix,*

$$\rho_{\mathrm{RSA}}(A,B) = \frac{\gamma_A^\top K \gamma_B}{\sqrt{\gamma_A^\top K \gamma_A}\sqrt{\gamma_B^\top K \gamma_B}}. \tag{64}$$

The score is one exactly when the two core geometries lie on the same positive ray in the cone.

Proof. By the definition of Pearson RSM–RSA as a cosine in projected-RSM space,

$$\rho_{\mathrm{RSA}}(A,B) = \cos(Q_f\gamma_A, Q_f\gamma_B).$$

Expanding the Frobenius inner products gives

$$\langle Q_f\gamma_A, Q_f\gamma_B\rangle_F = \sum_{i,j} \gamma_{Ai}\gamma_{Bj}\langle q_i, q_j\rangle_F = \gamma_A^\top K \gamma_B,$$

and similarly $\|Q_f\gamma_A\|_F^2 = \gamma_A^\top K \gamma_A$, which proves (64). Equality in Cauchy–Schwarz gives

$$Q_f\gamma_B = c\, Q_f\gamma_A \quad \text{for some } c > 0. \tag{65}$$

If the q_j are linearly independent, equality of the two weighted sums forces equality of their coefficients up to the same scalar c. □

The positive conclusion is that the represented function now determines a well-defined cone of core geometries. Different scale and multiplicity choices select different points in that cone, and RSA equals one precisely when the selected points lie on the same ray. This immediately exposes the remaining raw-RSA ambiguity.

Proposition 20 (Raw RSA is not forced by exact weak alignment)**.** *There is no universal lower bound on raw RSM-RSA that tends to one as adjacent weak-alignment errors tend to zero, even when the two reduced privileged-axis sets agree exactly. This is a direct consequence of the same scale, duplication, and feature-addition freedom identified by Theiss et al.*

Proof. Choose two nonzero feature vectors $f_1, f_2 \in \mathbb{R}^P$ such that

$$q_1 := \Pi_{\mathsf{H}}(f_1 f_1^\top), \qquad q_2 := \Pi_{\mathsf{H}}(f_2 f_2^\top)$$

are not collinear. Let network A contain one copy of each feature. Let network B_M contain M identical copies of f_1 and one copy of f_2. Split the original downstream readout vector of f_1 equally among the M copies. The sum of the M output contributions is unchanged, so A and B_M represent the same one-step function. Their minimal axis dictionaries are therefore the same $\{f_1, f_2\}$. Their weak error and reduced strong-axis error can therefore both be zero.

Their raw RSMs are

$$R_A = f_1 f_1^\top + f_2 f_2^\top, \qquad R_{B_M} = M f_1 f_1^\top + f_2 f_2^\top,$$

and hence

$$g_A = q_1 + q_2, \qquad g_{B_M} = M q_1 + q_2.$$

Because $M^{-1}g_{B_M} = q_1 + M^{-1}q_2 \to q_1$, continuity of normalization gives

$$\frac{g_{B_M}}{\|g_{B_M}\|_F} \longrightarrow \frac{q_1}{\|q_1\|_F}.$$

Therefore

$$\rho_{\mathrm{RSA}}(A, B_M) \longrightarrow \frac{\langle q_1 + q_2, q_1\rangle_F}{\|q_1 + q_2\|_F \, \|q_1\|_F}.$$

The limit is strictly less than one because $q_1 + q_2$ is not a positive multiple of q_1. Thus a sequence with identically zero weak error has raw RSA bounded away from one, which rules out any universal weak-error-only lower bound converging to one.

Duplication already proves the claim. Theiss feature addition gives still more freedom: if an added projected rank-one term lies outside the old cone, it creates a new RSM direction. Their Proposition B.4 shows that, as such generators are added, the attainable RSA scores against a fixed reference form nested intervals that can become very wide. Hence exact agreement of the essential axes does not determine a raw RSM. □

Proposition 20 is a no-go result for weak–strong equivalence for *raw* RSA, not for the core. Exact weak alignment, together with the usual minimality and identifiability hypotheses, determines the privileged-axis dictionary and the cone $\mathcal{C}_f$. It does not determine the weights γ or the redundant term a. A canonical RSA construction must therefore either remove this freedom or select a preferred point within it.

E.1.5 Two canonical RSM constructions

There are two natural ways to turn the now-well-defined essential core into one RSM. The *axis quotient* removes the remaining gauge: it gives each essential axis class one equal, unit-normalized vote. This is the simpler construction and is appropriate when the goal is to compare which privileged axes are present, independently of their scale, physical multiplicity, or downstream strength. *Minimum-norm balancing* instead selects a particular representative of the ReLU scale orbit. This is closer to the gauge-selection strategy of Theiss et al. (2026): it retains a gauge-invariant weighting based on how costly an axis is to generate and how strongly it is read out. The potential advantage is that two axes need not count equally when one plays a much larger role in the computation. The price is additional structure: one must choose a norm-based cost and common Euclidean coordinates in which that cost is evaluated.

Axis quotient: gauge removal.

Definition 33 (Axis-quotiented RSM and normalized RSA)**.** For each nonzero sampled feature vector $f_j \in \mathbb{R}^P$ associated with an essential axis class, define

$$P[f_j] := \frac{f_j f_j^\top}{\|f_j\|_2^2}.$$

If the essential core has m classes, define

$$R_{\mathrm{ax}}(G) := \frac{1}{m}\sum_{j=1}^{m} P[f_j], \qquad \rho_{\mathrm{ax}}(G, \widetilde{G}) := \rho_{\mathrm{RSA}}(R_{\mathrm{ax}}(G), R_{\mathrm{ax}}(\widetilde{G})). \tag{66}$$

The axis quotient gives one equal, scale-free contribution to each essential class, regardless of physical multiplicity.

Proposition 16 explains why the cataloged raw redundancies do not survive in this construction: duplicates represent one minimal class, zero groups disappear, constants enter the bias, and aligned/opposite groups leave at most one nonlinear class. The normalization in $P[f_j]$ removes the remaining positive ReLU gauge. Under Corollary 8, every minimal realization of the same generic function gives the same axis-quotiented RSM. A positive rescaling leaves each projector unchanged because

$$\frac{(\alpha f)(\alpha f)^\top}{\|\alpha f\|_2^2} = \frac{f f^\top}{\|f\|_2^2}.$$

Hence all minimal realizations produce the same average of projectors. The axis-quotiented RSM is thus well defined and may therefore write it as $R_{\mathrm{ax}}(f)$.

The key point is that R_{ax} does support Weak–Strong Equivalence:

Theorem 14 (Exact weak alignment implies axis-quotiented RSA equality)**.** *Suppose exact weak alignment holds at layers ℓ and $\ell+1$, and the induced one-step pair satisfies the minimality and identifiability hypotheses of Corollary 1. Then*

$$R_{\mathrm{ax}}(A) = R_{\mathrm{ax}}(B) \qquad \textit{and} \qquad \rho_{\mathrm{ax}}(A,B) = 1$$

whenever the common projected axis-quotiented RSM is nonzero. The same conclusion holds at every layer recovered by exact zippering.

Proof. Exact weak alignment at the two adjacent layers gives an equality of transported one-step functions by Lemma 1. Corollary 1 then gives a bijection of the minimal axes. After permutation, their sampled postactivation feature vectors satisfy

$$f_j^B = \alpha_j f_j^A, \qquad \alpha_j > 0,$$

with $\alpha_j = 1$ for Softplus. Therefore $P[f_j^B] = P[f_j^A]$ for every matched pair, and summing the projectors gives $R_{\mathrm{ax}}(A) = R_{\mathrm{ax}}(B)$. Identical nonzero projected RSMs have Pearson correlation one. Exact zippering supplies the same axis relation at every recovered layer. □

Minimum-norm balance: gauge selection. The axis quotient deliberately discards all relative weighting among essential axes. The minimum-representation-norm and minimum-weight-norm rules of Theiss et al. make a different choice: they select the ReLU scale at which the cost of producing an axis is balanced against the norm of the readout that uses it. The resulting RSM is more complicated, but it can retain information about computational importance. Its weights depend jointly on the size or cost of a generated feature and on the downstream readout that uses it, rather than giving every essential class one equal vote. This makes balancing attractive when the goal is not merely to compare the inventory of privileged axes, but to compare a norm-selected implementation of the computation. Unlike the axis quotient, however, it depends on the chosen MR or MW cost and on the Euclidean coordinates in which those norms are measured.

Definition 34 (Generation cost and minimum-norm balance)**.** For a ReLU minimal-core term (v_j, g_j), a *generation cost* is a positive number $s_j = s(g_j)$ satisfying

$$s(\alpha g_j) = \alpha^2 s(g_j), \qquad \alpha > 0. \tag{67}$$

It measures the squared norm assigned to producing the axis before the outgoing readout is applied. The two choices used by Theiss et al. are

$$s_j^{\mathrm{MR}} = \|h_j\|_2^2, \qquad h_j := \big(\rho(g_j(x^1)), \ldots, \rho(g_j(x^P))\big)^\top, \tag{68}$$

$$s_j^{\mathrm{MW}} = \|w_j\|_2^2 + b_j^2, \qquad g_j(x) = w_j^\top x + b_j. \tag{69}$$

The first is the feature-generation part of the minimum representation-norm parameterization (MRNP) objective; the second is the corresponding part of the minimum weight-norm parameterization (MWNP) objective.

Along the ReLU gauge $(v_j, g_j) \mapsto (\alpha_j^{-1} v_j, \alpha_j g_j)$, define

$$J_s(\alpha) := \sum_{j=1}^{m} \left(\alpha_j^2 s_j + \alpha_j^{-2} \|v_j\|_2^2\right). \tag{70}$$

A minimal core is *balanced* when its scales minimize J_s. All generation costs and readout norms are evaluated in the Euclidean coordinates in which the supplied one-step expansion is written.

Making an axis larger increases its generation cost but decreases the norm of its inverse-scaled readout; making it smaller does the opposite. Balance selects the unique point where these two pressures agree.

Theorem 15 (Balanced essential core and canonical RSM)**.** *Let f be a represented ReLU function whose minimal realizations form one permutation-and-positive-scale orbit, as Theorem 13 guarantees generically. Let $G = (b, \{(v_j, g_j)\}_{j=1}^m)$ be any minimal representative, with $s_j > 0$ and $v_j \neq 0$. Then* (70) *has a unique scale minimizer, with effective RSM weights*

$$\gamma_j^\star = (\alpha_j^\star)^2 = \frac{\|v_j\|_2}{\sqrt{s_j}}. \tag{71}$$

The resulting canonical RSM

$$R_s^{\mathrm{bal}}(f) := \sum_{j=1}^m \gamma_j^\star h_j h_j^\top \tag{72}$$

is independent of the representative chosen from the ReLU scale orbit.

Proof. By hypothesis, every minimal realization of f belongs to the same permutation-and-positive-scale orbit, so it is enough to minimize J_s along that orbit.

The objective separates over axes:

$$J_s(\alpha) = \sum_{j=1}^m J_j(\alpha_j), \qquad J_j(\alpha_j) = \alpha_j^2 s_j + \alpha_j^{-2}\|v_j\|_2^2.$$

Set $\gamma = \alpha_j^2 > 0$. Then

$$F_j(\gamma) = \gamma s_j + \gamma^{-1}\|v_j\|_2^2.$$

Because $s_j > 0$ and $v_j \neq 0$,

$$\lim_{\gamma\downarrow 0} F_j(\gamma) = +\infty, \qquad \lim_{\gamma\uparrow\infty} F_j(\gamma) = +\infty,$$

so a minimum exists in $(0, \infty)$. Differentiating gives

$$F_j'(\gamma) = s_j - \gamma^{-2}\|v_j\|_2^2, \qquad F_j''(\gamma) = 2\gamma^{-3}\|v_j\|_2^2 > 0.$$

The positive second derivative makes F_j strictly convex, hence its critical point is the unique minimizer. Solving $F_j'(\gamma) = 0$ yields

$$\gamma_j^\star = \frac{\|v_j\|_2}{\sqrt{s_j}},$$

which proves (71). Since the full objective is a sum of independent strictly convex terms, the vector of optimal scales is unique.

Under scale α_j, positive homogeneity gives the post-ReLU sample vector $\alpha_j h_j$. Its RSM contribution is

$$(\alpha_j h_j)(\alpha_j h_j)^\top = \alpha_j^2 h_j h_j^\top = \gamma_j h_j h_j^\top.$$

Substituting the optimal weights and summing proves (72).

It remains to show that the result does not depend on the initial representative of the ReLU scale orbit. Let another representative be

$$\widetilde{h}_j = \beta_j h_j, \qquad \widetilde{v}_j = \beta_j^{-1} v_j, \qquad \widetilde{s}_j = \beta_j^2 s_j, \qquad \beta_j > 0.$$

Its optimal effective weight is

$$\begin{aligned}\widetilde{\gamma}_j^\star &= \frac{\|\widetilde{v}_j\|_2}{\sqrt{\widetilde{s}_j}} \\ &= \frac{\beta_j^{-1}\|v_j\|_2}{\beta_j\sqrt{s_j}} \\ &= \beta_j^{-2}\gamma_j^\star.\end{aligned}$$

Therefore

$$\widetilde{\gamma}_j^\star \widetilde{h}_j \widetilde{h}_j^\top = \beta_j^{-2}\gamma_j^\star(\beta_j h_j)(\beta_j h_j)^\top = \gamma_j^\star h_j h_j^\top.$$

Every rank-one contribution is unchanged, and so is their sum. The single-orbit premise therefore makes this representative-independent matrix common to every minimal realization of f. □

Theorem 15 extends Theiss Proposition C.3 from a fixed cataloged orbit to the generically unique essential core. Its extra content is that the selected RSM no longer depends on assuming in advance that the chosen Theiss orbit exhausts all minimal realizations of the function.

Corollary 10 (Balanced RSM as a computationally weighted axis quotient)**.** *Under Theorem 15, the MR-balanced RSM is*

$$R_{\mathrm{BMR}}(f) = \sum_{j=1}^{m} \|v_j\|_2 \|h_j\|_2 \frac{h_j h_j^\top}{\|h_j\|_2^2}, \tag{73}$$

and the MW-balanced RSM is

$$R_{\mathrm{BMW}}(f) = \sum_{j=1}^{m} \frac{\|v_j\|_2 \|h_j\|_2^2}{\sqrt{\|w_j\|_2^2 + b_j^2}} \frac{h_j h_j^\top}{\|h_j\|_2^2}. \tag{74}$$

Thus R_{ax} gives one equal vote per essential axis, whereas the balanced RSM gives one gauge-invariant vote weighted by the axis's readout strength relative to its generation cost.

Proof. For MR balance, $s_j = \|h_j\|_2^2$, so Theorem 15 gives $\gamma_j^\star = \|v_j\|_2/\|h_j\|_2$. Therefore

$$\begin{aligned} \gamma_j^\star h_j h_j^\top &= \frac{\|v_j\|_2}{\|h_j\|_2} h_j h_j^\top \\ &= \|v_j\|_2 \|h_j\|_2 \frac{h_j h_j^\top}{\|h_j\|_2^2}. \end{aligned}$$

Summing proves (73).

For MW balance, $\gamma_j^\star = \|v_j\|_2/\sqrt{\|w_j\|_2^2 + b_j^2}$. Rewriting $h_j h_j^\top$ as $\|h_j\|_2^2 (h_j h_j^\top/\|h_j\|_2^2)$ gives (74). Gauge invariance follows from the representative-independence proved in Theorem 15. □

The preceding corollary makes the comparison with the axis quotient explicit. Both are averages of the same normalized rank-one projectors. The axis quotient gives each projector weight $1/m$; the balanced construction assigns a gauge-invariant weight determined by readout strength and generation cost. The latter can therefore retain a notion of computational importance that the equal-vote quotient intentionally removes.

Definition 35 (Common-coordinate one-step comparison)**.** Let G_A and G_B be adjacent one-step expansions, and let $E_\ell, E_{\ell+1}$ be their weak-comparison maps. Using the output action EG and input pullback E^*G from Definition 3, define

$$G_A^{\mathrm{com}} := E_{\ell+1} G_A, \qquad G_B^{\mathrm{com}} := E_\ell^* G_B. \tag{75}$$

These are the two one-step formulas written with the same layer-ℓ input coordinate and the same layer-$(\ell+1)$ output coordinate.

Theorem 16 (Exact weak alignment forces equality of balanced RSMs in common coordinates)**.** *Suppose exact weak alignment holds at layers ℓ and $\ell+1$. Let $(G_A^{\mathrm{com}}, G_B^{\mathrm{com}})$ be the common-coordinate pair in Definition 35. If this pair satisfies the minimality and identifiability hypotheses of Corollary 1, then, for either $s = s^{\mathrm{MR}}$ or $s = s^{\mathrm{MW}}$,*

$$R_s^{\mathrm{bal}}(G_A^{\mathrm{com}}) = R_s^{\mathrm{bal}}(G_B^{\mathrm{com}}). \tag{76}$$

Consequently their Pearson RSM–RSA score is one whenever the common projected RSM is nonzero.

Proof. Write the two original expansions and comparison maps as

$$G_A = \big(b_A, \{(v_j^A, g_j^A)\}_{j=1}^{m_A}\big), \qquad G_B = \big(b_B, \{(v_i^B, g_i^B)\}_{i=1}^{m_B}\big),$$
$$E_r(z) = T_r z + a_r, \qquad r \in \{\ell, \ell+1\}.$$

Unpacking Definition 35 gives

$$\begin{aligned} G_A^{\mathrm{com}} &= \big(T_{\ell+1} b_A + a_{\ell+1}, \{(T_{\ell+1} v_j^A, g_j^A)\}_{j=1}^{m_A}\big), \\ G_B^{\mathrm{com}} &= \big(b_B, \{(v_i^B, g_i^B \circ E_\ell)\}_{i=1}^{m_B}\big). \end{aligned}$$

For $z \in C$, exact weak alignment and Lemma 1 give

$$\begin{aligned} Y^{\rho}_{G^{\mathrm{com}}_B}(z) &= Y^{\rho}_{G_B}(E_\ell z) \\ &= E_{\ell+1}\big(Y^{\rho}_{G_A}(z)\big) \\ &= Y^{\rho}_{G^{\mathrm{com}}_A}(z). \end{aligned}$$

Hence

$$Y^{\rho}_{G^{\mathrm{com}}_A}(z) = Y^{\rho}_{G^{\mathrm{com}}_B}(z), \qquad z \in C.$$

The two expansions therefore represent exactly the same function of the same input variable and take values in the same output Euclidean space. Corollary 1 applies directly to this equality.

For brevity, write

$$\widehat{v}^A_j := T_{\ell+1} v^A_j, \qquad \widehat{g}^B_i := g^B_i \circ E_\ell.$$

Corollary 1 implies that $m_A = m_B =: m$ and, after a permutation of the B terms, there are positive scales λ_j such that

$$\widehat{g}^B_j = \lambda_j g^A_j, \qquad \lambda_j v^B_j = \widehat{v}^A_j. \tag{77}$$

Thus the matching of all counted axes and readouts is a conclusion of exact weak alignment plus one-step identifiability; it is not an additional assumption.

Let h^A_j and $\widehat{h}^B_j$ be the post-ReLU sample vectors generated by g^A_j and $\widehat{g}^B_j$ on the common input sample. Positive homogeneity and (77) give

$$\widehat{h}^B_j = \rho(\widehat{g}^B_j) = \rho(\lambda_j g^A_j) = \lambda_j h^A_j. \tag{78}$$

The readout relation gives

$$v^B_j = \lambda_j^{-1} \widehat{v}^A_j, \qquad \|v^B_j\|_2 = \lambda_j^{-1} \|\widehat{v}^A_j\|_2. \tag{79}$$

Both readout vectors in this equation lie in the common B-side output coordinates by construction.

We next compute the generation costs in the common input coordinates. For MRNP,

$$s^{\mathrm{MR}}_{Aj} = \|h^A_j\|_2^2, \qquad s^{\mathrm{MR,com}}_{Bj} = \|\widehat{h}^B_j\|_2^2 = \lambda_j^2 s^{\mathrm{MR}}_{Aj}.$$

For MWNP, write the original B-side affine argument as

$$g^B_i(y) = (w^B_i)^\top y + \beta^B_i.$$

Since $E_\ell(z) = T_\ell z + a_\ell$, its pullback appearing in G^{com}_B is

$$\begin{aligned} \widehat{g}^B_i(z) &= g^B_i(E_\ell z) \\ &= (w^B_i)^\top (T_\ell z + a_\ell) + \beta^B_i \\ &= (T_\ell^\top w^B_i)^\top z + \big((w^B_i)^\top a_\ell + \beta^B_i\big). \end{aligned}$$

Therefore its common-coordinate MW generation cost is

$$s^{\mathrm{MW,com}}_{Bi} = \|T_\ell^\top w^B_i\|_2^2 + \big((w^B_i)^\top a_\ell + \beta^B_i\big)^2. \tag{80}$$

Now write

$$g^A_j(z) = (w^A_j)^\top z + \beta^A_j, \qquad \widehat{g}^B_j(z) = (\widehat{w}^B_j)^\top z + \widehat{\beta}^B_j.$$

The first relation in (77) implies

$$(\widehat{w}^B_j, \widehat{\beta}^B_j) = \lambda_j (w^A_j, \beta^A_j).$$

Using (80), this yields

$$s^{\mathrm{MW,com}}_{Bj} = \|\widehat{w}^B_j\|_2^2 + (\widehat{\beta}^B_j)^2 = \lambda_j^2\big(\|w^A_j\|_2^2 + (\beta^A_j)^2\big) = \lambda_j^2 s^{\mathrm{MW}}_{Aj}.$$

Thus, for either balancing rule,

$$s^{\mathrm{com}}_{Bj} = \lambda_j^2 s_{Aj}. \tag{81}$$

The per-axis balancing calculation of Theorem 15 gives the effective weight

$$\gamma_j^\star = \frac{\|v_j\|_2}{\sqrt{s_j}}.$$

Define

$$\gamma_{Aj}^\star := \frac{\|\widehat{v}_j^A\|_2}{\sqrt{s_{Aj}}}, \qquad \gamma_{Bj}^\star := \frac{\|v_j^B\|_2}{\sqrt{s_{Bj}^{\mathrm{com}}}}.$$

Using (79) and (81),

$$\begin{aligned}\gamma_{Bj}^\star &= \frac{\lambda_j^{-1}\|\widehat{v}_j^A\|_2}{\sqrt{\lambda_j^2 s_{Aj}}} \\ &= \lambda_j^{-2}\gamma_{Aj}^\star.\end{aligned}$$

Combining this identity with (78), the matched B-side balanced contribution is

$$\begin{aligned}\gamma_{Bj}^\star \widehat{h}_j^B(\widehat{h}_j^B)^\top &= \lambda_j^{-2}\gamma_{Aj}^\star(\lambda_j h_j^A)(\lambda_j h_j^A)^\top \\ &= \gamma_{Aj}^\star h_j^A(h_j^A)^\top.\end{aligned}$$

The equality holds term by term after the identified permutation. Summing over $j = 1, \dots, m$ proves (76). If the common projected RSM is nonzero, its Pearson correlation with itself is one. □

The axis quotient and the balanced construction therefore solve different versions of the same problem. The quotient is the cleaner consequence of strong-axis identification and is insensitive to how heavily each axis participates in the computation. Minimum-norm balance is closer to the Theiss implementation-selection program: it chooses a particular gauge and preserves a norm-dependent weighting of computational importance, at the cost of a more elaborate and coordinate-dependent comparison.

E.1.6 Soft weak alignment, canonical RSA, and ridge prediction

Theorem 14 says that exact weak alignment determines the axis-quotiented RSM. As before, we would like to have a soft version of this statement. To achieve this, we need one elementary fact: nearby projected RSMs have nearby cosine similarity:

Lemma 21 (Cosine stability for projected RSMs)**.** *Let x and y be nonzero elements of a Frobenius inner-product space, with $\min\{\|x\|_F, \|y\|_F\} \geq \beta > 0$. Then*

$$1 - \frac{\langle x, y\rangle_F}{\|x\|_F\|y\|_F} \leq \frac{2\|x - y\|_F^2}{\beta^2}. \tag{82}$$

More specifically, if $g_A = g^\star + e_A$ and $g_B = g^\star + e_B$, and $r := \max\{\|e_A\|_F, \|e_B\|_F\} < \|g^\star\|_F$, then

$$\rho_{\mathrm{RSA}}(g_A, g_B) \geq 1 - \frac{2\|e_A - e_B\|_F^2}{\big(\|g^\star\|_F - r\big)^2}. \tag{83}$$

Proof. For nonzero vectors x and y in any Hilbert space,

$$1 - \frac{\langle x, y\rangle}{\|x\|\|y\|} = \frac{1}{2}\left\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\right\|^2. \tag{84}$$

We first bound the distance between the normalized vectors. Assume without loss of generality that $\|x\| \leq \|y\|$. By the triangle inequality,

$$\begin{aligned}\left\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\right\| &\leq \left\|\frac{x - y}{\|x\|}\right\| + \left\|y\left(\frac{1}{\|x\|} - \frac{1}{\|y\|}\right)\right\| \\ &= \frac{\|x - y\|}{\|x\|} + \frac{|\|y\| - \|x\||}{\|x\|}\end{aligned}$$

$$\leq \frac{2\|x-y\|}{\|x\|},$$

where the last line uses the reverse triangle inequality. Since $\|x\| = \min\{\|x\|, \|y\|\} \geq \beta$, substitution into (84) gives

$$1 - \cos(x,y) \leq \frac{1}{2}\left(\frac{2\|x-y\|}{\beta}\right)^2 = \frac{2\|x-y\|^2}{\beta^2},$$

which proves (82).

For the second statement, put $x = g_A = g^\star + e_A$ and $y = g_B = g^\star + e_B$. Then $x - y = e_A - e_B$. The reverse triangle inequality gives

$$\|g_A\|_F \geq \|g^\star\|_F - \|e_A\|_F \geq \|g^\star\|_F - r,$$

and the same lower bound holds for g_B. Apply the first part with $\beta = \|g^\star\|_F - r$ to obtain (83). □

Recall the proof of Appendix Theorem 3. For a threshold $\theta < 1$, it lets B_θ be the number of task-required axes that fail the best injective θ-match, and proves

$$B_\theta \leq \frac{e_{\mathrm{weak}}^2}{\kappa(\theta)^2}, \qquad e_{\mathrm{weak}} := \|W_{\ell+1}^B\|_{\mathrm{op}}\omega_\ell + \omega_{\ell+1}. \tag{85}$$

Theorem 17 (Soft weak alignment implies an axis-quotiented RSA bound)**.** *Fix a compact comparison family satisfying Appendix Theorem 3. Suppose the two reduced representations have the same m task-required axes, and compute the soft-axis score and R_{ax} from the same empirical axis vectors on Ω_P. If both projected axis-quotiented RSMs have Frobenius norm at least $\beta > 0$, then for every $\theta \in (0,1)$,*

$$\rho_{\mathrm{ax}}(A,B) \geq 1 - \frac{2}{\beta^2}\left[\frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}} + \frac{\sqrt{2}}{m}\frac{e_{\mathrm{weak}}^2}{\kappa(\theta)^2}\right]^2. \tag{86}$$

Proof. The recalled weak–strong estimate (85) controls how many axes can fail the threshold. We first show how one successful threshold match controls one normalized rank-one projector.

Let $u, v \in \mathbb{R}^P$ be a good matched pair. By the definition of the soft-axis score, there is an allowed nonzero scalar α such that

$$\frac{\|u - \alpha v\|_2}{\|u\|_2 + |\alpha|\|v\|_2} \leq \delta, \qquad \delta := 1 - \theta < 1. \tag{87}$$

Set $a = \|u\|_2$, $b = |\alpha|\|v\|_2$, and let c be the cosine between u and αv. Then

$$\|u - \alpha v\|_2^2 = a^2 + b^2 - 2abc.$$

Squaring (87) gives

$$a^2 + b^2 - 2abc \leq \delta^2(a+b)^2.$$

Also,

$$(a+b)^2 - \|u - \alpha v\|_2^2 = 2ab(1+c).$$

Using (87) and $1 + c \leq 2$,

$$4ab \geq 2ab(1+c) \geq (1-\delta^2)(a+b)^2,$$

so

$$\frac{(a+b)^2}{ab} \leq \frac{4}{1-\delta^2}. \tag{88}$$

Now

$$\begin{aligned} 1 - c &= \frac{\|u - \alpha v\|_2^2 - (a-b)^2}{2ab} \\ &\leq \frac{\|u - \alpha v\|_2^2}{2ab} \\ &\leq \frac{\delta^2(a+b)^2}{2ab} \end{aligned}$$

$$\leq \frac{2\delta^2}{1-\delta^2},$$

where the last line uses (88).

The normalized projector of a nonzero vector w is $P[w] = ww^\top/\|w\|_2^2$. A direct calculation gives

$$\|P[u] - P[v]\|_F^2 = 2(1-c^2) \leq 4(1-c).$$

Combining with the previous bound yields

$$\|P[u] - P[v]\|_F \leq \frac{2\sqrt{2}\delta}{\sqrt{1-\delta^2}} = \frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}}.$$

The sign of α is irrelevant because $P[\alpha v] = P[v]$.

Choose the injective matching used in Appendix Theorem 3. There are $m - B_\theta$ good pairs and B_θ bad pairs. For a bad pair, each normalized projector has Frobenius norm one, and

$$\|P[u] - P[v]\|_F^2 = 2\big(1 - \langle \widehat{u}, \widehat{v}\rangle^2\big) \leq 2,$$

so the distance is at most $\sqrt{2}$. Averaging the projector differences and using the triangle inequality gives

$$\begin{aligned}
\|R_{\mathrm{ax}}(A) - R_{\mathrm{ax}}(B)\|_F &\leq \frac{1}{m}\left[(m - B_\theta)\frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}} + B_\theta\sqrt{2}\right] \\
&\leq \frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}} + \frac{\sqrt{2}B_\theta}{m} \\
&\leq \frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}} + \frac{\sqrt{2}}{m}\frac{e_{\mathrm{weak}}^2}{\kappa(\theta)^2},
\end{aligned}$$

where the final line uses (85). Orthogonal projection is nonexpansive, so the same bound holds after applying Π_{H}.

Let $x = \Pi_{\mathsf{H}} R_{\mathrm{ax}}(A)$ and $y = \Pi_{\mathsf{H}} R_{\mathrm{ax}}(B)$. By assumption, both norms are at least β. Apply Lemma 21 to the preceding RSM-distance bound. This gives exactly (86). □

Corollary 11 (Asymptotic weak alignment implies normalized RSA convergence)**.** *Under the hypotheses of Theorem 17, if $\omega_\ell \to 0$ and $\omega_{\ell+1} \to 0$ along a sequence in the compact family, then*

$$\rho_{\mathrm{ax}}(A, B) \longrightarrow 1.$$

Proof. The adjacent weak error e_{weak} in (85) tends to zero because the compact family uniformly bounds $\|W_{\ell+1}^B\|_{\mathrm{op}}$. Fix $\theta < 1$. Then the integer B_θ is bounded above by a quantity tending to zero, so $B_\theta = 0$ for all sufficiently large members of the sequence. At those indices, the projected axis-quotient distance is at most

$$\frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}}.$$

Given any tolerance, first choose θ sufficiently close to one that this quantity is small enough, and then choose the sequence index large enough that $B_\theta = 0$. Lemma 21 then makes the RSA loss arbitrarily small. Hence $\rho_{\mathrm{ax}} \to 1$. □

A local ridge–canonical-RSA prediction. The soft theorem describes a one-sided frontier rather than a deterministic regression curve (see Fig. 17A). A simple local form is useful for empirical interpretation. Let R^2_{ridge} be the held-out squared correlation of a theorem-compatible linear comparison: the map is injective on the reduced core, or the comparison has been made effectively bidirectional or rank-matched. Write its normalized prediction residual as

$$R^2_{\mathrm{ridge}} = 1 - \varepsilon^2. \tag{89}$$

Corollary 12 (Local ridge–canonical-RSA relation)**.** *Let c_m and $c_\star$ be the projected canonical RSMs of a comparison model and a target. Suppose that, on a locally identifiable compact family,*

$$\|c_m - c_\star\|_F \le L_{\mathrm{can}} c_{\mathrm{w}} \varepsilon, \qquad \min\{\|c_m\|_F, \|c_\star\|_F\} \ge \beta > 0,$$

where $c_{\mathrm{w}}\varepsilon$ bounds the theorem-compatible adjacent weak defect and L_{can} converts that defect into canonical-RSM error. Then

$$1 - \rho_{\mathrm{canonical}}(m, \star) \le C\big(1 - R^2_{\mathrm{ridge}}\big), \qquad C := \frac{2L^2_{\mathrm{can}} c^2_{\mathrm{w}}}{\beta^2}. \tag{90}$$

Here $\rho_{\mathrm{canonical}}$ may be the axis-quotiented score, or the balanced score when its weights vary Lipschitz-continuously on the family.

Proof. By definition,

$$\rho_{\mathrm{canonical}}(m, \star) = \frac{\langle c_m, c_\star \rangle_F}{\|c_m\|_F \|c_\star\|_F}.$$

Lemma 21 gives

$$1 - \rho_{\mathrm{canonical}}(m, \star) \le \frac{2\|c_m - c_\star\|^2_F}{\beta^2}.$$

The assumed local weak-to-canonical estimate therefore implies

$$\begin{aligned} 1 - \rho_{\mathrm{canonical}}(m, \star) &\le \frac{2L^2_{\mathrm{can}} c^2_{\mathrm{w}}}{\beta^2} \varepsilon^2 \\ &= C\big(1 - R^2_{\mathrm{ridge}}\big), \end{aligned}$$

where the last line uses (89). □

Equation (90) becomes especially transparent when the canonical and raw RSMs are written side by side. Define

$$c_m := \Pi_{\mathsf{H}} R_{\mathrm{can}}(m), \qquad g_m := \Pi_{\mathsf{H}} R(H_m) = c_m + s_m, \tag{91}$$

where R_{can} is the axis-quotiented or balanced essential RSM. Then

$$\rho_{\mathrm{canonical}}(m, \star) = \cos(c_m, c_\star), \qquad \rho_{\mathrm{raw}}(m, \star) = \cos(c_m + s_m, c_\star + s_\star). \tag{92}$$

The residual s_m has two conceptually different pieces:

$$s_m = \underbrace{\big(g_m^{\mathrm{core}} - c_m\big)}_{\substack{\text{noncanonical scale} \\ \text{and multiplicity}}} + \underbrace{a_m}_{\substack{\text{redundant added-feature} \\ \text{geometry}}}. \tag{93}$$

The ridge bound controls the first cosine in (92) through the difference $c_m - c_\star$. It does not control $s_m - s_\star$. Hence canonical RSA has the one-sided prediction

$$1 - \rho_{\mathrm{canonical}} \lesssim C(1 - R^2_{\mathrm{ridge}}),$$

while raw RSA need not follow any single curve. At exact weak equivalence, the canonical cores can coincide, $c_m = c_\star$, so $\rho_{\mathrm{canonical}} = 1$. Theiss scaling and duplication can nevertheless change the first term of (93), and redundant feature addition can change the second. The same privileged axes can therefore support a broad horizontal range of raw RSA values at $R^2_{\mathrm{ridge}} = 1$.

When should models lie near the canonical frontier? The constant C in (90) should be understood as a worst-case upper envelope for the local comparison family, not as the slope that every empirical model is expected to attain. A sharper local description follows from ordinary smoothness. Let $\mathcal{M}$ be a finite-dimensional local family of comparison models with the target $\star$ as an interior point, and write

$$F(m) := 1 - R^2_{\mathrm{ridge}}(m, \star), \qquad c(m) := c_m.$$

Assume that F is C^2 and c is C^1 near $\star$. This is the relevant local regularity condition: it holds when the fitted linear comparison and its normalized prediction residual vary smoothly with the model, for example on a constant-rank least-squares family with a locally unique fitted map. It can fail at rank changes or when model selection causes a regularization parameter to jump.

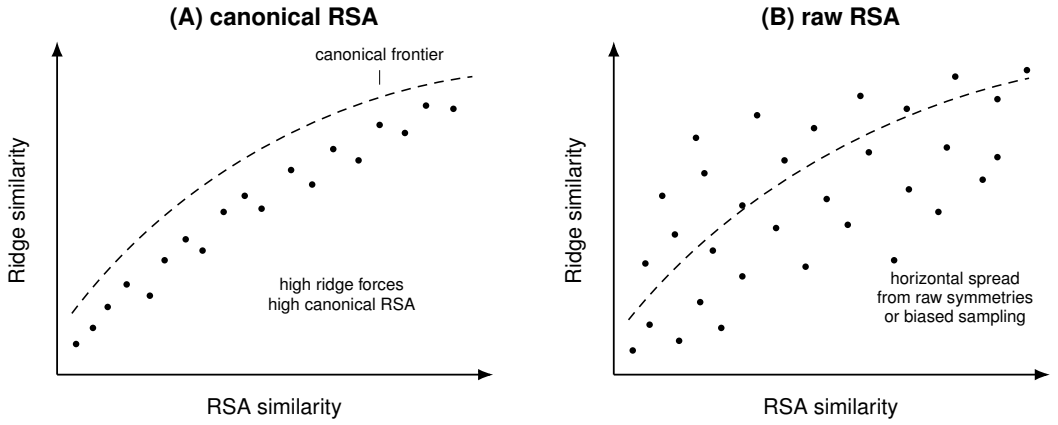


Figure 17: Schematic model-by-model scatter plots. **(A)** The canonical ridge–RSA relation is one-sided: sufficiently high ridge similarity forces high canonical RSA, so the comparison models lie below an increasing frontier. **(B)** Raw symmetries or unequal spatial sampling can reweight the observed geometry without comparably changing ridge predictivity, producing additional horizontal spread around the underlying trend.

In the present model–target setup, the target compared with itself has zero prediction residual, so $F(\star) = 0$; moreover $F \geq 0$. Hence $\star$ is a local minimum of F and $DF(\star) = 0$. For any smooth two-sided path $m(t) \in \mathcal{M}$ with $m(0) = \star$ and tangent $h := \dot{m}(0) \in T_\star\mathcal{M}$, Taylor's theorem therefore gives

$$1 - R^2_{\mathrm{ridge}}(m(t), \star) = a_{\mathrm{ridge}}(h)t^2 + o(t^2), \qquad a_{\mathrm{ridge}}(h) := \frac{1}{2}D^2F(\star)[h, h] \geq 0. \tag{94}$$

Thus the quadratic ridge expansion is a consequence of local smoothness, not an additional assumption. Strict positivity is a nondegeneracy property of the direction h: if $a_{\mathrm{ridge}}(h) = 0$, ridge is locally insensitive to that direction to second order.

Let

$$v := Dc(\star)[h],$$

so that the projected canonical RSM satisfies

$$c_{m(t)} = c_\star + tv + o(t).$$

Write $v = v_\parallel + v_\perp$, where $v_\parallel$ is parallel to $c_\star$ and $v_\perp$ is Frobenius-orthogonal to it. The local cosine expansion gives

$$1 - \rho_{\mathrm{canonical}}(m(t), \star) = \frac{\|v_\perp\|_F^2}{2\|c_\star\|_F^2}\, t^2 + o(t^2). \tag{95}$$

The local weak-to-canonical estimate used in Corollary 12 further implies that any direction with $v \neq 0$ must have $a_{\mathrm{ridge}}(h) > 0$. Indeed, substituting the local expansions into

$$\|c_{m(t)} - c_\star\|_F \leq L_{\mathrm{can}} c_{\mathrm{w}} \sqrt{1 - R^2_{\mathrm{ridge}}(m(t), \star)}$$

and dividing by $|t|$ gives

$$\|v\|_F \leq L_{\mathrm{can}} c_{\mathrm{w}} \sqrt{a_{\mathrm{ridge}}(h)}.$$

Thus a direction that changes the canonical RSM to first order cannot be invisible to ridge at quadratic order. Directions with $a_{\mathrm{ridge}}(h) = 0$ necessarily also have $v = 0$ and require a higher-order analysis.

For every direction with $a_{\mathrm{ridge}}(h) > 0$,

$$\lim_{t \to 0} \frac{1 - \rho_{\mathrm{canonical}}(m(t), \star)}{1 - R^2_{\mathrm{ridge}}(m(t), \star)} = \lambda(h) := \frac{\|v_\perp\|_F^2}{2\|c_\star\|_F^2 a_{\mathrm{ridge}}(h)}. \tag{96}$$

The sharp local envelope for the family is therefore governed by

$$C_{\mathrm{sharp}} := \sup_{\substack{h \in T_\star\mathcal{M} \\ a_{\mathrm{ridge}}(h) > 0}} \lambda(h), \qquad C_{\mathrm{sharp}} \leq C.$$

Literal equality with the proof constant C is not the empirical prediction. Rather, models lie near the observed frontier when their dominant directions of variation have $\lambda(h)$ close to C_{sharp}. If a model collection varies mainly along one common direction—for example, through training progress, bottleneck strength, or gradual degradation of one task-relevant feature family—then near the target it should approximately satisfy

$$1 - \rho_{\mathrm{canonical}} \approx \lambda(h)\big(1 - R^2_{\mathrm{ridge}}\big),$$

producing an approximately straight frontier ending at $(1, 1)$. A heterogeneous model collection explores several directions with different values of $\lambda(h)$ and should instead fill out a wedge below that frontier.

The distinction between $v_{\|}$ and $v_{\perp}$ has a direct interpretation. The perpendicular component changes the relative pattern of canonical stimulus similarities—making some stimulus pairs more similar and others less similar—and therefore lowers canonical RSA. The parallel component mainly rescales the entire canonical-RSM pattern, which cosine similarity largely ignores. Models can therefore fall well below the frontier when ridge error is produced mainly by such radial changes, by components removed by Π_{H}, or by axis-level errors that cancel when combined into the canonical RSM. Pure Theiss symmetries are different again: they leave $c_m = c_\star$ and hence $\rho_{\mathrm{canonical}} = 1$, while changing the raw residual s_m. They therefore generate the horizontal spread in raw RSA described by Proposition 20, rather than the nontrivial canonical frontier.

E.1.7 Sampling stability of RSA

The preceding results compare complete model layers. Neural experiments do not. Electrode arrays record units from a limited set of locations, while fMRI forms spatially local weighted averages and has uneven sensitivity and coverage across cortex. These effects are especially important when tuning is itself spatially organized, as it is in many real brain areas, or for example, in Topographical Deep Artificial Neural Networks (TDANNs), where units are placed on simulated cortical sheets and nearby units are encouraged to have similar response profiles (Margalit et al., 2024). A spatially localized sample from such a layer is therefore not merely smaller than the full layer: it is biased toward the functions represented in that part of the sheet. Figure 17 summarizes the resulting prediction. Canonical RSA retains the one-sided ridge frontier developed above, whereas raw RSA can be spread horizontally by unequal spatial sampling.

Spatial sampling produces a measurement-side version of the Theiss ambiguity. Let $H^N \in \mathbb{R}^{D_N \times P}$ be the complete response matrix of layer $N \in \{A, B\}$ on P stimuli, and let S_N describe what the experiment records from that layer. The measured responses and RSM are

$$Y_N = S_N H^N, \qquad R_N^{\mathrm{obs}} = Y_N^\top Y_N = H^{N\top} S_N^\top S_N H^N. \tag{97}$$

For simple unit selection, unequal recording gain, or repeated sampling of similar units, $S_N^\top S_N$ is diagonal. If $f_j^N \in \mathbb{R}^P$ is the response vector of unit j, then

$$R_N^{\mathrm{obs}} = \sum_j c_{Nj} f_j^N f_j^{N\top}, \qquad c_{Nj} \geq 0. \tag{98}$$

This is exactly the same RSM form that appears in the Theiss analysis: the observed geometry is a weighted sum of rank-one feature contributions. The source of the freedom is different—Theiss changes the network implementation while leaving observation fixed, whereas spatial sampling changes the observation while leaving the network fixed—but the concrete effects on the RSM closely parallel one another:

Spatial measurement effect	Theiss-like effect on the RSM
Unequal gain or unequal inclusion probability	Rescaling a feature contribution
Dense sampling of one functional patch	Giving those feature directions excess multiplicity, like duplication
A region represented in one sample but absent from the other	Adding rank-one terms on only one side
Failure to cover a region	Giving its feature directions zero weight
Local voxel pooling	Creating mixtures of nearby features, which is more general than simple scaling or duplication

Thus spatial bias is a *measurement-side analogue of the Theiss ambiguity*: the same underlying computation and the same privileged axes can yield different raw RSMs because the experiment assigns them different effective scales, multiplicities, and mixtures.

Topography turns spatial bias into functional bias. If m units are sampled uniformly without replacement from a layer of width D, then

$$\mathbb{E}[R^{\mathrm{obs}}] = \frac{m}{D} H^\top H. \tag{99}$$

The factor m/D does not affect Pearson RSM–RSA, so uniform sampling is unbiased in direction, although a finite sample still adds variability. If unit j is instead included with probability p_j, then

$$\mathbb{E}[R^{\mathrm{obs}}] = \sum_j p_j f_j f_j^\top, \tag{100}$$

which is generally not proportional to the full-layer RSM. In a topographic layer, p_j is correlated with tuning. Recording more units from the same patch therefore reduces uncertainty about that patch's RSM, but does not make the sample representative of the whole layer.

A simple example shows how large the effect can be. Suppose a layer contains K spatial patches dominated by orthogonal, equal-norm RSM directions $q_1, \ldots, q_K$. The whole layer has geometry $g_{\mathrm{full}} = \sum_{k=1}^K q_k$, while an array confined to one patch has geometry $g_{\mathrm{patch}} = q_1$. Then

$$\cos(g_{\mathrm{full}}, g_{\mathrm{patch}}) = \frac{1}{\sqrt{K}}. \tag{101}$$

The full and sampled systems have not acquired different computations; the sample has simply changed the relative representation of the same spatially organized feature families.

How electrophysiology and fMRI enter the thought experiment. For electrophysiology, a useful first approximation is to select units from a few localized neighborhoods of a TDANN-like sheet. Different array placements then sample different mixtures of tuning types. For fMRI, each measured channel is better modeled as a local weighted average of nearby units, together with uneven spatial coverage and sensitivity. TopoLM, for example, simulates the pooling component by Gaussian smoothing over nearby model units before analysis (Rathi et al., 2025); real fMRI additionally has spatially varying BOLD sensitivity, vascular filtering, distortion, and dropout (Kriegeskorte et al., 2010; Weiskopf et al., 2007). The relevant model comparison is therefore not only between two complete layers, but between the layers after applying the spatial selection and smoothing associated with the experiment.

RSA is specifically sensitive to sampling weights. Uniform and biased sampling have qualitatively different effects. Uniform sampling introduces variability that decreases as more units are recorded. Biased spatial sampling instead converges to a systematically reweighted representation. Ridge can absorb some such target-side reweighting, whereas raw RSA generally cannot.

Suppose the complete representations are exactly linearly related, $H^B = TH^A$, but only the target is spatially sampled. Then

$$Y_B = S_B H^B = S_B T H^A, \tag{102}$$

so a one-way ridge map from the complete source to the sampled target can remain exact. Reverse or bidirectional ridge is more demanding, because it also asks whether the sampled target retains every source direction (Thobani et al., 2025; Muzellec & Kar, 2025).

RSA can change even when both linear directions remain exact. If two measured representations differ only by an invertible diagonal reweighting, $Y_B = DY_A$, then each can linearly predict the other, but generically

$$R_B^{\mathrm{obs}} = Y_A^\top D^\top D Y_A \not\propto Y_A^\top Y_A = R_A^{\mathrm{obs}}. \tag{103}$$

Thus raw RSA depends not only on which feature directions are present, but also on how strongly the experiment weights them through gain, multiplicity, and spatial coverage. Figure 17(B) illustrates the resulting variation in RSA among representations with similar ridge scores.

What normalization can and cannot repair. Axis quotienting can remove unequal multiplicity among feature classes that were observed. It cannot recover a class that was never sampled or undo an unknown spatial mixture. Known inclusion probabilities can be corrected by inverse-probability weighting, although often at the cost of increased variance. Crossvalidated estimators such as crossnobis can remove finite-trial noise bias, but still estimate the geometry of the sampled population rather than that of the complete cortical sheet.

Predictions. For independently trained topographic networks, this analysis predicts:

(i) complete equivalent populations should have high ridge predictivity and high RSA;

(ii) under uniform sampling, RSA should approach the complete-population score as the number of measured units grows;

(iii) localized sampling can retain a systematic RSA bias even when many units are recorded;

(iv) a broadly sampled source can predict a localized target well by one-way ridge despite reduced raw RSA; and

(v) reverse or bidirectional ridge should reveal missing directions, whereas raw RSA can also change because directions present on both sides receive different weights (Thobani et al., 2025; Muzellec & Kar, 2025).

Thus, uniform subsampling mainly adds finite-sample variability, while persistent spatial bias changes the RSM obtained even with unlimited data. In a topographic system, this measurement bias acts like feature scaling or duplication and can separate raw RSA from ridge despite close alignment of the underlying feature directions.

E.2 CKA AND WEAK–STRONG EQUIVALENCE

Centered kernel alignment (CKA) is another widely used method for comparing representations (Kornblith et al., 2019). Like the raw RSM-RSA convention fixed above, raw linear CKA is sensitive to unequal feature scaling, duplication, redundant features, and biased measurement, so exact weak or strong alignment need not imply a high raw score. After removing those freedoms, however, exact and soft weak alignment imply exact and asymptotically perfect canonical CKA. We then distinguish the RSA conventions under which RSA and CKA coincide exactly (Williams, 2024) from those under which they can differ, show why raw RSM-RSA and raw CKA usually covary, sharpen the CKA-to-Procrustes implication using participation ratio (Harvey et al., 2024b), and explain how sampling can separate empirical scores.

E.2.1 LINEAR CKA AND THE CENTERED ESSENTIAL-CORE CONE

Keep the evaluation set $\Omega_P = \{x^1, \dots, x^P\}$ and representation orientation of (47). Let

$$C_P := I_P - \frac{1}{P}\mathbf{1}\mathbf{1}^\top, \qquad K(H) := C_P R(H) C_P = (HC_P)^\top (HC_P). \tag{104}$$

Thus $K(H)$ is the Gram matrix of the sample-centered population responses. When both centered Gram matrices are nonzero, define

$$\rho_{\mathrm{CKA}}(H_A, H_B) := \frac{\langle K(H_A), K(H_B)\rangle_F}{\|K(H_A)\|_F \, \|K(H_B)\|_F}, \qquad d_{\mathrm{CKA}} := 1 - \rho_{\mathrm{CKA}}. \tag{105}$$

This is linear CKA. Equivalently, with $\bar{H}_N := H_N C_P$,

$$\rho_{\mathrm{CKA}}(H_A, H_B) = \frac{\|\bar{H}_A \bar{H}_B^\top\|_F^2}{\|\bar{H}_A \bar{H}_A^\top\|_F \, \|\bar{H}_B \bar{H}_B^\top\|_F}. \tag{106}$$

The comparison with (48) is therefore explicit:

$$\rho_{\mathrm{CKA}} = \cos\big(C_P R_A C_P, C_P R_B C_P\big), \qquad \rho_{\mathrm{RSA}} = \cos\big(\Pi_{\mathsf{H}} R_A, \Pi_{\mathsf{H}} R_B\big). \tag{107}$$

Both scores globally normalize a transformed RSM, but the transformations are not nested. CKA double-centers the RSM, removing its row and column means and hence translations of the represented point cloud. The RSM-RSA convention in (48) instead discards the diagonal and subtracts one global off-diagonal mean.

Remark 9 (When RSA and CKA coincide exactly). The distinction in (107) is a distinction between *preprocessing conventions*, not between every method called RSA and every method called CKA. Let $d(H) := \operatorname{diag} R(H)$ and form the squared-Euclidean RDM

$$D(H) := d(H)\mathbf{1}^\top + \mathbf{1}d(H)^\top - 2R(H). \tag{108}$$

Double centering annihilates the first two terms, so

$$C_P D(H) C_P = -2C_P R(H) C_P = -2K(H). \tag{109}$$

Consequently,

$$\cos\big(C_P D(H_A) C_P, C_P D(H_B) C_P\big) = \rho_{\mathrm{CKA}}(H_A, H_B). \tag{110}$$

Thus cosine RSA on double-centered squared-Euclidean RDMs is exactly linear CKA. More generally, constructing an RDM from a positive-definite kernel and double centering yields the corresponding nonlinear CKA score (Williams, 2024, Proposition 1). By contrast, the score ρ_{RSA} in (48) uses Π_{H}, i.e. Pearson correlation on strict-upper-triangular RSM entries. All RSA–CKA differences discussed below refer to that RSM-RSA convention unless stated otherwise.

Remark 10 (Task-domain CKA). The finite-sample formulas have a direct task-domain form. Let $\mathcal{H} = L_0^2(\Omega)$ and let $X_N : \mathbb{R}^{d_N} \to \mathcal{H}$ send a to the centered scalar feature $a^\top(z_N - \mathbb{E}z_N)$. Then the centered kernel operator is the finite-rank operator $K_N = X_N X_N^*$, and linear CKA is the Hilbert–Schmidt cosine of K_A and K_B. Every factorization argument below is unchanged with matrices replaced by these finite-rank operators. A finite empirical CKA score certifies only the sampled representations unless the evaluation set is determining for the task patch. For the task-domain version, equality holds almost everywhere; when the task measure has full support on the patch, continuity of the network representations upgrades it to pointwise equality. The exact weak–strong and zippering corollaries below use one of these task-determining interpretations.

The Theiss decomposition and the generic uniqueness analysis in Appendix §E.1 are metric-independent until the final RSM projection is chosen. Applying double centering to Lemma 20 gives

$$K(H) = \sum_{j=1}^{m} \gamma_j \bar{f}_j \bar{f}_j^\top + \sum_{k=1}^{K} \eta_k \bar{u}_k \bar{u}_k^\top, \qquad \bar{f}_j := C_P f_j, \quad \bar{u}_k := C_P u_k, \tag{111}$$

with the same positive coefficients γ_j, η_k. Thus scaling and duplication reweight centered essential-axis contributions, while a redundant added feature can introduce a new centered-kernel direction.

Proposition 21 (The centered essential-core cone)**.** *Let f be a generic represented function, so that Corollary 8 supplies one essential axis dictionary. Choose one sampled feature vector f_j from each of its m classes and set*

$$k_j := (C_P f_j)(C_P f_j)^\top, \qquad \mathcal{C}_f^{\mathrm{CKA}} := \operatorname{cone}\{k_1, \dots, k_m\}. \tag{112}$$

Then $\mathcal{C}_f^{\mathrm{CKA}}$ is independent of the chosen minimal realization. Every representation in a cataloged Theiss orbit has

$$K(H) = \sum_{j=1}^{m} \gamma_j k_j + a, \qquad \gamma_j > 0, \tag{113}$$

where a is the centered contribution of redundant added features. If $a_A = a_B = 0$ and $G_{ij} := \langle k_i, k_j \rangle_F = (\bar{f}_i^\top \bar{f}_j)^2$, then

$$\rho_{\mathrm{CKA}}(A, B) = \frac{\gamma_A^\top G \gamma_B}{\sqrt{\gamma_A^\top G \gamma_A}\sqrt{\gamma_B^\top G \gamma_B}}, \tag{114}$$

and the score equals one exactly when the two weighted centered-core geometries lie on the same positive ray.

Proof. The orbitwise reduction, full-fiber generic uniqueness, and representative-independence arguments are exactly those of Proposition 16, Theorem 13, and Corollary 8; only the final linear map on the RSM changes from Π_{H} to $M \mapsto C_P M C_P$. Equation (111) therefore gives (113). Expanding the Frobenius cosine as in Proposition 19 gives (114), and equality follows from equality in Cauchy–Schwarz. □

The centered cone is function-intrinsic, but a raw representation still chooses arbitrary positive weights inside it. The next result is the CKA counterpart of Proposition 20; for CKA, duplication alone can make the score tend all the way to zero.

Proposition 22 (Raw linear CKA is not forced by exact weak alignment)**.** *There is no universal lower bound on raw linear CKA that tends to one as adjacent weak-alignment errors tend to zero, even when the represented one-step function and reduced essential core agree exactly. Indeed, there are exactly weakly aligned same-function pairs for which raw linear CKA tends to zero.*

Proof. Choose centered orthonormal vectors $u_1, u_2 \in \mathbb{R}^P$ with $P \geq 4$, and put $q_i = u_i u_i^\top$. Adding a sufficiently large constant multiple of $\mathbf{1}$ gives positive sampled preactivation vectors $g_i = u_i + c\mathbf{1}$ with $C_P g_i = u_i$. To realize them concretely, take P affinely independent task inputs and interpolate the prescribed values $g_i(x^\mu)$ by affine functionals. The centered rank-one terms satisfy $\langle q_1, q_2 \rangle_F = 0$ and $\|q_1\|_F = \|q_2\|_F = 1$.

Start from the reduced one-step function $F = v_1\varphi(g_1) + v_2\varphi(g_2)$. Let A_M contain M copies of g_1 and one copy of g_2, splitting v_1 equally among the copies. Let B_M contain one copy of g_1 and M copies of g_2, splitting v_2 equally among those copies. Both implementations compute exactly F, and minimal reduction returns the same two-axis core.

Write the physical population responses as

$$z_A(x) = g_1(x)a_1 + g_2(x)a_2, \qquad z_B(x) = g_1(x)b_1 + g_2(x)b_2,$$

where $a_1 = (\mathbf{1}_M, 0)$, $a_2 = e_{M+1}$, $b_1 = e_1$, and $b_2 = (0, \mathbf{1}_M)$. The pairs (a_1, a_2) and (b_1, b_2) are linearly independent, so a linear map sending a_i to b_i extends to an invertible map on $\mathbb{R}^{M+1}$. Thus the hidden representations are exactly weakly aligned, while the next-layer outputs are identical.

Their centered Gram matrices are

$$K(A_M) = Mq_1 + q_2, \qquad K(B_M) = q_1 + Mq_2.$$

Consequently,

$$\rho_{\mathrm{CKA}}(A_M, B_M) = \frac{2M}{M^2 + 1} \longrightarrow 0. \tag{115}$$

This rules out any weak-error-only lower bound for raw CKA. For activations with a function-preserving scale gauge, such as ReLU, featurewise rescaling produces the same reweighting mechanism; redundant feature addition supplies still more centered-kernel directions through (113). □

E.2.2 Canonical CKA: weak alignment implies CKA

The direct CKA analog of the canonical RSA results is obtained by removing the same scale, multiplicity, and redundant-addition freedoms. No new core-identifiability argument is needed: the canonical RSMs of Appendix §E.1.5 can simply be double-centered and compared by the CKA cosine.

Definition 36 (Axis-quotiented and balanced CKA)**.** For the axis-quotiented RSM of Definition 33, define

$$K_{\mathrm{ax}}(G) := C_P R_{\mathrm{ax}}(G) C_P, \qquad \rho_{\mathrm{CKA,ax}}(G, \widetilde{G}) := \frac{\langle K_{\mathrm{ax}}(G), K_{\mathrm{ax}}(\widetilde{G}) \rangle_F}{\|K_{\mathrm{ax}}(G)\|_F \, \|K_{\mathrm{ax}}(\widetilde{G})\|_F}. \tag{116}$$

For either balanced RSM R_s^{bal} from Theorem 15, define $K_s^{\mathrm{bal}} := C_P R_s^{\mathrm{bal}} C_P$ and its CKA score analogously.

Theorem 18 (Exact canonical CKA under weak alignment)**.** *Under the hypotheses of Theorem 14,*

$$K_{\mathrm{ax}}(A) = K_{\mathrm{ax}}(B), \qquad \rho_{\mathrm{CKA,ax}}(A, B) = 1, \tag{117}$$

whenever the common centered kernel is nonzero. Under the hypotheses of Theorem 16, the corresponding balanced CKA score is also one in the common coordinates. Both conclusions hold at every layer recovered by exact zippering.

Proof. Apply the linear map $M \mapsto C_P M C_P$ to the canonical RSM equalities in Theorems 14 and 16. Equality of the nonzero centered kernels makes their CKA cosine one. □

Theorem 19 (Soft canonical CKA under weak alignment)**.** *Under the hypotheses of Theorem 17, let*

$$D_{\mathrm{ax}}(\theta) := \frac{2\sqrt{2}(1-\theta)}{\sqrt{1-(1-\theta)^2}} + \frac{\sqrt{2}}{m}\frac{e_{\mathrm{weak}}^2}{\kappa(\theta)^2}. \tag{118}$$

If $\min\{\|K_{\mathrm{ax}}(A)\|_F, \|K_{\mathrm{ax}}(B)\|_F\} \geq \beta_{\mathrm{CKA}} > 0$*, then*

$$\rho_{\mathrm{CKA,ax}}(A,B) \geq 1 - \frac{2D_{\mathrm{ax}}(\theta)^2}{\beta_{\mathrm{CKA}}^2}. \tag{119}$$

Consequently, adjacent weak errors tending to zero imply $\rho_{\mathrm{CKA,ax}} \to 1$*. Soft zippering likewise implies upstream axis-quotiented CKA convergence after substituting the bad-axis bound from Theorem 8 into* (118)*.*

Proof. The proof of Theorem 17 gives $\|R_{\mathrm{ax}}(A) - R_{\mathrm{ax}}(B)\|_F \leq D_{\mathrm{ax}}(\theta)$. Double centering is a nonexpansive orthogonal projection in Frobenius space, so the same bound holds for the centered kernels; Lemma 21 then yields (119). The asymptotic and zippering conclusions follow from the same integer bad-axis argument as Corollary 11. □

Balanced CKA retains the computationally weighted votes of Corollary 10, while axis-quotiented CKA gives one scale-free vote per essential class. On compact families where the balanced weights vary continuously, the local argument of Corollary 12 gives the corresponding soft balanced-CKA bound.

E.2.3 How RSA, CKA, Procrustes, and ridge are related

Fix a target representation $\star$ and compare a collection of models m to it, as in Corollary 12. RSA and linear CKA both start from the same stimulus-by-stimulus RSM, but process it differently. Ridge is more flexible because it fits an affine map and can absorb changes that RSA and CKA retain. This section makes those statements precise.

Canonical RSA and canonical CKA agree near perfect alignment. Let R_m^{can} denote either the axis-quotiented RSM or a balanced RSM, using the same canonicalization for every model and the target. In the balanced case, assume as in Corollary 12 that the selected weights vary smoothly on the local comparison family. Put

$$r_m := \Pi_{\mathsf{H}} R_m^{\mathrm{can}}, \qquad k_m := C_P R_m^{\mathrm{can}} C_P.$$

Then

$$\rho_{\mathrm{RSA,can}}(m,\star) = \cos(r_m, r_\star), \qquad \rho_{\mathrm{CKA,can}}(m,\star) = \cos(k_m, k_\star).$$

Thus the two scores start from the same canonical RSM and differ only in the final operation applied to it. Under exact weak alignment both scores equal one by Theorems 14 and 18; under soft weak alignment both approach one.

The following local result says more: near a fixed target, small canonical RSA error and small canonical CKA error are comparable up to constants, provided the two scores are insensitive to the same local changes.

Proposition 23 (Local comparison of canonical RSA and CKA)**.** *Let* $R_\star^{\mathrm{can}}$ *lie in a finite-dimensional smooth local family of canonical RSMs, and suppose*

$$r_\star := \Pi_{\mathsf{H}} R_\star^{\mathrm{can}} \neq 0, \qquad k_\star := C_P R_\star^{\mathrm{can}} C_P \neq 0.$$

For a local perturbation Δ*, define*

$$Q_{\mathrm{RSA}}(\Delta) := \frac{1}{2\|r_\star\|_F^2}\left\|\Pi_{\mathsf{H}}\Delta - \frac{\langle \Pi_{\mathsf{H}}\Delta, r_\star\rangle_F}{\|r_\star\|_F^2} r_\star\right\|_F^2,$$

$$Q_{\mathrm{CKA}}(\Delta) := \frac{1}{2\|k_\star\|_F^2}\left\|C_P\Delta C_P - \frac{\langle C_P\Delta C_P, k_\star\rangle_F}{\|k_\star\|_F^2} k_\star\right\|_F^2.$$

If $R_m^{\mathrm{can}} = R_\star^{\mathrm{can}} + \Delta_m$*, then*

$$1 - \rho_{\mathrm{RSA,can}}(m,\star) = Q_{\mathrm{RSA}}(\Delta_m) + o(\|\Delta_m\|_F^2), \tag{120}$$

$$1 - \rho_{\mathrm{CKA,can}}(m, \star) = Q_{\mathrm{CKA}}(\Delta_m) + o(\|\Delta_m\|_F^2). \tag{121}$$

Let T be the tangent space of the family at the target. If Q_{RSA} and Q_{CKA} vanish on the same directions in T, then there are constants $0 < \lambda_- \le \lambda_+ < \infty$ such that

$$\lambda_- Q_{\mathrm{RSA}}(\Delta) \le Q_{\mathrm{CKA}}(\Delta) \le \lambda_+ Q_{\mathrm{RSA}}(\Delta), \qquad \Delta \in T. \tag{122}$$

Proof. For nonzero u and a small perturbation e,

$$1 - \cos(u + e, u) = \frac{1}{2\|u\|_F^2} \left\| e - \frac{\langle e, u\rangle_F}{\|u\|_F^2} u \right\|_F^2 + o(\|e\|_F^2).$$

Apply this first with $u = r_\star$ and $e = \Pi_{\mathsf{H}}\Delta_m$, and then with $u = k_\star$ and $e = C_P\Delta_m C_P$. This gives (120)–(121). After removing the directions on which both quadratic forms vanish, their square roots are norms on a finite-dimensional space, so they bound one another up to constants, giving (122). □

This is only a local statement for the Pearson RSM-RSA convention in (48). That score and CKA need not agree globally because their final operations discard different information. For every a,

$$C_P(\mathbf{1}a^\top + a\mathbf{1}^\top)C_P = 0,$$

whereas Π_{H} generally retains part of this term. Conversely, $\Pi_{\mathsf{H}} D = 0$ for every diagonal matrix D, whereas $C_P D C_P$ is generally nonzero. Canonicalization removes scale, multiplicity, and redundant-feature differences, but it does not make this RSM-RSA statistic identical to CKA. There is no contradiction with Remark 9, whose RSA statistic applies a different, double-centering operation to an RDM.

Raw RSA and raw CKA use the same feature weights. The raw relationship is clearest when the essential axes are fixed and there are no redundant added features. Let $f_1, \dots, f_m$ represent the common essential classes and define

$$q_j := \Pi_{\mathsf{H}}(f_j f_j^\top), \qquad k_j := (C_P f_j)(C_P f_j)^\top,$$

$$(G_{\mathrm{RSA}})_{ij} := \langle q_i, q_j\rangle_F, \qquad (G_{\mathrm{CKA}})_{ij} := \langle k_i, k_j\rangle_F.$$

Proposition 24 (Raw RSA and CKA use the same scale and multiplicity weights)**.** *Suppose the comparison model and target have no added-feature terms and have raw effective weights $\gamma_m, \gamma_\star \in \mathbb{R}^m_{>0}$ on the same essential classes. Then*

$$\rho_{\mathrm{RSA,raw}}(m, \star) = \frac{\gamma_m^\top G_{\mathrm{RSA}} \gamma_\star}{\sqrt{\gamma_m^\top G_{\mathrm{RSA}} \gamma_m}\,\sqrt{\gamma_\star^\top G_{\mathrm{RSA}} \gamma_\star}}, \tag{123}$$

$$\rho_{\mathrm{CKA,raw}}(m, \star) = \frac{\gamma_m^\top G_{\mathrm{CKA}} \gamma_\star}{\sqrt{\gamma_m^\top G_{\mathrm{CKA}} \gamma_m}\,\sqrt{\gamma_\star^\top G_{\mathrm{CKA}} \gamma_\star}}. \tag{124}$$

In particular, the same scale and multiplicity vectors enter both scores. If $G_{\mathrm{CKA}} = a G_{\mathrm{RSA}}$ for some $a > 0$ on the coefficient subspace explored by the model family, then the two raw scores are identical throughout that family.

Proof. Equation (123) is Proposition 19; equation (124) is Proposition 21. Both use the same coefficients $\gamma_m, \gamma_\star$; only the final matrix G_{RSA} or G_{CKA} differs. Multiplying that matrix by a positive scalar does not change the normalized score. □

This explains why raw RSA and raw CKA should usually covary rather than behave independently. If a feature is duplicated or rescaled, both metrics receive the same changed feature weight. They can still rank models differently because G_{RSA} and G_{CKA} process those shared weights differently. Added redundant features increase the disagreement because RSA and CKA retain different parts of the same added rank-one terms.

High linear CKA implies high Procrustes and ridge similarity. Near-perfect linear CKA means that the two centered Gram matrices are close after one global rescaling. The next lemma converts this into a concrete statement about the representations themselves: after one global scale and an orthogonal map, their centered activations are close. It also records the converse constraint and replaces a rank factor by the sharper participation ratio of the actual Gram-matrix discrepancy.

For the sampled columns $h_N^\mu := H_N e_\mu$, define

$$\mu_N := \frac{1}{P} H_N \mathbf{1}, \qquad \omega_P(E) := \left(\frac{1}{P} \sum_{\mu=1}^{P} \|E(h_A^\mu) - h_B^\mu\|_2^2 \right)^{1/2}.$$

Normalize the centered empirical factors by

$$X := P^{-1/2} C_P H_A^\top, \qquad Y := P^{-1/2} C_P H_B^\top, \qquad K_A^c := XX^\top, \qquad K_B^c := YY^\top.$$

The factor P^{-1} in $K_N^c = P^{-1} K(H_N)$ does not change CKA.

Lemma 22 (Sharp participation-ratio bounds for CKA and scaled Procrustes)**.** *Assume $d_A \le d_B$, let $c = \rho_{\mathrm{CKA}}(H_A, H_B)$, and set*

$$\tau^2 := \frac{\|K_B^c\|_F}{\|K_A^c\|_F}, \qquad \Delta := K_B^c - \tau^2 K_A^c, \qquad D_\Delta := \|\Delta\|_F^2 = 2\|K_B^c\|_F^2 (1 - c). \tag{125}$$

For $\Delta \neq 0$, let

$$\mathcal{R}_\Delta := \frac{\|\Delta\|_*^2}{\|\Delta\|_F^2}, \qquad r := \mathrm{rank}(\Delta), \tag{126}$$

and use the convention $\mathcal{R}_\Delta = 0$ when $\Delta = 0$. Finally put

$$a := \mathrm{tr}(K_B^c) + \tau^2 \,\mathrm{tr}(K_A^c), \qquad p := \min_{QQ^\top = I_{d_A}} \|Y - \tau X Q\|_F.$$

Then

$$a - \sqrt{a^2 - \mathcal{R}_\Delta D_\Delta} \;\le\; p^2 \;\le\; \sqrt{\mathcal{R}_\Delta D_\Delta}. \tag{127}$$

Both inequalities are sharp over positive-semidefinite Gram-matrix pairs. Moreover, some minimizer $Q \in \mathbb{R}^{d_A \times d_B}$ with $QQ^\top = I_{d_A}$ makes

$$E_{\mathrm{Proc}}(z) := \tau Q^\top (z - \mu_A) + \mu_B$$

satisfy $\omega_P(E_{\mathrm{Proc}}) = p$, and hence

$$\omega_P(E_{\mathrm{Proc}}) \le \Gamma_{\mathrm{PR}}(c) := (2\mathcal{R}_\Delta)^{1/4} \|K_B^c\|_F^{1/2} (1 - c)^{1/4} \le \Gamma(c) := (2r)^{1/4} \|K_B^c\|_F^{1/2} (1 - c)^{1/4}, \tag{128}$$

because $\mathcal{R}_\Delta \le r \le d_A + d_B$.

The same statements hold on the task domain with Hilbert–Schmidt and trace norms and finite operator rank. On a regular fixed-rank family whose nonzero Gram eigenvalues are uniformly bounded away from zero, local Lipschitz stability of the matrix square root gives the stronger estimate

$$\omega_P(E_{\mathrm{Proc}}) = O(\|\Delta\|_F) = O\big((1 - c)^{1/2}\big). \tag{129}$$

Proof. Apply Proposition 3 of Harvey et al. (2024b) to the two factor matrices τX and Y. Their Gram matrices are $\tau^2 K_A^c$ and K_B^c, their squared kernel-space Euclidean discrepancy is D_Δ, their normalization sum is a, and their Procrustes distance is p. This gives (127); the same paper proves that both bounds are attained. Equivalently, the identification

$$p = d_{\mathrm{Bures}}(\tau^2 K_A^c, K_B^c)$$

is the Procrustes–Bures duality of Harvey et al. (2024a).

The identity for D_Δ follows by expanding the Frobenius square and using the definition of τ:

$$\begin{aligned} \|K_B^c - \tau^2 K_A^c\|_F^2 &= \|K_B^c\|_F^2 + \tau^4 \|K_A^c\|_F^2 - 2\tau^2 \langle K_A^c, K_B^c \rangle_F \\ &= 2\|K_B^c\|_F^2 (1 - c). \end{aligned}$$

Since $\mathcal{R}_\Delta \le \mathrm{rank}(\Delta) = r$, the upper half of (127) gives

$$p^2 \le \sqrt{\mathcal{R}_\Delta}\, \|\Delta\|_F = \sqrt{2\mathcal{R}_\Delta}\, \|K_B^c\|_F (1 - c)^{1/2},$$

and taking square roots proves (128). Restoring the means identifies p with the displayed empirical RMS error. The final local estimate is the usual Lipschitz stability of the positive-semidefinite square root on a fixed-rank stratum separated from zero. □

CKA compares centered Gram matrices; Procrustes compares the centered activations after choosing the best scale and orthogonal map. If $\widehat{K}_N := K(H_N)/\|K(H_N)\|_F$, then

$$\|\widehat{K}_A - \widehat{K}_B\|_F^2 = 2\big(1 - \rho_{\mathrm{CKA}}(H_A, H_B)\big),$$

so CKA near one means that the normalized centered Gram matrices are close. Lemma 22 then shows that the activations themselves are close after the best scale and orthogonal map, while its lower bound quantifies the converse constraint. The scores are not identical away from one. If s_i are the singular values of the centered cross-covariance, the CKA alignment numerator is $\sum_i s_i^2$, whereas the Procrustes/NBS alignment numerator is $\sum_i s_i$; equivalently, CKA uses a Frobenius norm and Procrustes/NBS a nuclear norm (Harvey et al., 2024a). This difference can be practically important rather than merely cosmetic: Cloos et al. (2025) derive quadratic sensitivity of CKA, versus linear sensitivity of NBS, to the variance of a perturbed principal component, and observe that CKA can be high while lower-variance task-relevant components remain poorly captured. Their jointly optimized examples also show high angular-Procrustes similarity implying high CKA but not the converse; that empirical implication is dataset-specific, whereas (127) is the general finite-dimensional statement.

Ridge is more flexible than either metric because it can fit any affine map, including anisotropic reweightings. Therefore high ridge similarity need not imply high raw CKA. The converse does hold near one: the scaled-Procrustes map above is one possible ridge map, so high raw linear CKA forces high ridge prediction.

Corollary 13 (Consequences for ridge)**.** *Use the fixed-target/model-zoo setup and the normalized prediction-error convention*

$$R^2_{\mathrm{ridge}} = 1 - \varepsilon^2_{\mathrm{ridge}}$$

from Corollary 12.

(i) Let

$$K_m^{\mathrm{can}} := C_P R_m^{\mathrm{can}} C_P, \qquad K_\star^{\mathrm{can}} := C_P R_\star^{\mathrm{can}} C_P$$

be the centered canonical kernels. Suppose, on a locally identifiable compact family,

$$\|K_m^{\mathrm{can}} - K_\star^{\mathrm{can}}\|_F \le L_{\mathrm{CKA}} c_{\mathrm{w}} \varepsilon_{\mathrm{ridge}}, \qquad \min\{\|K_m^{\mathrm{can}}\|_F, \|K_\star^{\mathrm{can}}\|_F\} \ge \beta_{\mathrm{CKA}} > 0.$$

Then

$$1 - \rho_{\mathrm{CKA,can}}(m, \star) \le C_{\mathrm{CKA}}\big(1 - R^2_{\mathrm{ridge}}\big), \qquad C_{\mathrm{CKA}} := \frac{2L^2_{\mathrm{CKA}} c_{\mathrm{w}}^2}{\beta^2_{\mathrm{CKA}}}. \tag{130}$$

Thus high theorem-compatible ridge prediction forces high canonical CKA locally.

(ii) Let $c_{\mathrm{raw}} := \rho_{\mathrm{CKA,raw}}(m, \star)$, orient the comparison from model to target as in Lemma 22, and suppose

$$Y_\star := P^{-1/2} C_P H_\star^\top, \qquad \|Y_\star\|_F \ge \beta_Y > 0.$$

Then the best affine ridge fit satisfies

$$1 - R^2_{\mathrm{ridge}} \le \frac{\Gamma(c_{\mathrm{raw}})^2}{\beta_Y^2}. \tag{131}$$

Consequently, on a regular fixed-rank family,

$$1 - R^2_{\mathrm{ridge}} \le C_{\mathrm{CR}}\big(1 - c_{\mathrm{raw}}\big), \tag{132}$$

whereas the global rank-changing bound gives

$$1 - R^2_{\mathrm{ridge}} \le C^{\mathrm{glob}}_{\mathrm{CR}} \sqrt{1 - c_{\mathrm{raw}}}. \tag{133}$$

There is no converse lower bound on raw CKA in terms of ridge: Proposition 22 gives pairs with $R^2_{\mathrm{ridge}} = 1$ but raw CKA tending to zero.

Proof. For part (i), apply Lemma 21 to the centered-kernel bound and use $\varepsilon^2_{\mathrm{ridge}} = 1 - R^2_{\mathrm{ridge}}$. For part (ii), ridge optimizes over all affine maps, while the scaled-Procrustes map of Lemma 22 is one member of that class. Its normalized error is at most $\Gamma(c_{\mathrm{raw}})/\|Y_\star\|_F$, proving (131). The fixed-rank and global forms follow from the two bounds in that lemma. The final statement is the duplication construction of Proposition 22. □

Applying part (ii) at both adjacent layers and then Corollary 12 also gives a one-way relation from raw CKA to canonical RSA. If

$$\delta_{\mathrm{CKA}} := \max_{r\in\{\ell,\ell+1\}} \big(1 - \rho^{(r)}_{\mathrm{CKA,raw}}(m,\star)\big),$$

then, under the same local identifiability assumptions,

$$1 - \rho^{(\ell)}_{\mathrm{RSA,can}}(m,\star) \le \widetilde{C}_{\mathrm{RC}}\,\delta_{\mathrm{CKA}} \tag{134}$$

on regular fixed-rank families, while the global rank-changing estimate has right-hand side $\widetilde{C}^{\mathrm{glob}}_{\mathrm{RC}}\sqrt{\delta_{\mathrm{CKA}}}$. At CKA equal to one, the Procrustes error is zero. These implications rely on the Gram-factor structure of *linear* CKA; nonlinear-kernel CKA equal to one need not imply an affine map between the original representations.

What this predicts in empirical model comparisons. The main conclusions are:

- After canonicalization, weak alignment makes both RSA and CKA approach one; near-perfect alignment, their errors are comparable when they are sensitive to the same local changes.
- Raw RSA and raw CKA usually covary because the same feature scales and multiplicities enter both, but they can rank models differently because they process the resulting RSM differently.
- High raw linear CKA forces high Procrustes and ridge similarity. High ridge similarity does not force high raw CKA, and can coexist with a wide range of raw RSA scores.
- Under the assumptions of Corollary 12, high theorem-compatible ridge prediction also forces high canonical RSA and canonical CKA locally.
- High raw CKA at adjacent layers can be passed through the Procrustes map and the weak–strong results to force high canonical RSA upstream; low raw CKA by itself proves little.
- Scale, duplication, redundant features, and biased neural sampling can change raw RSA and CKA even when the task-relevant representation is shared.

These distinctions also qualify the interpretation of Bo et al. (2024). Their comparison pools trained and randomized networks in the same model zoo. Trained and untrained networks are *not* two minimal solutions to the same hard task, so this comparison lies outside the main contravariance regime. Training produces large representational changes that raw RSA, CKA, and Procrustes retain, while ridge can ignore some of them by linearly reweighting and mixing features. Because trained-versus-randomized differences also strongly affect behavior, this setup can make RSA, CKA, and Procrustes appear more behaviorally informative than ridge. This is consistent with our theory, rather than evidence that those metrics are generally preferable.

E.2.4 Sampling and empirical metric disagreement

Raw CKA is sensitive to sampling. Uniform sampling introduces error that shrinks with sample size, while biased sampling converges to a systematically reweighted representation. Ridge can absorb some target-side reweighting, whereas raw RSA and CKA retain it.

Unit and spatial sampling. Let $Y_N = S_N H_N$ be the measured representation. Its centered CKA kernel is

$$K_N^{\mathrm{obs}} = C_P H_N^\top S_N^\top S_N H_N C_P. \tag{135}$$

Thus unequal inclusion probabilities or gains reweight feature contributions. Localized sampling over-represents some tuning classes and misses others, while voxel pooling mixes them. In a topographic representation, spatial sampling bias is therefore also functional sampling bias.

Lemma 23 (Unit subsampling)**.** *Write the complete centered kernel as*

$$K(H) = \sum_{j=1}^{d} k_j, \qquad k_j = \bar{f}_j \bar{f}_j^\top, \qquad \bar{f}_j = C_P f_j.$$

If q units are sampled independently with probabilities $p = (p_1, \dots, p_d)$, let

$$\widehat{K}_q = \frac{1}{q}\sum_{a=1}^{q} k_{J_a}, \qquad K_p = \sum_{j=1}^{d} p_j k_j.$$

Then

$$\mathbb{E}\widehat{K}_q = K_p, \qquad \mathbb{E}\|\widehat{K}_q - K_p\|_F^2 = \frac{1}{q}\left(\sum_{j=1}^{d} p_j \|k_j\|_F^2 - \|K_p\|_F^2\right). \tag{136}$$

Uniform sampling therefore converges to $K(H)/d$, which gives the same CKA score as the complete kernel, while biased sampling converges to a differently weighted kernel.

Proof. The expectation is immediate, and independence gives the displayed $1/q$ variance. □

Repeatedly sampling one tuning class therefore has the same effect as increasing its multiplicity. Known sampling probabilities can be corrected by inverse-probability weighting, but an unobserved feature class cannot be recovered. RSA and CKA may also respond differently to the same bias because they process the sampled Gram matrix differently.

Stimulus sampling. CKA is likewise sensitive to which stimuli are sampled. If S_I selects a stimulus subset, then

$$K_{N,I} = C_p S_I H_N^\top H_N S_I^\top C_p. \tag{137}$$

Independent samples from the same task distribution converge to the population score. A biased stimulus set instead converges to the score for a different task distribution. Because centering is recomputed on the subset, $K_{N,I}$ is not generally a principal submatrix of the complete centered kernel.

Repeated measurement. Repeated measurements remove trial-noise bias, not sampling bias. If $\widetilde{H}_N^{(r)} = H_N + \Xi_N^{(r)}$ with independent mean-zero noise, then for $r \neq s$,

$$\mathbb{E}\Big[C_P \widetilde{H}_N^{(r)\top} \widetilde{H}_N^{(s)} C_P\Big] = K(H_N).$$

Cross-repeat estimators therefore recover the clean kernel in expectation, but cannot recover units or stimuli that were never sampled.

Thus uniform sampling produces shrinking error, while persistent sampling bias changes the limiting Gram matrix. Raw CKA can consequently disagree with ridge, RSA, or Procrustes even when the complete representations satisfy the canonical relationships derived above.

F Empirical protocols, controls, and additional results

Table 3: Our theoretical predictions and corresponding empirical tests.

Result	Formal conclusion	Measured consequence
Zippering	Terminal agreement propagates upstream in minimal, generically null-trivial solutions.	Ordered correspondences across depth and across ResNet–ViT hierarchies.
Weak–strong equivalence	Adjacent weak alignment exposes distinct task-used nonlinear axes.	Pre-/post-nonlinearity zipper pattern; learned channels and MLP gates beat mixed-axis controls.
Transformer components	Branch identifiability forces task-used gates and whole heads to match, up to exact reparameterizations.	Pre-GELU gates and whole-head contributions recur; we make no claim for raw $Q/K/V/O$ axes.
Used-axis budget	Required usedness lower-bounds the alignable fraction.	Source-only usedness predicts cross-seed recurrence, cross-family vision transfer, and Pythia causal identity.

Data and models. Vision comparisons use independently trained AlexNet, ResNet-18/34/50, and ViT models evaluated on a common held-out ImageNet sample. AlexNet comparisons use 10 seeds and 12,000 images (45 seed pairs and 120 seed triples); ResNet comparisons use two seeds at each of three depths and 12,000 images; ViT comparisons use three seeds and approximately 49,997 validation images. Cross-architecture comparisons use 18 ResNet–ViT model pairs and 12,000 images. Language comparisons use 10 independently trained Pythia-160M models and 12,000 token observations (45 pairs and 120 triples). Causal head interventions use 128 contexts at layers 2, 5, 6, 8, 11, and 12.

Global weak alignment. We fit affine ridge maps on training observations and report held-out coefficient of determination, without PCA, CCA, or an imposed rank restriction. To compare preactivations with post-activation representations, we also fit the map to the target preactivation and then apply the target model's nonlinearity. Thus Figure 2's green curves mean "target ReLU/GELU of the affinely predicted target preactivation." For ResNets, "before final ReLU" means the sum of the residual and skip branches immediately before the stage-ending ReLU. We evaluate every contrast on held-out inputs rather than infer it from training fit. For unequal depths, an ordered dynamic-programming path maximizes layer similarity subject to monotonicity; this produces the depth warp in Figure 3. For ViTs, the mean-pooled patch-token representation averages all image-patch tokens and excludes the class token.

Local privileged components. We choose a one-to-one component assignment by Pearson correlation on training observations and report the mean correlation of those fixed pairs on held-out observations. For an MLP, candidate axes are the scalar coordinates immediately before the coordinatewise nonlinearity; for a convolutional network, they are pre-nonlinearity channels. Random orthogonal mixtures preserve the full representational span while removing the identity of each learned coordinate. The shuffled-input controls independently permute image or token observations, destroying cross-model stimulus correspondence. We do not treat transformer residual coordinates as privileged. We represent a learned head by its whole task-visible contribution O_h, or equivalently by the quotient objects $\Gamma_h = W_Q W_K^\top/\sqrt{d_h}$ and $C_h = W_V W_O$ at the parameter level. The routing correlation in Figure 4c compares matched heads' attention patterns $P_h(X)$ on held-out inputs.

For the head-mixture control, draw Q uniformly from the orthogonal transformations that fix the all-ones vector and mix the whole-head contributions:

$$Q^\top Q = I, \qquad Q\mathbf{1} = \mathbf{1}, \qquad \widetilde{O}_j(X) = \sum_{h=1}^{H} Q_{hj} O_h(X), \qquad \sum_{j=1}^{H} \widetilde{O}_j(X) = \sum_{h=1}^{H} O_h(X). \quad (138)$$

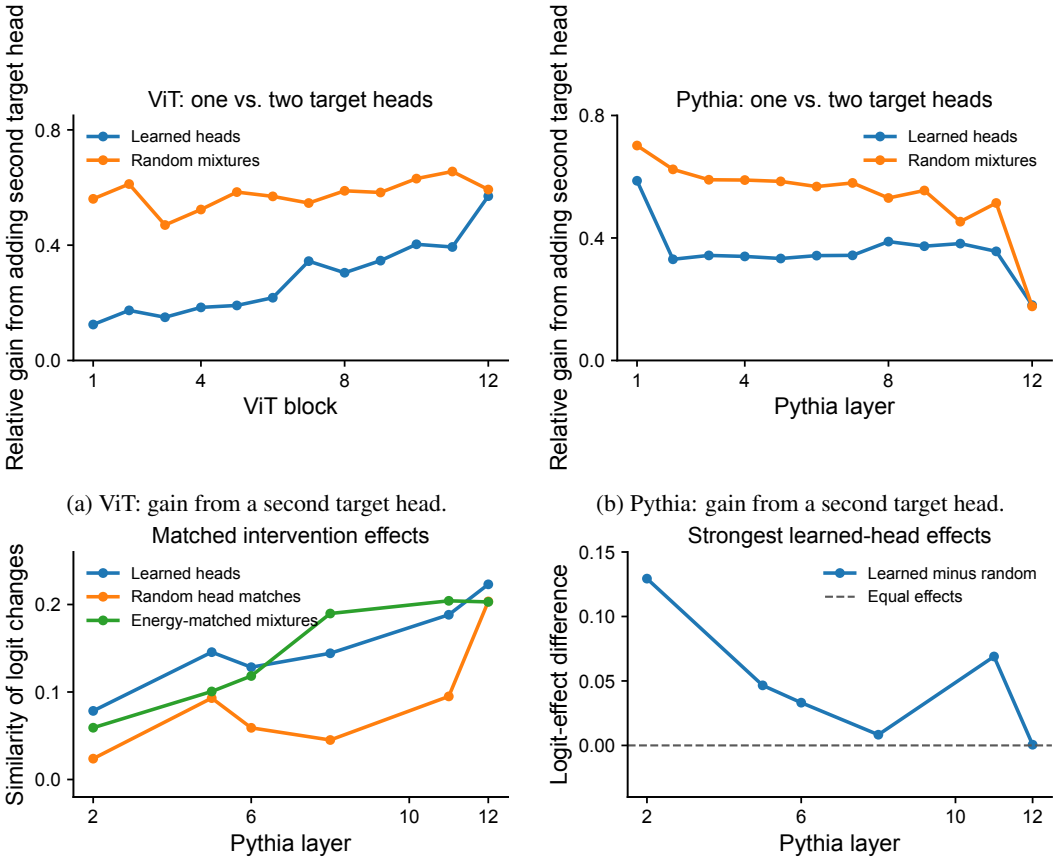
This sum-preserving random mixture leaves the aggregate attention update unchanged while erasing individual head identity. A three-model cycle closes when the fixed one-to-one matches $A \to B \to C \to A$ return a head to itself; chance is $1/H$.

To test the formal no-splitting requirement in Definition 26, item A3, let $R^2_{\text{best } k}$ be the best held-out prediction of one source head using k target heads. Figure 18a,b plots

$$g_{1\to 2} = \frac{R^2_{\text{best } 2} - R^2_{\text{best } 1}}{\max\{R^2_{\text{best } 1}, 10^{-12}\}}. \quad (139)$$

A small gain means that one target head already captures the source head, rather than several target heads being needed together.

Usedness and causal identity. The source-only channel or MLP-gate score is $u_j = \mathrm{Var}[\varphi(z_j)]\|W_{2,:,j}\|_2^2$ (with the analogous outgoing convolutional norm); the attention score is $u_h = \mathbb{E}\|O_h(X)\|_2^2$. We use these finite-sample scores to rank component use within a trained network and test whether more task-used components recur more strongly. For the ResNet stage curves in Figure 5, we average blocks within each architecture and then give ResNet-18, ResNet-34, and ResNet-50 equal weight. The top- and bottom-used quartiles in panel (c) are the highest and lowest 25% of heads under the source-only score. We mean-ablate each matched head at one query token and compare the resulting mean next-token-logit changes across models by cosine similarity. Correctly matched learned Pythia heads exceed random head matches in 248 of 270 pair/layer comparisons. Energy-matched controls rescale each random sum-preserving head mixture so that mean



(a) ViT: gain from a second target head.

(b) Pythia: gain from a second target head.

(c) Matched intervention effects.

(d) Strongest learned-head effects.

Figure 18: **Learned attention heads map one-to-one and produce reproducible changes in next-token predictions.** (a,b) Adding a second target head improves learned-head matches much less than random sum-preserving mixtures; the plotted relative gain is Eq. 139, and smaller values mean that one target head already captures the source head. This directly tests Supplement Definition 26, item A3. (c) Correctly matched Pythia heads produce more similar next-token-logit changes than random head matches in $248/270$ model-pair/layer comparisons across ten seeds and 128 held-out contexts at six layers. (d) At every tested layer, the strongest learned head exceeds the average maximum among energy-matched random mixtures and beats an individual random-mixture maximum in about 67% of comparisons. The sum-preserving mixtures are defined in Eq. 138.

ablation has the same RMS residual-stream perturbation as its learned-head counterpart. At layers 2, 6, 8, and 11, the mean top-used versus shuffled causal cosines are respectively $.169/.006$, $.366/.022$, $.404/.023$, and $.388/.033$.