PRODUCT SPECTRUM NEURONS: MAKING INTERACTION ORDER ADDRESSABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks multiply features everywhere, yet interaction order usually remains entangled with width, depth, and gating rather than freely addressable. The Product Spectrum Neuron (PSNeuron) computes, in each of its R learned bases, elementary symmetric polynomials of degree 2 through K over G disjoint slots. Every order becomes its own channel, with three exact properties: on Rademacher inputs channel j carries Walsh degree j; increasing K at fixed geometry appends orders without disturbing lower ones; and any order can be masked in a trained model without retraining or modifying learned parameters. Across three tabular tasks, one common PSNeuron architecture with $K = 4$ fixed in advance, a validation-selected layout, and task-specific learning rates removes 34.7% to 51.7% of the trainable parameters of the selected dense references while remaining within 0.44 points of their accuracy. At $K = 2$, narrow DINOv3 heads outperform an exact-parameter additive control in 69 of 75 paired seeds across 15 dataset and width settings, and on 20 fresh seeds a $K = 4$ DeltaNet mixer beats an exact same-slot additive twin with 48 fewer order coefficients and a conditional GELU control on associative recall after Holm correction. On controlled teachers, exact PSNeuron witnesses establish representability independently of training, order 6 content is recovered reliably near its target degree, and same-model spectra and masks verify the planted decomposition in recovered models and localize missing upper-order content in failed runs. PSNeuron turns interaction order into an addressable experimental coordinate of a neural site whose representability, recovery, and functional use can be examined within the same trained model.

1 INTRODUCTION

Neural networks multiply features in gates, attention, and cross layers (Shazeer, 2020; Vaswani et al., 2017; Wang et al., 2021), but practitioners rarely control interaction order because it shifts with width, depth, and gating. This raises two questions: can structured interactions replace dense capacity at a site, and does a trained model represent, recover, and use a particular order?

We introduce Product Spectrum Neurons (PSNeurons), a neural-site operator building on higher order ANOVA factorization and vector valued polynomial models (Blondel et al., 2016; 2017). Each basis reads learned linear slots on a fixed partition; orders 2 through K are exact elementary-symmetric channels of the same slots, mixed by a vector readout. Elementary symmetric polynomials, ANOVA interactions over shared factors, and separate polynomial-order branches are established; PSNeuron's novelty is a replaceable site with one learned grouped geometry shared across orders: raising K preserves lower channels, adds one coefficient per basis beyond $K = 3$, and leaves every order intervenable after training (Section 3).

The evidence tests these two parts separately: sites with $K = 2$ carry one order and isolate the grouped product against an additive head of exactly equal size, and the spectrum, used at $K = 4$ in the common replacement of tabular, vision, and recurrent sites, is tested by order masks on trained models and on controlled teachers up to order 8. Addressable order separates representability from recovery, enables same-model order interventions, and makes shared geometry a testable inductive bias; parameter efficiency is measured across tabular, vision, sequence, and language hosts (Section 4). This lets order-specific behavior be tested within one trained model rather than inferred only from aggregate accuracy.

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.

Contributions.

1. **Interaction order as an addressable part of a neural site.** Built on ANOVA and elementary symmetric polynomial algebra, PSNeuron gives each order from 2 to K its own exact channel over grouped slots of one shared factor bank: raising K leaves the lower channels unchanged, the learned order coefficient matrix contains $K-1$ coefficients per basis when $K \geq 3$, any order can be masked after training without changing a parameter, and on Rademacher inputs each channel is exactly one Walsh degree. A vector readout lets the operator replace a dense module (Section 3).
2. **Measured parameter efficient operating regimes.** We measure site replacements across tabular, recurrent, vision, and language-model hosts, together with a depth-wise order-allocation study on HIGGS Small. A common $K = 4$ tabular replacement makes models 34.7 to 40.8% smaller than their dense hosts within half a point of accuracy (Section 5.1).
3. **A gain from the grouped product over an additive control.** At $K = 2$ and exactly equal parameters, narrow heads on frozen DINOv3 features lead the additive head by 5.70 to 7.36 points on average at widths 2 to 4, in 69 of 75 seed pairs over five seeds, though not at width 8 in a three seed ladder; on Boolean cube probes an order matched $K = 8$ unit reaches 99.4% on the full eight bit parity truth table and $K = 4$ and 6 units solve XOR whenever the partition separates it, where a larger additive control stays at chance; and on 20 fresh DeltaNet seeds a $K = 4$ mixer exceeds an exact additive twin on the same slots and a conditional GELU control on MQAR after Holm correction (Section 5.2).
4. **Representability, recovery, and use of each order, measured separately.** On controlled teachers with exact witnesses, high order failures arise in recovery despite established representability: order 6 targets are recovered reliably by students of order 6 and 7 and order 8 targets in about a quarter of runs. Masks and delivered spectra confirm the planted decomposition in recovered models and locate the missing high orders in failed ones (Section 5.3).

2 Related Work

Multiplicative computation. Gates, attention, and recurrent and state space models multiply features (Dauphin et al., 2017; Shazeer, 2020; Vaswani et al., 2017; Gu & Dao, 2024; Yang et al., 2024), and multiplicative interactions have been studied as a general architectural primitive (Jayakumar et al., 2020). Product and sigma-pi units make the multiplication an explicit neuron (Rumelhart et al., 1986; Durbin & Rumelhart, 1989; Giles & Maxwell, 1987; Shin & Ghosh, 1991), ridge polynomial networks grow by adding pi-sigma units of increasing degree, whose biased factors mix all lower degrees (Shin & Ghosh, 1995), and products of many terms are hard to optimize (Leerink et al., 1994). Product networks feed pairwise products to a network, and Pi-Nets, Multilinear Operator Networks, DCN and DCN-V2, xDeepFM, and AutoInt raise interaction degree with depth or crossings (Qu et al., 2016; Chrysos et al., 2020; 2022; Cheng et al., 2024; Wang et al., 2017; 2021; Lian et al., 2018; Song et al., 2019). Bilinear MLPs make a quadratic layer interpretable from its weights (Pearce et al., 2025), and SPAM learns interpretable additive models with low rank polynomial interactions (Dubey et al., 2022); the first carries one order and the second is a standalone predictor, so neither offers a neural site whose exact orders share one geometry and can be intervened on after training. PSNeuron computes every order from 2 to K at one site, each in its own channel, so its orders are set by construction and can be read or removed.

Explicit interaction models. Factorization machines and HOFM are the algebraic foundation of PSNeuron (Rendle, 2010; Blondel et al., 2016). They established higher order ANOVA interactions, factors shared across degrees, weights for each degree, and efficient training by dynamic programming; since each ANOVA kernel is homogeneous of a single degree, a scalar HOFM can already weight or drop its degrees separately. Vector valued polynomial prediction extends the construction to shared output bases (Blondel et al., 2017), neural factorization machines feed a vector of pairwise interactions to a network (He & Chua, 2017), field aware factorization machines give each feature a factor per field (Juan et al., 2016), NIT limits interaction order (Tsang et al., 2018), Deep Interaction Machine computes factorization machine interactions of every order up to a maximum with Newton's identities beside a deep network (Yu et al., 2020), and kVNN replaces convolution kernels by parallel polynomial branches, one per order, each with its own learnable centers (Yun et al., 2026).

What PSNeuron adds. With singleton slots, no centering, and the identity coordinate frame, one PSNeuron basis at one order reduces to a term of the ANOVA kernel up to a frozen scalar calibration. PSNeuron partitions the coordinates of a neural site into groups, like feature fields, reads one learned factor bank at every order, and mixes the orders of each basis before a vector readout, so the operator can replace a dense module; when $K \geq 3$ its learned order coefficient matrix contains $K - 1$ coefficients per basis, raising K leaves the lower channels unchanged, each square free channel carries exactly one degree, which on Rademacher inputs is one Walsh degree, and any order can be masked in a trained model. Sharing the bank couples the order tensors through shared factors, as in shared HOFM variants and tensor decompositions (Hitchcock, 1927; Kolda & Bader, 2009; Novikov et al., 2015), the opposite of the order specific centers of kVNN, a choice Section 5.4 tests; Table 7 compares the closest models. A shared degree singleton control with equal parameters has no resolved accuracy difference on three tabular tasks, which isolates the grouping choice at that operating point (Table 60); the contribution is the grouped neural site and its order interventions and representation, recovery, and use semantics, not superiority over HOFM algebra. Separating what a model can represent from what training recovers follows work on optimization failure and spectral bias (Shalev-Shwartz et al., 2017; Rahaman et al., 2019; Barak et al., 2022); exact witnesses let us measure it for each order.

3 Product Spectrum Neurons

Let a PSNeuron site have input width d, output width o, R bases of G slots each, and maximum order K, with $2 \leq K \leq G \leq d$ (notation: Table 4). Every order reads one learned factor bank $W_0 \in \mathbb{R}^{R \times d}$, so one pass through the bank feeds the whole spectrum (Figure 1, with a worked example).

Linear slots and exact order channels. Basis r uses row W_r of W_0, without a bias. A seeded permutation of the d input coordinates is cut into G disjoint nonempty blocks $\mathcal{G}_{r,1}, \ldots, \mathcal{G}_{r,G}$ whose sizes differ by at most one; the partition is fixed at construction and never learned. Feedforward sites share one draw across bases, $\mathcal{G}_{r,g} = \mathcal{G}_g$. Cross head sites use one contiguous block per attention head, the recurrent mixer uses separate blocks for state and input, and the frozen DINOv3 heads and the controlled students draw one partition per basis. No product takes two coordinates from the same block. Slot g is

$$q_{r,g}(x) = |\mathcal{G}_{r,g}|^{-1/2} \sum_{i \in \mathcal{G}_{r,g}} W_{r,i} x_i, \tag{1}$$

and for $j \in \{2, \ldots, K\}$ the order j channel is the elementary symmetric polynomial of the slots,

$$e_{r,j}(x) = \sum_{g_1 < \cdots < g_j} \prod_{t=1}^{j} q_{r,g_t}(x), \qquad \hat{e}_{r,j}(x) = c_j \binom{G}{j}^{-1/2} e_{r,j}(x), \tag{2}$$

with a frozen calibration constant c_j. With frozen center μ_j, frozen gauge factor ζ_j, order mask $m_j \in \{0, 1\}$, and order coefficient $\gamma_{r,j}$,

$$\tilde{e}_{r,j}(x) = \zeta_j\big(\hat{e}_{r,j}(x) - \mu_j\big), \qquad s_r(x) = \sum_{j=2}^{K} \gamma_{r,j} m_j \tilde{e}_{r,j}(x), \qquad P(x) = B_{\text{core}} s(x) + b_{\text{core}}. \tag{3}$$

The site output $P(x) \in \mathbb{R}^o$ is a vector with core readout $B_{\text{core}} \in \mathbb{R}^{o \times R}$, not a scalar activation. In every reported configuration $\mu_j = 0$ and $\zeta_j = 1$; the general form allows optional centering and a function preserving rescaling of the bank (Section A.3). The coefficients $\gamma_{r,j}$ are learned when several orders are carried and fixed at 1 when one order is carried, and $b_{\text{core}} = 0$ except in the frozen DINOv3 heads, which train a class bias. Every raw channel is homogeneous of degree j, $\hat{e}_{r,j}(ax) = a^j \hat{e}_{r,j}(x)$ for real a in the coordinates that enter P (Proposition 1); a nonlinear conditioner moves this statement to the conditioned coordinates, and a host residual makes the enclosing block inhomogeneous. For Rademacher inputs to P each channel is exactly one Walsh degree, so orders are orthogonal (Proposition 3). Each basis forms products of G linear factors: with nonzero slot rows an order 2 basis has Hessian rank G, a smooth ridge unit $\phi(u^\top x)$ at most one (Proposition 2).

Initialization, calibration, and training. We draw $W_0 \sim \mathcal{N}(0, 1)$ and $B_{\text{core}} \sim \mathcal{N}(0, 1/R)$, and set $\gamma_{r,j} = |\mathcal{J}|^{-1/2}$ over the set $\mathcal{J}$ of carried orders. Before the first step the initialized host runs once

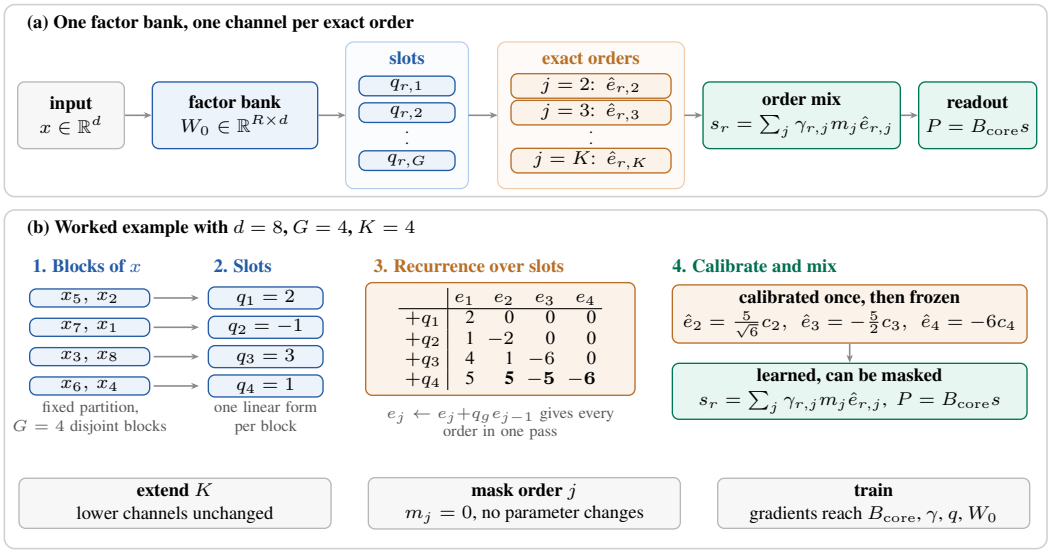

Figure 1: PSNeuron computation. (a) Each of R bases r, a row of W_0, forms G slots on disjoint input blocks; its order j channel $\hat{e}_{r,j}$ is the calibrated elementary symmetric polynomial of degree j in the slots, and learned coefficients $\gamma_{r,j}$ and masks $m_j \in \{0, 1\}$ mix the orders. (b) Illustrative values for one basis: the recurrence starts at $e_0 = 1$, higher orders first, and $\hat{e}_j = c_j \binom{G}{j}^{-1/2} e_j$.

in evaluation mode, without gradients, on a fixed training slice, and each site is calibrated on the tensor it actually receives. The bank W_0 is rescaled once so that its slots have unit root mean square (RMS) jointly over examples, positions, bases, and slots, and then $c_j = 1/\operatorname{RMS}\big(\binom{G}{j}^{-1/2} e_{r,j}\big)$ is fitted for each order on the rescaled bank. The DINOv3 heads keep their GELU control's initialization scale (Section G.1), and training differentiates the recurrence of Eq. (16) directly (Section A.10).

Addressability, nesting, and cost. Masking order j sets $m_j = 0$ and changes no parameter or frozen constant. Raising K to $K' > K$ at fixed G, bank, partition, and calibration leaves channels 2 through K unchanged (Remark 1). The recovery study at fixed G relies on this nesting (Section 4). With shared geometry the operator has

$$P_{\text{shared}}(R, K) = Rd + oR + \begin{cases} 0, & K = 2, \\ R(K-1), & K \geq 3, \end{cases} \tag{4}$$

trainable parameters, because the single coefficient at $K = 2$ is redundant with the readout; placement parameters are counted separately. A basis therefore costs $d + o$ at $K = 2$ and only $K - 1$ more at $K \geq 3$. The operator ties the factor bank across orders,

$$W^{(2)} = W^{(3)} = \cdots = W^{(K)} = W_0, \tag{5}$$

an inductive bias whose value depends on whether the orders of a target share directions (Section 5.4).

Replacement and augmentation. Placement is kept separate from the spectrum. A *replacement* removes dense trainable capacity at a site; with conditioner A_{host} and outer readout B_{host},

$$y = B_{\text{host}}\, P(A_{\text{host}} x). \tag{6}$$

An *augmentation* keeps a conventional carrier and adds the branch beside it,

$$y = C_{\text{host}}(x) + \lambda B_{\text{host}}\, P(A_{\text{host}} x), \tag{7}$$

and is compared at equal or larger capacity, not by size. The main tabular site is the feedforward sublayer of the second of two residual blocks of width 64 that apply LayerNorm before each sublayer, $h \leftarrow h + B_{\text{host}} P(\text{LN}(h))$, with identity A_{host}, $d = o = 64$, $G = 8$, $R = 8$, $K = 4$, and an outer readout of rank 8, $B_{\text{host}} = U_B V_B$. This gives $1{,}048 + 1{,}024 = 2{,}072$ trainable site parameters, 24 of them order coefficients; the block skip is shared with the dense control, and the FineWeb-Edu

feedforward placements and CIFAR100 feedforward sites use the same replacement form. Since nothing nonlinear separates B_{host} and B_{core}, a trained site folds exactly into a $1{,}048$ parameter operator through $B_{\text{core}} \leftarrow U_B V_B B_{\text{core}}$; we nevertheless count the $2{,}072$ parameters that are trained. The cross head site deletes an attention output projection and returns $y = [o_1; \ldots; o_H] + P(z)$ with $z_h = P_h o_h$, which is Eq. (7) with an identity carrier and one slot per head; since the carrier has no parameters and the projection is removed, this site is compared by size. The recurrent mixer replaces the SwiGLU channel mixer of the last DeltaNet block, which reads only the normalized input u_t, by a PSNeuron that reads u_t and also the state readout $r_t = S_t q_t$, using either the joint spectrum over state and input slots or only its cross terms $e_j^{\text{cross}} = e_j(\text{all}) - e_j(\text{state}) - e_j(\text{input})$. The DINOv3 heads apply P directly to frozen CLS features, and the latent head used on EuroSAT (Helber et al., 2019) places it between two learned maps, $y = U\,P(Dz)$. Section E gives the equations and counts for every placement.

4 Experimental Design

We separate three claims. ***Representability*** means that an exact witness exists, that is, a parameter setting of the realized geometry that computes the target exactly. ***Recovery*** means that training reaches a prescribed held out criterion, and ***functional use*** means that an intervention on the trained model changes held out quality. Section F maps every experiment to the claim it tests.

Hosts, objectives, and controls. Tabular studies use HIGGS Small, Jannis, and Helena (Baldi et al., 2014; Feurer et al., 2021) in the residual MLP of Section 3; features are imputed and standardized transductively, with statistics of all rows including test rows but no labels, identically for every model (Section G.2). Sequence studies use MQAR (Arora et al., 2024) and multi-needle retrieval in an attention host with two blocks and in a DeltaNet with four layers (Yang et al., 2024), also trained on state tracking. Vision studies use a ViT with six blocks and $d_{\text{model}} = 192$ (Dosovitskiy et al., 2021) trained on CIFAR100 (Krizhevsky, 2009), and heads on frozen DINOv3 ViT-B/16 CLS features (Siméoni et al., 2025); FineWeb-Edu (Penedo et al., 2024) tests single-FFN and separate all-six-FFN replacements (Section E). Dense controls are GELU (Hendrycks & Gimpel, 2016) and SwiGLU (Shazeer, 2020) networks resized in hidden width, with a low rank input projection where width alone cannot reach a budget. The additive control sums a GELU of each slot. Attention score corrections, residual adapters, and frozen heads that keep a linear path are augmentations.

Selection and training. Unless Section G states otherwise, learning rates are selected on tuning seeds 101 and 102 by validation quality at the final step over $\{2, 4, 8, 16, 32\} \times 10^{-4}$, and a model selecting an edge gets one extension to 10^{-4} or 6.4×10^{-3}, shared by the common tabular replacement and its references. The site layout, placement, and dense references are then chosen by mean validation quality of the five final seeds 0 to 4, whose test values we report; reported test values are not inputs to these selection rules. For the DeltaNet mixer a validation rule fixed before the choice found the joint and cross term layouts too close to call, and the authors chose the joint layout, which led on validation of these seeds for MQAR and multi-needle retrieval; comparisons of this selected layout, including strict ones, carry post selection optimism that prospectively fixed comparisons do not. The common replacement uses $K = 4$, fixed in advance for the tabular, recurrent, and CIFAR100 feedforward sites; capacity sweeps add $K = 2$, 6, and 8, and validation selects the cross head order $K = 2$. Models train in fp32 with AdamW, weight decay 0.01, no gradient clipping, and the study-specific schedules in Section G.

Measuring parameter efficiency. A *named comparison* of a PSNeuron configuration O with a dense model D gives, at a stated scope (the replaced site or the whole model), $r = P_O/P_D$, the factor $C = 1/r$ by which the dense model is larger, and the saving $S = 100(1-r)\%$, always beside the quality difference; for a replacement, D is the same host with the dense module restored. For a set A of complete configurations of one task, host, and protocol, with mean quality $\bar{q}_i$ and trainable count P_i,

$$\begin{aligned} P_A^{\text{obs}}(q) &= \min_{i \in A:\, \bar{q}_i \geq q} P_i, & \eta(q) &= P_{\text{D}}^{\text{obs}}(q)/P_{\text{PS}}^{\text{obs}}(q), \\ F_A^{\text{obs}}(p) &= \max_{i \in A:\, P_i \leq p} \bar{q}_i, & \Delta_F(p) &= F_{\text{PS}}^{\text{obs}}(p) - F_{\text{D}}^{\text{obs}}(p), \end{aligned} \tag{8}$$

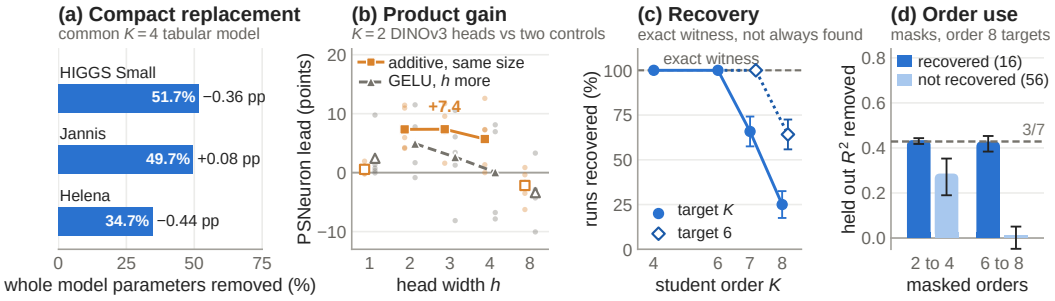

Figure 2: (a) Compact tabular replacement. The common $K = 4$ PSNeuron model removes 51.7, 49.7, and 34.7% of whole model trainable parameters relative to the validation selected dense references on HIGGS Small, Jannis, and Helena; labels show the corresponding test accuracy changes (-0.36, $+0.08$, and -0.44 points; Table 28). (b) DINOv3 head accuracy differences against equal parameter additive and slightly larger GELU heads; dots are datasets. Filled markers: five seeds, widths 2 to 4; open markers: a separate three seed study, widths 1 and 8. (c) Recovery at held out $R^2 \geq 0.9$ in 120 runs per setting; bars are bootstrap 95% intervals. (d) Held out R^2 lost on masking orders; medians and interquartile ranges over a separate 72 run study; dashed: $3/7$ for exact recovery.

so $\eta > 1$ means that PSNeuron reaches a common threshold with fewer parameters, $S_Q = 100(1 - 1/\eta)\%$ is that saving, and $\Delta_F > 0$ means higher quality under a common budget. O is *smaller and better* than D when $P_O < P_D$ and $\bar{q}_O > \bar{q}_D$, and O is non dominated in a named set when no member is at least as small and at least as good while strictly better on one of the two. Frontiers and thresholds are descriptive; only fixed comparisons paired by training seed receive a nominal 95% interval, Bonferroni checked within each stratum. The strict tier holds configurations whose learning rate search stayed inside its grid and whose tuning records agree across seeds; the descriptive tier adds the others. All views use final test means; head counts exclude the frozen backbone.

Controlled teachers. Inputs are Rademacher vectors in $\{-1, +1\}^{32}$. Each teacher is an elementary symmetric target on 8 fixed coordinates, without trainable parameters, in which each of its orders 2 to K contributes an equal share $1/(K-1)$ of the variance, $1/7$ for the order 8 teacher. Students use $R = 64$, $G = 8$, and one balanced partition per basis whose first basis separates the teacher coordinates, so Proposition 4 gives an exact witness of width one. A student of order K carries orders 2 to K, each starting at coefficient $1/\sqrt{7}$, with the calibration of the order 8 reference (Section I.1). Recovery means held out $R^2 \geq 0.9$. Order masks remove exact Walsh degrees only under these controlled assumptions; on real hosts they measure predictive dependence on a channel.

5 Results

5.1 Parameter efficiency across hosts

At $K = 4$, tabular replacements save 34.7–40.8% of whole-model parameters against restored hosts and 51.7, 49.7, and 34.7% against validation-selected dense references on HIGGS Small, Jannis, and Helena within 0.44 points; ViT feedforward replacement loses accuracy (Tables 1 and 28 and Figure 2). On HIGGS Small, progressive depth-wise $2 \to 4$ reaches 70.78% at 4,188 parameters versus 70.20% at 4,212 for uniform $4 \to 4$; secondary metrics favor it, but without an eligible fixed-pair interval this is descriptive (Table 40). On MQAR a 5,632-parameter cross-head site reaches 99.97% recall versus a 16,512-parameter projection; equal-size linear correction reaches 99.98%, low-rank correction 12.3–12.6% at the lowest grid rate, and identity carrier alone $99.96 \pm 0.03\%$, within the registered one-point allowance, making the projection dispensable; separate shared-rate PSNeuron seeds stall (Tables 22 and 30). DeltaNet's native mixer also reads state, making comparisons with it descriptive; earlier five-seed same-input GELU and SwiGLU comparisons on MQAR, state tracking, and multi-needle retrieval are unresolved (Table 32 and Section K).

With tabular rates bracketed, Jannis $K = 4$ gains 0.55 points over its larger native host (paired 95% interval $[0.19, 0.91]$); same-budget GELU is unresolved there and on HIGGS Small but wins on Helena (Section K.3). Against resized dense networks, HIGGS Small PSNeuron has higher point

Table 1: Measured replacement size and parameter efficiency. (a) C: dense/PSNeuron replaced-site count; r: PSNeuron/dense whole-model count; five-seed means; $\pm$ s.d. where shown. Quality is accuracy/recall (%) except FineWeb NLL (↓). (b) Smallest measured models reaching post hoc accuracy q; η, S_Q: Eq. (8). †PSNeuron or ‡both chose a grid edge; descriptive. §MQAR carrier alone reaches 99.96% (Table 30). FineWeb all-block controls: Table 39.

(a) Named replacement comparisons: one site unless all 6 FFNs is stated

host, task	replaced params, dense / PS	C	r	dense quality	PSNeuron quality	relation
MLP, HIGGS Small	16,576 / 2,072	8.0×	0.59	71.24 ± 0.37	70.91 ± 0.62	smaller, worse
MLP, Jannis	16,576 / 2,072	8.0×	0.61	69.68 ± 0.19	70.07 ± 0.27	smaller, better
MLP, Helena	16,576 / 2,072	8.0×	0.65	36.09 ± 0.31	35.65 ± 0.33	smaller, worse
ViT, CIFAR100	295,872 / 5,019	59×	0.89	40.62 ± 0.75	39.98 ± 0.71	smaller, worse
attention, MQAR	16,512 / 5,632	2.9×	0.96	83.15 ± 37.60	99.97 ± 0.01	projection dispensable§
DeltaNet, MQAR	130,944 / 4,400	30×	0.85	43.40 ± 48.82	77.98 ± 23.87	smaller, better
DeltaNet, multi-needle retrieval	130,944 / 4,400	30×	0.85	98.42 ± 1.64	99.61 ± 0.17	smaller, better
DeltaNet, state tracking	130,944 / 4,400	30×	0.85	84.12 ± 28.04	80.18 ± 19.60	smaller, worse
FineWeb-Edu, all 6 FFNs	3,142,656 / 392,286	8.0×	0.70	5.308 NLL	5.263 NLL	smaller, lower NLL

(b) Parameters needed to reach a common accuracy threshold q, set after testing

task, PSNeuron order	q (%)	P_{PS}	P_{D}	accuracy, PS / D (%)	η	saving S_Q (%)
HIGGS Small, $K = 2$‡	71.10	19,970	27,202	71.12 / 71.16	1.362	26.59
HIGGS Small, $K = 4$‡	71.17	19,982	35,458	71.18 / 71.24	1.774	43.65
HIGGS Small, $K = 2$‡	71.20	21,762	35,458	71.23 / 71.24	1.629	38.63
EuroSAT latent head, $K = 4$†	95.50	6,376	7,690	95.88 / 97.21	1.206	17.09
CIFAR100 ViT, $K = 2$	39.90	2,416,612	2,416,192	40.00 / 40.15	1.000	−0.02
Helena, $K = 2$	35.50	26,276	25,796	35.64 / 35.88	0.982	−1.86
Jannis, $K = 8$‡	70.00	23,547	21,578	70.06 / 70.14	0.916	−9.13

Table 2: Pareto inventory: strict / descriptive configurations. Whole model counts except frozen heads. Non dominated counts comparable strict configurations; *smaller than every control. Smaller and better uses counts and seed means; CI > 0: strict paired intervals, after Bonferroni in parentheses. Savings and controls beaten are descriptive. Definitions: Section 4 and Table 62. §Not read as PSNeuron efficiency evidence, since the identity carrier alone reaches 99.96% (Table 30).

host, site (tasks)	configs	non dominated	smaller, better	CI > 0	saving %, model (site)	controls beaten
synthetic teachers, standalone unit (11)	35 / 44	35 of 35*	34 / 46	32 (18)	39.4 to 86.8	GELU, additive
Boolean cube probes, standalone unit (3)	8 / 20	no control	0 / 19	0	15.0 to 72.5	GELU, additive
tabular MLP, feedforward (3)	48 / 98	4 of 47	8 / 311	0 (0)	0.0 to 51.9 (0.3 to 95.7)	GELU, SwiGLU
MLP on a mixed order target	2 / 4	2 of 2*	0 / 2	0	43.5 to 53.8 (93.7 to 95.7)	GELU, SwiGLU
CIFAR100 ViT, last feedforward	48 / 48	0 of 41	18 / 18	0 (0)	0.0 to 2.2 (0.1 to 79.5)	GELU
CIFAR100 ViT, all six feedforward	4 / 4	4 of 4*	0 / 0	0	none	none
CIFAR100 ViT, attention output	10 / 16	1 of 10	4 / 8	0 (0)	0.1 to 0.3 (50.0 to 87.5)	GELU, additive
MQAR, attention output§	1 / 4	not read	–	0	no size claim	none claimed
DeltaNet, channel mixer (3)	119 / 120	55 of 105	468 / 1079	66 (0)	0.0 to 14.7 (0.0 to 96.7)	GELU, SwiGLU, additive, native
FineWeb-Edu, single FFN	21 / 21	1 of 21	0 / 0	0	none	none
heads on frozen DINOv3 (2)	80 / 99	3 of 59	0 / 0	0	none	none
augmentations, 5 placements	29 / 54	4 of 18	1 / 27	0 (0)	no size claim	none claimed

estimates at 41 of 45 shared budgets by at most 0.17 points; at post hoc 71.10–71.20% thresholds it uses 26.6–43.6% fewer parameters (Table 1(b), Figure 8). Yet every HIGGS Small configuration is dominated in the strict tier. The EuroSAT latent head uses 17.09% fewer trainable parameters than a more accurate linear head. Broader dense frontiers favor Jannis, Helena, and the ViT feedforward site, whose gap is 0.15–0.52 points (Section L).

In the Pareto inventory, 18 synthetic teacher pairs beat additive after correction, while GELU also fits coherent teachers. Real hosts have 1418 descriptive smaller-and-better pairs; 66 strict DeltaNet pairs have nominal positive intervals, none after within-stratum correction, and none of the 315 advance size-matched pairs survive joint correction (Table 2 and Section L). Three $K = 2$ DeltaNet mixers (746,436–756,084 parameters) average 81.6–87.2% MQAR versus the best control's 79.65% at 772,420 parameters, descriptively, with bimodal seeds.

Site parameter and MAC ratios favor PSNeuron at tabular and MQAR sites (Table 3(a)), but 48 matched whole model time ratios span 0.83 to 20.2 (Table 18); the fused slot projection speeds up large inputs but slows nine other trained configurations, so component times show no whole model speedup. PSNeuron can launch 16× as many kernels and adds 0.25 to 1.42 ms per inference batch; the tabular MLP (0.99 versus 0.65 ms) and frozen heads (0.82 versus 0.14 ms) lag at small batches.

Table 3: Resources and runtime. (a) Site parameters and forward MACs, PSNeuron/dense (inverse in parentheses). (b) Matched whole model A100 times, PSNeuron/control; rung n has about $1/n$ of the dense budget. (c) Reference/fused projection times; (d, R, G): input width, bases, slots. All 48 pairs: Table 18.

(a) site	dense module	parameters	forward MACs
tabular feedforward	GELU, width 128	2,072 / 16,576 (8.0×)	6.06k / 16.38k (2.7×)
tabular feedforward	SwiGLU, width 128	2,072 / 24,576 (11.9×)	6.06k / 24.70k (4.1×)
MQAR attention output	linear projection	5,632 / 16,512 (2.9×)	13.54k / 16.38k (1.2×)
(b) host, site, order	batch, inputs	inference	training step
FineWeb-Edu, single FFN, $K = 4$, rung 16	16, task data	0.99 to 1.00×	0.98 to 1.01×
CIFAR100 ViT, attention output, $K = 2$	8,192, synthetic	1.03×	1.02×
CIFAR100 ViT, feedforward, $K = 2$, rung 16	8,192, synthetic	1.00×	1.02×
(c) geometry (d, R, G)	input rows	time (μs)	speedup
FineWeb-Edu, single FFN, as trained $(128, 349, 8)$	8,192	405.8 / 274.0	1.48×
frozen head, more rows $(768, 1463, 8)$	8,192	8,269.8 / 5,635.8	1.47×
frozen head, more rows $(768, 32, 8)$	65,536	1,463.0 / 776.4	1.88×
FineWeb-Edu, single FFN, more rows $(128, 349, 8)$	262,144	10,752.8 / 5,381.2	2.00×
nine other configurations, as trained	256 to 16,384	86.8 to 408.0 / 116.8 to 558.0	0.73 to 0.75×

5.2 What the product contributes

Sites with $K = 2$ isolate the grouped product against an exact parameter additive control. In a five seed study on frozen DINOv3 features, PSNeuron heads at widths 2 to 4 have higher means than the additive heads in 14 of 15 dataset and width settings and win 69 of 75 seed pairs; mean gains by width are 7.34, 7.36, and 5.70 points. Nominal paired intervals exclude zero in 11 of 15, 8 after Bonferroni correction, none favoring the additive head. Every corrected positive comparison involves a head selected at the upper grid edge (Figure 2(b), Section J.10). GELU comparisons are mixed; three corrected intervals favor GELU. The separate three-seed ladder shows the same narrow-width lead but a width-8 reversal to GELU and additive (Tables 51 and 53). A bilinear-head comparison leaves grouping-specific attribution unresolved (Section J.10).

With architecture and rates frozen, 20 fresh disjoint DeltaNet seeds give joint $K = 4$ MQAR $68.22 \pm 31.84\%$, above an exact same-slot additive twin with 48 fewer order coefficients ($24.15 \pm 31.96\%$; $+44.08$ points, nominal $[22.39, 65.77]$, $16/20$ seeds, Holm-adjusted exact sign-flip $p = 0.0020$) and conditional GELU ($30.84 \pm 37.39\%$; $+37.38$, $[15.69, 59.08]$, $14/20$, $p = 0.0027$; Table 57). Outcomes are bimodal, SwiGLU is unresolved, and the earlier joint-over-cross layout lead fails to replicate (-1.14, $[-22.15, 19.86]$); cohorts stay separate. In Boolean cube probes, $K = 8$ reaches $99.4 \pm 1.4\%$ on the entire 256-point eight-bit parity truth table versus 52.1% for a larger additive control. Order-matched $K = 4$ and 6 units solve four- and six-coordinate XOR on every seed whose partition separates the target, while additive controls stay at chance; inadmissible six-coordinate partitions leave PSNeuron at chance too (Table 41). Across 30 solved models, masking the target order lowers accuracy to 27–52%; other single-order masks leave it unchanged in $134/140$ cases.

5.3 Representable orders are not always recovered

Exact witnesses establish representability before training; spectra and masks separate recovery from use. At fixed $G = 8$, matched or nearby students recover order 4 and 6 in all 120 runs, versus 65.8 and 25% at orders 7 and 8 (Figure 2(c), Table 23). An order 8 student recovers order 6 in 64.2%; initialization and rate partly confound the declines (Sections I.1 to I.3). An aligned study varying slots and target scale recovers $44/48$ order-6 and $16/72$ order-8 runs; a smaller 2,177-parameter GELU fits both in a separate probe (mean R^2 0.999, 0.998), locating failure in recovery. Gauge fixation helps but enlarges the effective step size (Section I.2).

In the aligned teacher, removing three of seven orthogonal planted orders from exact recovery removes $3/7$ of held-out R^2 (Corollary 1); the 16 recovered models realize this decomposition (medians 0.4346, 0.4286; Figure 2(d)). In the other 56 runs, masking orders 2–4 costs median 0.29 versus only 0.01 for orders 6–8; delivered shares agree (Figure 4).

On real tasks, masking order 4 costs direct heads 41.67 to 48.91 points on EuroSAT and 0.51 to 2.09 on ImageNet100; pooled tabular masks cost 0.11 and 0.36 points on HIGGS Small and Jannis. Retraining without B_{host} lowers only Jannis accuracy; all are descriptive (Section J.6).

5.4 WHICH TARGETS AND PLACEMENTS BENEFIT

On synthetic teachers, shared geometry needs 2.00–$2.25\times$ fewer factor parameters for aligned orders but $3.43\times$ more for unaligned ones (Section I.5 and Table 27). Adverse placements include attention-score corrections (12.6–50.5% MQAR versus 99.97–99.99% bilinear), CIFAR100 cross-head (-1.06 points), and all six ViT feedforward sites (64.2% fewer parameters, 7.26–9.13 points lost; Tables 29, 34, and 49; augmentations in Tables 35 and 38). FineWeb-Edu is mixed: single-block runs trail budget-matched SwiGLU by 0.006–0.072 NLL; an all-block $K = 4$ model (6.315M parameters) scores 5.263 NLL and 193.09 perplexity versus 5.308 and 202.10 for a 9.065M-parameter SwiGLU (30.3% fewer parameters). A compatible 6.314M-parameter SwiGLU scores 5.020 NLL, so the favorable named-reference point is descriptive, not a language-family frontier (Tables 1, 37, and 39). At equal whole-model size (9.065M parameters), the completed $K = 2$ all-block PSNeuron also has slightly lower NLL than SwiGLU (5.3046 versus 5.3081), descriptively.

6 LIMITATIONS AND FUTURE WORK

Hosts remain small and teachers synthetic; DINOv3 planted features validate mask reading on continuous representations but, with no failed runs, not failure localization (Section I.4). The shared degree HOFM style control (Section 2) resolves neither equivalence nor superiority at $n = 5$. Fresh DeltaNet MQAR remains bimodal; SwiGLU and layout are unresolved, kernel variants stay separate, and the identity carrier explains the cross head saving (Table 30). Same budget GELU is unresolved or better in bracketed $K = 4$ tabular comparisons (Section 5.1); some HIGGS Small frontier and narrow DINOv3 selections remain at grid edges. Training only preprocessing preserves Helena but shifts HIGGS Small/Jannis; retuning and numerical variation preclude a sole preprocessing explanation. In the jointly corrected 315-pair family, fixed at an earlier cutoff, 113 losses and no gains survive (Section L). Savings rarely yield whole model speedups; future work should test the breadth of these parameter efficient regimes across host families at matched budgets and pursue reliable high order recovery, fused kernels, and bracketed product comparisons.

7 CONCLUSION

PSNeurons make interaction order addressable at a neural site: orders 2 through K are exact channels on shared grouped geometry, preserved as K grows and maskable after training without parameter changes. On Rademacher inputs each channel has one Walsh degree, and a vector readout permits dense-module replacement.

Common $K = 4$ tabular replacements use 34.7–40.8% fewer whole-model parameters within half a point; Jannis beats its larger host by 0.55 points but is unresolved against same-budget GELU. Progressive depth-wise order favors HIGGS Small descriptively.

At equal parameters, narrow DINOv3 PSNeuron heads lead additive heads in $14/15$ means (11 positive paired intervals, 8 corrected). On 20 fresh DeltaNet seeds, joint $K = 4$ beats exact additive and conditional GELU on MQAR after Holm correction; outcomes are bimodal, while SwiGLU, layout, and other recurrent tasks remain unresolved.

Exact witnesses separate representability from recovery: order-6 succeeds reliably, order-8 in about a quarter of runs though GELU fits both. Masks locate planted orders, including on frozen DINOv3 features, and upper-order losses in failed runs. A FineWeb-Edu all-block point beats a larger SwiGLU but trails a smaller one; broader dense frontiers favor Jannis, Helena, and ViT, while some DINOv3 GELU and width-8 additive heads lead. Made addressable, interaction order becomes an experimental coordinate of a trained neural site, not a property inferred only from aggregate accuracy.

Reproducibility Statement

The supplementary material provides the definitions, experimental protocols, complete positive and adverse results, and deterministic analyses needed to reproduce the paper's reported quantities. The operator and mathematical properties are specified in Sections A and 3; task placements and training and evaluation protocols are given in Sections 4, E, and G. Complete task results and ablations are reported in Section J, while parameter, arithmetic, timing, and memory measurements are defined with their measurement protocols in Section H. Random seeds, data splits, learning rate selection, metrics, and the separation of validation from test are stated for each study. An anonymized implementation archive is provided in the supplementary material for peer review.

AI Use Statement

The research questions, core method, experimental objectives, and scientific hypotheses were developed by the authors. Public benchmark datasets were selected and used by the authors. Generative AI tools were used for implementation support, experiment tooling, analysis support, literature retrieval, manuscript drafting and editing under author direction, and checking mathematical arguments and proofs. Reported empirical values come from recorded experiment outputs, not from a generative model. All AI assisted code, analyses, mathematical claims, proofs, citations, and manuscript text were reviewed and verified by the authors, who take responsibility for the final content.

References

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LY3ukUANko.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014. doi: 10.1038/ncomms5308.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764, 2022. doi: 10.52202/068431-1581.

Mathieu Blondel, Akinori Fujino, Naonori Ueda, and Masakazu Ishihata. Higher-order factorization machines. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3351–3359, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/158fc2ddd52ec2cf54d3c161f2dd6517-Abstract.html.

Mathieu Blondel, Vlad Niculae, Takuma Otsuka, and Naonori Ueda. Multi-output polynomial networks and factorization machines. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3349–3359, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dea9ddb25cbf2352cf4dec30222a02a5-Abstract.html.

Yixin Cheng, Grigorios Chrysos, Markos Georgopoulos, and Volkan Cevher. Multilinear operator networks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bbCL5aRjUx.

Grigorios G. Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Yannis Panagakis, Jiankang Deng, and Stefanos Zafeiriou. Π-nets: Deep polynomial neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7325–7335, 2020. doi: 10.1109/CVPR42600.2020.00735.

Grigorios G. Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Jiankang Deng, Yannis Panagakis, and Stefanos Zafeiriou. Deep polynomial neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4021–4034, 2022. doi: 10.1109/TPAMI.2021.3058891.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 933–941. PMLR, 2017. URL https://proceedings.mlr.press/v70/dauphin17a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Abhimanyu Dubey, Filip Radenovic, and Dhruv Mahajan. Scalable interpretability via polynomials. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2663. URL https://arxiv.org/abs/2205.14108.

Richard Durbin and David E. Rumelhart. Product units: A computationally powerful and biologically plausible extension to backpropagation networks. *Neural Computation*, 1(1):133–142, 1989. doi: 10.1162/NECO.1989.1.1.133.

Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Müller, Joaquin Vanschoren, and Frank Hutter. OpenML-Python: an extensible python api for OpenML. *Journal of Machine Learning Research*, 22(100):1–5, 2021. URL https://www.jmlr.org/papers/v22/19-920.html.

C. Lee Giles and Tom Maxwell. Learning, invariance, and generalization in high-order neural networks. *Applied Optics*, 26(23):4972–4978, 1987. doi: 10.1364/AO.26.004972.

Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973. doi: 10.1137/0710036.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 355–364, 2017. doi: 10.1145/3077136.3080777.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. doi: 10.48550/arXiv.1606.08415. URL https://arxiv.org/abs/1606.08415.

Frank L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927. doi: 10.1002/sapm192761164.

Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rylnK6VtDH.

Yu-Chin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 43–50, 2016. doi: 10.1145/2959100.2959134.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500, 2009. doi: 10.1137/07070111X.

Alex Krizhevsky. Learning multiple layers of features from tiny images. MSc thesis, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Laurens R. Leerink, C. Lee Giles, Bill G. Horne, and Marwan A. Jabri. Learning with product units. In *Advances in Neural Information Processing Systems 7*, 1994. URL https://proceedings.neurips.cc/paper/1994/hash/4fac9ba115140ac4f1c22da82aa0bc7f-Abstract.html.

Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. xDeepFM: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1754–1763, 2018. doi: 10.1145/3219819.3220023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011. URL https://aclanthology.org/P11-1015/.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL https://research.google/pubs/reading-digits-in-natural-images-with-unsupervised-feature-learning/.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P. Vetrov. Tensorizing neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 442–450, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/6855456e2fe46a9d49d3d3af4f57443d-Abstract.html.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Michael T. Pearce, Thomas Dooms, Alice Rigg, Jose Oramas, and Lee Sharkey. Bilinear MLPs enable weight-based mechanistic interpretability. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gI0kPklUKS.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-0970.

Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. Product-based neural networks for user response prediction. In *IEEE International Conference on Data Mining*, pp. 1149–1154, 2016. doi: 10.1109/ICDM.2016.0151. URL https://arxiv.org/abs/1611.00144.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Steffen Rendle. Factorization machines. In *IEEE International Conference on Data Mining*, pp. 995–1000, 2010. doi: 10.1109/ICDM.2010.127.

David E. Rumelhart, Geoffrey E. Hinton, and James L. McClelland. A general framework for parallel distributed processing. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pp. 45–76. MIT Press, 1986. doi: 10.7551/mitpress/5236.003.0005.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3067–3075. PMLR, 2017. URL https://proceedings.mlr.press/v70/shalev-shwartz17a.html.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. doi: 10.48550/arXiv.2002.05202. URL https://arxiv.org/abs/2002.05202.

Y. Shin and J. Ghosh. The pi-sigma network: an efficient higher-order neural network for pattern classification and function approximation. In *International Joint Conference on Neural Networks*, volume 1, pp. 13–18, 1991. doi: 10.1109/IJCNN.1991.155142.

Yoan Shin and Joydeep Ghosh. Ridge polynomial networks. *IEEE Transactions on Neural Networks*, 6(3):610–622, 1995. doi: 10.1109/72.377967.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. doi: 10.48550/arXiv.2508.10104. URL https://arxiv.org/abs/2508.10104.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. AutoInt: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 1161–1170, 2019. doi: 10.1145/3357384.3357925.

Trond Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983. doi: 10.1137/0720042.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, volume 12356 of *Lecture Notes in Computer Science*, pp. 776–794. Springer, 2020. doi: 10.1007/978-3-030-58621-8_45.

Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. Neural interaction transparency (NIT): Disentangling learned interactions for improved interpretability. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5804–5813, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/74378afe5e8b20910cf1f939e57f0480-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & Cross Network for ad click predictions. In *Proceedings of the ADKDD Workshop*, pp. 12:1–12:7, 2017. doi: 10.1145/3124749.3124754.

Ruoxi Wang, Rakesh Shivanna, Derek Zhiyuan Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. DCN V2: Improved Deep & Cross Network and practical lessons for web-scale learning to rank systems. In *The Web Conference*, pp. 1785–1797, 2021. doi: 10.1145/3442381.3450078.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, pp. 115491–115522, 2024. doi: 10.52202/079017-3668.

Feng Yu, Zhaocheng Liu, Qiang Liu, Haoli Zhang, Shu Wu, and Liang Wang. Deep interaction machine: A simple but effective model for high-order feature interactions. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 2285–2288, 2020. doi: 10.1145/3340531.3412077.

Haoyu Yun, Hamid Krim, and Yufang Bao. kVNN: Learnable multi-kernel Volterra neural networks. *arXiv preprint arXiv:2604.15141*, 2026. doi: 10.48550/arXiv.2604.15141. URL `https://arxiv.org/abs/2604.15141`.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. doi: 10.48550/arXiv.1910.04867. URL `https://arxiv.org/abs/1910.04867`.

A Mathematical Properties of Product Spectrum Neurons

This appendix collects the algebraic statements behind Section 3. Unless stated otherwise they concern the raw uncentered spectrum of one site, before a host adds any residual path, because a host block can add linear or constant terms that lie outside the spectrum. Most of the facts are elementary; we state them with the exact hypotheses under which the paper uses them.

A.1 Notation and the raw spectrum

For basis r, the d input coordinates are partitioned into disjoint nonempty blocks $\mathcal{G}_{r,1}, \dots, \mathcal{G}_{r,G}$, shared by all bases or drawn separately for each basis (Section 3); every statement below holds in both cases unless it says otherwise. The slots, channels, basis values, and site output are defined in Eqs. (1) to (3). We use $e_{r,j}$ for every j from 0 to G, with $e_{r,0} = 1$, although the operator uses only orders 2 to K. Every reported configuration has $\mu_j = 0$ and $\zeta_j = 1$ (Table 4), so there the delivered channel $\tilde{e}_{r,j}$ equals $\hat{e}_{r,j}$.

A.2 Forward computation, tensor shapes, and differentiation

For shape bookkeeping, flatten any leading batch, token, or spatial axes into N positions; the implementation preserves those leading axes rather than materializing this flattening. The core therefore maps

$$x \in \mathbb{R}^{N\times d} \longrightarrow q \in \mathbb{R}^{N\times R\times G} \longrightarrow (\tilde{e}_{r,j}) \in \mathbb{R}^{N\times R\times (K-1)} \longrightarrow s \in \mathbb{R}^{N\times R} \longrightarrow P(x) \in \mathbb{R}^{N\times o}.$$

Indices range over $r \in \{1, \dots, R\}$, $g \in \{1, \dots, G\}$, $j \in \{2, \dots, K\}$, and output coordinate $u \in \{1, \dots, o\}$. When a placement uses a linear conditioner from a host width d_{h}, $A_{\mathrm{host}} : \mathbb{R}^{d_{\mathrm{h}}} \to \mathbb{R}^{d}$; a linear outer readout to host width o_{h} has $B_{\mathrm{host}} : \mathbb{R}^{o} \to \mathbb{R}^{o_{\mathrm{h}}}$. Nonlinear conditioners have the same input/output shape contract but move the exact-degree statement to their output coordinates (Proposition 1).

One forward pass has five stages. First, the placement supplies the coordinates seen by the neuron, either x directly or $A_{\mathrm{host}}(x)$. Second, each row W_r forms the G slot values of Eq. (1); within a basis the blocks are disjoint, nonempty, and cover the coordinates assigned to that partition. Third, the recurrence of Eq. (16) produces every ESP $e_{r,j}$ through order K from the same slots. Fourth, frozen calibration/centering/gauge constants produce $\tilde{e}_{r,j}$ and the mask and order coefficients form $s_r = \sum_j \gamma_{r,j} m_j \tilde{e}_{r,j}$. Fifth, B_{core} maps the R basis values to o outputs, after which the host applies

Table 4: Notation for the PSNeuron of Section 3 and for this appendix. The middle column gives the shape of the whole tensor at one site, or marks a quantity as frozen, that is, fixed before training and never updated; a symbol indexed by r, g, or j denotes a single entry. Rows below cover the core forward computation and the scalar-output diagnostics. A few letters take a second meaning where they are defined: r, C, and S for the parameter ratio, the factor by which the dense model is larger, and the saving of Section 4; C also for a number of classes, λ_j for the coefficients of a teacher, S for a set of coordinates and for the rank of an order specific correction, whose factors are U, V, and A and whose rows have Σ norm n (Section D), s_j for the delivered share of order j (Eq. (20)), m for a degree bound (Corollary 2), and β for the moment coefficients of Adam or the scale of a product target.

symbol	shape or status	meaning
d, o	integers	input and output widths of the site
N	integer	number of leading positions after flattening batch/token/spatial axes for shape bookkeeping; the implementation preserves those leading axes
R, G, K	integers	bases, slots per basis, and maximum order, with $2 \le K \le G \le d$
r, g, j, u	indices	basis $1:R$, slot $1:G$, carried order $2:K$, and output coordinate $1:o$
x	d	input of the PSNeuron; a placement with a conditioner feeds it $z = A_{\mathrm{host}}(x)$ instead (Eq. (6))
W_0	$R \times d$	trained factor bank shared by every order, $W^{(j)} = W_0$ in Eq. (5); row W_r serves basis r
$\mathcal{G}_{r,g}$	partition	coordinates of slot g of basis r; disjoint and nonempty within a basis, fixed at construction
$q_{r,g}$	scalar	slot value, a linear form without bias (Eq. (1))
$e_{r,j}$	scalar	elementary symmetric polynomial of order j in the slots of basis r
$E_j^{(g)}$	scalar	running order-j ESP after the first g slots in Eq. (16); $E_0^{(g)} = 1$ and $E_j^{(0)} = 0$ for $j > 0$
c_j	frozen	calibration constant of order j, shared by all bases (Eq. (10))
$\hat{e}_{r,j}$	scalar	$c_j \binom{G}{j}^{-1/2} e_{r,j}$; calibrated raw channel of exact order j
μ_j	frozen	optional order center; 0 in every reported configuration
ζ_j	frozen	gauge factor for a common rescaling of the bank (Section A.3); 1 in every reported configuration
$\tilde{e}_{r,j}$	scalar	$\zeta_j(\hat{e}_{r,j} - \mu_j)$; delivered channel
m_j	$\{0, 1\}$	order mask; setting it changes no parameter or frozen constant
$\mathcal{J}$	set	carried orders, a subset of $\{2, \dots, K\}$; $m_j = 0$ for every order outside it
γ	$R \times (K-1)$	order coefficients $\gamma_{r,j}$; trained when several orders are carried, except in the fixed mixture control (Section A.14), and fixed at 1 for a single order
s	R	basis values $s_r = \sum_j \gamma_{r,j} m_j \tilde{e}_{r,j}$
$B_{\mathrm{core}}, b_{\mathrm{core}}$	$o \times R, o$	core readout with columns b_r; optional bias, used only by the DINOv3 heads of the ladder and its five seed study
P	o	site output $P(x) = B_{\mathrm{core}} s(x) + b_{\mathrm{core}}$
$A_{\mathrm{host}}, B_{\mathrm{host}}$	placement	optional conditioner and outer readout of a placement, both outside the core (Eq. (6))
$C_{\mathrm{host}}, \lambda$	placement	carrier kept by an augmentation and the scale of the added branch (Eq. (7))
$C^{(r)}$	$o \times (K-1)$	output by order coefficients $b_r \gamma_r^\top$ of basis r (Eq. (17))
β	$R \times (K-1)$	combined coefficients $B_{\mathrm{core},1r} \gamma_{r,j}$ of a scalar output; derived, not a parameter
κ_j	scalar output	contribution of order j to a scalar prediction, $\kappa_j(x) = \sum_r B_{\mathrm{core},1r} \gamma_{r,j} m_j \tilde{e}_{r,j}(x)$; derived, not a parameter
$y, \hat{y}, s_j$	scalars	target, scalar prediction, and signed delivered share $s_j = \widehat{\mathrm{Cov}}(y, \kappa_j)/\widehat{\mathrm{Var}}(y)$ in Eq. (20); s_j is distinct from the basis value s_r

the replacement or augmentation map of Eqs. (6) and (7). Thus a channel of order j is a product of j distinct slots; expanding the slots can select at most one coordinate contribution from each selected block. Two raw coordinates placed in the same block of one basis cannot therefore occur as two separate factors of one monomial from that basis. If a conditioner mixes coordinates first, this statement applies to the conditioned coordinates.

Training uses ordinary reverse-mode differentiation through these additions and multiplications. For the raw ESP,

$$\frac{\partial e_{r,j}}{\partial q_{r,g}} = e_{j-1}(q_{r,1}, \dots, \widehat{q_{r,g}}, \dots, q_{r,G}), \tag{9}$$

where the hat means that slot g is omitted. Hence

$$\frac{\partial s_r}{\partial \gamma_{r,j}} = m_j \tilde{e}_{r,j}, \qquad \frac{\partial s_r}{\partial e_{r,j}} = \gamma_{r,j} m_j \zeta_j c_j \binom{G}{j}^{-1/2},$$

and the chain rule continues through $q_{r,g}$ to W_r and to a learned host conditioner. The constants c_j, μ_j, and ζ_j, the partition, and an inference-time order mask are not optimizer parameters. At $K = 2$ the single carried coefficient is fixed to 1, so no gradient or trainable parameter is assigned to γ; the readout supplies the per-basis scale. For $K \ge 3$ with a learned mixture, gradients also update γ.

Masking an order changes only the multiplicative selector m_j used in the forward pass: it neither retrains nor mutates W_0, γ, B_{core}, the partition, or the frozen calibration state. Section B.1 checks the recurrence against direct subset enumeration in both values and gradients.

A.3 Calibration and frozen constants

The constants c_j, μ_j, and ζ_j do not depend on the input of a forward pass and are never trained. Calibration fixes c_j once, before the first optimizer step, on the fixed slice $\mathcal{D}_{\mathrm{cal}}$ of training data described in Sections 3 and G.1. It first multiplies the factor bank by the single scalar $1/\operatorname{RMS}(q)$, with the root mean square taken jointly over every example and position of $\mathcal{D}_{\mathrm{cal}}$, the R bases, and the G slots. On the rescaled bank it then sets one scalar per order, shared by all bases,

$$c_j = 1 \Big/ \operatorname{RMS}_{x \in \mathcal{D}_{\mathrm{cal}},\, r \le R}\Big(\tbinom{G}{j}^{-1/2} e_{r,j}(x) \Big), \tag{10}$$

so that the calibrated channels of each order have unit root mean square jointly over $\mathcal{D}_{\mathrm{cal}}$ and the R bases. Every such statistic is required to be finite and nonzero and is never floored. Since c_j is computed from order j alone, it does not depend on the maximum order K. The PSNeuron heads of the DINOv3 ladder and of its five seed study keep the measured bank scale as a frozen multiplier on the slots instead (Section G.1).

Centering. Optional centering instead sets c_j to the reciprocal standard deviation of the same quantity and fixes the frozen center μ_j, equal to c_j times its mean, so that the centered channels of each order have zero mean and unit root mean square jointly over $\mathcal{D}_{\mathrm{cal}}$ and the R bases.

Gauge factor. The factor ζ_j compensates an exact common rescaling of the factor bank. If every row of W_0 is multiplied by the same $a \neq 0$ while c_j stays fixed, every slot is multiplied by a, because it is linear in the row, and $\hat{e}_{r,j}$ by a^j. Replacing ζ_j by $a^{-j}\zeta_j$ then leaves every delivered channel $\tilde{e}_{r,j}$, and hence the site output, unchanged with γ and B_{core} held fixed, provided $\mu_j = 0$; a nonzero center does not scale, so the compensation applies only to uncentered channels. The factor is neither trained nor fitted to data. It equals one in every reported model; only the optional power of two rescaling of the factor bank in Section D, which one configuration of the optimizer selection round there used, changes it during training.

A.4 Exact order homogeneity

Proposition 1 (Exact homogeneity of each raw order channel)**.** *Assume every slot $q_{r,g}$ is a linear form without bias in the input to the Product Spectrum Neuron. Then, for every $j \in \{2, \dots, K\}$ and every real scalar a,*

$$\hat{e}_{r,j}(ax) = a^j \hat{e}_{r,j}(x). \tag{11}$$

Proof. Linearity gives $q_{r,g}(ax) = a q_{r,g}(x)$. Every monomial in $e_{r,j}$ contains exactly j slot values, so every monomial acquires a factor a^j. The constants c_j and $\binom{G}{j}^{-1/2}$ do not depend on the input and therefore preserve the identity. □

The channel is thus homogeneous of degree j, and its polynomial degree is exactly j unless it vanishes identically, which happens exactly when fewer than j of its slots are nonzero linear forms.

Reference frame of the degree statement. Equation (11) holds in the coordinates entering the neuron. If a host applies a nonlinear conditioner $z = A_{\mathrm{host}}(x)$ before the neuron, the identity remains true in z, but the composition $\hat{e}_{r,j}(A_{\mathrm{host}}(x))$ need not be a polynomial of degree j in the original x. Linear conditioners without bias preserve homogeneity in x, although a nonzero channel can vanish after a map of deficient rank. A dense conditioner can also mix the supports of different slots, so squarefree monomials and exact Walsh degree in the original coordinates need an additional condition of disjoint support; homogeneity alone does not imply them.

Centering and finite precision. The centered channel $\hat{e}_{r,j}(x) - \mu_j$ adds a term of degree zero and is generally not homogeneous, and the factor ζ_j, which does not depend on the input, cannot restore homogeneity. Every statement about exact order therefore refers to the uncentered spectrum. Proposition 1 is an identity in real arithmetic. Rounded addition and multiplication commute with scaling by a power of two while every intermediate value stays in the normal range, so repeating the same operations in the same order reproduces the identity bit for bit, as Section B.1 checks for $a \in \{1/2, 2, 4\}$.

A.5 Grouped quadratic structure at order two

For a fixed basis r, collect the slot maps into $Q_r \in \mathbb{R}^{G \times d}$ with $(Q_r)_{g,i} = |\mathcal{G}_{r,g}|^{-1/2} W_{r,i}$ for $i \in \mathcal{G}_{r,g}$ and 0 otherwise, so that $q_r(x) = Q_r x$ in the coordinates x entering the neuron. Here $\mathbf{1} \in \mathbb{R}^G$ is the all-ones vector and I_G the $G \times G$ identity. The quadratic channel obeys

$$e_{r,2}(x) = \frac{1}{2}\left[(\mathbf{1}^\top q_r)^2 - q_r^\top q_r\right] = \frac{1}{2} x^\top Q_r^\top (\mathbf{1}\mathbf{1}^\top - I_G) Q_r x. \tag{12}$$

Proposition 2 (Hessian rank of one grouped quadratic basis)**.** *If $G \geq 2$ and W_r has a nonzero entry in every block $\mathcal{G}_{r,1}, \ldots, \mathcal{G}_{r,G}$, then*

$$\nabla_x^2 e_{r,2} = Q_r^\top (\mathbf{1}\mathbf{1}^\top - I_G) Q_r$$

has rank G, and every entry whose row and column index lie in the same block is zero.

Proof. The rows of Q_r are nonzero and have disjoint supports, so they are linearly independent and Q_r has full row rank G. Hence Q_r maps $\mathbb{R}^d$ onto $\mathbb{R}^G$ and $Q_r^\top$ is injective. The matrix $M = \mathbf{1}\mathbf{1}^\top - I_G$ has eigenvalue $G - 1$ along $\mathbf{1}$ and eigenvalue -1 on its orthogonal complement, so it is invertible for $G \geq 2$, and $Q_r^\top M Q_r$ has rank G. For $i, k \in \mathcal{G}_{r,g}$ only row g of Q_r is nonzero at i and at k, so the (i, k) entry equals $(Q_r)_{g,i} M_{gg} (Q_r)_{g,k} = 0$. □

With $K = 2$ and $\mu_2 = 0$ the basis value is $s_r = \gamma_{r,2} m_2 \zeta_2 c_2 \binom{G}{2}^{-1/2} e_{r,2}$, so its Hessian is the same matrix times a scalar that is nonzero whenever order 2 is carried with a nonzero coefficient. A scalar ridge unit $\phi(u^\top x)$ that is twice differentiable reads the same d coordinates but has Hessian $\phi''(u^\top x) u u^\top$, of rank at most one. A sum of G ridge units applied block by block, $\sum_g \phi(u_g^\top x_{\mathcal{G}_{r,g}})$, can reach rank G with the same weights and partition, but its Hessian is block diagonal; with $\phi = \mathrm{GELU}$ it is one basis of the additive control of Section A.14. The grouped product has the complementary pattern, since all of its curvature couples different blocks. This is a local structural distinction. It is not a theorem that PSNeuron represents more functions at fixed width, optimizes more easily, or explains a score on a real task.

A.6 Nesting with respect to the maximum order

Remark 1 (Lower order channels are preserved when K grows). Let $2 \leq K < K' \leq G$. At fixed G, factor bank, partition, and frozen constants, each channel $\hat{e}_{r,j}$ with $j \leq K$ depends only on the slots, on j, and on c_j, and c_j does not depend on the maximum order (Section A.3). Exposing the channels of orders $K + 1, \ldots, K'$ therefore leaves every $\hat{e}_{r,j}$ and $\tilde{e}_{r,j}$ with $j \leq K$ unchanged. The recurrence of Eq. (16) computes each E_j from orders no higher than j, so the identity also holds exactly in floating point arithmetic (Section B.1). Constants fitted separately at K and K' agree in real arithmetic; the fixed G study installs the constants of its order 8 reference so that they also agree bit for bit (Section I.1).

Nesting concerns the channels available to a model, not the functions that trained models compute. It does not state that $e_{r,K'}$ contains $e_{r,K}$ as a component, since the two are distinct homogeneous polynomials over the same slots. Nor does it make the basis values of two models equal. The added orders enter s_r with coefficients that are nonzero at initialization, so models trained at K and at K' start from different functions, which is why degree and initialization remain partly confounded in the recovery study (Section 5.3).

A.7 WALSH LEVELS ON THE BOOLEAN CUBE

Substituting the slots into $e_{r,j}$ gives

$$e_{r,j}(x) = \sum_{S} \prod_{i \in S} |\mathcal{G}_{r,g(i)}|^{-1/2} W_{r,i}\, x_i, \tag{13}$$

where S runs over the sets of j coordinates that meet every block of basis r at most once and $g(i)$ is the block that contains i. This is the degree j ANOVA kernel of higher order factorization machines (Blondel et al., 2016) with its sum restricted to such sets, and the restriction disappears when every block is a singleton. Summing all orders gives $\sum_{j=0}^{G} e_{r,j} = \prod_{g=1}^{G}(1 + q_{r,g})$, the product of a pi-sigma unit (Shin & Ghosh, 1991) with unit biases and summing units restricted to disjoint blocks, before its output nonlinearity.

For x uniform on $\{-1, +1\}^d$, the Walsh characters $\chi_S(x) = \prod_{i \in S} x_i$ form an orthonormal basis of the functions of x. The *Walsh level* k of a function is its component of Walsh degree k, the sum of its terms on characters with $|S| = k$, and by Parseval's identity $\mathbb{E}[(y - f)^2] = \sum_k \mathbb{E}[(t_k - f_k)^2]$ whenever y and f have levels t_k and f_k.

Proposition 3 (Walsh level of each order and orthogonality across orders)**.** *Let the coordinates that enter the neuron be x uniform on $\{-1, +1\}^d$, read without a conditioner through disjoint blocks. Then each raw channel $\hat{e}_{r,j}$ is a linear combination of Walsh characters χ_S with $|S| = j$, and for $j \neq j'$ and any bases r, r',*

$$\mathbb{E}\big[\hat{e}_{r,j}(x)\hat{e}_{r',j'}(x)\big] = 0. \tag{14}$$

Proof. By Eq. (13), every monomial of $e_{r,j}$ takes one coordinate from each of j distinct blocks. The blocks are disjoint, so no coordinate appears twice and the monomial is the character χ_S of a set with $|S| = j$. Distinct characters are orthogonal under the uniform distribution, and characters of different sizes are distinct. The argument uses each basis separately, so it covers partitions shared by all bases and partitions drawn for each basis. □

Orthogonality holds for this distribution. On a general continuous or empirical distribution different degrees need not be uncorrelated, and a nonzero center also breaks it. Since $\mathbb{E}[\hat{e}_{r,j}] = 0$ for every $j \geq 2$, the centered channels of orders $j \neq j'$ satisfy $\mathbb{E}[(\hat{e}_{r,j} - \mu_j)(\hat{e}_{r',j'} - \mu_{j'})] = \mu_j \mu_{j'}$. The controlled experiments use Rademacher inputs because the decomposition by order is then exact without any further projection.

Corollary 1 (Order masks remove exact Walsh levels)**.** *Assume the hypotheses of Proposition 3, a scalar output, and $\mu_j = 0$. Write the prediction as $\hat{y}(x) = b_{\text{core}} + \sum_{j=2}^{K} \kappa_j(x)$ with order contributions $\kappa_j(x) = \sum_{r=1}^{R} B_{\text{core},1r} \gamma_{r,j} \tilde{e}_{r,j}(x)$, as in Eq.* (19)*, and let $\hat{y}_{-J}$ be the prediction with $m_j = 0$ for every j in a set J of orders. Each κ_j lies in Walsh level j, so the mask removes exactly the levels in J and leaves every other level of $\hat{y}$ unchanged. For a target y with Walsh levels t_k and $\mathrm{Var}(y) > 0$, and with the population $R^2(f) = 1 - \mathbb{E}[(y - f)^2]/\mathrm{Var}(y)$ of a predictor f, the R^2 lost by the mask is*

$$R^2(\hat{y}) - R^2(\hat{y}_{-J}) = \sum_{j \in J} \frac{\mathbb{E}[t_j^2] - \mathbb{E}[(t_j - \kappa_j)^2]}{\mathrm{Var}(y)} = \sum_{j \in J} \frac{2\,\mathbb{E}[y\kappa_j] - \mathbb{E}[\kappa_j^2]}{\mathrm{Var}(y)}. \tag{15}$$

The loss never exceeds $\sum_{j \in J} \mathbb{E}[t_j^2]/\mathrm{Var}(y)$, the share of target variance at the masked levels, equals that share exactly when $\kappa_j = t_j$ for every $j \in J$, and is at least that share minus $1 - R^2(\hat{y})$.

Proof. By Proposition 3, κ_j is its own Walsh level j and $\hat{y}$ has no other nonconstant terms. Parseval's identity gives $\mathbb{E}[(y - \hat{y})^2] = (t_0 - b_{\text{core}})^2 + \sum_{k \geq 1} \mathbb{E}[(t_k - \kappa_k)^2]$, where $t_0 = \mathbb{E}[y]$ and $\kappa_k = 0$ outside $\{2, \ldots, K\}$. The mask replaces $\mathbb{E}[(t_j - \kappa_j)^2]$ by $\mathbb{E}[t_j^2]$ for each $j \in J$ and leaves every other term unchanged. Subtracting and dividing by $\mathrm{Var}(y)$ gives the first form, and $\mathbb{E}[t_j \kappa_j] = \mathbb{E}[y\kappa_j]$, which holds because κ_j is orthogonal to every other level, gives the second. Each summand of the first form is at most $\mathbb{E}[t_j^2]/\mathrm{Var}(y)$, with equality exactly when $t_j = \kappa_j$. The terms $\mathbb{E}[(t_j - \kappa_j)^2]$ with $j \in J$ sum to at most $\mathbb{E}[(y - \hat{y})^2] = (1 - R^2(\hat{y}))\,\mathrm{Var}(y)$, which gives the lower bound. □

Only the masked orders enter Eq. (15), so errors at the other orders do not affect the predicted loss. The controlled teacher of order 8 has the form of Proposition 4 with $G = 8$ and $\lambda_j^2 = 1/7$; since $\binom{G}{j}^{-1/2} e_j$ of G Rademacher signs is a normalized sum of $\binom{G}{j}$ distinct characters, it plants variance $\mathrm{Var}(y)/7$ at each of the orders 2 to 8 and nothing elsewhere. Masking orders 2 to 4, or 6 to 8, therefore lowers the R^2 of a model with $\kappa_j = t_j$ at the masked orders by exactly $3/7$, and leaves that of a model with $\kappa_j = 0$ at the masked orders unchanged. Held out R^2 estimates the population quantity on a finite sample, and a model counted as recovered (held out $R^2 \geq 0.9$) need not satisfy $\kappa_j = t_j$ exactly. The value $3/7$ is therefore the reference for the medians measured in Section 5.3. By the lower bound of Corollary 1, a model whose population R^2 is at least 0.9 loses between $3/7 - 0.1$ and $3/7$ under either mask, so a loss near $3/7$ follows from recovery and is not separate evidence for it, and a held out estimate can exceed $3/7$ only through sampling error. The ratio $\mathbb{E}[y\kappa_j]/\mathrm{Var}(y)$ is the population value of the delivered share in Eq. (20).

A.8 Exact representational witness for the controlled teachers

Write $e_j(v_1, \ldots, v_G)$ for the elementary symmetric polynomial of order j in G variables.

Proposition 4 (Width one witness for the planted teachers)**.** *Fix a basis $r_\star$ of the student and one coordinate $i_g \in \mathcal{G}_{r_\star,g}$ in each of its G blocks, and let the target be*

$$y(x) = \sum_{j=2}^{K_T} \lambda_j \binom{G}{j}^{-1/2} e_j(x_{i_1}, \ldots, x_{i_G})$$

with real coefficients λ_j and $2 \leq K_T \leq G$. Consider a PSNeuron with scalar output, any width $R \geq 1$, maximum order $K \geq K_T$, orders 2 to K_T carried with learned coefficients, $\mu_j = 0$, and any frozen nonzero constants c_j and ζ_j. Some setting of its trainable parameters gives $P(x) = y(x)$ for every input x.

Proof. Set $W_{r_\star,i_g} = |\mathcal{G}_{r_\star,g}|^{1/2}$ for $g = 1, \ldots, G$ and every other entry of row $r_\star$ to zero, so that $q_{r_\star,g}(x) = x_{i_g}$ and $e_{r_\star,j}(x) = e_j(x_{i_1}, \ldots, x_{i_G})$. Set $\gamma_{r_\star,j} = \lambda_j/(c_j\zeta_j)$ for $2 \leq j \leq K_T$ and $\gamma_{r_\star,j} = 0$ for every other carried order, the readout entry of $r_\star$ to 1, and every other readout entry and b_{core} to 0. Then

$$P(x) = s_{r_\star}(x) = \sum_{j=2}^{K_T} \gamma_{r_\star,j}\, \zeta_j c_j \binom{G}{j}^{-1/2} e_{r_\star,j}(x) = y(x).$$

The frozen constants are absorbed into $\gamma_{r_\star,j}$, and the other $R - 1$ bases are switched off by their readout entries. □

The teachers of the fixed G study have exactly this form, with $G = 8$, a support that places one coordinate in each block of the student's first basis, and $\lambda_j = \pm 1/\sqrt{7}$ (Section I.1); the aligned diagnostic uses the same form with $G = K_T$ and coefficients of magnitude $1/\sqrt{K_T - 1}$ (Section I.2). On Rademacher inputs e_j of G signs is a sum of $\binom{G}{j}$ distinct characters, so each planted order contributes variance λ_j^2: $1/7$ in the fixed G study and $1/(K_T - 1)$ in the aligned diagnostic, whose targets have unit variance. In both, every student of order at least K_T contains an exact witness of width one; for $K < K_T$, Corollary 2 bounds what any setting can reach. The construction also needs one basis whose partition separates the support. When every basis shares one partition, the input is not conditioned, and the readout is linear, a Walsh character with two support coordinates in the same block has zero coefficient in the output at every order and every width R (Section J.3). The witness is specific to this family and is not a universality statement. An arbitrary polynomial of degree K over d coordinates can require many bases, because one basis ties all of its monomials through the same slot weights.

A.9 A conditional degree ceiling

The Boolean construction also bounds what a model of lower order can reach when its realized predictor is genuinely limited in degree.

Corollary 2 (R^2 ceiling of a predictor of limited degree)**.** *Let x be uniform on $\{-1,+1\}^d$, let f have a Walsh expansion on characters of degree at most m, which any polynomial of total degree at most m has after reducing $x_i^2 = 1$, and let y have Walsh levels t_k and $\mathrm{Var}(y) > 0$. Then $\mathbb{E}[(y-f)^2] \geq \sum_{k>m} \mathbb{E}[t_k^2]$, so the population R^2 of f is at most $\sum_{1\leq k\leq m} \mathbb{E}[t_k^2]/\mathrm{Var}(y)$. A PSNeuron with scalar output that reads x without a conditioner and has maximum order K is such a predictor with $m = K$, so for a planted teacher through order K_T whose orders carry equal shares of the target variance, as in Table 24, its ceiling is $(\min(K, K_T) - 1)/(K_T - 1)$, the ceiling column of that table.*

Proof. Split y into its levels up to m and above m. Characters of different sizes are orthogonal, so the cross term vanishes, so $\mathbb{E}[(y-f)^2] = \mathbb{E}[(y_{\leq m} - f)^2] + \sum_{k>m} \mathbb{E}[t_k^2]$. The teacher puts variance λ_j^2 at each level $2, \ldots, K_T$ with λ_j^2 constant, which gives the stated ratio. □

The hypothesis concerns the realized function, not one layer. Composing several layers of limited degree can raise the total degree, so these statements do not apply to a deep network merely because each of its layers has limited degree. Zero correlation of a real valued score with a parity target also does not impose a chance level ceiling on its thresholded classifier, since thresholding is a separate nonlinear operation; the counterexample on scores versus decisions in Section J.3 makes this explicit.

A.10 Elementary symmetric recurrence

The elementary symmetric polynomials (ESPs) of the slots are evaluated without enumerating the $\binom{G}{j}$ subsets, by the classical recurrence that Blondel et al. (2016) use for ANOVA kernels. For one basis r write q_g for $q_{r,g}$. Set $E_0 = 1$ and $E_j = 0$ for $0 < j \leq K$, and process slot q_g, for $g = 1, \ldots, G$, by

$$E_j \leftarrow E_j + q_g E_{j-1}, \qquad j = \min(g, K), \ldots, 1, \tag{16}$$

with j descending.

Remark 2 (Correctness, and why the order of the sweep matters only in place). After the first g slots, E_j is the ESP of order j in those slots, because a subset of order j either excludes q_g, contributing the previous E_j, or includes it, contributing $q_g E_{j-1}$. Each E_j must therefore read the value of E_{j-1} from the previous slot. When E is updated in place, the descending sweep guarantees this. Applying Eq. (16) to the whole vector at once, out of place, reads only values from the previous slot and so needs no ordering over j. Every multiplication and addition of the descending sweep appears in the out of place form with the same operands in the same order; the out of place form also multiplies and adds the exact zeros $q_g \cdot 0$ for orders that the descending sweep skips, so the two forms give the same values.

Every reported model was trained with the descending form under automatic differentiation; the released implementation evaluates without gradients in the out of place form. The out of place form launches three kernels per slot, a multiplication, a concatenation, and an addition, instead of a pair per slot and order, 37 rather than 83 kernels for the recurrence at $G = K = 8$, and in exchange executes GK multiplications per basis, $\binom{K}{2}$ more than the $\sum_{g=1}^{G} \min(g, K)$ of the descending sweep. At $G = 8$ and every tested K the two forms give bitwise equal forward values, and on an A100 GPU their gradients, summed in a different order and compared at $K = 4$, 6, and 8, agree bitwise at $K = 4$ and differ by at most 1.6×10^{-7} in relative terms (Section B.1).

Evaluating orders through K therefore needs a number of scalar updates proportional to GK per basis, in addition to the linear factor projection. One basis represents $N(G, K) = \sum_{j=2}^{K} \binom{G}{j}$ products of distinct slots, and more monomials in the input coordinates (Eq. (13)); $N(G, K)$ can grow quickly with K, but it is not a parameter count, since all of these products are tied by the shared factor weights and the order coefficients.

A.11 Parameter count of the shared spectrum

With the full spectrum carried and no core bias, the operator has the $P_{\mathrm{shared}}(R, K)$ parameters of Eq. (4): Rd in the factor bank, oR in the readout, and $R(K-1)$ order coefficients when $K \geq 3$. At the tabular site, where $d = o = 64$ and $R = 8$, one basis therefore costs 128 parameters at $K = 2$,

131 at $K = 4$, and 135 at $K = 8$, and the $K = 4$ operator has $8 \cdot 64 + 8 \cdot 64 + 8 \cdot 3 = 1{,}048$ parameters. For a general carried order set $\mathcal{J} \subseteq \{2, \dots, K\}$ the count is

$$Rd + oR + \delta_\gamma R(K-1) + \delta_b\, o,$$

where $\delta_\gamma = 1$ when γ is trained and 0 otherwise, and $\delta_b = 1$ when the core bias is used. The coefficients γ are trained exactly when $|\mathcal{J}| \geq 2$ and the mixture is not fixed by design. A single carried order keeps the fixed coefficient 1 even when $K > 2$, and the fixed mixture control of Section A.14 has equal coefficients that are not trained. With learned coefficients the matrix γ keeps all $R(K-1)$ entries, including those of masked orders and those of orders outside $\mathcal{J}$, which stay at zero, so neither a mask nor a smaller carried set changes the parameter count.

The readout rescaling $(b_r, \gamma_r) \mapsto (a\, b_r, a^{-1}\gamma_r)$, with b_r and γ_r as in Section A.12, leaves the output unchanged, which is why a learned coefficient at $K = 2$ would only duplicate the readout. The same rescaling leaves one coefficient per basis redundant at $K \geq 3$ as well, and the factor rescaling of Remark 4 adds one further redundant direction per basis at every K, absorbed by γ_r at $K \geq 3$ and by b_r at $K = 2$; these directions are kept, trained, and counted, so Eq. (4) does not understate the size of a PSNeuron. The recurrence arithmetic grows with K while the factor bank does not, which is why parameter cost and compute cost are reported separately (Section H).

A.12 Coupling of outputs and orders

Before an optional output bias, output $u \in \{1, \dots, o\}$ is

$$P_u(x) = \sum_{r=1}^{R} \sum_{j=2}^{K} B_{\mathrm{core},ur}\, \gamma_{r,j}\, m_j \tilde{e}_{r,j}(x),$$

where the frozen calibration, center, gauge factor, and mask all enter through the masked delivered channel $m_j \tilde{e}_{r,j}$. For one basis r, let $b_r \in \mathbb{R}^o$ be column r of B_{core} and $\gamma_r = (\gamma_{r,2}, \dots, \gamma_{r,K})^\top$, and define the matrix of output by order coefficients

$$C^{(r)}_{u,j} = B_{\mathrm{core},ur}\gamma_{r,j}, \qquad C^{(r)} = b_r \gamma_r^\top \in \mathbb{R}^{o \times (K-1)}. \tag{17}$$

Remark 3 (Rank restriction on output by order coefficients). Each $C^{(r)}$ has rank at most one. A PSNeuron with several outputs therefore has no independent coefficient for every combination of output, basis, and order; within a basis the output and order coefficients factor through b_r and γ_r. Masks act on the channels $\tilde{e}_{r,j}$ and leave this restriction in place.

For a scalar output each $C^{(r)}$ is the single row of combined coefficients $\beta_{r,j} = B_{\mathrm{core},1r}\gamma_{r,j}$, and the restriction is vacuous: with learned order coefficients every $\beta \in \mathbb{R}^{R \times (K-1)}$ is attained, for instance by $B_{\mathrm{core},1r} = 1$ and $\gamma_{r,j} = \beta_{r,j}$. With several outputs the restriction matters, because an arbitrary output by order coefficient matrix can have rank above one.

A.13 Scale redundancy of the shared geometry

Remark 4 (Scale gauge within a basis). For the uncentered spectrum with learned order coefficients, any $a \neq 0$ permits $W_r \mapsto aW_r$ and $\gamma_{r,j} \mapsto a^{-j}\gamma_{r,j}$ without changing $s_r(x)$, since scaling W_r multiplies order j by a^j and the coefficient contributes a^{-j}. If the only carried coefficient is fixed, the same rescaling of W_r is compensated by $b_r \mapsto a^{-j} b_r$, which preserves the site output but scales the basis value. With a frozen nonzero center the same map shifts s_r by the constant $\sum_j \gamma_{r,j} m_j \zeta_j \mu_j (1 - a^{-j})$, so it preserves the site output only up to a constant, which an output bias can absorb.

Because factor and coefficient amplitudes trade against one another without changing the function, coefficient magnitude does not measure how much the function uses an order; the paper measures that use with masks inside the same model. The gauge factor ζ_j of Section A.3 is the same identity applied to all rows at once, with the compensation held in a frozen constant instead of in γ.

Gauge fixed reparameterization. Remark 4 yields the exact reparameterization compared in Section I.2. Write each row as $W_r = \rho_r \bar{W}_r$ with $\rho_r = \|W_r\|_2 > 0$ and $\|\bar{W}_r\|_2 = 1$. For fixed

calibration constants $\hat{e}_{r,j}(x; W_r) = \rho_r^j \hat{e}_{r,j}(x; \bar{W}_r)$, so with $\tilde{\gamma}_{r,j} = \rho_r^j \gamma_{r,j}$, $\mu_j = 0$, and $\zeta_j = 1$, as in every reported configuration, the basis value is exactly $s_r(x) = \sum_j \tilde{\gamma}_{r,j} m_j \hat{e}_{r,j}(x; \bar{W}_r)$. The gauge fixed model trains $(\bar{W}, \tilde{\gamma}, B_{\text{core}})$ in place of $(W, \gamma, B_{\text{core}})$, with parameters of the same shapes, and renormalizes every row to unit norm, to floating point precision, after each optimizer step; the forward computation is unchanged. Its calibration normalizes the rows and then fits its own constants c_j^{gf} on $\bar{W}$, skipping the bank rescale, which would move the rows off the unit sphere. Each pair of runs starts from one draw with $\tilde{\gamma}_{r,j} = \rho_r^j \gamma_{r,j}\, c_j / c_j^{\text{gf}}$, where ρ_r is the norm of the corresponding row of the unconstrained model, so both compute the same function at the first step. The function class and the parameter count are unchanged (2,560 parameters at $R = 64$, $d = 32$, $K = 8$, and $o = 1$), and no frozen constant is added. Only the optimization coordinates change, and with them the effective Adam step in each direction.

A.14 Controls derived from the spectrum

Several experimental controls keep the slot construction and change one modeling assumption. Each has $Rd + oR$ trainable parameters, plus o with a core bias, which is the count of Eq. (4) without the $R(K-1)$ order coefficients.

Single order. A single order control carries one order j with the fixed coefficient $\gamma_{r,j} = 1$. At $G = K = j$ its channel is c_j times the product $\prod_{g=1}^{G} q_{r,g}$ of all slots, and the readout absorbs the constant (checked at $G = K \in \{2, 3, 4\}$ in Section B.1). Single order heads of orders 3 and 4 are matched controls in Table 53.

Fixed mixture. A fixed mixture carries several orders with equal coefficients that are not trained, orders 2, 3, and 4 in Table 53. It differs from a spectrum carrying the same orders only in that the mixture is not learned.

Independent order geometry. An independent control removes the sharing constraint. Its R bases are split among the $|\mathcal{J}|$ carried orders, $R = \sum_j R_j$ with $R_j = \lfloor R/|\mathcal{J}| \rfloor$ and the remainder assigned to the lowest order. The R_j bases of order j have their own factor bank $W^{(j)} \in \mathbb{R}^{R_j \times d}$, their own partitions, their own bank rescale, and their own frozen calibration constants, and each emits only its own order channel with the fixed coefficient 1. The basis values share one readout, so the control has no order coefficients and the same count as a single order control of the same total R. Table 27 compares the two geometries at their recovery widths, with R counted over all orders.

Additive control. The *additive control* keeps the grouped factor projections, the partitions, the bank rescale, and the readout, except in the recurrent mixer, where it reads the pair $[r_t; u_t]$ without the conditioners and on a partition of its own (Section E) and where an exact additive twin keeps them (Section K.2), and replaces the product aggregation by a sum of pointwise nonlinearities, $s_r(x) = \sum_g \text{GELU}(q_{r,g}(x))$. It has no order channels, no order coefficients, and no calibration constants c_j. At equal R the grouped factorization is kept and only the aggregation changes, so the comparison isolates what the product adds over a sum of the same slots; where the additive control is wider or reads slots of its own, as in the Boolean cube probes, the narrow mixed order probe, and the recurrent mixer, the width or the slots change as well. A linear sum of the slots is not used as a control, because summing linear factors gives again a linear map.

B PSNeuron Configuration and Implementation

Tables 5 and 6 list every setting of the operator and of its placement in a host, its effect and parameter cost, and its values in this paper; Section E specifies each placement.

Partition semantics. When one draw of the balanced partition serves every basis, two coordinates entering the operator that the draw places in one block never enter a common product, at any order and in any basis; when each basis draws its own partition, they can enter a product as soon as one basis separates them. A conditioner that mixes coordinates, such as the low rank part of the direct replacement head, moves this statement to its output. The Boolean cube probes show the first case:

Table 5: PSNeuron operator settings and their values in this paper, for the operator of Eq. (3); Table 6 continues with its placement in a host and its training.

setting	effect and parameter cost	value in this paper
maximum order K	exposes the exact orders 2 to K, with $K \leq G$; raising K at fixed G, bank, partition, and calibration leaves lower channels unchanged (Remark 1); carries $R(K-1)$ order coefficients when $K \geq 3$ and the coefficients are learned and none at $K = 2$, so the step from $K = 2$ to 3 adds $2R$ and each later order R; a higher K also adds recurrence work	4 at the replaced sites of Table 1(a) other than the attention output and in the latent, direct, and augmented heads of Table 35; 2 at the main cross head sites and in the DINOv3 heads of the ladder and its five seed study, wider heads included; 2 to 8 in the capacity sweeps, in every other placement (Table 9), and in the controlled studies
slot count G	fixes the factor grouping and the highest attainable order; adds no parameters at fixed R, d, o, and K, but the masked reference projection executes RGd multiply accumulates and the recurrence work grows with G and K (Sections A.10 and H.1)	8 in every task placement, the Boolean probes, and the fixed G study ($G_s = G_x = 4$ in the recurrent mixer); $G = K$ in the aligned recovery diagnostic and the sharing study
basis count R	adds bases, and therefore width, without changing the nominal order; each basis costs $d + o$ parameters, plus $K - 1$ order coefficients when $K \geq 3$ and the coefficients are learned (Eq. (4), Section A.11)	per placement in Table 9, set by the budget in the capacity sweeps and the direct replacement heads; 8 in the Boolean probes; 64 in the fixed G and aligned students; searched in the sharing study, whose confirmed recovery widths range from 2 to 30 (Table 27)
partition $\mathcal{G}_{r,g}$	assigns input coordinates to slots; has no parameters and is never learned, but is a strong inductive bias	balanced in every placement; one seeded draw shared by all bases at the feedforward and post attention sites, in the direct replacement heads, and in the Boolean probes; one contiguous block per head at the cross head sites and contiguous state and input blocks in the recurrent mixer, both shared by all bases and fixed without a seed, except the seeded free slot variant of Table 31; one draw per basis in the DINOv3 heads of the ladder and its five seed study and in the students of the fixed G, aligned, and sharing studies; redrawn with each seed in the narrow probe of Table 24; one coordinate per slot wherever $d = G = 8$
order coefficients $\gamma_{r,j}$	weight the carried orders inside each basis	learned when several orders are carried, except in the fixed mixture control, whose equal coefficients are not trained; fixed at 1 for a single order
initialization	starting values of W_0, B_{core}, and γ; no parameters	$W_0 \sim \mathcal{N}(0, 1)$, $B_{\mathrm{core}} \sim \mathcal{N}(0, 1/R)$, and $\gamma_{r,j} = \lvert\mathcal{J}\rvert^{-1/2}$ (Section 3), except the DINOv3 heads of the ladder and its five seed study, which start at the scale of the GELU control (Section G.1), and the students of the fixed G study, which start every carried coefficient at $1/\sqrt{7}$, with other starts as secondary conditions (Sections I.1 and I.2)
calibration c_j	frozen scale for each order, never learned; conditions the channels without changing their degree (Sections A.3, 3, and G.1)	fitted once before the first step on the first 128 training examples; the fixed G students take that of the order 8 reference, fitted on 2,048 training rows; 512 rows in the aligned diagnostic (Section I.2); the first 2,048 training rows in the factor sharing study (Section I.5); in the DINOv3 ladder the first 4,096 fitting rows on ImageNet100 and all 800 fitting rows of each VTAB-1k dataset (Section G)
centering μ_j, gauge factor ζ_j	frozen for each order (Section A.3); μ_j subtracts a mean and adds a term of degree zero; ζ_j compensates an exact common rescaling of the bank, changing no degree	$\mu_j = 0$ and $\zeta_j = 1$ in every reported run
order mask m_j	removes order j at inference without changing any parameter or frozen constant (Section B.1)	$m_j = 1$ except in mask interventions
core readout B_{core}, bias b_{core}	B_{core} maps the R basis values to the output with oR parameters and oR multiply accumulates, changing no degree; b_{core} adds a term of degree zero outside the homogeneous spectrum	B_{core} always present; b_{core} zero, except the class bias of the DINOv3 heads of the ladder and its five seed study

on seeds whose shared partition leaves two target coordinates in one block, the target lies outside the represented family and training stays at chance (Section J.3). Structural partitions, such as one block per attention head, are architectural choices, not preprocessing.

Additive control. At equal R, input, and readout the additive control has exactly the parameters of a $K = 2$ PSNeuron (Section A.14). The DINOv3 heads of the ladder and its five seed study and the CIFAR100 cross head site compare the two at equal counts (Figure 2(b), Table 50). The Boolean cube probes give the additive control 16 bases, twice as many as the PSNeuron, so it has twice the parameters of a $K = 2$ PSNeuron and more than every order matched PSNeuron (Table 41); the narrow mixed order probe and the synthetic teachers of the inventory likewise give it 16 bases, 528 parameters against 264 to 320 (Table 24, Section L). In the recurrent mixer the additive control does not share the slots of the PSNeuron mixer: it reads the pair $[r_t; u_t]$ without the conditioners A_s and

Table 6: PSNeuron placement and training settings and their values in this paper, continuing Table 5; the placement is that of Eqs. (6) and (7).

setting	effect and parameter cost	value in this paper
conditioner A_{host}	optional map before the operator; a linear map without bias keeps every channel homogeneous of its degree in the input of the placement, although one Walsh degree there needs slots of disjoint support (Proposition 1), and a nonlinear one moves the degree statement to the conditioned coordinates	per placement in Table 9
outer readout B_{host}	optional linear map after P, full or low rank; changes the order of no channel, and when nothing nonlinear lies between it and B_{core} it folds into B_{core}, so it adds parameters without enlarging the functions the site can compute, and restricts them when its rank is below $\min(o, R)$	per placement in Table 9; absent where none is listed
carrier C_{host}, λ	a path beside the branch (Eq. (7)); a carrier alone does not make a placement an augmentation	per placement in Table 9: none at the feedforward sites and in the recurrent mixer, an identity at the cross head sites; λ learned for the attention scores, in the direct replacement head, and in the residual tabular anatomy of Table 42, fixed at 1 elsewhere
geometry	shared geometry is the method, every order reading W_0 (Eq. (5)); the order specific geometries GFT, SH-GFT, and SGA give chosen orders a corrected copy of W_0 at extra parameter and arithmetic cost	shared; GFT, SH-GFT, and SGA are implemented but not selected (Section D); independent banks for each order (Table 27) and the additive control are controls (Section A.14)
order subset	exposes an explicit subset of the orders 2 to K; a single carried order drops the $R(K-1)$ coefficients, while a subset of two or more keeps all $R(K-1)$, those of the omitted orders included (Section A.11)	all orders 2 to K, except the single order controls (Table 53)
seed	fixes the factor initialization and the partition; no parameters	set per training seed, except that the DINOv3 heads of the ladder and its five seed study and the students of the fixed G, aligned, and sharing studies keep one fixed partition in every run, and the contiguous partitions of the cross head sites and the recurrent mixer use no seed
training procedure	optimizer of all trained parameters; an optional intrinsic optimizer for SGA changes training only and adds nothing at inference (Section D)	AdamW in every run outside Section D

A_x, through 12 bases of eight slots on a balanced partition of its 256 coordinates, and has $4{,}608$ parameters against $4{,}352$ to $4{,}464$ (Section J.2).

B.1 Numerical verification

Elementary symmetric recurrence. At $G = K \in \{2, 4, 6, 8\}$, direct enumeration of the subsets agrees with the recurrence in double precision at every exposed order, in values and in gradients, with relative discrepancies below 10^{-15}. At $G = K \in \{2, 3, 4\}$ the uncalibrated single order channel, the product of all slots, agrees with an independent implementation of that product to a relative error below 10^{-6} in single precision, and at $G = K = 2$ the operator reproduces an independent two slot product layer exactly once the factor bank and readout are copied; calibration multiplies the channel by the constant c_K, which the readout absorbs. The descending update used in training and the out of place update used for evaluation without gradients (Section A.10) give bitwise identical forward values at $G = 8$ and $K \in \{2, 3, 4, 6, 8\}$. On an A100 GPU their backward passes sum in different orders; the gradients, compared there at $K = 4$, 6, and 8, agree bitwise at $K = 4$ and differ by 8.8×10^{-8} to 1.6×10^{-7} in relative terms at $K = 6$ and 8.

Homogeneity and nesting. Scaling the input by $a \in \{1/2, 2, 4\}$ scales every tested raw channel of order j by exactly a^j while the intermediate values remain representable (Proposition 1). Raising the maximum order at fixed G, factor bank, partition, and calibration leaves every channel already exposed at the smaller K unchanged in the tested models (Remark 1).

Order masks. In the released implementation an order mask multiplies channel j after it is computed (Eq. (3)); while a mask is applied, every parameter, frozen constant, calibration included, and stored state is bytewise unchanged, and lifting the mask restores the original output bit for bit. The aligned order 8 runs applied the mask instead by zeroing the coefficients $\gamma_{\cdot,j}$ and restoring them afterwards, which gives the same predictions (Section G.1).

Fused slot projection. For a partition shared by all bases with blocks of equal size, the fused slot projection multiplies each slot row only by the coordinates of its own block, Rd multiply accumulates per input row instead of the RGd of the masked reference projection that computes every reported model (Section H.1). With integer weights and inputs and unit block scales, where no rounding occurs, the two agree bit for bit at six geometries with d from 16 to 768, so they compute the same function; with general floating point values they sum the same products in different orders and can differ in the last bits, which is why Section H.2 reports functional equivalence rather than bitwise identity.

C Comparison with the Closest Models

Each entry of Table 7 records what the cited paper states or tests, not what its construction might allow: an entry reads not established when the paper is silent on a property that its construction could still have, and a yes that follows from the stated construction without a test carries a note that says so. The Deep Interaction Machine column is graded from its abstract alone (note g), because its full text was not available to us.

Table 7: Properties of interaction order in the closest models, as stated or tested in each paper. Partial entries hold in a narrower or different sense, given in the notes below the table.

property	HOFM (Blondel et al., 2016)	multi output polynomial networks (Blondel et al., 2017)	Deep Interaction Machine (Yu et al., 2020)	kVNN (Yun et al., 2026)	PSNeuron
several orders as separate terms	yes	no, degree 2 only	partial[g]	yes	yes (Eq. (3))
factors shared across orders	yes, as a variant[a]	not applicable	not established	no, separate centers	yes (Eq. (5))
vector map inside a network	not established[h]	partial[b]	not established	yes, convolution layer	yes (Eq. (6))
lower orders unchanged as the maximum order grows	yes[c]	not applicable	not established	yes[d]	yes (Remark 1)
single orders masked in a trained model	not established	not applicable	not established	not established	yes (Sections 5.2 and 5.3)
representation, recovery, and use per order	partial[e]	not established	not established	partial[f]	yes (Section 5.3)

[a] HOFM models each degree with its own factor matrix and adds two kernels that share parameters across degrees, the inhomogeneous ANOVA kernel and the all subsets kernel. [b] The outputs share one basis of hidden units, but the vector is the output of the whole predictor rather than of a site inside a network. [c] The dynamic program returns the lower degree kernels together with the highest one; holding the factors and degree weights fixed leaves these terms unchanged when the maximum grows. The paper does not test this intervention. [d] The third order configuration adds a cubic branch beside the lower order branches, which have separate parameters and so do not change; the paper does not test this intervention, and its second and third order models are trained separately. [e] Test AUC is reported for models retrained at each maximum degree from 2 to 5. [f] A representation theorem is given for each order; recovery and use of single orders are not measured. [g] The abstract states that interactions of arbitrary order are computed through power sums with Newton's identities; how the orders are parameterized and combined, and the other properties, are not stated there. [h] HOFM is evaluated as a scalar predictor, which does not rule out a vector extension.

D Order Specific Geometries and an Optional Optimizer

Every PSNeuron outside this section uses the shared geometry $W^{(j)} = W_0$ of Eq. (5) and is trained with AdamW. With one or two families of directions, sharing lowers the width needed for recovery when the orders of a target reuse factor directions and raises it when they do not (Sections 5.4 and I.5), so the implementation also offers, through one setting of the operator, three order specific geometries that give chosen orders a corrected copy of W_0, aimed at targets whose orders need different directions and at high orders, where recovery with the shared geometry falls (Section 5.3), and an optional optimizer for one of them. Here S is the rank of a correction and Σ a frozen second moment matrix of the inputs, the identity unless stated otherwise. We write $\|w\|_\Sigma^2 = w^\top \Sigma w$ for a row viewed as a column vector and $\|Z\|_\Sigma^2 = \mathrm{tr}(Z\Sigma Z^\top)$ for a matrix of rows Z.

Definitions. GFT and SH-GFT store one low rank correction $\Delta W_j = U\,\mathrm{diag}(\tilde{a}_j)V^\top$, with $U \in \mathbb{R}^{R\times S}$, $V \in \mathbb{R}^{d\times S}$, and $A \in \mathbb{R}^{J\times S}$ for the J corrected orders, that is $S(R+d+J)$ parameters, where $\tilde{a}_j$ is the row of $\tilde{A}$ for order j. GFT corrects every order, $J = K-1$, with centered coefficients $\tilde{A} = A - \frac{1}{J}\mathbf{1}\mathbf{1}^\top A$ (where $\mathbf{1} \in \mathbb{R}^J$ is the all-ones vector), since a component common to all orders, taken alone, turns every order alike, which a change of W_0 together with a rescaling of the order

coefficients reproduces. SH-GFT corrects only the high orders $\mathcal{H} = \{4\}$, $\{5, 6\}$, and $\{6, 7, 8\}$ at $K = 4$, 6, and 8, with $\tilde{A} = A$, since its uncorrected orders pin W_0. The correction is not added to W_0: with $\bar{W}_r$ row r of W_0 projected to unit Σ norm and v row r of ΔW_j, all rows in the following equation viewed as column vectors, order j reads

$$W_r^{(j)} = \cos\theta\, \bar{W}_r + \frac{\sin\theta}{n}\,(v - \chi \bar{W}_r), \qquad n^2 = v^\top \Sigma v, \quad \chi = v^\top \Sigma \bar{W}_r, \quad \theta = \theta_{\max} \tanh n, \tag{18}$$

with $\sin\theta/n$ at its analytic limit at $n = 0$. This turning map changes the direction of a row but bounds its radial change, $\cos^2\theta_{\max} \le \|W_r^{(j)}\|_\Sigma^2 \le 1$ for $0 < \theta_{\max} < \pi/2$, the range the implementation accepts, so a radial factor ν of a row, which reaches order j as ν^j, is confined to $[\cos\theta_{\max}, 1]$ rather than removed, and ν^8 can still fall to $\cos^8\theta_{\max} \approx 3 \times 10^{-4}$ at the default given below, and it reads parameters only, never the input, so every channel keeps its exact degree (Proposition 1). The defaults are $S = 1$ and $\theta_{\max} = 1.2$. Initialization sets $A = 0$ and draws U and V with normal entries of standard deviation 0.1, so the model starts at exactly the shared function of the projected bank, with a nonzero gradient in A; the base rows are projected to unit Σ norm after the rescaling of the bank in Section 3, which the projection supersedes, and before c_j is fitted, and again after every optimizer step, at no cost at inference.

Cost of GFT and SH-GFT. At $K = 8$ and $S = 1$ the corrections add only 103 and 99 parameters, about 4%, for totals of 2,663 and 2,659, but each corrected order needs its own slot values and recurrence instead of reading the slots of W_0. With one partition for all bases, executed multiply accumulates per token, which include the masked reference projection of Section H.1, are 2.19 times those of the shared operator for GFT and 1.58 times for SH-GFT, and 3.89 times for GFT at $S = 8$; under the compact count of Table 16, which charges one projection multiply accumulate per learned weight, the ratios are 5.12, 2.97, and 10.76. On an A100 at batch 512 in fp32, the unfused implementation took 5.5 and 4.0 times as long per training step and 6.4 and 4.1 times per inference call, in a timing run that did not check for contention on the host.

SGA and an optional optimizer. SGA, spectral geometry allocation, groups the orders $2, \ldots, K$ into contiguous bands. A band $\mathcal{B}$ with $t_\mathcal{B}$ private directions stores $H_\mathcal{B} \in \mathbb{R}^{Rd \times t_\mathcal{B}}$ and $\alpha_\mathcal{B} \in \mathbb{R}^{|\mathcal{B}| \times t_\mathcal{B}}$, that is $t_\mathcal{B}(Rd + |\mathcal{B}|)$ parameters, and each of its orders reads $W^{(j)} = W_0 + \Delta W_j$ with $\Delta W_j = \sum_{s=1}^{t_\mathcal{B}} \alpha_{j,s} \Psi_{\mathcal{B},s}$. The $R \times d$ matrices $\Psi_{\mathcal{B},s}$, with $\mathrm{tr}(\Psi_{\mathcal{B},s} \Sigma \Psi_{\mathcal{B},s'}^\top) = \delta_{ss'}$, come from the columns of $H_\mathcal{B}$ by a QR factorization in whitened coordinates; an optional truncation to spatial rank $p_\mathcal{B}$ bounds their rank but need not keep them orthonormal. The orders of a band with $t_\mathcal{B} = 0$ read W_0. SGA has no turning map, and its random initialization does not start at the shared function. The allocation, the bands with their ranks, is fixed before training: for $K \in \{2, 3, 4, 5, 7\}$ it is the cheapest contiguous band plan, charging $\min\{p_\mathcal{B}(R + d), Rd\}$ parameters for a direction of spatial rank $p_\mathcal{B}$, whose closed form error stays below 1% of the target variance on coherent, partially coherent, and incoherent synthetic corrections at once, computed at $R = 8$ and $d = 8K$, and for $K \in \{6, 8\}$ the cheapest such plan that reproduces a trained GFT model with $S = 6$ to within 1% of its output variance on that model's held out inputs. Every resulting band holds one order with $t = 1$, except order 5 at $K = 6$, which has none. The spatial rank is 4 for every order at $K = 4$, 2 for order 2 at $K = 8$, and full for the other corrected orders at $K = 6$ and 8, each of which can therefore read an arbitrary factor bank; the implementation stores every direction densely, as counted below. With a scalar output at $R = 64$ and $d = 32$, SGA therefore has 8,451, 10,628, and 16,903 parameters at $K = 4$, 6, and 8, which is 3.67, 4.37, and 6.60 times the 2,304, 2,432, and 2,560 of the shared operator.

The optional intrinsic optimizer trains SGA with such single order bands and a scalar output. It writes $\Delta W_j = \rho_j Z_j$ with $\|Z_j\|_\Sigma = 1$, which removes the scale redundancy of $(H_\mathcal{B}, \alpha_\mathcal{B})$; starts every amplitude at zero, with random orientations or, under spectral initialization, the leading singular directions of the whitened loss gradient with respect to the weights of order j; eliminates the linear readout by variable projection (Golub & Pereyra, 1973); and takes trust region Gauss Newton steps in the metric of the realized order weights, solved by truncated conjugate gradients (Steihaug, 1983). Three optional extensions, all off by default, change this procedure: a functional trust rule shortens a step when the output of an order moves too far, a second rule changes how a step leaves a zero amplitude, and a rescaling of the whole factor bank by powers of two is compensated exactly, in real arithmetic, through the gauge factors ζ_j of Section A.3. This rescaling changes ζ_j during training and is used only by one configuration of the selection round below, the only runs in the paper whose

ζ_j departs from 1. After every accepted step the state is written back as ordinary SGA parameters, which reproduce the realized order weights of the optimizer up to rounding, checked against a fixed multiple of machine precision; the trained model therefore adds no parameter, buffer, or operation at inference.

Table 8: Exploratory screen of GFT and SH-GFT: parameters added at $K = 8$ to the 2,560 of the shared operator, learning rate, and runs reaching held out $R^2 \geq 0.9$. Learning rates were selected on the held out test R^2 of two tuning seeds, and the corrected settings ran at two weight decays and so have twice as many runs; the columns do not compare geometries at matched protocol.

	shared	GFT				SH-GFT			
		$S = 1$	$S = 2$	$S = 4$	$S = 8$	$S = 1$	$S = 2$	$S = 4$	$S = 8$
added parameters	0	103	206	412	824	99	198	396	792
learning rate	0.0256	0.0064	0.0128	0.0256	0.0256	0.0128	0.0256	0.0064	0.0128
recovered, $K = 4$	9/9	18/18	18/18	18/18	18/18	18/18	18/18	18/18	18/18
$K = 6$	31/36	66/72	64/72	70/72	69/72	69/72	71/72	60/72	63/72
$K = 8$	5/36	9/72	12/72	43/72	46/72	11/72	39/72	4/72	14/72

Exploratory screen. Before the shared geometry was fixed, a screen trained GFT and SH-GFT beside the shared geometry with AdamW and a cosine schedule (Table 8), on coherent, partially coherent, and incoherent teachers built like those of Section I.5, with $M = 1$, $d = 32$, $R = 64$, $G = K$, and $K = 4$, 6, and 8; at $K > 4$ the partially coherent teacher puts orders 2 and 3 on one set of directions and the higher orders on another. The gains of GFT at $S \geq 4$ and SH-GFT at $S = 2$ at $K = 8$ come from the coherent and partially coherent teachers: on the incoherent teacher, whose orders need different directions, no setting recovers more than 2 of 24 runs at $K = 8$, and the shared geometry 0 of 12. The same screen also trained dense GELU and SwiGLU networks with 2,551 and 2,547 parameters at $K = 8$, which the table omits: they recovered 24 and 24 of 36 runs at $K = 6$, against 31 for the shared geometry, and 23 and 13 of 36 at $K = 8$, against 5, none of them on the incoherent teacher. Slot gains and independent banks for each order, also omitted, recovered at most 11 of 72 runs at $K = 8$. Besides the protocol differences in the caption, parameters and arithmetic differ between settings, the corrected settings also project the base rows to unit norm after every step, which the shared column does not, although a shared geometry with this projection recovered a similar 29 of 36 runs at $K = 6$ and 6 of 36 at $K = 8$, and the corrections used a separate partition for each basis instead of the shared partition that the implementation requires, so the screen is not evidence that the corrections help.

Prespecified tests. A development round after the screen, at $K = 8$, compared GFT with $S = 6$ and SH-GFT with $S = 3$, each with and without three changes to training that preserve the function class, and carried SH-GFT, with a sign initialization and a normalization of its correction factors after every step, to a confirmation fixed in advance: 72 runs over 12 seeds, two weight decays, and the three teacher families, required to recover at least 90% of the runs of every family and to use its high orders; its learning rate was retuned to the lower edge of its grid, and its seeds included the four on which the development round had selected it. It recovered 8 of 24 coherent, 12 of 24 partially coherent, and 0 of 24 incoherent runs, and failed; it included no shared geometry model, so it measures the variant alone. Like the screen, the development round, the confirmation, the two trained GFT references of the SGA allocation, and their retraining below used a separate partition for each basis, a research version of the corrections, whereas the implementation described and costed above requires one partition for all bases.

Installed without training, the SGA allocations decoded from the two trained GFT references, both trained on the incoherent teacher, with $R = 128$ and $G = 6$ at $K = 6$ and $R = 64$ and $G = 8$ at $K = 8$, reproduce the reference outputs with $R^2 = 1.000000$ at $K = 6$ and 0.995829 at $K = 8$. The decoding was selected on the held out inputs on which this R^2 is computed, so these are reconstructions of known solutions, not generalization. Trained from scratch with four seeds on each teacher family, $R = 64$, $G = K$, $\Sigma = I$, Adam without weight decay at an untuned constant learning rate of 3×10^{-3}, batch 512, and 3,000 steps extended to 6,000 or 10,000 while validation R^2 still rose, SGA recovers all 12 runs at $K = 4$, with a lowest family median held out R^2 of 0.99999 and every order used, that is, masking it lowers held out R^2 by at least 10^{-3}. At $K = 6$ it recovers 7 of 12: 3 of 4 on the coherent and partially coherent families, with medians 0.99998 and 0.999995, and 1 of 4 on the incoherent family, with median 0.4077; orders 2 to 4 are used in all 12 runs, order 5 in

10, and order 6 in 7. A gate fixed before these runs required, among other criteria, a lowest family median of at least 0.80 before the confirmation at $K = 6$ and any training at $K = 8$. It failed on that criterion alone, so neither stage was run, and the allocation, learning rate, and schedule were left unchanged. GFT with $S = 6$ and $R = 64$, retrained on the same teachers with its own optimizer and learning rates, recovered 7 of 9 runs at $K = 6$, again 1 of 3 on the incoherent family, and 9 of 9 at $K = 8$. No shared geometry model was trained in these runs. Earlier SGA prototypes, trained with a sampled loss or through a placeholder for the band bases, do not use the implemented geometry and are not reported.

The intrinsic optimizer passed its numerical checks, which compare its Jacobian products with their adjoints, its variable projection with finite differences, and its matrix free solve with an explicit one, and, with spectral initialization, its prespecified sanity stage at $K = 4$, recovering all 9 runs, three seeds on each teacher family. At $K = 6$, on the teachers and splits of the SGA runs, it recovered 3 of 12 runs from random orientations, or 4 when a run that stopped early at held out $R^2 = 0.903$ is counted, and 0 of 12 after spectral initialization, whereas Adam recovered 6 and 8 of 12 from the same two starts, against 7 from the default initialization of SGA above. From each optimizer start, 5 of the 12 runs stopped early through a defect in the evaluation of trial steps that was later repaired; the repaired optimizer reproduced the 7 other failures after spectral initialization, and their end points are not stationary, so these are failures of optimization, not of capacity. With the functional trust rule it recovered none of the 4 runs, the early stopped one included, that it had recovered from random orientations without the rule. Spectral initialization therefore helped only under Adam, and there the incoherent family still recovered 1 of 4 runs. A later round with a protocol fixed in advance compared seven configurations of this optimizer, all started from random orientations and using the functional trust rule, which differ in the policy of that rule, in the treatment of zero amplitudes, and in the rescaling of the bank, over 126 runs at $K = 4$, 6, and 8, and selected none: no configuration met the prespecified eligibility criteria at $K = 6$ or $K = 8$, and at $K = 8$ most runs fit the training data with $R^2 \geq 0.999$ while none reached a validation R^2 of 0.90.

Why the shared geometry is kept. No matched comparison with the shared geometry was run, so no alternative lost one. The shared geometry remains the reported method because no alternative passed the test that would have promoted it, the alternatives cost the parameters or arithmetic that the paper sets out to save, and the headline sites use $K = 2$ or 4, where the shared geometry recovers the controlled targets reliably; no variant was tried at the task sites with higher K. The intrinsic optimizer was not adopted because it failed at $K = 6$, where SGA trained with Adam had also failed its gate. Whether order specific geometry is worth its cost remains open; three steps would address it: a matched comparison at equal parameters and arithmetic with learning rates selected on validation data, a training procedure that reaches the SGA allocations reliably from scratch at $K \geq 6$, and a kernel that fuses the recurrences of the corrected orders.

E PSNeuron Architectures for Each Task

Table 9 specifies every placement and Table 10 every trainable host; Figure 3 draws the principal placements, Figure 5 the adapter, the attention score correction, and the heads with $K = 4$ on frozen features, and Figure 6 the tabular anatomies. Each placement is compared with its own matched controls, and no placement serves as the control of another.

The partitions of the feedforward, cross head, recurrent mixer, and DINOv3 ladder placements are listed in Table 5; in the adapters, the latent and augmented heads, and the attention score correction $d = G = 8$, so every balanced partition places one coordinate in each slot; the direct replacement heads, with $d = 768$, share one seeded balanced draw across bases, as the feedforward sites do. The controlled teachers and the Boolean probes use the bare operator $P(x)$ without a host (Sections 4 and J.3).

Cross head site. Each head output $o_h \in \mathbb{R}^{16}$ is mapped by its own bias free $P_h \in \mathbb{R}^{4\times16}$, drawn from $\mathcal{N}(0, 1/16)$, and the eight results are concatenated head by head into $z \in \mathbb{R}^{32}$, so the maps P_h form a block diagonal conditioner A_{host}. The contiguous partition places the four coordinates of head h in slot h, so every term of order 2 multiplies two distinct heads and no product stays within one head. The branch scale is fixed at $\lambda = 1$, and no learned matrix follows the core readout of P. The 5,632 site parameters are 512 in the head maps ($8 \times 4 \times 16$), 1,024 in the factor bank (32×32), and

Table 9: PSNeuron placements in the reported experiments. A *replacement* (Eq. (6)) removes trainable parameters that its control holds at the compared site, whatever its carrier, and supports size comparisons; an *augmentation* (Eq. (7)) keeps all of them and adds the branch, so its parameter ratio to the control is at least one by construction and it supports attribution at equal or larger capacity. The post attention site keeps every host parameter, as an augmentation does, but inserts the block in series with no carrier beside it. d and o are the input and output widths of P, G the slots, R the bases, and K the maximum order; numbers in parentheses are trainable parameters of the site, which include its conditioners, down and up maps, and outer readouts, none of them attributed to the operator; the kept carrier of an augmentation is not included.

setting	placement and configuration	matched control
tabular feedforward: HIGGS Small, Jannis, Helena	replacement of the feedforward network of the second of two residual blocks of width 64, $h \leftarrow h + B_{\text{host}} P(\text{LN}\, h)$; identity A_{host}, $d = o = 64$, $G = 8$, $R = 8$, $K = 4$, B_{host} of rank 8 (2,072; 1,048 folded); other orders and widths in Figure 8, other anatomies in Table 42, and both blocks replaced in Table 48	the host with its GELU network restored (16,576); resized GELU and SwiGLU networks; the reference selected on validation, a bias free SwiGLU (24,576) on HIGGS Small and Jannis and the GELU host on Helena
ViT feedforward: CIFAR100	replacement of the GELU network ($192 \to 768 \to 192$, 295,872) of the last of six blocks by the same form with $d = o = 192$, $G = 8$, $K \in \{2, 4, 6, 8\}$; R and the rank of B_{host} set the budget (4,992 to 147,944; 5,019 at $K = 4$, the smallest configuration at that order, with $R = 9$ and rank 4 in Table 1(a), and 6,168 at $K = 4$ with $R = 8$ and rank 8 in the order and placement sweeps); a variant replaces all six networks	GELU networks resized at the site; the native model
post attention site: CIFAR100	insertion in series: the block of the ViT feedforward site, with $R = 8$ and an outer readout of rank 8 (6,144 to 6,200 parameters for $K = 2$ to 8), maps the attention output a of the last block to $B_{\text{host}} P(a)$ before the residual addition; the host keeps every parameter, so the ratio is at least one, but no carrier is kept beside the branch, and that attention output reaches the residual stream only through $B_{\text{host}} P$, whose outputs lie in a subspace of dimension 8	the same host with the site empty
language model feedforward: FineWeb-Edu	replacement of the feedforward network of the second block of a two block language model by the same form with $d = o = 128$, $G = 8$, $K \in \{2, 4, 6, 8\}$; R and the rank of B_{host} set the budget (3,328 to 98,304; the $K = 8$ site with 98,124 parameters has $R = 340$ and rank 34)	SwiGLU networks rebuilt at six budgets from 3,342 to 98,304 site parameters
language model all-block: FineWeb-Edu	separate six-layer width-256 causal TransformerHost with eight attention heads and tied token embeddings; all six feedforward networks are replaced by six independently parameterized copies of the same $B_{\text{host}} P(x)$ FFN-role realization, with $d = o = 256$, $G = 8$, readout rank 8, common K per run, and R set by the capacity rung. At $K = 4$, pe08 the six replacement sites sum to 392,286 parameters and the whole model has 6,315,102	six rebuilt SwiGLU feedforward networks; the rung-1 stack has 3,142,656 site parameters and 9,065,472 whole-model parameters, while the compatible pe08 SwiGLU has 391,680 site and 6,314,496 whole-model parameters
cross head: MQAR	replacement of the output projection (16,512) of the last of two blocks with eight heads of dim 16: $y = [o_1; \ldots; o_8] + P(z)$, $z_h = P_h o_h$, $P_h \in \mathbb{R}^{4\times 16}$, identity carrier; $d = 32$, $o = 128$, one slot per head ($G = 8$), $R = 32$, $K = 2$ (5,632); $K = 4, 6, 8$ in Table 32	the native output projection; at the site count, a linear map of the head latents and a low rank output projection; the carrier alone
cross head: CIFAR100	the same site in the last of six ViT blocks with eight heads of dim 24: $P_h \in \mathbb{R}^{4\times 24}$, $o = 192$, $K = 2$ (7,936 in place of 37,056); $K = 4$, 6, and 8, compact sites, and a free slot partition in Table 31	the native output projection; the additive control at 7,936
recurrent mixer: MQAR, multi-needle, state tracking	replacement of the SwiGLU channel mixer of the last of four DeltaNet blocks by $P_{\text{SI}}(r_t, u_t)$; $A_s, A_x \in \mathbb{R}^{8\times 128}$, $G_s = G_x = 4$, $R = 16$, the joint spectrum or its cross terms only, $K \in \{2, 4, 6, 8\}$ (4,352 to 4,464)	the native DeltaNet, whose mixer (130,944) reads u_t only; dense GELU and SwiGLU mixers and an additive control that read $[r_t; u_t]$ with 4,288 to 4,624 parameters; on fresh seeds an additive twin on the same slots (4,352)

4,096 in the core readout (128×32); at $K = 2$ no order coefficient is trained. The deleted projection has $128 \times 128 + 128 = 16{,}512$ parameters. On CIFAR100 the heads have dim 24, and the site has $768 + 1{,}024 + 6{,}144 = 7{,}936$ parameters in place of a projection of 37,056.

Recurrent mixer. In the last of the four DeltaNet blocks, let x_t be the residual stream and $h_t = \text{LN}_1(x_t)$. The kept DeltaNet path computes $q_t = W_q h_t$, $v_t = W_v h_t$, and

$$k_t = W_k h_t / \|W_k h_t\|_2, \quad \omega_t = \text{sigmoid}(w_\omega^\top h_t + b_\omega), \quad S_t = S_{t-1} + \omega_t \big(v_t - S_{t-1} k_t\big) k_t^\top,$$
$$r_t = S_t q_t, \quad x_t^\Delta = x_t + W_o\big(r_t \odot \text{SiLU}(W_g h_t)\big), \quad u_t = \text{LN}_2(x_t^\Delta),$$

Here t indexes tokens; $S_t \in \mathbb{R}^{128\times 128}$ is the recurrent state; $q_t, k_t, v_t, r_t \in \mathbb{R}^{128}$ are the query, normalized key, value, and state readout; $W_q, W_k, W_v, W_o, W_g \in \mathbb{R}^{128\times 128}$ and $w_\omega \in \mathbb{R}^{128}, b_\omega \in \mathbb{R}$ are trained. The scalar ω_t is the write strength of the delta rule, and $\odot$ denotes elementwise

Table 9: PSNeuron placements in the reported experiments (continued).

setting	placement and configuration	matched control
narrow frozen heads: DINOv3, five datasets	the head is P itself on the CLS feature $z \in \mathbb{R}^{768}$: $d = 768$, $o = C$, $G = 8$, $K = 2$, $R = h \in \{1, 2, 3, 4, 8\}$ in the ladder and $h \in \{2, 3, 4, 8, 16, 32, 64, 128\}$ in the five seed studies, one partition per basis drawn deterministically from a seed that the additive head shares, a class bias; a comparison at matched width rather than a size reduction	GELU head $768 \to h \to C$ (h more parameters); additive head with the same count; single order heads of orders 3 and 4 and a fixed mixture of orders 2 to 4 at every width of the ladder and at $h = 3$ in the five seed study
frozen heads at $K = 4$: ImageNet100, EuroSAT	augmentation $Wz + b + U\,P(Dz)$ with $D \in \mathbb{R}^{8\times768}$, $U \in \mathbb{R}^{C\times8}$, $d = o = 8$, $G = 8$, $R = 8$ (branch 7,096 or 6,376); two replacements drop $Wz + b$: the latent head $U\,P(Dz)$, whose count is that of the branch, and the direct head $\lambda B_{\text{host}} P(A_{\text{host}} z)$ with $d = 768$, $o = C$, and $R = 8$ (20,857 or 18,697)	the linear head $Wz + b$ (76,900 or 7,690)
wider frozen heads: ImageNet100, EuroSAT	the direct replacement head at twelve budgets (39,381 to 1,333,917), with $K \in \{2, 4, 6, 8\}$ selected on validation at each	GELU and SwiGLU heads built at each budget; the linear head
residual adapters: ImageNet100, EuroSAT, IMDb	augmentation $h + U\,P(Dh)$ after the trunk of a trainable MLP of width 128 on frozen features (Table 10); $D \in \mathbb{R}^{8\times128}$, $U \in \mathbb{R}^{128\times8}$, $d = o = 8$, $G = 8$, $R = 8$, $K \in \{2, 4, 6, 8\}$ (128 to 184 more than the GELU adapter)	the same adapter with GELU in place of P (2,048)
attention scores: MQAR, multi-needle	augmentation of the logits of the last of two blocks with four heads of dim 32 by $\lambda\, \tilde{q}_t^\top M \tilde{k}_s$, $\tilde{q}_t = P_q(A_q q_t)$, $\tilde{k}_s = P_k(A_k k_s)$, $A_q, A_k \in \mathbb{R}^{8\times32}$, $d = o = 8$, $G = 8$, $K \in \{2, 4, 6, 8\}$; the heads share the parameters of one correction ($R = 8$ and 833 to 945 parameters with a full M; the diagonal M selected on validation is rebuilt with 16, 13, 12, and 11 bases at $K = 2$, 4, 6, and 8, giving 1,015 to 1,033)	the host without the correction; low rank bilinear corrections of rank 16, 8, and 4 on both tasks

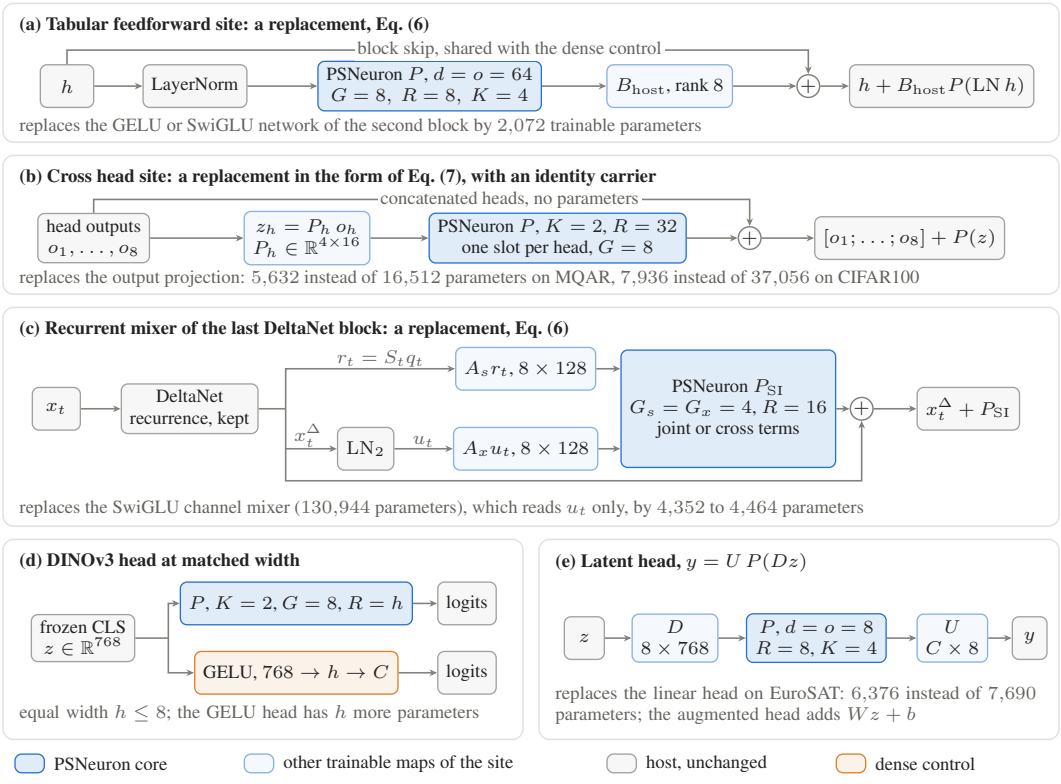

Figure 3: Principal placements; counts are trainable parameters of the site. The ViT, single-block FineWeb-Edu, and all-block FineWeb-Edu feedforward placements have the form of (a), with $d = o =$ 192, 128, and 256, respectively. In (d) the additive head, not drawn, has exactly the parameters of the PSNeuron head.

Table 10: Trainable parameters of each host with its native module at the compared site and with the PSNeuron site in place. For a replacement, r is the second total divided by the first (Section 4); both models are instantiated and counted, and frozen features are excluded. The output projection is a small part of each host, so the cross head sites make the whole model only 3.76% smaller on MQAR and 1.08% smaller on CIFAR100, and their reduction is best read at the site. Every block applies LayerNorm before each sublayer, and linear maps carry biases unless they are marked bias free. Splitting the width into attention heads changes no count, because the query, key, value, and output maps act on the full width. The ViT of the cross head site and the ViT of the other sites, like the attention hosts of the cross head and attention score sites, therefore have equal counts but are separate models trained separately. F is the raw feature width and C the number of classes.

host	components	native total	with the PSNeuron site
tabular MLP, width 64, two blocks	encoder $64F + 64$; per block a LayerNorm (128) and a GELU network $64 \to 128 \to 64$ (16,576), or in the second block a bias free SwiGLU of hidden width 128 (24,576); final LayerNorm (128); readout $65C$	35,522, 37,316, 41,828 with two GELU networks (HIGGS Small, Jannis, Helena, with $(F, C) = (28, 2), (54, 4)$, and $(27, 100)$)	21,018, 22,812, 27,324; $r = 0.59, 0.61, 0.65$
attention host, $d_{\text{model}} = 128$, two blocks, for the cross head and attention score sites	token embedding and learned positions ($2 \times 64 \times 128$); per block bias free query, key, and value maps (49,152), output projection (16,512), two LayerNorms (512), GELU network $128 \to 256 \to 128$ (65,920); final LayerNorm (256); readout $128 \to 64$ (8,256)	289,088	cross head site 278,208, $r = 0.9624$
ViT, $d_{\text{model}} = 192$, six blocks, patch 4, three heads of dim 64, or eight of dim 24 for the cross head site	patch embedding $48 \times 192 + 192$; learned positions 64×192; per block bias free query, key, and value maps (110,592), output projection (37,056), LayerNorms (768), GELU network $192 \to 768 \to 192$ (295,872); final LayerNorm (384); mean pooling over the 64 patches; readout $192 \to 100$ (19,300)	2,707,108	cross head site 2,677,988, $r = 0.9892$; $K = 4$ feedforward site of Table 1(a) 2,416,255, $r = 0.8926$, or 2,417,404, $r = 0.8930$, with the 6,168 parameter site; all six feedforward networks replaced 968,740 to 969,076, $r = 0.36$
DeltaNet, $d_{\text{model}} = 128$, four blocks	token embedding (8,192); per block two LayerNorms (512), five bias free 128×128 maps for query, key, value, gate, and output (81,920), write strength (129), bias free SwiGLU channel mixer of hidden width 341 (130,944); final LayerNorm (256); readout $128 \to 64$ (8,256)	870,724 on MQAR and multi-needle retrieval, 858,788 on state tracking, whose embedding (3×128) and readout ($128 \to 32$) follow its vocabulary	joint $K = 4$ mixer 744,180, or 732,244 on state tracking, $r = 0.85$ on both
language model, $d_{\text{model}} = 128$, two blocks	word vocabulary of 16,384 fitted on the training documents only, context of 512 tokens; word embedding ($16{,}384 \times 128$) and learned positions (512×128); per block two LayerNorms (512) and a gated linear recurrence whose gate, bias free value map, and bias free mix of state and input have 65,664 parameters; GELU network $128 \to 256 \to 128$ (65,920) in the first block; final LayerNorm (256); readout $128 \to 16{,}384$ (2,113,536)	4,474,752 without the second feedforward site; 4,478,094 to 4,573,056 with the rebuilt SwiGLU networks	4,478,080 to 4,573,056
language transformer, $d_{\text{model}} = 256$, six blocks	causal TransformerHost with eight attention heads, tied token embedding/readout weights, and six native SwiGLU feedforward networks of hidden width 682; the fixed non-FFN host contributes 5,922,816 trainable parameters and the native six-FFN stack 3,142,656	9,065,472 at the native/rung-1 SwiGLU stack; the compatible pe08 SwiGLU totals 6,314,496	all-block $K = 4$ pe08 PSNeuron stack 392,286, whole model 6,315,102; the six sites are independently parameterized
MLP on frozen DINOv3 or MPNet features, width 128, two blocks	encoder $768 \to 128$ (98,432); per block a LayerNorm (256) and a GELU network $128 \to 256 \to 128$ (65,920); final LayerNorm (256); readout $129C$; the standard residual adapter (2,048)	245,988 on ImageNet100, 234,378 on EuroSAT, 233,346 on IMDb	246,116 to 246,172 on ImageNet100, 234,506 to 234,562 on EuroSAT, and 233,474 to 233,530 on IMDb, an augmentation by the adapter branch

multiplication. The native block returns $x_t^{\Delta} + \mathrm{SwiGLU}(u_t)$, where the channel mixer is a bias free SwiGLU of hidden width $341 = \lfloor 2 \cdot 512/3 \rfloor$ ($3 \times 128 \times 341 = 130{,}944$ parameters). The PSNeuron mixer returns $x_t^{\Delta} + P_{\mathrm{SI}}(r_t, u_t)$ instead. It also reads the state readout r_t, though not the state matrix S_t, and has no carrier or recurrence of its own, so the comparison with the native model measures replacing the channel mixer by a PSNeuron with an additional state input. Its bias free conditioners, drawn from $\mathcal{N}(0, 1/128)$, form $z_s = A_s r_t$ and $z_x = A_x u_t$; the joint mixer evaluates the ordinary spectrum on $[z_s; z_x]$, and the cross term mixer keeps only the terms that touch both families,

$$e_j^{\mathrm{cross}} = e_j(z_s, z_x) - e_j(z_s, 0) - e_j(0, z_x) = \sum_{a=1}^{j-1} e_a(z_s) e_{j-a}(z_x).$$

Here $e_j(z_s, z_x)$ is $e_{r,j}$ over all eight slots of basis r, and $e_a(z_s)$ and $e_a(z_x)$ are the elementary symmetric polynomials of its four state and four input slots, zero for $a > 4$; the cross term mixer uses e_j^{cross} in place of $e_{r,j}$ in Eq. (2), with the same factor $\binom{8}{j}^{-1/2}$ and a constant c_j fitted on the cross terms themselves. At $K = 4$ the site has $4{,}400$ trainable parameters: $2{,}048$ in the two conditioners, 256 in the 16×16 factor bank, 48 order coefficients (16×3), and $2{,}048$ in the 128×16 core readout; the branch scale is fixed at 1, and neither it nor the masks add parameters. The same decomposition gives $4{,}352$, $4{,}432$, and $4{,}464$ at $K = 2$, 6, and 8, and the host outside the site has $739{,}780$ parameters on MQAR and multi-needle retrieval and $727{,}844$ on state tracking.

Dense controls in the same slot read the pair as $[r_t; u_t]$: a GELU network $256 \to 11 \to 128$ with biases ($4{,}363$ parameters), a GELU network of hidden width 12 behind a rank 11 input map ($4{,}624$), bias free SwiGLU networks of hidden width 8 behind a rank 12 or 13 input map ($4{,}288$ or $4{,}560$), and the additive control, which has no conditioners and sums a GELU of each slot over $R = 12$ bases drawn on the 256 coordinates of $[r_t; u_t]$, with one seeded balanced partition into eight slots of 32 coordinates that mixes state and input coordinates ($12 \cdot 256 + 128 \cdot 12 = 4{,}608$); Section J.2 reports them, and a separate same input study includes complete state tracking controls (Section K). The additive twin of the fresh seed holdout instead copies the conditioners, factor bank, partition, bank rescale, and readout of the PSNeuron mixer at the same seed and sums a GELU of each of its slots, so it lacks only the 48 order coefficients ($4{,}352$ parameters; Section K.2). Some runs evaluate the delta rule chunkwise, which reorders floating point operations and leaves the model unchanged: every MQAR run, the native model included, and the runs marked in Table 32.

Frozen feature heads and adapters. In the residual adapter $h + U\,P(Dh)$ and the augmented head $Wz + b + U\,P(Dz)$, the up map U starts at exactly zero, so each block equals its carrier at step zero. The gradient into D and into the PSNeuron parameters is then zero, because the branch is multiplied by U, while U itself receives a nonzero gradient; one optimizer step can open the branch, but zero initialization alone does not guarantee that training moves it. The latent replacement head $U\,P(Dz)$ has the same construction without $Wz + b$, so its logits are zero until the first update of U. The down map D has no bias and is drawn from $\mathcal{N}(0, 1/128)$ in the adapters and $\mathcal{N}(0, 1/768)$ in the heads, and P carries no readout bias. The adapter control replaces P by a GELU, so the two differ by exactly the parameters of the core: 128 at $K = 2$ and $128 + 8(K - 1)$ at $K \geq 3$. The direct replacement head $\lambda B_{\mathrm{host}} P(A_{\mathrm{host}} z)$ learns the scalar λ, which starts at 1. At $K = 4$ its conditioner is $A_{\mathrm{host}} = I + U_A V_A$ with factors of rank 8 and $U_A = 0$ at initialization, P has $d = 768$, $o = C$, $G = 8$, and $R = 8$, and $B_{\mathrm{host}} \in \mathbb{R}^{C \times C}$ has rank 8, so the head has $12{,}288 + 6{,}144 + 8C + 24 + 16C + 1$ parameters, $20{,}857$ on ImageNet100 and $18{,}697$ on EuroSAT; its number of bases and the two ranks set the budget of the wider heads in Table 36.

Attention scores. The query and key branches $\tilde{q}_t$ and $\tilde{k}_s$ run once per token and side, never once per pair of tokens, with separate banks and A_q, A_k drawn from $\mathcal{N}(0, 1/32)$. The logit becomes $q_t \cdot k_s / \sqrt{d_h} + \lambda\, \tilde{q}_t^{\top} M \tilde{k}_s$, with head dim $d_h = 32$, a full learned $M \in \mathbb{R}^{8 \times 8}$, and a learned scalar λ that starts at 1. The correction enters before the softmax and leaves the scaled dot product untouched, so the control is this placement with the second term deleted. It adds $2(256 + 128) + 64 + 1 = 833$ parameters at $K = 2$, and $16(K - 1)$ more at $K \geq 3$. Each order j term of $\tilde{q}_t$ and $\tilde{k}_s$ is homogeneous of degree j in its own token, so at $K = 2$ the correction is homogeneous of degree 2 in q_t and of degree 2 in k_s, while at $K \geq 3$ it mixes the bidegrees (i, j) with $2 \leq i, j \leq K$ and is not homogeneous. Selection on validation chose a diagonal M, rebuilt to the budget of the rank 16 bilinear correction by raising R to 16, 13, 12, and 11 at $K = 2$, 4, 6, and 8 (Table 34). The low rank

controls replace the correction by $\lambda\,(q_t^\top U')(V'^\top k_s)$ with $U', V' \in \mathbb{R}^{32\times r}$, which adds $2 \cdot 32\,r + 1$ parameters at ranks $r = 16$, 8, and 4.

F Map of Experiments and Claims

Section 4 separates three claims about order, namely representability, recovery, and functional use, from the size and attribution comparisons that a placement supports (Section E). Tables 11 and 12 list each experiment family in the order of Sections 5.1 to 5.4, followed by the resource and training speed measurements, and point to the section or table that reports it in full. Results that bound or contradict a claim appear beside those that support it.

Table 11: Map of experiments: the question, primary control, and main outcome of each experiment family, favorable or not. The location holds the full numbers, intervals, and caveats.

experiment	scientific question	primary control	main outcome	location
Replacement and size (Section 5.1)				
common $K = 4$ tabular replacement	size at host quality	GELU host; references selected on validation	Smaller than dense hosts and selected references at near-host quality; Jannis gains 0.55 points nominally over its host after learning rate bracketing, while same-budget GELU is unresolved on HIGGS Small and Jannis and higher on Helena; a smaller SwiGLU has a higher mean on HIGGS Small and Helena before bracketing, but only on Helena after bracketing, with an unresolved interval	Tables 28 and 58
every tabular configuration	size envelopes	resized dense networks	The PSNeuron envelope leads narrowly at some HIGGS Small budgets, but all its undominated configurations chose the learning-rate grid edge, so none is strict; envelopes cross on Jannis and dense leads on Helena	Section L.5
tabular replacements selected on validation	size and quality	host, paired by seed	37–44% fewer parameters; Jannis gains 0.44 points nominally but not after correction, HIGGS Small is unresolved, and Helena loses	Section L
preprocessing fitted on the training split	sensitivity of the tabular comparisons	the same comparisons with preprocessing on all rows	Same-budget GELU remains unresolved on HIGGS Small and Jannis and higher on Helena; the host contrast becomes positive on HIGGS Small and unresolved on Jannis, where SwiGLU leads; retuning confounds the change	Table 59
thresholds set after testing	parameters for a common accuracy	smallest dense network reaching it	Post-hoc HIGGS Small and EuroSAT size gains, with both HIGGS sides at grid edges and the EuroSAT linear head more accurate; dense is smaller on Jannis, Helena, and the CIFAR100 site	Table 1(b), Table 61
ViT feedforward site	size at native quality	native; resized GELU networks	The $K = 4$ site uses fewer parameters but loses accuracy to native; the resized GELU frontier has higher quality	Table 1, Figure 8(d)
cross head site	replacing the attention output projection	native projection; the host without it; linear and low rank maps; additive control	MQAR reaches near-ceiling recall at $K = 2$, 4, but a same-size linear correction and even the identity carrier do too, so no product advantage is isolated; one PSNeuron seed stalls at a shared rate. CIFAR100 loses 1.06 points to native, is unresolved against additive, and compact sites remain worse	Tables 22, 29 to 31, and 50
DeltaNet channel mixer	a mixer reading state and input	native DeltaNet; same input dense and additive mixers	Several mixers exceed native descriptively on multi-needle and state tracking, the latter partly through a stalled native seed; on MQAR only joint $K = 4$ leads. Same-input GELU and SwiGLU are unresolved on MQAR and multi-needle in the earlier five-seed cohort; some $K = 2$ rungs lead descriptively with bimodal outcomes	Table 32, Section J.2, Section K
wider DINOv3 heads	quality at equal budget	GELU, SwiGLU, and linear heads	below the GELU head at every budget, although the two round to the same accuracy at one EuroSAT budget; no PSNeuron head reaches the linear head	Table 36
frozen heads at $K = 4$	replacing or augmenting a linear head	linear head	EuroSAT latent head 17.1% smaller and 1.33 points below; ImageNet100 replacements 11.3 and 14.0 points below; a branch beside the linear head costs 0.20 and 0.40 points	Table 35
Pareto inventory	smaller and better pairs	every control of a task and protocol	Corrected strict wins occur only for synthetic teachers against additive, which GELU can also fit; DeltaNet wins are nominal only, and among matched-size pairs no gain but 113 losses survive correction	Tables 2 and 62, Section L

Table 12: Map of experiments, continued. Columns as in Table 11.

experiment	scientific question	primary control	main outcome	location
What the product contributes (Section 5.2)				
DINOv3 heads, $K = 2$	the grouped product at matched width	additive head of equal count; GELU, single order, mixture, and bilinear heads	At narrow $h = 2$–4, PSNeuron leads additive in $14/15$ five-seed means with 8 corrected positive intervals; at larger widths GELU and additive often lead. Bilinear controls do not resolve grouping-specific attribution, and many selected rates reach the grid edge	Figure 7 and Tables 52 to 55
DeltaNet mixer, joint $K = 4$	the product against controls on the same inputs	additive twin; conditional GELU with the same inputs	On 20 fresh MQAR seeds, gains of 44.08 and 37.38 points over the additive twin and conditional GELU survive Holm correction; outcomes are bimodal, SwiGLU is unresolved, and the joint-over-cross layout lead does not replicate; other recurrent tasks lack a gain	Section K.2, Section J.2, Section K
Boolean cube probes	representability, recovery, use	additive control; GELU networks	Order-matched units solve parity and admissible XOR while additive stays near chance; GELU fits admissible XOR and partly fits parity. Inadmissible six-coordinate partitions fail, and masking the target order damages all solved models	Table 41
singleton geometry, $K = 4$	grouped slots at equal count	shared degree control on single coordinates	No resolved accuracy difference on three tabular tasks at five seeds; singleton training steps are 3.8–4.0 times slower	Table 60
Represented order and recovered order (Section 5.3)				
teachers at fixed $G = 8$	recovery with exact witnesses	students of several orders	Orders 4 and 6 recover in all matched or nearby runs; recovery deteriorates at orders 7 and 8, even for some order-6 targets with $K = 8$; degree and initialization partly confounded	Table 23
aligned diagnostic, $G = K$	recovery of representable targets	GELU network	$44/48$ recoveries at order 6, $16/72$ at order 8; a separate GELU probe fits these targets, while slot count and scale change with order	Section I.2
gauge fixed coordinates	optimization at equal count	original coordinates	Order-8 recovery rises from 22.2% to 48.6% with equal initial coefficients, confounded by effective step size; starting only order 2 changes little	Section I.2
masks on teachers	functional use	loss of $3/7$ under exact recovery	Recovered models lose approximately the predicted $3/7$ under each mask; failed runs retain lower-order effects but lack upper-order use	Section I.3
planted interactions in DINOv3 features	order diagnosis on real continuous inputs	absent orders; GELU network for learnability	All 120 students recover on continuous features at widths 64 and 768; masks locate planted orders with negligible absent-order effects, largely implied by recovery and silent on natural labels	Section I.4
masks on real tasks	reliance on a channel	the same model unmasked	Direct frozen heads depend on order 4; pooled tabular mask losses are small, and removing the host branch upon retraining harms only Jannis; all are descriptive	Section J.6

Table 13: Map of experiments, continued. Columns as in Table 11.

experiment	scientific question	primary control	main outcome	location
Targets and placements (Section 5.4)				
sharing across orders	factor width for recovery	a bank for each order	Shared factors save 2.00–$2.25\times$ on coherent targets, give mixed results on partially coherent targets, and cost $3.43\times$ more on incoherent targets	Table 27
order specific geometries	recovery at high order	shared geometry	Unmatched, test-tuned screens favor corrections on coherent or partially coherent high-order targets, but not incoherent ones; SH-GFT confirmation and the SGA gate fail	Section D
attention scores	augmentation	uncorrected host; bilinear corrections	MQAR recall 12.6 to 50.5% against 99.94% native and 99.97 to 99.99% bilinear; on multi-needle retrieval the two corrections overlap	Table 34
six ViT feedforward sites	size	native model	64.2% fewer parameters; accuracy 31.49 to 33.36% against 40.62%	Table 49
tabular anatomy, exposed order	host pieces; choice of K	other anatomies; bare PSNeuron	The validation-selected anatomy is never test-best, though within 0.27 points; bare PSNeuron peaks at $K = 4$ on three tasks and at $K = 2$ on EuroSAT, all at grid-edge rates	Tables 42 and 43
both tabular feedforward sites	size	single site replacement	$4{,}466$ to $10{,}358$ parameters; below the single site replacement on HIGGS Small and Jannis	Table 48
residual adapters	augmentation	standard adapter	every PSNeuron adapter is larger, by 128 to 184 parameters, and less accurate on ImageNet100 and EuroSAT; on IMDb three have higher means, none with a resolved interval	Table 38, Section J.2
language model feedforward	replacement	SwiGLU at six budgets	held out NLL above SwiGLU at every budget, by 0.006 to 0.072 nats	Table 37
post attention site	insertion in series	host with the site empty	the empty site is ahead at every order	Section J.5
Resources and training speed (Sections 5.1 and H)				
parameters, MACs, latency	cost	dense module; matched control	Site parameters and MACs fall in tabular and MQAR placements, while narrow DINOv3 heads use more MACs; 48 matched GPU timings range from 0.83 to $20.2\times$, often slower with up to $16\times$ more launches; fused projection helps large inputs but is slower in nine of ten trained settings	Tables 3, 16, and 18
training speed	steps to a threshold	GELU and SwiGLU; primary dense controls	One controlled target reaches threshold sooner than GELU at similar time, while SwiGLU is faster; Jannis gains require larger rates, and frozen heads, cross-head sites, and controlled targets mostly take longer or are censored	Table 21, Section H.3

G EXPERIMENTAL PROTOCOLS AND DATA

Architectures are defined in Section E and PSNeuron configurations in Section B. Losses are cross entropy, on the queried positions for retrieval, and mean squared error for the controlled teachers; no host uses dropout or data augmentation.

G.1 COMMON TRAINING AND EVALUATION PROTOCOLS

The rules below hold for every study unless its paragraph in Section G.2 states otherwise.

Calibration once, before optimization. Every PSNeuron is calibrated once, before the first optimizer step, on a fixed slice of training data, the first 128 examples unless a study states otherwise; validation and test data are never used. Section A.3 gives the procedure. The constants c_j are frozen and never recomputed during training or inference, and every reported run leaves the channels uncentered ($\mu_j = 0$), which preserves exact degree.

The PSNeuron heads of the DINOv3 ladder and of its five seed study (Section J.10) depart from this procedure and from the initialization below in two respects: their bank and class readout are drawn from normal distributions with standard deviations $1/\sqrt{3 \cdot 768}$ and $1/\sqrt{3h}$, the initialization scale of the linear layers of the GELU control, and the measured rescale is kept as a fixed multiplier on the slots rather than multiplied into the bank. Both forms compute the same function at the first step, but folding the scale into the bank would change the magnitude of the parameters that AdamW updates; with the fixed multiplier, the PSNeuron head and its GELU control present the same parameter scale to the optimizer.

Initialization and partition. The bank W_0, the partition, and the readout B_{core} of Section 3 are drawn in that order from random streams derived from one seed, so one seed fixes all three; the partition of each placement is fixed at construction (Section E), and the heads of the DINOv3 ladder and its five seed study and the controlled students draw it from one fixed seed in every run. Calibration rescales the bank but never the readout, so the readout sets the magnitude of the branch at the first step. No activation, clipping, or normalization that depends on the input sits between the layer input and the raw product channels. The order coefficients start equal at $|\mathcal{J}|^{-1/2}$ (Section 3), which is $1/\sqrt{K-1}$ when the full spectrum is carried; the students of the fixed G study instead start every carried coefficient at $1/\sqrt{7}$ (Section 4). Under the initialization model with equal variance and orthogonal channels that we use to interpret the controlled studies, equal coefficients give each calibrated order the same nominal share; they do not guarantee equal shares of output variance on arbitrary inputs. A layer carrying a single order, including every $K = 2$ layer such as the cross head site and the DINOv3 ladder heads, fixes $\gamma_{r,j} = 1$, which Eq. (4) therefore does not count; the tabular site learns an $R \times (K-1)$ coefficient matrix over orders 2 to 4.

Optimizer and learning rates. Models train with AdamW (Loshchilov & Hutter, 2019), $\beta = (0.9, 0.95)$, and weight decay 0.01 on every tensor with two or more dimensions (weight matrices, factor banks, readouts, and the $R \times (K-1)$ order coefficient matrix), with no decay on biases and normalization parameters. The learning rate is constant without warmup, every model runs in fp32, and no study uses gradient clipping. Learning rates are selected by the mean validation quality of the two tuning seeds at the final step, over the base grid

$$\{2 \times 10^{-4},\ 4 \times 10^{-4},\ 8 \times 10^{-4},\ 1.6 \times 10^{-3},\ 3.2 \times 10^{-3}\}.$$

When a selected rate lies at a grid edge, one predefined extension round adds 1×10^{-4}, 6.4×10^{-3}, or both, to the grid of that model; the common tabular replacement and its GELU and SwiGLU references share one extension for each task. The matched control studies of Section K and the linear controls of Table 30 instead fix one grid for every compared model in advance and never extend it: the base grid for the linear controls, the base grid with both extension rates for the same input dense mixers, and eight rates from 2×10^{-4} to 2.56×10^{-2} for the tabular studies. In these studies and in the DINOv3 studies a tie on validation goes to the smaller rate.

Selection and final test discipline. Learning rates, and then the site layout, placement, and dense references, are selected on validation as stated in Section 4. Reported test values are not inputs to

these selection rules; test quality is reported at the final optimizer step of each final run, although the tuning runs of the DINOv3 ladder also recorded it, and validation never selects a checkpoint or stops training early. Two controlled studies read test quality during selection: the factor sharing study selects its learning rates on the held out test R^2 of its tuning seeds, where validation gives the same rates, and brackets its widths with a pilot that reads the held out test R^2 of reported seeds (Section I.5), and the exploratory geometry screen selects learning rates on the held out test R^2 of tuning seeds (Table 8). The tabular architecture (placement, K, R, and readout rank) was fixed before any of the three tabular test values was examined, and the DINOv3 ladder follows the same rule over its 3 final seeds.

Inference interventions. In a trained model, a temporary forward mask sets selected order channels to zero while parameters, calibration constants, inputs, and all other channels stay fixed, and nothing is retrained. In the aligned order 8 runs of Section I.3 the mask was applied by zeroing the coefficients $\gamma_{\cdot,j}$ and restoring them afterwards, which gives the same predictions. The effect of masking a set of orders is therefore the quality of the full model minus that of the same model with those orders masked, an intervention effect within one model rather than a conventional ablation with retraining.

Budget discipline. The common replacement fixes the interaction order at $K = 4$ in advance, and each threshold and Pareto comparison holds the order of its PSNeuron configuration fixed; the envelopes of Figure 8 range over every measured order, and the tabular replacements ranked by validation accuracy in Section L, the cross head site, and the wider frozen heads of Table 36 select the order on validation. Replacement families vary other capacity axes, such as basis, latent, bottleneck, or readout width, while the dense references vary their hidden width and, where width alone cannot reach a budget, the rank of a low rank input projection. Table 9 states which comparison each placement supports.

Statistical units and procedures. Final seed counts are those of Table 14, except that the separately trained structural comparisons use three seeds at the learning rate of the intact model (Table 47) and some tabular order mask conditions are evaluated on seeds 0 to 2 only (Section J.6); the factor sharing study uses ten at each width (Table 27), and the fixed G study uses 120 matched tuples per setting, paired across settings. Values after $\pm$ are sample standard deviations over final seeds, not confidence intervals. Paired comparisons use the matched seed structure. Intervals for fixed pairs are nominal two sided 95% Student t intervals over seed differences, Bonferroni checked within each stratum and reported beside a prespecified sensitivity family corrected jointly (Section L and Table 50); order masks are summarized by Hodges-Lehmann estimates with bounds from the exact signed rank distribution (Section J.6); and the fixed G study reports bootstrap intervals over runs and over teacher classes, because tuples that share a class are not independent (Section I.1).

G.2 Study settings

Fixed G recovery protocol. The fixed G recovery study trains on a CPU with AdamW, weight decay 0, a cosine schedule without warmup, and fp32. Its learning rate is tuned on validation data alone for each setting (k, t), a student that carries orders 2 to k trained on the target through order t, with three tuning tuples per rate, each a triple of teacher, initialization, and data seeds (Section I.1), over a base grid of eight rates spaced by factors of two from 2×10^{-4} to 2.56×10^{-2}. The rates 1×10^{-4} and 5.12×10^{-2}, one step beyond each end of the grid, were also trained but are consulted only when the winner of the base grid lies at an edge, and a second upper step at 0.1024 was added for the three settings whose selection stayed at the upper edge, $(6, 6)$, $(7, 6)$, and $(7, 7)$; the selected rates are in the LR column of Table 23. The prespecified edge rule keeps the setting $(8, 6)$ at 0.0128, although the rate 5.12×10^{-2}, which the rule did not consult, reached the recovery criterion on all three tuning tuples against two of three; the recovery reported for that setting may therefore understate what it reaches at its best tested rate. The aligned $G = K$ diagnostic, its gauge fixed reparameterization, and the factor sharing study of Section I also train with AdamW, weight decay 0, and a cosine schedule without warmup; their step counts and learning rates are given there.

Five of the nine original tabular selections in Table 15 (all three models on HIGGS Small, and PSNeuron and SwiGLU on Jannis) remain at the upper edge of the extended grid, so the reference selected on validation in Table 28 shares the tuning status of PSNeuron: both at the upper edge on HIGGS Small and Jannis, both interior on Helena. A separate common $K = 4$ cohort brackets every

Table 14: Training budgets and seeds of each study. The metric column names the reported quantity; higher is better for all of them except held out NLL. Tuning seeds are 101 and 102, and final seeds are 0 to 4, or 0 to 2 in the DINOv3 ladder, unless stated otherwise. For the DINOv3 ladder and the five seed study at $h = 2$ to 128 the first value is ImageNet100 and the second each task of the Visual Task Adaptation Benchmark (VTAB-1k). The FineWeb-Edu batch holds 16 sequences of 512 tokens. The synthetic teachers, mixed order targets, and product targets of Section J.6 train for 2,000 steps at batch 512 with five final seeds; the other controlled studies of Section I state their own budgets.

study	primary metric	steps	batch	tune seeds	final seeds
tabular replacement	accuracy	3000	256	2	5
cross head associative recall	recall@1	4000	64	2	5
recurrent mixer (DeltaNet)	recall@1	4000	64	2	5
attention score augmentation	recall@1	4000	64	2	5
CIFAR100 ViT	accuracy	8000	128	2	5
FineWeb-Edu language model	held out NLL	4000	16	2	5
FineWeb-Edu all-block transformer	held out NLL	7000	16	2	5
frozen feature adapters and heads	accuracy	2000	256	2	5
DINOv3 narrow head ladder	accuracy	12400 / 200	512 / 256	2	3
DINOv3 five seed study, $h = 2$ to 128, and linear probe	accuracy	12400 / 200	512 / 256	2	5
parity and XOR	accuracy	2000	256	2	5
fixed G recovery	held out R^2 and recovery	3000	512	3 tuples	120 tuples

Table 15: Principal configurations and selected learning rates. Every rate is chosen on validation data, never on test scores. Tabular totals are the trainable parameters of the complete models and include the unchanged host around the replacement site of width 64. Every tabular model differs from the GELU host only in the feedforward network of the second block, and the first block keeps its GELU network.

setting	exact configuration	selected learning rate
tabular PSNeuron	$B_{\mathrm{host}}(P(x))$, width 64, $G = 8$, $K = 4$, $R = 8$, host readout of rank 8; totals 21,018, 22,812, 27,324 on HIGGS Small, Jannis, Helena	HIGGS Small 6.4×10^{-3}, Jannis 6.4×10^{-3}, Helena 1.6×10^{-3}
tabular dense references	GELU or SwiGLU, hidden 128, same host site; selected SwiGLU on HIGGS Small and Jannis (totals 43,522, 45,316) and GELU on Helena (41,828)	GELU 6.4×10^{-3}, 4×10^{-4}, 8×10^{-4} and SwiGLU 6.4×10^{-3}, 6.4×10^{-3}, 8×10^{-4} on HIGGS Small, Jannis, Helena
MQAR cross head	eight heads, head dimension 16, head latent 4,	4×10^{-4}; the native control selected 1.6×10^{-3}
PSNeuron	$d = 32$, $G = 8$, $K = 2$, $R = 32$	
DINOv3 ladder	$d = 768$, $G = 8$, $K = 2$, $R = h$; $h \in \{1, 2, 3, 4, 8\}$	3.2×10^{-3}, the upper grid edge, for the PSNeuron head at every $h \geq 2$ and for the GELU head at every $h \geq 2$ except Flowers102 at $h = 2$ (4×10^{-4}); spread over the grid at $h = 1$; no extension round
fixed G recovery	$G = 8$, student and target orders of Table 23	tuned for each setting, 0.0128 to 0.1024

relevant tabular selection without changing the original table (Section K). The original grid edge configurations remain descriptive in the Pareto inventory (Section 4); neither cohort shows PSNeuron ahead of the dense frontier at every budget.

DINOv3 protocol and datasets. The heads, their matched controls, and the extraction of the frozen class token (CLS) features are described in Section J.10. The downstream datasets are the Describable Textures Dataset (DTD) (Cimpoi et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), ImageNet100 derived from ImageNet (Deng et al., 2009), Oxford-IIIT Pets (Parkhi et al., 2012), and Street View House Numbers (SVHN) (Netzer et al., 2011). ImageNet100 carves a validation set of 5,000 examples from its training split for learning rate selection and reports accuracy on its held out evaluation split. The four VTAB-1k tasks follow the VTAB-1k protocol (Zhai et al., 2019): a training set of one thousand examples, from which a validation split of 200 leaves 800 for fitting, and accuracy on the held out split of each task. Every head, the GELU and additive controls included, uses the base grid of five rates without an extension round. Optimization departs from the general protocol only in weight decay 10^{-4} on every parameter and a cosine schedule without warmup. Calibration uses the first 4,096 fitting rows, or all of them when fewer exist, and the same rows serve every model in a setting; each VTAB-1k task therefore calibrates on its 800 fitting rows, with no padding, repetition, or resampling.

The upper grid point, 3.2×10^{-3}, is selected in 128 of the 150 settings (five datasets, five widths, six head types). Besides the PSNeuron and GELU selections of Table 15, the additive head selects it in all but two settings with h from 2 to 4 (Flowers102 at $h = 2$ with 4×10^{-4}, SVHN at $h = 3$ with 1.6×10^{-3}). The ladder comparisons at narrow width are therefore unbracketed for PSNeuron and its controls alike. The five seed study of Section J.10 shares one grid of eight rates, 2×10^{-4} to 2.56×10^{-2}, across every head and never extends it; its largest rate is selected in 32 of the 60 settings, so its comparisons remain unbracketed as well. Its extension to $h = 8$ to 128 selects an interior rate for every head outside ImageNet100; on ImageNet100 the GELU head at $h = 8$ and PSNeuron at $h = 16$ select the largest rate, and eight of the nine heads at $h \geq 32$ and the linear probe the smallest.

Other heads and adapters on frozen features. The latent, direct, augmented, and wider replacement heads and the residual adapters also read frozen DINOv3 features of EuroSAT (Helber et al., 2019) and ImageNet100, but unlike the ladder heads they follow the general optimizer protocol of Section G.1. EuroSAT, which is distributed without a split, is divided once by a seeded random permutation (seed 0) into 21,600 training and 5,400 test images; a validation set of 5,400 images is carved from the training part, the remaining 16,200 are used for fitting, and accuracy is reported once on the test images. ImageNet100 carves a validation set of 26,689 images from its training split, fits on the remaining 100,000, and reports accuracy on the same 5,000 held out images as the ladder. The residual adapters on IMDb read frozen sentence embeddings of width 768 (Section J.2); a validation set of 5,000 reviews is carved from the 25,000 training reviews, the remaining 20,000 are used for fitting, and accuracy is reported once on the 25,000 test reviews.

Tabular architecture and data. The host and the common $B(P(x))$ site are those of Section 3: PSNeuron replaces only the feedforward network of the second of two residual blocks, and the site has 1,048 core and 1,024 host readout parameters. On HIGGS Small the GELU host has 35,522 parameters: 16,576 in the replaced feedforward network, 16,576 in the identical network of the first block, which stays dense, and 2,370 in the encoder, the readout, and three LayerNorms. The replacement therefore covers 46.7% of the host's parameters, and replacing both feedforward networks would cover 93.3% (accuracy tradeoffs in Table 48). The SwiGLU references selected on validation have 24,576 parameters in the second feedforward network, and Table 15 gives the totals of the complete models.

HIGGS Small follows the HIGGS benchmark of Baldi et al. (2014) and is accessed as OpenML dataset 23512; Jannis and Helena are OpenML datasets 41168 and 41169. We use OpenML through its reproducible dataset interface (Feurer et al., 2021). Each task is split once by a seeded random permutation of the full pool (split seed 0, not stratified), identical for every model and seed, into 62,000/15,000/21,050 training, validation, and test rows for HIGGS Small, 50,000/10,000/23,733 for Jannis, and 39,000/9,000/17,196 for Helena, from 28, 54, and 27 raw features. Categorical columns are one hot encoded; numeric columns are imputed with their median and standardized. These statistics come from the full unlabeled feature pool before the split, and labels play no part in them. The preprocessing is therefore transductive rather than fitted on the training split alone, and every compared model receives identical tensors and indices. A sensitivity analysis that fits these statistics on training rows and retunes every comparator is in Section K.

Vision and language data. The CIFAR100 ViT trains on 45,000 images and selects on 5,000 validation images, both carved from the official training set, and reports accuracy once on the official 10,000 test images; pixels are scaled to $[0, 1]$ and normalized per channel with the standard CIFAR100 mean and standard deviation, identically for every model. The FineWeb-Edu language model trains on sequences of 512 tokens drawn from 4,000 documents with a vocabulary of 16,384; one tenth of the documents forms the test split, and one tenth of the training sequences is carved for validation. The earlier single-FFN host follows the general optimizer protocol of Section G.1. The separate all-block transformer uses the same FineWeb-Edu loader, splits, and batch of 16×512 tokens, but trains for 7,000 steps; each cell selects its base learning rate on tuning seeds 101 and 102 from $\{2, 4, 8, 16, 32\} \times 10^{-4}$, warms up linearly for 350 steps (5%), then decays by a cosine schedule to zero, and evaluates final seeds 0 to 4 at the selected rate.

Sequence data and metric. MQAR contains 8,192/1,024/1,024 generated training, validation, and test sequences of length 64 over a vocabulary of 64. Each sequence begins with eight distinct key and value pairs stored as 16 interleaved tokens. The remaining 48 positions are query keys sampled

from those pairs; the targets are the associated values, positions that are not queries are ignored, and the metric is recall@1. The cross head study and the attention score study use different placements and are reported separately in Tables 29 and 34. The multi-needle task has the split sizes and sequence length of MQAR, and the attention score study draws it from the same generator. A haystack of 60 positions holds a filler symbol everywhere except four needles at distinct random positions, whose values are drawn from 62 value symbols; four final positions carry one reserved query symbol, and the targets there are the needle values in order of appearance. The vocabulary of 64 therefore holds the 62 values, the filler, and the query symbol. State tracking uses the same split sizes and sequences of 64 tokens over three operation symbols, and the target, read at the last position, is the state of a counter modulo 32. Recall@1 is the metric of all three tasks.

Recurrent mixer protocol. The recurrent MQAR, multi-needle, and state tracking comparisons use a DeltaNet host with four layers and $d_{\text{model}} = 128$, in which PSNeuron replaces the channel mixer of the last block and reads the recurrent state readout together with the input. Learning rates follow the general protocol, so each model received the extension round only when its own selection lay at an edge, and the native model searched a wider grid than most mixers: up to 6.4×10^{-3} on MQAR and multi-needle retrieval and down to 10^{-4} on state tracking, whereas every PSNeuron mixer of Table 32 searched 2×10^{-4} to 3.2×10^{-3} except the cross term mixer with $K = 4$ on multi-needle retrieval and the cross term mixer with $K = 8$ on state tracking, which were extended in the same way, as were 15 of the 48 capacity rungs on multi-needle retrieval; the GELU and SwiGLU mixers near 4,400 parameters that read the state were extended to 6.4×10^{-3} on MQAR, as were one GELU and the additive mixer on multi-needle retrieval, where the additive mixer still selected the largest rate (Section J.2). The recurrent comparisons across these grids are therefore descriptive.

Each mixer is calibrated on the two tensors that reach it when the initialized host runs once in evaluation mode on its calibration batch (Section 3); the joint mixer is calibrated like any PSNeuron. The cross term mixer, which keeps only the product terms that touch both the state and the input (Section E), first rescales its factor bank and then fits its frozen order constants on the cross term channels themselves; fitting them on the joint spectrum would calibrate a quantity that its forward pass does not use.

Choice of the DeltaNet layout. A validation rule, fixed before the choice but after part of its validation input existed, scored each layout by its validation recall relative to the native model, averaged over $K = 2$, 4, 6, and 8 and over MQAR, multi-needle retrieval, and state tracking. It ranked the cross term layout ahead by 0.041, no more than its threshold, the largest population standard deviation of the native model's validation recall over seeds (0.436, on MQAR), and so found the layouts too close to call. The authors then chose the joint layout, which leads on validation on both primary benchmarks, with relative scores of 0.636 against 0.600 on MQAR and 1.007 against 0.994 on multi-needle retrieval, while the cross term layout leads on the supporting state tracking task, 1.083 against 0.910. The joint lead on MQAR comes from $K = 4$, with validation recall of 77.8 against 29.9%; the cross term layout leads at $K = 2$, 6, and 8. These validation values come from the five final seeds whose test values are reported, not from the tuning seeds. Test values of both layouts had already been computed when the choice was made; neither the rule nor the recorded basis of the choice uses them, so the choice was not blind to them, but on test the rule also finds the layouts too close to call, and the joint layout leads on both primary benchmarks there as well. The order $K = 4$ was fixed in advance, and the capacity rungs of Section J.2 repeat the joint layout. Comparisons of this layout are therefore selected layout comparisons: their test values were not used for selection, but a choice made on the same seed identities adds post selection optimism that a replication on fresh seeds would remove. On 20 fresh seeds the joint layout does not lead the cross term layout on MQAR at $K = 4$ (-1.14 points, nominal interval $[-22.15, 19.86]$; Section K.2), so its MQAR lead does not replicate; multi-needle retrieval was not retested.

Probes of exact order. Eight bit parity draws $x \in \{-1, +1\}^8$, predicts the sign of $\prod_i x_i$, and evaluates $K \in \{2, 4, 6, 8\}$. The nominal split sizes are 8,192/2,048/2,048, but the training draw covers all 256 points of the cube. One bit sits in each slot, so the partition always separates the target, and parity tests whether a high exact order can be trained on the truth table in an aligned setting. The XOR tasks draw $x \in \{-1, +1\}^{16}$ and predict a fixed product of four or of six coordinates with a PSNeuron whose order matches the target; held out rows that also appear in another split are removed deterministically, so held out evaluation is disjoint by row. Their shared partition can make the target

unrepresentable, so a second cohort uses the first five seeds whose partition separates the target, identified before training (Section J.3).

H Compute and Systems Behavior

Parameter count, executed arithmetic, latency, throughput, and memory are separate resource axes and are reported separately; a smaller parameter count is not a claim about speed, and no speedup of a complete model is claimed. Arithmetic is counted in multiply accumulate operations (MACs), one per scalar multiplication, elementwise products included.

Hardware and software. The matched timings of Tables 17 to 19 use one NVIDIA A100 80GB PCIe with PyTorch 2.6.0 and CUDA 12.4, in fp32 with TF32 matrix multiplication disabled. The tabular, sequence, CIFAR100, language model, and EuroSAT runs and the ImageNet100 heads outside the DINOv3 ladder and its five seed study trained on NVIDIA GPUs with PyTorch 2.6.0 in fp32, and the DINOv3 ladder and the exact order probes also trained on NVIDIA GPUs; the five seed DINOv3 study trained on a CPU with two to eight threads (PyTorch 2.6.0). The fixed G study trained on a CPU (AMD EPYC 7742, one thread, PyTorch 2.6.0). The learning curves of Tables 21 and 22 and the CPU timings of the isolated sites in Section H.2 were measured on a CPU with four threads (PyTorch 2.14.0), and the order cost profile of Table 20 on a CPU with eight threads (PyTorch 2.6.0).

H.1 Structural cost

Trainable parameters follow Eq. (4). All reported models compute the grouped slots with a masked dense projection, the reference projection: every basis multiplies the whole input by G masked copies of its weight row, so the Rd learned weights cost RGd projection MACs, of which $(G-1)Rd$ multiply structural zeros. The compact count of Table 16 charges one projection MAC per learned weight and keeps every other multiply, including the recurrence, order weighting, readout, and the masking and scaling of the weights once per call; both counts spread that per call work over the batch of the measurement, 256 rows at the tabular site, $4{,}096$ tokens at the MQAR site, and 512 rows for the DINOv3 heads.

Table 16: Trainable parameters and forward MACs at representative sites; lower is better in the three count columns. Executed counts use the reference masked projection and compact counts its blockwise equivalent, per example (tabular, DINOv3) or per token (MQAR) at the site, for the descending recurrence evaluated in training. The out of place form that the released implementation evaluates without gradients (Section A.10) executes $\binom{K}{2}$ more multiplies per basis, giving the 6.06k and 13.54k forward MACs of Table 3(a) at the tabular and MQAR sites. The DINOv3 rows are the ImageNet100 heads with $C = 100$ classes; they carry only order 2 but, as in training, evaluate the recurrence to order 4 and scale every slot at run time, and both counts include this work.

setting	compared module	trainable params	executed MACs	compact MACs
tabular	PSNeuron $B_{\mathrm{host}}P(x)$	2,072	6.01k	2.42k
	GELU $h = 128$	16,576	16.38k	16.38k
	SwiGLU $h = 128$	24,576	24.70k	24.70k
MQAR output projection	PSNeuron cross head site	5,632	13.51k	6.34k
	native linear projection	16,512	16.38k	16.38k
DINOv3 $h = 2$	PSNeuron head, $K = 2$	1,836	12.62k	1.86k
	GELU head	1,838	1.74k	1.74k
DINOv3 $h = 3$	PSNeuron head, $K = 2$	2,704	18.92k	2.80k
	GELU head	2,707	2.60k	2.60k

Structural zeros, the whole difference between the executed and compact counts, make up 60% of the executed MACs at the tabular site ($3{,}584$ multiplies), 53% at the MQAR site, and 85% in both DINOv3 heads. The compact count at the tabular site is 0.15 and 0.10 times that of the GELU and SwiGLU networks. In the narrow DINOv3 heads the masked projection makes the executed arithmetic 7.27 times that of the GELU head of the same width, while the compact count is 1.07 times. These gaps are properties of the masked projection, not extra parameters; the fused projection below skips the structural zeros only when all bases share one partition into blocks of equal size, so it does not apply to the narrow DINOv3 heads, whose bases draw separate partitions.

At the tabular site ($d = o = 64$, $G = 8$) with an outer readout B_{host} of rank 8 (1,024 parameters), Eq. (4) gives $128R + 1{,}024$ trainable parameters at $K = 2$ and $R(K-1)$ more at $K \geq 3$: 2,048, 2,072, 2,088, and 2,104 for the trained sites with $R = 8$ and $K = 2$, 4, 6, and 8. An additional basis costs $d + o = 128$ parameters, and $K - 1$ more at $K \geq 3$, whereas raising K from 2 to 8 at $R = 8$ adds $7R = 56$, less than one basis. These counts do not show that a higher order substitutes for width at matched quality; Sections 5 and L make that comparison only where measured configurations allow it.

H.2 Measured runtime and memory

Table 17 fixes the protocol for timing complete models on one HIGGS Small pair, and Table 18 applies it to 48 pairs of a PSNeuron site and a control with matched parameters, at $K = 2$ and $K = 4$ and up to three matched sizes per task: rungs 1, 4, and 16, at which the site holds about 1, $1/4$, and $1/16$ of the dense budget, as in Table 3(b). For each K there are 3 FineWeb-Edu language model pairs, with PSNeuron in the feedforward network of the second block against SwiGLU in the same place (quality in Table 37); 1 pair at the attention output of the CIFAR100 ViT, against the additive cross head of matched size; 3 pairs at the feedforward site of the same ViT, 9 tabular pairs on HIGGS Small, Jannis, and Helena, and 6 pairs of replacement heads with a conditioner on frozen DINOv3 features of EuroSAT and ImageNet100 (Table 36), all against GELU networks or heads; and 2 pairs of attention score corrections on MQAR and multi-needle retrieval, against a low rank bilinear correction. Training batches are each host's own: 16 sequences of 512 tokens for the language model, 256 rows for tabular models and frozen heads, 128 images for the ViT, and 64 sequences for attention scores. The largest measured batches are 256 sequences for the language model, where 1,024 ran out of memory, 65,536 rows for tabular models and frozen heads, 8,192 images for the ViT (2,048 for one ViT feedforward pair at $K = 4$), and 4,096 sequences for attention scores.

Table 17: Timing protocol, shown on one HIGGS Small pair and applied at the training batch to all 48 pairs of Table 18. This pair, rung 4 of the HIGGS Small pairs at $K = 4$, is matched in size and is not the smaller replacement of Table 1. Ratios above one mean that PSNeuron is slower.

item	protocol
Task and scope	HIGGS Small classification; the complete model, with only the feedforward network of the second block changed.
PSNeuron geometry	Two residual blocks of width 64; the first feedforward network is GELU with hidden width 128, the second is $B_{\text{host}}P(x)$ with $d = o = 64$, $G = 8$, $K = 4$, $R = 16$, and an outer readout B_{host} of rank 16. Site: 4,144 parameters; whole model: 23,090.
Control geometry	The same host; the second feedforward network is $64 \to 31 \to 64$ with GELU and biases. Site: 4,063 parameters; whole model: 23,009.
Data and loss	One resident batch of 256 training examples; cross entropy on the task logits; both models are calibrated on 128 training examples before timing.
Hardware and precision	One NVIDIA A100 80GB PCIe, PyTorch 2.6.0, CUDA 12.4; fp32 with TF32 matrix multiplication disabled; no mixed precision, compilation, or fused block.
Timing	Synchronized wall clock medians: inference over 50 timed calls after 20 warmup calls, training over 25 timed steps after 5 warmup steps. Inference is one evaluation forward without gradients; a training step clears the gradients, evaluates the task loss, backpropagates, and applies AdamW at learning rate 10^{-3}, used only for timing. Three separate executions; the median and range of their ratios are given below.
Memory	Peak allocated CUDA memory during the timed training loop. Both models and the batch stay on the device, so this is the allocation of the process rather than the footprint of one model.

metric	median ratio P/C	range over executions
inference	1.49×	1.35 to 2.01×
full training step	2.36×	1.72 to 2.49×
peak training memory, MiB, P/C	18.9/18.4 (first execution)	18.9 to 42.4 / 18.4 to 41.9; ratio 1.01 to 1.03

Over all 48 pairs the time ratios range from 0.83 to 20.16, and the HIGGS Small pair of Table 17 is one member of the tabular range, not a representative value. At the training batch the language model runs within 3% of its control and the ViT attention output within 6% at inference; the ViT feedforward site is slower at inference in every configuration, and the training steps of both ViT sites range from faster to slower than those of their controls. At the training batch, every tabular, attention score, and frozen head configuration is slower than its control at inference and per training step. Per forward pass PSNeuron launches 53 to 55 kernels against 13 or 14 for tabular models, 139 to 143 against 55 for attention scores, 46 to 48 against 3 for frozen heads, 154 to 156 against 114 at the ViT feedforward site, 158 or 160 against 123 at the ViT attention output, and 6,212 to 6,214 against 6,174 in the language model. These counts come from a recount of the same 48 pairs on a CPU with the same builders and input shapes. The counter of the timing jobs skipped every

Table 18: Whole model A100 latency of PSNeuron over its matched control (P/C) for all 48 configurations; values below one favor PSNeuron. Each entry spans the configurations of one host, site, and order K and the executions of each. The training batch columns use a real batch of the host's own size and its task loss; the largest batch columns use synthetic inputs of the same shape and a shared surrogate loss. At the training batch, the median inference time over all configurations of both orders and all executions is 0.99 ms for PSNeuron against 0.65 ms for the control in the tabular MLP and 0.82 against 0.14 ms for the frozen heads; across the six sites, the difference of these medians, the time PSNeuron adds per inference batch, ranges from 0.25 ms at the ViT attention output to 1.42 ms at the attention scores.

			training batch		largest measured batch	
host	site	K	inference	training step	inference	training step
language model, FineWeb-Edu	feedforward	2	1.00 to 1.03×	0.99 to 1.03×	1.01 to 1.15×	1.00 to 1.05×
language model, FineWeb-Edu	feedforward	4	0.99 to 1.02×	0.98 to 1.01×	1.00 to 1.27×	1.00 to 1.07×
ViT	attention output	2	1.04 to 1.05×	0.88 to 1.13×	1.03×	1.02×
ViT	attention output	4	1.03 to 1.06×	1.17 to 1.20×	1.04×	1.03×
ViT	feedforward	2	1.04 to 1.41×	0.92 to 1.20×	1.00 to 1.45×	1.02 to 1.46×
ViT	feedforward	4	1.04 to 1.69×	0.83 to 1.25×	1.02 to 1.75×	1.02 to 1.62×
tabular MLP	feedforward	2	1.37 to 3.23×	1.20 to 2.13×	0.99 to 3.11×	1.55 to 3.80×
tabular MLP	feedforward	4	1.27 to 3.30×	1.48 to 2.70×	1.01 to 4.98×	1.64 to 4.99×
attention host	attention scores	2	2.09 to 2.13×	1.74 to 1.77×	1.78 to 1.80×	1.67×
attention host	attention scores	4	2.12 to 2.23×	1.31 to 2.15×	2.10 to 2.18×	1.81 to 1.83×
frozen DINOv3 features	prediction head	2	4.43 to 10.18×	2.18 to 4.50×	5.99 to 11.62×	7.40 to 16.87×
frozen DINOv3 features	prediction head	4	4.49 to 10.23×	2.81 to 6.10×	4.54 to 15.89×	4.89 to 20.16×

operator whose name contained that of the transpose, so it omitted the eight concatenations of each recurrence call of a PSNeuron site and, on both sides, two negations at the attention score site and 1,024 host concatenations in the language model; with that counter the recount reproduces all 96 counts of the first timing job, while seven arm counts of a later job are two higher because an out of memory error left those models in training mode. At the largest batches the ViT attention output is within 4% of parity, while the attention score corrections remain 1.67 to 2.18 times slower and the frozen heads become slower still, up to 20.16 times per training step; one tabular pair (Jannis, $K = 2$, rung 16) reaches 0.99 at inference with 65,536 rows in its only execution at that batch, so it is not distinguishable from parity, while its training step there is 1.77 times slower.

In the first execution at the training batch, peak memory allocated during the training step, which includes the parameters of both models held on the device, is 1.01 to 1.18 times that of the control for the language model, 1.01 to 1.03 for the ViT attention output, 1.01 to 1.64 for the ViT feedforward site, 1.01 to 1.22 for tabular models, 1.30 to 1.46 for attention scores, and 1.10 to 2.04 for frozen heads. Later executions hold different resident allocations, so their peaks are not pooled.

Exact compared sites. The tabular and MQAR attention output sites of Table 1(a) and the narrow DINOv3 heads are not among the 48 pairs. Timed in isolation on an A100 of the same type (the site alone on inputs of its shape, a mean squared output as loss, AdamW at learning rate 10^{-3}, medians of 200 calls after 50 warmup calls and of 100 steps after 20, two separate runs), the tabular site takes 6.32 to 6.90 times the inference time of the GELU site and 4.42 to 4.46 times its training step (5.00 to 5.25 and 4.18 to 4.28 times those of SwiGLU), and the MQAR cross head site takes 6.99 to 10.95 and 4.33 to 5.44 times those of the native projection; the two runs disagree most for this site. The narrow DINOv3 heads take 4.73 to 4.84 and 4.05 to 4.10 times those of the GELU heads. On a CPU with four threads, with operator calls rather than GPU kernels counted by the profiler on a single thread, one forward of these sites issues 143 to 157 operator calls against 14 to 26 for the dense modules, and the time follows the number of calls rather than the arithmetic: the tabular site executes 0.37 times the MACs of the GELU site but takes 4.7 times its inference time.

Fused slot projection. Removing the multiplies by structural zeros speeds up the projection only for the widest geometries and does not shorten a complete forward pass at the sizes trained here, because at the sizes trained here the work is small and kernel launches, synchronization, and intermediate memory traffic dominate. Table 19 compares a fused projection that skips the zero blocks with the reference masked projection on identical weights and inputs; the two agree bitwise on small integer weights and inputs with unit block scales, where no rounding occurs, and differ on random inputs by at most 9.5×10^{-7} in absolute value, so they compute the same function and differ only in the order of floating point additions. At nine of the ten geometries and row counts of trained

configurations the fused projection reaches only 0.73 to 0.75 times the speed of the reference; the exception is the feedforward geometry of the language model at $8{,}192$ rows, where it is 1.48 times faster, and raising the row count of wide trained geometries gives 1.47 to 2.00. A narrow geometry with $(d, R) = (32, 32)$ stays slower at $8{,}192$ rows, so no single row count marks the break even point; the effective row count combines the batch with sequence or patch positions.

Table 19: Reference masked slot projection against the fused projection, forward only, with $G = 8$; medians of 100 synchronized calls after 20 warmup calls, in microseconds (lower is better). The last column is the reference time over the fused time, so values above one favor the fused projection. The first ten rows are geometries and row counts of trained configurations; the last three raise the row count of trained geometries.

d	R	G	input rows	reference (μs)	fused (μs)	ref./fused
8	8	8	256	87.5	118.2	0.74
8	8	8	16,384	86.9	117.6	0.74
64	8	8	256	87.5	117.4	0.75
16	8	8	256	87.1	117.4	0.74
32	8	8	512	87.8	117.6	0.75
16	16	8	4,096	88.2	117.1	0.75
32	32	8	8,192	86.8	116.8	0.74
32	32	8	4,096	88.1	117.3	0.75
128	349	8	8,192	405.8	274.0	1.48
768	1463	8	256	408.0	558.0	0.73
768	32	8	65,536	1,463.0	776.4	1.88
128	349	8	262,144	10,752.8	5,381.2	2.00
768	1463	8	8,192	8,269.8	5,635.8	1.47

In complete PSNeuron models at their training batch (forward only, 27 configurations, with a fused kernel tiled differently from the one in Table 19), the projection takes 0.3 to 14.8% of the forward time, so even a free projection could not shorten the forward pass by more than about 15%; substituting the fused projection for the reference one changes that time by factors of 0.93 to 1.00 (reference over fused). A fused kernel for the whole PSNeuron block, timed forward only against the reference layer on five geometries at 256, $4{,}096$, and $16{,}384$ rows, is slower in 14 of the 15 settings, at 0.28 to 0.85 times the reference speed, and 1.40 times faster only for the language model geometry $(d, R) = (128, 349)$ at $16{,}384$ rows. Its outputs differ from those of the reference by up to 1.5×10^{-5}, so it is not bitwise identical to the reference layer and is not used for any other result.

Cost of additional orders. At fixed $R = 505$, $d = o = 128$, and $G = 8$, raising K from 2 to 8 adds only $7R = 3{,}535$ parameters but increases the work and the activation state kept for training, and with them the batch 1 latency, the training step time, and the peak training memory of the layer (Table 20). The step time is not monotone in K, so these values describe this measurement rather than a smooth law.

H.3 Learning dynamics and steps to a quality threshold

On the controlled tabular target of Table 21, PSNeuron reaches $R^2 \geq 0.999$ in fewer steps than the GELU network of the same host, but its steps are more expensive, so the estimated times are close; SwiGLU reaches the threshold in fewer steps and less time. At the MQAR cross head site, at one shared learning rate (Table 22), PSNeuron reaches recall 0.99 before the native projection on seed 1, after it on seed 2, and never on seed 0. Neither comparison supports a general claim of faster learning.

Steps to a common threshold. Four studies compare the training speed of their PSNeuron models with that of a primary dense control, paired by seed: the tabular replacement against the GELU network of its host, the cross head sites of MQAR and the CIFAR100 ViT at $K = 2$, 4, 6, and 8 against the native output projection, the heads on frozen DINOv3 features against GELU heads, and a controlled suite of synthetic targets against a GELU network. For each seed, T_{95} is the first evaluated step at which the validation quality of the PSNeuron model reaches

$$q_{95} = Q_0^{\mathrm{PS}} + 0.95 \left(Q_{\mathrm{final}}^{\mathrm{C}} - Q_0^{\mathrm{PS}} \right),$$

Table 20: Cost of raising K at fixed geometry ($R = 505$, $d = o = 128$, $G = 8$), measured on a CPU with eight threads; lower is better in every column except tokens per second. Parameters and peak memory (resident set size above its value after construction, measured in a fresh process) are those of the layer alone. Times are those of a two layer DeltaNet sequence model (vocabulary 256, width 128) with the layer as the channel mixer of both blocks: a training step at batch 32 and sequence length 128, an inference forward at the same batch, and latency at batch 1. Each time is the median of nine repeats after two warmups, taken round robin across K so that process warmup is not charged to whichever K runs first. In this measurement both training and inference evaluate the descending recurrence of Eq. (16), which executes 15, 26, 33, and 36 multiplies per basis and token at these K, whereas the out of place form that the released implementation evaluates without gradients (Section A.10) executes GK, that is 16, 32, 48, and 64. These are CPU results and are not offered as a proxy for GPU cost.

		training			inference		
K	params	step, ms	tokens/s	peak RSS, MB	batch 1, ms	batch fwd., ms	peak RSS, MB
2	129,280	423.7	9,667	373.8	23.03	97.4	124.1
4	130,795	479.0	8,551	605.8	24.46	117.2	182.0
6	131,805	802.4	5,105	771.2	25.35	124.3	322.6
8	132,815	727.9	5,627	853.7	25.57	133.7	321.9

Table 21: Learning dynamics at the tabular site, measured on a CPU with four threads; fewer is better in the step and time columns. The task is a controlled regression target on 32 Rademacher inputs whose orders 2 to 6 carry equal shares of the variance, with a scalar output, placed in the tabular feedforward host, so the parameter counts differ from Table 15. It trains for 3,000 steps at rates selected on validation with tuning seeds 101 and 102 over the five point grid, without an extension round (SwiGLU selects the upper edge, so its curve may be pessimistic). Entries are medians over final seeds 0 to 4 of the first recorded step, every 50 steps, at which validation R^2 reaches the threshold. "Run" is the median wall time of a complete training run including evaluation; "est. to 0.999" scales it by the fraction of steps needed to reach the threshold, an estimate rather than a timed crossing.

model	params	LR	median steps to R^2			run, s	est. to 0.999, s
			0.9	0.99	0.999		
PSNeuron $B_{\mathrm{host}}(P(x))$, $K = 4$	21,209	0.0016	249	299	599	21.5	4.3
GELU, hidden 128	35,713	0.0016	249	349	949	14.4	4.6
SwiGLU, hidden 128	43,713	0.0032	149	249	449	15.2	2.3

Table 22: Steps to a recall threshold at the MQAR cross head site, from separate runs of both models on a CPU at one shared learning rate of 8×10^{-4}, batch 64, and 4,000 steps, with validation recall@1 recorded every 50 steps. Each entry is the first recorded step at or above the threshold, not an interpolated crossing, and n/a marks a run that never reaches it. The final recall of PSNeuron seed 0 is close to the 0.125 obtained by guessing among the eight values present in a sequence. The tuned study with five seeds, which selects a rate for each model, is a different cohort; there one seed of the native control stalls.

model	seed	final recall	steps to 0.5	to 0.95	to 0.99
PSNeuron cross head $K = 2$	0	0.1285	n/a	n/a	n/a
	1	0.9998	1,999	2,049	2,099
	2	0.9996	2,749	2,799	2,849
native output projection	0	0.9997	1,899	1,949	2,049
	1	0.9995	2,149	2,149	2,199
	2	0.9995	1,949	1,999	1,999

where Q_0^{PS} is its initial quality and $Q_{\mathrm{final}}^{\mathrm{C}}$ the final quality of the control on the same seed; for the control, T_{95} is the first evaluated step at which it reaches $Q_0^{\mathrm{C}} + 0.95\left(Q_{\mathrm{final}}^{\mathrm{C}} - Q_0^{\mathrm{C}}\right)$, where Q_0^{C} is its own initial quality. The threshold measures progress from the start of the PSNeuron model toward the final quality of the control, so it is not the step at which the two match, and it differs from the fixed thresholds of Table 21. A seed that never reaches the threshold is right censored. A PSNeuron model counts as later when the paired 95% interval of the difference in T_{95}, PSNeuron minus control, lies above zero or when it is censored on at least one seed, and as earlier when the interval lies below zero and its terminal quality is not significantly below the control's. A comparison in which the PSNeuron model is never censored counts as neither when the two sides differ in horizon, batch size, optimizer, schedule, evaluation split, or evaluation interval, or when a control seed diverges, does not improve on its initial quality, or ends more than a quarter of the control's median gain below its median final quality; the learning rates may differ. On each tabular task the ten primary comparisons are the five anatomies of Table 42 at $K = 4$ against the host with its GELU network, and a ladder of budgets, $K = 2$ at rungs 2, 4, and 8 and $K = 4$ at rungs 2 and 4, in which rung n rebuilds both the PSNeuron site and the GELU network at about $1/n$ of the 16,576 parameters of the host's GELU site; the same five anatomies on the controlled target of Table 21 add five comparisons.

On Jannis, six of ten primary comparisons reach the threshold 155 to 370 steps earlier than their dense controls, but the selected PSNeuron learning rates are 2 to 16 times those of the controls, and the intervals of terminal quality cover zero in all six. No primary comparison in any other study is earlier. Comparisons counted as later are four of ten on Helena, by 240 to 380 steps, among them the $K = 2$ and $K = 4$ models at rung 2 and the $K = 4$ model at rung 4, which are 380, 370, and 240 steps later and end with lower accuracy; two of ten on HIGGS Small; one of five on the controlled target with equal shares through order 6 of Table 21; 31 of 54 comparisons of heads on frozen DINOv3 features, 27 with the paired interval above zero and 4 with the PSNeuron head censored; and 6 of 8 comparisons at the cross head sites, 4 and 2 in the same two ways. In the controlled suite of teachers, parity, XOR, and mixed order targets, the PSNeuron model is censored in 51 of 56 comparisons. A two sided F test on the variances of T_{95} across seeds applies only to comparisons in which the PSNeuron model is never censored, its terminal quality is not significantly below the control's, and the protocol and control checks pass, 33 of the 153 primary comparisons, each with $n = 5$ seeds; it favors PSNeuron in none of them and finds a larger variance for PSNeuron in 4, all on the controlled target with equal shares through order 6. The evidence therefore supports a local step advantage on Jannis under its selected learning rates, not a general claim of faster optimization or of lower training cost in wall clock time.

I Controlled Mechanistic Studies

This appendix reports the controlled studies behind Sections 5.3 and 5.4. All use analytic targets whose content at each order is known, on Rademacher inputs except in Section I.4, which plants them in frozen DINOv3 features.

I.1 Recovery at fixed slot count and the statistical unit

Every teacher of the fixed G study is an analytic target on one support S of eight input coordinates, the same for every teacher, that places one coordinate in each slot of the students' first basis. The target through order K is

$$T_K(x) = \sum_{j=2}^{K} \lambda_j \binom{8}{j}^{-1/2} e_j(x_S), \qquad \lambda_j = \sigma_j/\sqrt{7}, \qquad \sigma_j \in \{-1, +1\},$$

which has the form of Proposition 4. A teacher has no trainable parameters; its seed draws only the seven signs σ_2 to σ_8, of which the target through order K keeps the first $K - 1$. On Rademacher inputs the components of different orders are orthogonal and each has variance exactly $1/7$. Lower orders are identical across K and are not renormalized, so the target through order K has variance $(K-1)/7$ and each of its orders carries the share $1/(K-1)$. Recovery means final held out $R^2 \geq 0.9$; we report recovery only at this threshold and not the final R^2 of each run, so we make no claim of robustness to the threshold, for example at 0.8 or 0.95.

Students read $d = 32$ inputs with $R = 64$, $G = 8$, and one balanced partition per basis, the same in every run, whose first basis separates the support, so an exact witness of width one exists (Proposition 4). Every run trains on $8{,}192$ training rows, with $4{,}096$ validation and $4{,}096$ test rows, for the $3{,}000$ steps at batch 512 of Table 14, and recovery is read once on the test rows after the last step. A setting (k, t) is a student that carries orders 2 to k trained on the target through order t. In the matched initialization of every reported setting (Section 4), a student starts each carried coefficient at the teacher's scale $1/\sqrt{7}$ and takes the rescaled bank and the calibration constants of the $K = 8$ reference built from the same initialization seed and calibrated on $2{,}048$ rows, so the channels that two students share are identical at the first step. A *tuple* is one triple of teacher, initialization, and data seeds, where the data seed fixes the samples, the batch order, and the calibration rows. The ith tuple uses the ith initialization and data seeds, and the same 120 tuples serve every setting. Two sign vectors pose the same learning problem in distribution when they differ by a global negation, which the readout absorbs, or by the reflection $x \mapsto -x$, which flips the signs of the odd orders. These symmetries group the teachers into *classes*, counted in Table 23, so the 120 tuples of a setting are repeated draws within a few classes, not 120 independent teachers. The teacher of the ith tuple is the ith admitted by a scan of teacher seeds fixed before training, which alternates the two order 4 classes and skips every teacher that shares an order 6 class with one of the three tuning teachers. The scan admits 5 of the 8 classes at order 6, and hence 10 of 16 at order 7 and 20 of 32 at order 8, and the final teachers occupy all of them and both classes at order 4.

Learning rates were tuned for each setting on validation data alone (Section G). From order 6 upward, the tuning teachers and the final teachers belong to disjoint classes. At order 4 there are only two classes, so the settings $(4, 4)$ and $(6, 4)$ were tuned on teachers from the classes of their final tuples and are descriptive. The two settings with a student of order 7 are descriptive because their selected rate, 0.1024, is still the upper edge of the tested range. That rate comes from a second upward extension of the grid, beyond the single extension that the tuning protocol provides, and was added for $(6, 6)$, $(7, 6)$, and $(7, 7)$ before any final or test result was seen.

Table 23: All settings (k, t) of the fixed G study under the matched initialization, each over the same 120 tuples; "classes" counts the classes of the 120 final teachers at the target order t. "Recovery" is the fraction of tuples reaching held out $R^2 \geq 0.9$; higher is better. The run interval is a percentile bootstrap 95% interval that resamples tuples, and the class interval resamples the five order 6 classes of the final teachers, the level at which tuning and final teachers are disjoint; both use $10{,}000$ draws, and Figure 2(c) shows the run interval. Neither turns the tuples into independent teachers, and a recovery of one with interval $[1, 1]$ does not establish perfect recovery in the population. The settings $(4, 4)$ and $(6, 4)$ and the two at the upper edge are descriptive.

setting (k, t)	LR	classes	recovery	run interval	class interval	tuning
$(4, 4)$	0.0128	2	1.000	$[1.000, 1.000]$	$[1.000, 1.000]$	bracketed
$(6, 4)$	0.0128	2	1.000	$[1.000, 1.000]$	$[1.000, 1.000]$	bracketed
$(6, 6)$	0.0512	5	1.000	$[1.000, 1.000]$	$[1.000, 1.000]$	bracketed
$(7, 6)$	0.1024	5	1.000	$[1.000, 1.000]$	$[1.000, 1.000]$	upper edge
$(7, 7)$	0.1024	10	0.658	$[0.575, 0.742]$	$[0.633, 0.688]$	upper edge
$(8, 6)$	0.0128	5	0.642	$[0.558, 0.725]$	$[0.554, 0.719]$	bracketed
$(8, 8)$	0.0256	20	0.250	$[0.175, 0.325]$	$[0.211, 0.283]$	bracketed

The three primary contrasts, from $(6, 6)$ to $(8, 6)$, from $(8, 6)$ to $(8, 8)$, and from $(6, 6)$ to $(8, 8)$, which changes student and target together, lower recovery by 0.358, 0.392, and 0.750, with run intervals $[0.275, 0.442]$, $[0.283, 0.500]$, and $[0.675, 0.825]$ and class intervals $[0.281, 0.446]$, $[0.312, 0.461]$, and $[0.717, 0.789]$. The class interval, which resamples only five classes, is nowhere wider than the run interval, in the table or for these contrasts, so it cannot serve as the more conservative of the two. Exact McNemar tests on the paired tuples give very small p values, but they too treat the 120 tuples as independent, and we do not use them for inference. Both contrasts that involve $(8, 6)$ carry a tuning caveat: the edge rule kept the rate of $(8, 6)$ at 0.0128, although the unconsulted rate 0.0512 reached the recovery criterion on more tuning tuples (Section G). If the recovery of $(8, 6)$ is therefore understated, the first contrast overstates the change and the second understates it.

Only the second contrast keeps the student fixed and changes the content of the target. The first also starts orders 7 and 8 with nonzero coefficients, so under equal and orthogonal initial channels a nominal $2/7$ of the initial output variance of the order 8 student lies at orders that the order 6 target lacks (Section G.1). The study has no control that starts orders 7 and 8 at zero and keeps them

trainable. Disabling the two orders permanently is not such a control: when the order 8 student takes the factors, partition, calibration, readout, coefficients, and optimizer state of the order 6 student, disabling orders 7 and 8 reproduces the $(6, 6)$ run exactly, in forward values, gradients, and optimizer steps. The closest measured condition is in the aligned order 8 runs of Section I.2, with other sign patterns and a different protocol: starting the order 2 coefficient at one and every higher coefficient at zero, with all orders trainable, recovers 13 of 72 runs against 16 of 72 with equal starting coefficients $1/\sqrt{7}$. The effect of adding orders 7 and 8 therefore remains confounded with their initialization. A secondary *native* initialization calibrates each student itself, which yields the same constants because neither c_j nor the bank rescale depends on the maximum order (Section A.3), and starts every coefficient at $1/\sqrt{k-1}$, so for the students of order 8 it coincides with the matched initialization. It gives the same recovery fraction in all seven settings, full recovery in four of them and 0.658 at $(7, 7)$, with $(7, 6)$ then selecting the rate 0.0512. Because the two initializations differ in the initial scale of each order below order 8, we do not compare them for inference.

I.2 Representability and recovery in two geometries

The PSNeuron rows of the narrow probe of Table 24 describe a geometry that mostly cannot express its targets. For the order 6 target on four of the five seeds, and for the order 8 target on all five, no basis places the coordinates of the single monomial of top degree in distinct slots, so no setting of the weights expresses the target, while the one order 6 seed whose partition does is solved. The probe therefore measures this geometry, not the value of covering the target orders: raw R^2 rises with K up to the order of each target, but so does the analytic ceiling, and as a fraction of the ceiling the $K = 4$ student is ahead of the students that cover the target on both targets.

An aligned geometry removes this obstruction and makes every target representable, so only optimization remains. It uses $d = 32$, $R = 64$, and G equal to the target order and to the size of its support. Every basis keeps its own fixed partition, and the partition of the first basis places one support coordinate in each slot, so that basis alone separates every planted monomial and an exact witness of width one exists (Proposition 4). Training runs for 3,000 steps at batch 512, on as many training, validation, and test rows as in the fixed G study, at the learning rate 3.2×10^{-3}, the upper edge of the general grid of five rates in Section G. Each target has unit variance and the share $1/(K-1)$ at each of its orders, so the order 8 target has the teacher construction and the student geometry of the fixed G study, with coefficients of magnitude $1/\sqrt{7}$ and $G = 8$, whereas the order 6 target has coefficients of magnitude $1/\sqrt{5}$ and $G = 6$: the slot count and the coefficient scale change together with the order. The seeds are fully crossed: four initialization seeds for the order 6 target and six for the order 8 target, each with four sign patterns and three data streams, where a data stream fixes the rows of every split and the batch order. The sign patterns follow a rule fixed before training: all positive, alternating, and two approximately balanced vectors drawn at random. No all negative pattern is included, because a sign vector and its negation pose the same problem under $B_{\text{core}} \mapsto -B_{\text{core}}$. The reflection $x \mapsto -x$ of Section I.1 maps the all positive pattern onto the alternating one, so these two also pose the same problem in distribution, and the four patterns represent three classes. Each student starts every coefficient at $1/\sqrt{K-1}$, the magnitude of the planted coefficients, and calibrates on 512 training rows rather than the 2,048 of the fixed G study. Because the learning rate is fixed rather than tuned for each setting, the recovery rates describe that rate. Runs that share a seed of any kind are not independent, and neither set is pooled with the fixed G study.

On the order 6 target, 44 of 48 runs recover, with a mean held out R^2 of 0.9606 and a median of 1.0000, at 2,432 parameters, slightly more than the 2,177 of the GELU network that fits targets of the same construction in the narrow probe (Table 24). On the order 8 target, with 2,560 parameters, 16 of 72 runs recover, with mean 0.4956 and median 0.4459. The sample standard deviation of held out R^2 over runs rises from 0.1239 to 0.3308, because the outcomes split into recovered models and fits of low quality. Every value in these runs stayed finite, which rules out divergence but not other numerical or optimization difficulties.

An exact reparameterization. The gauge fixed reparameterization of Section A.13 constrains every factor row to unit norm, moves its radius into the order coefficients, and keeps the function class, the 2,560 parameters at order 8, and the initial function of the unconstrained model. On the same 72 order 8 runs it raises recovery from 0.2222 to 0.4861: 4 runs recover only in the original coordinates and 23 only in the new ones, with an exact two sided McNemar $p = 3.1 \times 10^{-4}$. Different

pairs share initialization seeds, sign patterns, and data streams, so this p value treats dependent pairs as independent and is nominal. By initialization seed, recovery out of 12 runs changes from 0, 9, 3, 4, 0, and 0 in the original coordinates to 6, 8, 9, 7, 2, and 3: five seeds improve, one loses a run, and the three seeds that never recover in the original coordinates recover 11 of 36 runs after reparameterization. At order 6 the reparameterized model recovers all 48 runs against 44 ($p = 0.125$). In the paired comparison that starts the order 2 coefficient at one and every higher coefficient at zero, with every order trainable, the condition cited in Section I.1, it recovers 14 of the 72 order 8 runs against 13 ($p = 1.0$). The learning rate was not retuned for either parameterization, so both recovery rates hold at the shared rate and not at the best rate of each. At that rate a row of unit norm also takes a larger effective Adam step than its unconstrained twin, which confounds the gain with step size. A weight normalized control on the same 72 runs, which starts from the same function with the scale invariance and starting coefficients of the reparameterized model but trains rows of free norm, recovers 12 of them, so scale invariance alone does not produce the gain. The comparison therefore shows that the optimization coordinates can matter for recovery at order 8, not that they are what limits it.

Table 24: Narrow mixed order probe: Rademacher inputs, $d = 32$, and scalar targets with equal shares through order 6 or 8; five seeds, 2,000 training steps, held out $R^2\pm$ seed standard deviation. The PSNeuron students have $R = 8$, $G = 8$, and a balanced partition redrawn for every seed, which mostly cannot express the target, so these rows are not used for conclusions about the operator (Section I.2). The ceiling is the share of target variance at orders 2 through K, the population bound of Corollary 2 for a student of order K, attained when some basis separates the support; the realized partitions lower what is attainable. By seed, the $K = 6$ student reaches 0.508, 1.000, 0.310, 0.282, and 0.499 on the order 6 target, where 1.000 ($R^2 = 0.9999992$) is the one seed with an admissible partition, and the $K = 8$ student 0.224, 0.546, 0.256, 0.182, and 0.270 on the order 8 target. The additive control is defined in Section A.14.

		target through order 6		through order 8	
model	params	ceiling	R^2	ceiling	R^2
PSNeuron $K = 2$	264	0.20	0.123 ± 0.003	0.14	0.061 ± 0.006
PSNeuron $K = 4$	288	0.60	0.409 ± 0.116	0.43	0.242 ± 0.056
PSNeuron $K = 6$	304	1.00	0.519 ± 0.288	0.71	0.285 ± 0.120
PSNeuron $K = 8$	320	1.00	0.516 ± 0.291	1.00	0.296 ± 0.144
additive control	528	n/a	-0.002 ± 0.010	n/a	-0.005 ± 0.005
GELU MLP	2,177	n/a	0.999 ± 0.0002	n/a	0.998 ± 0.0004

I.3 Spectral failure and functional use within one model

The order spectrum and the masks use the 72 aligned order 8 runs of Section I.2. The spectrum of the failed runs (Figure 4 and Table 25) is a structured pattern of failure, not an explanation of why optimization ends in it.

Delivered share. Targets with equal shares allow a diagnostic resolved by order, because Walsh degrees are orthogonal under the input distribution and the planted content at each order is known. For a PSNeuron with scalar output, collect the delivered channels of order j over the bases as $\tilde{e}_j(x) \in \mathbb{R}^R$, which equal $\hat{e}_j(x)$ here because $\mu_j = 0$ and $\zeta_j = 1$ (Table 4), and the order coefficients as $\gamma_j \in \mathbb{R}^R$; $\odot$ denotes elementwise multiplication. The controlled students have no host readout and no core bias, so with every mask $m_j = 1$ the prediction splits exactly into the contributions of the orders,

$$\kappa_j(x) = B_{\text{core}}\big(\gamma_j \odot \tilde{e}_j(x)\big), \qquad \hat{y}(x) = \sum_{j=2}^{K} \kappa_j(x). \tag{19}$$

For held out samples $\{(x_n, y_n)\}_{n=1}^N$ with $\bar{y} = N^{-1}\sum_n y_n$, the delivered share of order j is the signed covariance of κ_j with the target, normalized by the target variance,

$$s_j = \frac{N^{-1}\sum_{n=1}^N (y_n - \bar{y})\kappa_j(x_n)}{N^{-1}\sum_{n=1}^N (y_n - \bar{y})^2} = \frac{\widehat{\mathrm{Cov}}(y, \kappa_j)}{\widehat{\mathrm{Var}}(y)}. \tag{20}$$

A negative alignment keeps its sign, and no absolute value or renormalization is applied. The shares therefore need not lie in $[0, 1]$, and their sum is

$$\sum_{j=2}^{K} s_j = \frac{\widehat{\mathrm{Cov}}(y, \hat{y})}{\widehat{\mathrm{Var}}(y)}. \tag{21}$$

A sum of one establishes this covariance alignment, not zero prediction error. If a centered $u \neq 0$ satisfies $\mathrm{Cov}(y, u) = 0$ and $\hat{y} = y + u$, the sum is one while the mean squared error is $\mathbb{E}[u^2] > 0$. Delivered share is thus neither an explained variance partition nor a causal attribution.

In the controlled teacher of order 8, whose planted orders each carry $1/7$ of the target variance, a model that reproduces every planted component has population delivered share $1/7$ at each order (Corollary 1). Measured shares need not equal $1/7$, because learned components are approximate and held out samples finite, and matching the shares alone does not prove exact recovery. This reference belongs to the controlled construction and does not transfer to other input distributions.

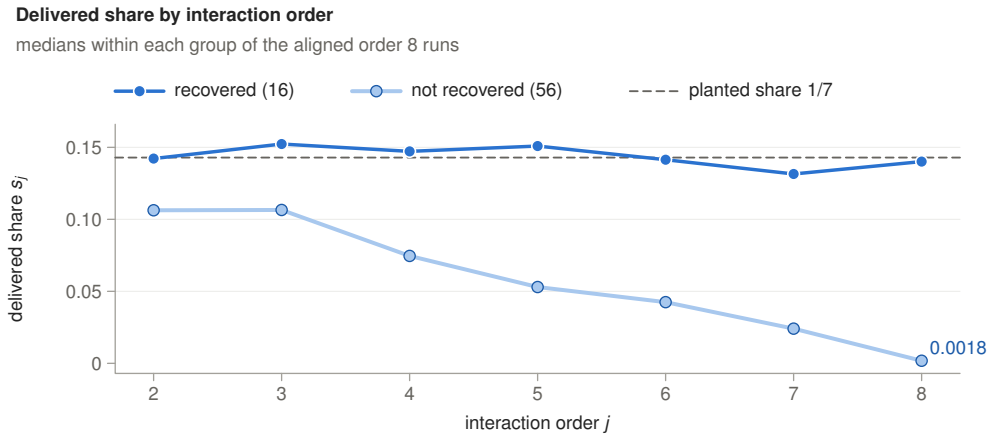

Figure 4: Delivered share s_j of Eq. (20) by interaction order in the 72 aligned order 8 runs, as medians over the 16 recovered models and the 56 models that did not recover; the dashed line is the planted share $1/7$. Recovered models stay near the planted share at every order, while the others keep about three quarters of it at orders 2 and 3 and progressively lose the upper orders. Table 25 lists the values and the spread among recovered models.

Table 25: Order spectrum of the 72 aligned order 8 runs, of which 16 recovered the target and 56 did not. Every entry is a median over the models of its group. The delivered share s_j is defined in Eq. (20), and its planted population value is $1/7 \approx 0.1429$ at every order. The mask loss ΔR_j^2 is the held out R^2 lost when order j alone is masked. Channel RMS is the root mean square of $\hat{e}_{r,j}(x)$ over held out inputs and the 64 bases, and coefficient RMS the root mean square of $\gamma_{r,j}$ over the bases. Over all orders of the recovered models, s_j lies between 0.0991 and 0.1761 and ΔR_j^2 between 0.0987 and 0.1850.

	delivered share s_j		mask loss ΔR_j^2		RMS, recovered models	
order j	recovered	not recovered	recovered	not recovered	channel	coefficient
2	0.1422	0.1063	0.1422	0.1108	0.7908	0.1303
3	0.1523	0.1065	0.1517	0.0913	0.6888	0.1320
4	0.1472	0.0746	0.1562	0.0884	0.5849	0.1750
5	0.1509	0.0530	0.1517	0.0445	0.4949	0.2442
6	0.1414	0.0425	0.1422	0.0266	0.4179	0.2986
7	0.1315	0.0241	0.1407	−0.0005	0.3131	0.3466
8	0.1401	0.0018	0.1403	−0.0172	0.2673	0.3687

Functional use is measured by the mask $m_j = 0$ of Section 3 at inference in the same model, applied by zeroing the order coefficients and restoring them exactly afterwards (Section G.1). In the 16 recovered models, removing orders 2 to 4 lowers held out R^2 by a median of 0.4346, with observed range 0.3891 to 0.4557, and removing orders 6 to 8 lowers it by 0.4286, with range 0.3107 to 0.4764;

the ranges are minima and maxima, not confidence intervals. Both medians are close to the population value $3/7 \approx 0.4286$ that Corollary 1 gives for a model reproducing the planted components. The signed rank bounds of the paired difference between the two masks include zero (Table 45, which also gives the Hodges-Lehmann estimates).

These 16 models are not independent confirmations. They come from three initialization seeds and nine distinct pairs of sign pattern and data stream, and a model that reproduces the teacher computes the same function whatever its initialization; in the three of these pairs where two or three models fit the teacher exactly, their mask losses agree to within 10^{-6}. For such models Corollary 1 fixes the mask loss through the teacher alone, so these masks confirm that recovered models realize the planted decomposition rather than testing use independently of recovery. The informative contrast is with the 56 failed models, whose intact held out R^2 ranges from -0.03 to 0.85. Removing orders 6 to 8 costs them a median of only 0.0128, against 0.2863 for orders 2 to 4, and masking order 7 or order 8 alone gives a negative median loss (Table 25).

In the recovered models the channel RMS falls and the coefficient RMS rises from order 2 to order 8 (Table 25), so, as the scale redundancy of Remark 4 implies, channel and coefficient magnitudes must be read together.

I.4 Planted interactions in frozen DINOv3 features

The studies above use Rademacher inputs, and the aligned study also uses a partition matched to the target. A registered external validity test asks whether recovery and mask reading carry over to real continuous representations without such a partition. The inputs are frozen DINOv3 ViT-B/16 class token features of ImageNet100 images (Section J.10), standardized, either a fixed subset of 64 coordinates, the primary configuration, or all 768, with 121,689 fitting, 5,000 validation, and 5,000 test rows. Each target sums sparse monomials with disjoint supports planted in these features at order 2, at order 4, or at both, so order 3 is absent from every target and serves as an absence control. The student is a PSNeuron with $K = 4$, $R = 16$, $G = 8$, learned order coefficients, and a balanced partition fixed at random without reference to the target; a GELU network of hidden width 16 checks only whether each target is learnable and never enters the verdict. Learning rates and step counts were selected on validation in 768 tuning runs, and 20 evaluation seeds per family, each with its own target, were then trained once each, 240 evaluation runs of both networks in all. A run is included, and counts as recovering its target, when it finishes with validation R^2 of at least 0.80, a lower bar than the held out $R^2 \geq 0.9$ that defines recovery for the controlled teachers; masks are read on test. The registered criteria require that the planted orders be exactly the orders with the largest mask effects, tested by an exact binomial test against its chance of $1/3$, and that masking each planted order lower held out R^2 more than masking each absent order, by exact sign tests with a Holm correction.

Both configurations pass and agree (Table 26). Every family includes all 20 runs, the planted orders carry the largest mask effects in 20 of 20 runs of every family ($p = 2.87 \times 10^{-10}$), every planted order loses more than every absent order (Holm corrected $p = 1.91 \times 10^{-6}$ for every pair), and masking an absent order changes test R^2 by at most 1.16×10^{-3} at width 64 and 4.7×10^{-4} at 768; the lowest test R^2 is 0.9389 and 0.8390. These p values sit at the floor of their exact tests and record unanimity, not effect size. The median absent order effect is below 10^{-11} in magnitude in every family; the largest values above are small leakage between degrees in a few runs, too small to mislead the masks. Recovery was not free of tuning: 53 of the 384 PSNeuron tuning runs missed validation R^2 of 0.80, and every PSNeuron selection chose 49,600 steps, the upper edge of the step grid. The GELU network fits the order 2 targets in all 20 runs at both widths and none of the order 4 or mixed targets, and at width 768 its selected rates for those targets lie at the lower edge of its grid.

Because every channel has exact degree, correct identification is close to implied once a model recovers its target. The test therefore shows that recovery and mask reading transfer from the Rademacher teachers to real continuous representations without a partition aligned with the target. It is not an independent test of order use and says nothing about natural ImageNet labels or about orders present in DINOv3 representations; it did not test the detection of a present order 3, robustness to partition misalignment in general, any advantage over GELU, or parameter efficiency, and its statements hold in the standardized feature frame. No run failed, so the contrast between recovered and failed models of Section I.3 has no counterpart here.

Table 26: Sparse interactions planted in standardized DINOv3 features: 20 targets per row, each fitted once by a PSNeuron student with $K = 4$, $R = 16$, $G = 8$, and a partition drawn without reference to the target (1,089 parameters at input width 64, 12,353 at 768). Included: runs that finish with validation $R^2 \geq 0.80$, the registered inclusion rule; lowest: the lowest test R^2 among them. Top orders: runs for which the planted orders have exactly the largest mask effects (one order for single-order targets, both for the mixed target); this is the registered event with chance probability $1/3$. ΔR^2_j: test R^2 lost when order j is masked, median over included runs, above 1 when the masked model predicts worse than the mean; absent: the largest absolute effect of masking an order that the target does not contain. GELU: runs in which a GELU network of nearly equal size reaches validation $R^2 \geq 0.8$, a check of learnability only.

width	planted	included	lowest	top orders	ΔR^2_2	ΔR^2_3	ΔR^2_4	absent	GELU
64	order 2	20/20	1.0000	20/20	1.0007	0.0000	0.0000	0.0000	20/20
	order 4	20/20	1.0000	20/20	0.0000	0.0000	1.0002	0.0000	0/20
	orders 2 and 4	20/20	0.9389	20/20	0.4971	0.0000	0.5169	0.0012	0/20
768	order 2	20/20	1.0000	20/20	1.0019	0.0000	0.0000	0.0000	20/20
	order 4	20/20	0.9589	20/20	0.0000	0.0000	1.0006	0.0005	0/20
	orders 2 and 4	20/20	0.8390	20/20	0.4924	0.0000	0.4952	0.0003	0/20

I.5 Factor sharing as an inductive bias

A separate width study asks how sharing one factor bank across orders changes the width needed for recovery. It has its own targets and students: $d = 32$ Rademacher inputs, a scalar output, and $G = K$. Each target plants M families of directions. A planted basis puts one support coordinate in each slot of a partition row of the student's bank, and the order coefficients are solved so that every order from 2 to K carries exactly $1/(K-1)$ of the target variance, with no content at degrees 0 and 1. A *coherent* target lets all orders of a family use one planted basis, an *incoherent* target plants a separate basis for each order, and a *partially coherent* target at $K = 4$ puts orders 2 and 3 on one basis and order 4 on another. The shared geometry is the reported PSNeuron. The independent geometry of Section A.14 splits its R bases among the $K - 1$ orders and gives each order's bank the partition rows of the planted bases that carry that order, so that at a sufficient width both geometries can represent every target; R counts the bases of all orders. Every run trains for 3,000 steps at batch 512 on 8,192 training, 4,096 validation, and 4,096 test rows, with final seeds 0 to 9, each of which also draws the signs of the planted coefficients; calibration uses the first 2,048 training rows. Each geometry uses one learning rate, selected on tuning seeds 101 and 102 by the median held out test R^2 over the coherent and incoherent targets at $K = 6$ with one family and the widths 16 and 128, and then held fixed across all targets and widths: 0.0256 for the shared and 0.1024 for the independent geometry; selection on their validation R^2 gives the same two rates. The independent rate is the upper edge of the tuning grid, and a check at 0.2048 gave a lower median held out R^2 on the tuning seeds. The width of each geometry was bracketed by a pilot on seeds 0 to 5, which reads their held out test R^2 and accepts a width when five of the six recover. A width counts as confirmed when at least 8 of the 10 final seeds recover; the six pilot seeds are among the ten, so the confirmation is not independent of the search.

Sharing saves factor parameters on the coherent targets, costs them on the incoherent one, and on the partially coherent targets saves them with one family of directions but costs them with two (Table 27). With one or two families of directions, sharing can therefore substitute for factor width when the orders of a target reuse directions and becomes restrictive when they do not. The printed widths are recovery widths on the searched grid, not globally minimal widths. Settings without a confirmed width for both geometries are not printed, among them four whose pilot width recovered on only 6 or 7 of the 10 final seeds. The pilots of the settings that are not printed, which are not confirmed, point both ways. At $K = 6$ with one family, the shared geometry reaches the pilot criterion at one basis on the coherent target, against 15 for the independent geometry, but needs 64 bases against 16 on the incoherent target; with two families on the incoherent target it reaches the criterion at no width up to 512, against 45. With four families on the coherent target the two geometries need 14 and 16 bases at $K = 4$, and at $K = 6$ the shared geometry first reaches the criterion at 512 bases, against 181. At $K = 2$ the two geometries coincide and are not compared. This study does not explain the recovery gap at order 8 in the fixed G study.

Table 27: Confirmed recovery widths for factor sharing. M counts the families of planted directions, so a coherent target plants M bases, a partially coherent target at $K = 4$ plants $2M$, and an incoherent target $(K-1)M$, six in the incoherent row. "Factor params" count the learned factor directions, Rd, and total parameters add the readout and, in the shared geometry, the order coefficients. "Recovery" counts final seeds reaching held out $R^2 \geq 0.9$ out of ten; higher is better. The ratio divides the factor parameters of the larger geometry by those of the smaller, so it gives the size of the advantage; "reversal" marks rows where the independent geometry is the smaller one.

target	geometry	R	total params	factor params	recovery	ratio
coherent, $K = 4$, $M = 1$	shared	2	72	64	9/10	$2.00\times$
	independent	4	132	128	8/10	
coherent, $K = 4$, $M = 2$	shared	4	144	128	10/10	$2.25\times$
	independent	9	297	288	10/10	
coherent, $K = 6$, $M = 2$	shared	14	532	448	9/10	$2.14\times$
	independent	30	990	960	8/10	
partially coherent, $K = 4$, $M = 1$	shared	3	108	96	8/10	$1.33\times$
	independent	4	132	128	9/10	
partially coherent, $K = 4$, $M = 2$	shared	14	504	448	9/10	$1.17\times$ reversal
	independent	12	396	384	8/10	
incoherent, $K = 4$, $M = 2$	shared	24	864	768	8/10	$3.43\times$ reversal
	independent	7	231	224	8/10	

I.6 What the controlled evidence does not identify

Recovery depends strongly on the initialization but is not determined by it. In the 72 aligned order 8 runs, a one way decomposition attributes 0.47 of the variance of held out R^2 to the initialization seed, 0.16 to the sign pattern, and 0.004 to the data stream, and 0.42, 0.08, and 0.02 of the variance of the recovery indicator. A separate design of 27 runs on the same target splits the data stream into a sample seed and a batch order seed. In four of its pairs the data, the teacher, and the initialization are identical and only the batch order differs, yet the two runs fall on opposite sides of the recovery criterion, for example $R^2 = 1.0000$ against 0.3039. We therefore stop at the measured distinction between representability, recovery, spectral failure, and functional use and do not propose an optimization mechanism.

J Extended Results and Ablations

This appendix gives the evidence behind Section 5 and the adverse results that bound it, followed by the ablations (Sections J.4 to J.6), the budget studies (Sections J.7 to J.9), and the DINOv3 width ladder (Section J.10).

J.1 Tabular reference comparisons

Every tabular replacement in Table 1 uses one common $K = 4$ architecture (Table 28). The dense reference of each task was chosen on validation before any test evaluation from two evaluated dense families, a GELU and a SwiGLU network at the same site, so the reduction does not rest on a comparator chosen afterwards. Data, splits, and the transductive preprocessing shared by every tabular model are given in Section G.

The selection score is the mean over the five seeds of the validation accuracy at the end of training: 71.884, 72.047, and 72.077% for PSNeuron, GELU, and SwiGLU on HIGGS Small, 69.912, 69.872, and 70.048% on Jannis, and 35.109, 35.529, and 35.396% on Helena. The GELU network is the restored host of Table 28; the SwiGLU network, whose site has $24{,}576$ parameters, is the selected reference on HIGGS Small and Jannis, and on Helena, where GELU was selected, it has $49{,}828$ parameters and $35.91 \pm 0.24\%$ test accuracy. Against GELU, the smaller dense family, the whole model reductions are 40.8% and 38.9% on HIGGS Small and Jannis and 34.7% on Helena, where GELU is also the selected reference; against the selected SwiGLU references on HIGGS Small and Jannis they are 51.7% and 49.7%. All compare whole models with the same input and output. The native host retrained under the protocol of the replacement is a separate ablation comparison (Table 44) and supplies no dense network at the budget of the replacement. Five of these nine models selected 6.4×10^{-3}, the largest rate of the extended grid, so those selections are not bracketed (Section G.2); a separate cohort brackets every relevant rate (Section K), and Figure 8 places these

Table 28: The common $K = 4$ replacement of Table 1(a) on the three tabular tasks, also shown with its trained site folded into a $1{,}048$ parameter operator, against every dense reference: its own host with the GELU network restored (r is the parameter ratio), GELU at the budget of the replacement, SwiGLU with a 544 parameter site, and the reference selected on validation (SwiGLU on HIGGS Small and Jannis, the GELU host itself on Helena). Entries are mean test accuracy over five seeds $\pm$ seed standard deviation and whole model parameters.

	HIGGS Small		Jannis		Helena	
model	acc. (%)	params	acc. (%)	params	acc. (%)	params
PSNeuron $K = 4$, as trained	70.91 ± 0.62	21,018	70.07 ± 0.27	22,812	35.65 ± 0.33	27,324
folded, same predictions	70.91 ± 0.62	19,994	70.07 ± 0.27	21,788	35.65 ± 0.33	26,300
dense host, GELU restored	71.24 ± 0.37	35,522	69.68 ± 0.19	37,316	36.09 ± 0.31	41,828
PSNeuron change; r	-0.33	$r = 0.59$	$+0.39$	$r = 0.61$	-0.44	$r = 0.65$
GELU at the replacement budget	71.02 ± 0.43	21,010	69.85 ± 0.18	22,804	35.95 ± 0.32	27,316
SwiGLU, site of 544 parameters	71.05 ± 0.37	19,490	69.98 ± 0.09	21,284	35.88 ± 0.19	25,796
reference selected on validation	71.27 ± 0.34	43,522	69.99 ± 0.31	45,316	36.09 ± 0.31	41,828

points descriptively among every measured tabular configuration and the dense networks resized to the same budgets.

J.2 Secondary placements and adverse results

Cross head site. Table 29 compares the selected $K = 2$ cross head site with the native output projection on MQAR and CIFAR100. On MQAR the native mean contains one collapsed run, seed 0 at 15.902%, which reproduces when the run is repeated; the other four native runs reach 99.949 to 99.988%, and all five PSNeuron seeds at least 99.953%. This is a pattern over five seeds, not an estimate of how often the native model fails. In the Pareto inventory, fixed at an earlier cutoff, this $K = 2$ model and the native model fall in the descriptive tier and the $K = 4$ cross head model in the strict tier. Because the identity carrier alone reaches 99.96% with no site parameters (Table 30), Tables 2 and 62 do not read the MQAR attention output rows as PSNeuron parameter efficiency evidence and keep their arithmetic for provenance only. On CIFAR100 the replacement is 1.062 points below the native projection (five paired seed differences, standard deviation 0.465, paired $t = -5.11$ with 4 degrees of freedom). The native MQAR model here is not the native model of the attention score study (Table 34): both have $289{,}088$ parameters, but the two hosts split their width into different heads (Section E) and selected their own learning rates, 1.6×10^{-3} and 8×10^{-4}.

Table 29: Cross head replacement of the attention output projection selected on validation, with $K = 2$, on MQAR and CIFAR100 (Krizhevsky, 2009); entries are test means $\pm$ seed standard deviation over five seeds, in percent, higher is better. The PSNeuron site is a replacement on both tasks, but the projection is a small part of each model, so the whole model shrinks only to $r = 0.9624$ on MQAR and 0.9892 on CIFAR100, and the reduction is stated at the site. On MQAR the host with the projection deleted reaches 99.96%, within the registered allowance of one point (Table 30).

	MQAR, recall@1			CIFAR100, accuracy		
model	model params	site params	test, %	model params	site params	test, %
PSNeuron cross head $K = 2$	278,208	5,632	99.971 ± 0.015	2,677,988	7,936	38.288 ± 0.495
native output projection	289,088	16,512	83.154 ± 37.595	2,707,108	37,056	39.350 ± 0.637

Linear controls at the cross head site. Table 30 tests whether the MQAR result of the $K = 2$ site needs the product. Its six models other than the identity carrier are trained together on one GPU type with the grid $\{2, 4, 8, 16, 32\} \times 10^{-4}$ and no extension, tuning seeds 101 and 102, and final seeds 0 to 4; the PSNeuron and native rows are those of Table 29. The linear correction has its own head maps P_h, of the shape and initial distribution of those of the PSNeuron site, keeps its identity carrier, and replaces $P(z)$ by $U(Vz)$ without activation, so the site is affine in the head outputs and cannot multiply two heads. At rank 32 without biases it has the parameter shapes and count of the PSNeuron site, and UV can be any linear map of the 32 latent coordinates; at rank 31, with biases, any map of rank at most 31. Both corrections therefore spend more parameters than one dense map of z would need. The low rank output projection factors the native projection without a carrier. Under the prespecified reading, a control matches PSNeuron when the upper end of the paired 95% interval of their difference in test recall is at most one point. Both linear corrections match, and PSNeuron leads

by more than one point on no seed. Both low rank projections stay below 13% on every seed and at every rate of the grid, near the 12.5% of guessing among the eight stored values. The $K = 2$ cross head site uses a third of the parameters of the $16{,}512$ parameter projection, but on MQAR it shows no advantage of the product, and the carrier alone below explains its size result; the $K = 4$ site was not compared with these controls. The linear corrections also solve the task at more learning rates: like the native model they reach validation recall of at least 99% in eight of ten tuning runs, against five of ten for PSNeuron, and at the rate PSNeuron selected the rank 32 correction first reaches it at steps $1{,}674$ to $1{,}849$ of the final seeds, against $2{,}024$ to $2{,}874$.

The carrier alone. A registered control deletes the output projection of the last block and adds no learned correction, so the heads reach the residual stream through the identity carrier, $y = [o_1; \ldots; o_8]$; the site has no parameters and the whole model $272{,}576$, and the first block keeps its native projection. With the grid, tuning seeds, and final seeds of Table 30, it reaches $99.96 \pm 0.03\%$ test recall, at least 99.92% on every seed, within the registered allowance of one recall point of both the PSNeuron site (PSNeuron minus carrier 0.01 points, nominal interval $[-0.04, 0.06]$) and the rank 32 linear correction (0.02 points, $[-0.02, 0.07]$). The verdict rests on the carrier's own final seeds alone, a mean 0.96 points above the registered 99% bar, so it does not depend on whether the selected rate is bracketed; that rate is the smallest of the grid, 2×10^{-4}, and won on validation by 1.02×10^{-3} recall points over the tied 4×10^{-4} and 8×10^{-4}. The PSNeuron, native, and rank 32 models retrained beside it reproduce their earlier values bit for bit. With five seeds, exact paired tests cannot go below $p = 0.0625$, so these pairs are read only as lying within the allowance, not as equivalent. In this host and task the $16{,}512$ parameter output projection of the last block is therefore unnecessary: deleting it already solves MQAR, so the compact cross head result shows that the projection can be dropped, not that the PSNeuron site or the linear correction is what solves the task. The control was run for one host, one task, and five seeds, not at the CIFAR100 site or for the $K = 4$ site.

Table 30: Controls at the MQAR cross head site: test recall@1 in percent, mean $\pm$ seed standard deviation over final seeds 0 to 4, higher is better. All rows share one learning rate grid; the whole models hold $278{,}080$ to $278{,}336$ parameters, the native model $289{,}088$, and the model with the identity carrier only $272{,}576$, trained in a later registered study (Section J.2). $o = [o_1; \ldots; o_8]$ are the head outputs, $z_h = P_h o_h$ the head latents of the PSNeuron site, and U, V learned factors. Difference: PSNeuron minus the row, mean over paired seeds with its nominal 95% t interval, in points. †The smallest rate of the grid; lower rates were not tried.

model	site map	site params	rate	recall	lowest seed	difference
PSNeuron cross head, $K = 2$	$o + P(z)$	5,632	4×10^{-4}	99.97 ± 0.01	99.95	
native projection	$Wo + b$	16,512	1.6×10^{-3}	83.15 ± 37.60	15.90	$16.82\ [-29.86, 63.50]$
identity carrier only	o	0	$2 \times 10^{-4\dagger}$	99.96 ± 0.03	99.92	$0.01\ [-0.04, 0.06]$
linear correction, rank 32	$o + UVz$	5,632	4×10^{-4}	99.98 ± 0.01	99.97	$-0.01\ [-0.04, 0.01]$
linear correction, rank 31, biases	$o + U(Vz + c) + b$	5,631	1.6×10^{-3}	99.98 ± 0.01	99.97	$-0.01\ [-0.04, 0.02]$
low rank projection, rank 21	$UVo + b$	5,504	$2 \times 10^{-4\dagger}$	12.56 ± 0.13	12.43	$87.41\ [87.23, 87.58]$
low rank projection, rank 22	$UVo + b$	5,760	$2 \times 10^{-4\dagger}$	12.25 ± 0.30	11.95	$87.72\ [87.34, 88.10]$

CIFAR100 cross head capacity, slot assignment, and masks. Two compact operating points of the cross head site, $K = 4$ and $K = 6$ with 995 and 997 site parameters, stay below the native projection of Table 29 (Table 31); they extend the measured range of this site and are separate from the feedforward capacity study of Table 49. Two configurations with $K = 8$ are adverse: with 999 site parameters ($2{,}671{,}051$ in the model) accuracy stays at chance, 1.00% on all five seeds, and the model with $8{,}160$ site parameters trained at a learning rate of 8.5×10^{-4} reaches only 29.687% (Table 31).

In the per head form, slot g holds exactly the latent coordinates of head g, so every product term of order j multiplies j different heads; the free slot form also admits products within a head. At $K = 4$ free slots are higher on all three seeds, by $+1.297$ points on average (Table 31), so these runs give no support for restricting the products to distinct heads on CIFAR100; at $K = 8$ the two forms are indistinguishable ($+0.180$ points). Both comparisons are local evidence from three seeds, not a statement about partitions in general.

Order masks at evaluation ask how much one trained model relies on one order (Table 31). Only two masks cost more than half a point, order 2 of the $K = 2$ model, which removes its whole branch and leaves the identity carrier, and order 8 of a $K = 8$ model; the latter reaches only 29.687% intact,

so its dependence on order 8 says nothing about whether a good cross head model needs that order. Nine further masks, of orders 2 and 4 at $K = 4$, orders 2, 4, and 6 at $K = 6$, and orders 2, 4, 6, and 7 at $K = 8$, are scored on the validation split and pooled in Table 46; together the two sets cover every order from 2 to K, although on different models and splits.

Table 31: CIFAR100 cross head site; accuracy is held out test mean ± seed standard deviation. Upper block: two compact capacity points with five seeds, and the free slot retraining. In the free slot rows a balanced partition drawn from the seed replaces the assignment of one slot per head; each is a three seed training paired with the per head model of the same seeds, learning rate, and data split, and the rate is the one the per head model selected, interior to its grid at $K = 4$ and at the lower edge at $K = 8$. Δ is free slot minus per head accuracy in percentage points, with the observed paired range, not a confidence interval. Lower block: order masks at evaluation, three seeds per row. Masked and intact accuracies are test accuracies of the same trained model evaluated one after the other; nothing is retrained or changed. The effect is intact minus masked accuracy in percentage points, with the observed range over the seeds. Rows with the same K and learning rate mask identically trained models and share their intact scores. At $K = 8$ the masks of orders 3 and 5 use the per head model of the upper block, while order 8 is masked in a separate model trained at 8.5×10^{-4}. The $K = 2$ model has the architecture and count of the cross head row of Table 29 but is a separate three seed training at 4×10^{-4} instead of 2×10^{-4}.

comparison	setting	model and site params	learning rate	test accuracy, %	paired Δ, points
capacity	$K = 4$	2,671,047 and 995	2×10^{-4}	38.506 ± 0.680	n/a
capacity	$K = 6$	2,671,049 and 997	8×10^{-4}	30.434 ± 1.545	n/a
free slots	$K = 4$	2,678,084 and 8,032	2×10^{-4}	39.000 ± 0.720 against 37.703 ± 0.609	$+1.297$ [0.480, 2.830]
free slots	$K = 8$	2,678,212 and 8,160	10^{-4}	37.777 ± 0.576 against 37.597 ± 0.337	$+0.180$ [-0.320, 0.580]

K	masked order	learning rate	masked accuracy, %	intact accuracy, %	effect, points [range]
2	2	4×10^{-4}	36.377 ± 0.327	38.697 ± 0.417	$+2.320$ [1.870, 3.150]
4	3	2×10^{-4}	37.400 ± 0.641	37.703 ± 0.609	$+0.303$ [0.270, 0.340]
6	3	10^{-4}	36.903 ± 0.565	37.277 ± 0.647	$+0.373$ [0.050, 0.560]
6	5	10^{-4}	36.940 ± 0.625	37.277 ± 0.647	$+0.337$ [0.020, 0.530]
8	3	10^{-4}	37.473 ± 0.273	37.597 ± 0.337	$+0.123$ [-0.060, 0.260]
8	5	10^{-4}	37.530 ± 0.349	37.597 ± 0.337	$+0.067$ [0.050, 0.090]
8	8	8.5×10^{-4}	3.167 ± 1.140	29.687 ± 2.044	$+26.520$ [24.710, 27.610]

Sequence families. Tables 32 and 33 give every configuration and every seed behind the sequence rows of Table 1. The native DeltaNet searched a grid extended to 6.4×10^{-3} on MQAR and multi-needle retrieval and to 10^{-4} on state tracking, while all eight PSNeuron mixers on MQAR and seven of eight on each other task used the base grid of five rates, so these comparisons are descriptive (Section G). On MQAR the native model selected 3.2×10^{-3}, and the joint mixers with $K = 2$, 4, and 8 and the cross term mixers with $K = 4$, 6, and 8 selected 8×10^{-4}, the joint $K = 6$ mixer 4×10^{-4}, and the cross term $K = 2$ mixer 1.6×10^{-3}; on multi-needle retrieval the cross term mixer with $K = 4$ selected 3.2×10^{-3}. On multi-needle retrieval only the cross term mixer with $K = 6$, which has one seed at 70.2%, falls below the native mean. On MQAR only the joint $K = 4$ mixer exceeds it; the mixer and the native model are both bimodal across seeds, with 3 and 2 of five seeds above 90%. On state tracking the four cross term mixers and the joint $K = 6$ mixer have higher means and three to four times smaller seed spread, because none of their seeds stalls like the native seed at 34.08%; the native median of 96.48% is still higher than theirs, except for the cross term mixer with $K = 2$ at 96.78%.

Dense mixers that read the state. Dense GELU and SwiGLU mixers in the slot of the PSNeuron mixer read the same pair $[r_t; u_t]$ (Section E), as does an additive control that splits the raw pair into eight blocks of its own, with twelve bases and no conditioners, and sums a GELU of each of its slots. All are trained with the same seeds, base grid, and extension rule, evaluating the delta rule chunkwise, as does the joint $K = 4$ row of the selected layout on MQAR (Section G). On MQAR the GELU mixers with 4,363 and 4,624 parameters reach 42.38 ± 47.87 and $60.92 \pm 49.34\%$, and the SwiGLU mixers with 4,288 and 4,560 parameters 25.79 ± 39.02 and $58.86 \pm 48.18\%$. All four selected 3.2×10^{-3} after 6.4×10^{-3} was evaluated and found worse on validation, whereas the joint $K = 4$ mixer ($77.98 \pm 23.87\%$) selected 8×10^{-4} inside the base grid. It leads each in mean, but because the grids differ these pairs are descriptive, and the lead rests on one configuration: a joint $K = 4$ mixer

Table 32: Every configuration of the two sequence families behind the sequence rows of Table 1, five seeds each: mean test recall (%) ± seed standard deviation. The DeltaNet rows replace the channel mixer of the last block, the attention rows the output projection of the last block; site parameters are those of the replaced module. * marks a run that evaluated the delta rule chunkwise while the native model of the same task evaluated the same recurrence one step at a time, which changes only the order of floating point operations; on MQAR the native model and every mixer evaluated it chunkwise.

DeltaNet mixer	site parameters	MQAR	multi-needle	state tracking
native model	130,944	43.40 ± 48.82	98.42 ± 1.64	84.12 ± 28.04
cross terms, $K = 2$	4,352	42.35 ± 48.79	98.69 ± 1.07	92.54 ± 8.45
cross terms, $K = 4$	4,400	30.12 ± 33.06	$99.82 \pm 0.10^*$	$89.90 \pm 7.07^*$
cross terms, $K = 6$	4,432	23.71 ± 37.86	93.66 ± 13.09	91.29 ± 7.35
cross terms, $K = 8$	4,464	9.28 ± 8.18	99.32 ± 0.45	$88.93 \pm 7.27^*$
joint, $K = 2$	4,352	24.38 ± 41.51	99.26 ± 0.70	$59.18 \pm 49.09^*$
joint, $K = 4$	4,400	77.98 ± 23.87	99.61 ± 0.17	80.18 ± 19.60
joint, $K = 6$	4,432	2.29 ± 0.28	$98.93 \pm 0.76^*$	88.59 ± 7.25
joint, $K = 8$	4,464	6.56 ± 4.76	99.28 ± 0.61	$77.70 \pm 25.85^*$
attention output	site parameters	MQAR		
native model	16,512	83.15 ± 37.60		
cross head, $K = 2$	5,632	99.97 ± 0.01		
cross head, $K = 4$	5,728	99.97 ± 0.02		
cross head, $K = 6$	5,792	96.21 ± 8.45		
cross head, $K = 8$	5,856	96.32 ± 5.02		

Table 33: Test recall (%) of each final seed for the sequence rows of Table 1. Site parameters are those of the replaced module in the last block.

host, task	model	site parameters	seed 0	seed 1	seed 2	seed 3	seed 4	mean
attention, MQAR	native	16,512	15.90	99.96	99.95	99.99	99.98	83.15
	PSNeuron, $K = 2$	5,632	99.97	99.99	99.96	99.99	99.95	99.97
DeltaNet, MQAR	native	130,944	7.59	4.95	98.35	10.81	95.30	43.40
	PSNeuron, $K = 4$	4,400	96.90	92.80	95.28	45.38	59.56	77.98
DeltaNet, multi-needle	native	130,944	99.71	95.85	99.41	97.71	99.41	98.42
	PSNeuron, $K = 4$	4,400	99.39	99.78	99.61	99.78	99.51	99.61
DeltaNet, state tracking	native	130,944	96.48	98.83	93.46	97.75	34.08	84.12
	PSNeuron, $K = 4$	4,400	91.89	45.90	81.84	89.26	91.99	80.18

with $4{,}694$ parameters reaches $61.83 \pm 40.08\%$, level with the larger dense mixers, and joint mixers with $K = 2$, 6, and 8 at $4{,}313$ to $4{,}730$ parameters reach 2.29 to 38.97%. The additive control, with $4{,}608$ parameters and its learning rate inside the grid, stalls on all five MQAR seeds at $7.66 \pm 0.75\%$. Against it the joint $K = 4$ mixer is a strict pair of the inventory, smaller by 208 parameters and ahead by 70.3 points, with a nominal paired 95% interval of 41.0 to 99.6 points; a Bonferroni correction over the 504 comparisons of its stratum, which includes the capacity rungs below, lowers the lower end to -94.2 points, and a capacity rung that repeats this mixer exactly counts as a second pair in Table 2. On multi-needle retrieval the SwiGLU mixer with $4{,}288$ parameters reaches $99.57 \pm 0.45\%$, the GELU mixer with $4{,}624$ parameters $99.10 \pm 0.87\%$, and the additive control $98.69 \pm 1.53\%$ at the largest rate evaluated, level with the joint $K = 4$ mixer at $99.61 \pm 0.17\%$. Retrained chunkwise, that mixer selects 1.6×10^{-3} and reaches $99.16 \pm 0.92\%$, with paired intervals of $[-1.9, 1.1]$ and $[-1.8, 1.9]$ points against the SwiGLU mixer and the larger GELU mixer. The GELU mixer with $4{,}363$ parameters has two seeds at 75.56 and 72.19% ($89.42 \pm 14.24\%$). A separate exploratory state tracking control selected on validation a same input SwiGLU mixer with $4{,}288$ site parameters over a GELU mixer (66.76%); its mean of $80.04 \pm 10.74\%$ stays below the native model and the cross term mixers of Table 32, whose kernels differ in part, and with five seeds it is descriptive and not pooled with the registered same input study, which finds no resolved advantage for either side on state tracking (Section K). Together, the dense mixers stay unresolved against the product on multi-needle retrieval and establish only a higher PSNeuron point estimate on MQAR, where the product exceeds an additive control on the same inputs with a nominal interval. These are seeds 0 to 4; on 20 fresh seeds the joint $K = 4$ mixer exceeds an exact additive twin on its own slots and the conditional GELU control after Holm correction, and this additive control only at the nominal level (Section K.2).

Capacity rungs of the recurrent mixers. Each PSNeuron order and the GELU mixer are also trained at twelve sizes, the SwiGLU mixer at twelve on multi-needle retrieval and eleven on MQAR,

and the additive mixer at eleven, from about 4,300 to 131,000 site parameters, with five final seeds (Table 62). The PSNeuron rungs are all in the selected joint layout. On MQAR they select from a learning rate grid up to 3.2×10^{-3}, and the native model, every SwiGLU mixer, and 10 of the 12 GELU mixers from one up to 6.4×10^{-3}, so a strict pair there joins a PSNeuron rung only with controls of its own grid, mostly additive mixers, and every comparison with the SwiGLU and native mixers below is descriptive. On multi-needle retrieval the grid reaches 6.4×10^{-3} for the native model, 15 of the 48 PSNeuron rungs, 3 of the 12 GELU mixers, 4 of the 12 SwiGLU mixers, and 7 of the 11 additive mixers, and strict pairs join configurations of the same grid. On MQAR the rungs reach 4.78 to 87.23% at $K = 2$, 28.66 to 77.98% at $K = 4$, 2.29 to 73.17% at $K = 6$, and 3.90 to 20.64% at $K = 8$, against 8.05 to 79.65% for SwiGLU, 7.73 to 77.36% for GELU, 7.19 to 25.30% for the additive mixer, and 43.40% for the native model; stalled seeds occur in every family. Against all of them nine distinct PSNeuron measurements are non dominated: three $K = 2$ rungs with 6,656, 8,096, and 16,304 site parameters (746,436, 747,876, and 756,084 in the model) reach 81.60 ± 21.94, 87.23 ± 23.62, and 83.08 ± 31.35%, a mean that no control configuration reaches (the best is a SwiGLU mixer with 32,640 site parameters at 79.65 ± 39.35%), and $K = 2$ rungs with 5,360 and 10,832 site parameters (78.46 and 78.86%), $K = 4$ mixers with 4,400, 4,694, and 8,075 (77.98, 61.83, and 62.83%), and the cross term $K = 2$ mixer with 4,352 (42.35%) complete the set. The $K = 6$ rung with 21,716 site parameters reaches 73.17%, above the native model, but a smaller SwiGLU mixer reaches 73.41%. On multi-needle retrieval the rungs reach 94.71 to 99.84% and the dense mixers 89.08 to 99.84%; $K = 2$, 4, and 6 rungs with 4,640 to 6,458 site parameters (99.60 to 99.81%) and the cross term $K = 4$ mixer (99.82%) are non dominated, by less than 0.3 points and 1% of the model's parameters against the nearest dense mixers, and the six mixers that evaluate the delta rule one step at a time are non dominated only because the larger native model is their sole comparator. No nominal positive interval of a strict pair survives correction within its stratum (Table 62), and these frontiers are descriptive and lie outside the jointly corrected family of Section L.

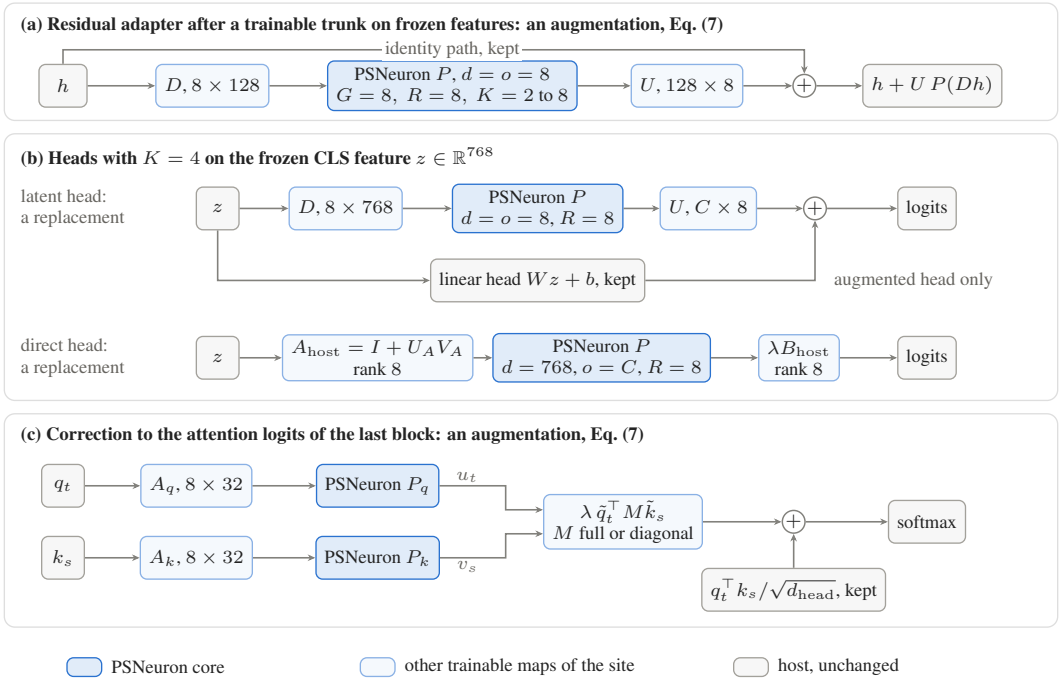

Figure 5: Placements used for attribution and on frozen features. (a) The residual adapter keeps the identity path and adds $U\,P(Dh)$; the standard adapter has the same maps with a GELU in place of P. (b) The latent head $U\,P(Dz)$ and the direct head $\lambda B_{\text{host}} P(A_{\text{host}} z)$ replace the linear head, and the augmented head adds the latent branch to the kept linear head $Wz + b$. (c) The correction is computed token by token, the four heads of dimension 32 share the parameters of one correction, each applying it to its own queries and keys, and $\lambda = 0$ returns the native scores. C is the number of classes; trainable parameter counts are in Tables 34, 35, and 38.

Corrections to the attention scores. Every PSNeuron correction $\lambda\, \tilde{q}_t^\top M \tilde{k}_s$ of the attention logits (Figure 5(c)) lowers MQAR recall, while all three bilinear corrections keep it (Table 34). With the full kernel, the five seeds at $K = 2$ split into two near 99.9% and three near 12.7%; the diagonal form chosen on validation solves the task on one or two of five seeds at $K = 2$ to 6 and on none at $K = 8$. On multi-needle retrieval every correction stays below the native model, the PSNeuron corrections by 1.4 to 12.9 points, so on that task the two kinds of correction overlap.

Table 34: Corrections $\lambda\, \tilde{q}_t^\top M \tilde{k}_s$ to the attention logits of the last block, with $\tilde{q}_t = P_q(A_q q_t)$ and $\tilde{k}_s = P_k(A_k k_s)$ (Section E); test recall@1 as mean $\pm$ seed standard deviation over five seeds, in percent, higher is better. The first PSNeuron block uses a full 8×8 kernel M. Selection on validation chose a diagonal M; the selected block rebuilds that form to the budget of the rank 16 bilinear correction by giving each of P_q and P_k 16, 13, 12, and 11 bases at $K = 2$, 4, 6, and 8, while the masks of Table 46 and the kernel comparisons of Table 47 use it at its own size (777 to 889 correction parameters). A bilinear correction adds a learned bilinear form of rank 16, 8, or 4 to the same logits. Multi-needle is a different retrieval task on the same host, not MQAR with several queries (Section G); it has no keys, so it measures retrieval by position rather than by association. Every correction is added to the preserved scaled dot product, so each corrected model has more parameters than the native model ($r = 1.0029$ to 1.0033 for the full kernel) and supports attribution, not a size claim; $\lambda = 0$ returns that product exactly.

correction		model params	correction params	MQAR	multi-needle
none (native)		289,088	0	99.941 ± 0.060	97.412 ± 0.400
PSNeuron, full M	$K = 2$	289,921	833	47.520 ± 47.832	95.410 ± 1.545
	$K = 4$	289,969	881	12.802 ± 1.161	93.984 ± 4.192
	$K = 6$	290,001	913	12.635 ± 0.477	84.507 ± 15.509
	$K = 8$	290,033	945	14.478 ± 2.885	88.843 ± 10.217
PSNeuron, diagonal M, selected	$K = 2$	290,121	1,033	50.543 ± 45.196	96.016 ± 1.557
	$K = 4$	290,103	1,015	29.891 ± 39.157	95.654 ± 0.447
	$K = 6$	290,113	1,025	30.695 ± 38.661	91.128 ± 10.890
	$K = 8$	290,115	1,027	14.545 ± 2.250	93.584 ± 2.168
bilinear	rank 16	290,113	1,025	99.974 ± 0.019	96.997 ± 1.371
	rank 8	289,601	513	99.967 ± 0.029	94.131 ± 6.064
	rank 4	289,345	257	99.986 ± 0.011	95.435 ± 1.335

Heads with $K = 4$ on frozen features. The augmented head of Table 35 keeps the linear head, so it is larger by construction; it retains 99.8% and 99.6% of the linear head's accuracy on ImageNet100 and EuroSAT, which measures what the branch adds at equal or larger capacity, not a reduction. Only the two replacement heads can be smaller. Averaged over the two benchmarks, validation accuracy relative to the linear head was 94.0% for the direct head and 92.6% for the latent head, and this ratio selected the direct form. On the test split the averages are 93.5% and 92.0%, and they hide a split verdict: on EuroSAT, where the direct head is larger than the linear head, the two heads retain 98.9% and 98.6% of its accuracy, but on ImageNet100 only 88.2% and 85.3%, so the latent head, 10.8 times smaller than the linear head, trails it by about 14 points on the harder benchmark.

Table 35: Heads with $K = 4$ on frozen DINOv3 features (Figure 5(b)). The linear head is the reference and r is each head's trainable count divided by the linear head's; every count is exact. The augmented head keeps the linear head and adds the latent branch $U\,P(Dz)$ beside it. The two replacement heads drop the linear head: the latent head is $U\,P(Dz)$, and the direct head is $\lambda B_{\text{host}} P(A_{\text{host}} z)$ with a rank 8 conditioner and readout. Accuracy is held out test mean $\pm$ seed standard deviation over five seeds.

	ImageNet100			EuroSAT		
head	parameters	r	accuracy, %	parameters	r	accuracy, %
linear	76,900	1	95.440 ± 0.105	7,690	1	97.207 ± 0.048
linear plus latent branch	83,996	1.0923	95.240 ± 0.170	14,066	1.8291	96.811 ± 0.146
direct replacement	20,857	0.2712	84.148 ± 0.589	18,697	2.4313	96.096 ± 0.303
latent replacement	7,096	0.0923	81.452 ± 0.651	6,376	0.8291	95.878 ± 0.228

Wider replacement heads on frozen features. The three seed ladder of Section J.10 stops at $h = 8$, and the separate five seed study reaches $h = 128$ in its registered extension (Section J.10). The frozen head study also trained direct replacement heads at twelve larger budgets (Table 36; protocol

in Section G); their runtime against GELU heads of matched size enters Table 18. At every budget the PSNeuron head is below the GELU head, by 2.42 to 4.94 points on ImageNet100 and by less than 0.01 to 1.14 points on EuroSAT, where the two heads round to the same accuracy at 598,263 parameters, and no PSNeuron head reaches the linear head. Together with the ladder at $h = 8$, the five seed study, in which the GELU head has the higher mean on ImageNet100 at every width and on DTD and Pets at $h = 4$, and its wider heads, which trail the GELU head in all 25 settings from $h = 8$ to 128 and the additive head in 21, with four unresolved, this confines any benefit of PSNeuron heads on frozen features to narrow widths and, against GELU heads, to some datasets.

Table 36: Replacement heads on frozen DINOv3 features at larger budgets, five seeds each. At every budget rung the PSNeuron head is the order $K \in \{2, 4, 6, 8\}$ with the best validation accuracy, and the GELU and SwiGLU heads are the dense heads built for the same rung, also selected on validation, with parameter counts within 2.2% of the PSNeuron head. Entries are mean test accuracy (%); rung n has about $1/n$ of the largest budget. The linear head reaches 95.44% with 76,900 parameters on ImageNet100 and 97.21% with 7,690 on EuroSAT.

	ImageNet100					EuroSAT				
rung	parameters	K	PSNeuron	GELU	SwiGLU	parameters	K	PSNeuron	GELU	SwiGLU
1	1,333,917	2	92.98	95.40	95.30	1,195,767	2	96.93	97.12	97.00
2	667,293	2	92.70	95.24	95.21	598,263	2	96.89	96.89	97.00
4	333,581	2	92.58	95.22	94.86	298,733	2	96.79	97.06	97.09
6	222,395	8	90.52	95.35	94.83	199,249	2	96.22	97.24	97.13
8	166,657	2	92.80	95.36	94.82	149,377	2	96.61	97.28	96.94
10	133,461	4	91.78	95.29	94.67	119,653	2	96.50	97.21	96.94
12	111,105	2	92.58	95.16	94.66	99,585	2	96.65	97.19	97.00
16	83,425	4	92.14	95.22	94.31	74,689	2	96.59	97.21	97.08
20	66,241	2	91.63	94.99	94.14	59,518	8	96.02	97.17	96.87
24	55,553	2	89.93	94.87	94.03	49,793	2	96.40	97.22	97.03
28	47,498	6	89.74	94.66	93.88	42,687	4	96.26	97.24	97.02
30	44,913	4	90.57	94.76	93.75	39,381	2	96.25	97.23	96.96

Residual adapters. The standard residual adapter is smaller and more accurate than every PSNeuron adapter on both DINOv3 benchmarks (Table 38): at $K = 2, 4, 6, 8$ the PSNeuron adapters change mean accuracy by -0.060, -0.160, -0.156, and -0.268 points on ImageNet100 and by -0.230, -0.422, -0.107, and -0.333 points on EuroSAT. On IMDb sentiment classification (Maas et al., 2011), with frozen sentence embeddings of width 768 from a pretrained MPNet encoder of the Sentence Transformers family (Reimers & Gurevych, 2019) and the official split of 25,000 training and 25,000 test reviews, with 5,000 of the training reviews held out for validation, the standard adapter reaches 86.53%, and the same PSNeuron adapters change it by -0.12, $+0.22$, $+0.05$, and $+0.48$ points; the paired intervals at $K = 2$, 4, and 6 include zero, and at $K = 8$, whose learning rate grid was extended, the comparison is descriptive and has no interval. These adapters keep the identity path and make no size claim.

Language model. In the single-block FineWeb-Edu placement, each of the 21 complete PSNeuron feedforward configurations has a higher held out NLL than the SwiGLU network rebuilt at its budget (Table 37). The gap is smallest for $K = 2$ at every budget, and even the best PSNeuron configuration (5.0585 nats, $K = 2$ with 3,328 site parameters) is worse than the worst SwiGLU network (5.0527 nats).

All-block FineWeb-Edu replacement. A separate six-layer width-256 TransformerHost replaces all six feedforward networks with independently parameterized copies of the FFN-role PSNeuron realization (Table 39 and Section E); its completed $K = 2$ and $K = 4$ (two capacity rungs) and $K = 6$ configurations each have five final seeds. At the $K = 4$ smaller rung, 6,315,102 model-trainable parameters reach NLL 5.2631 and mean perplexity 193.09, against 5.3081 and 202.10 for a 9,065,472-parameter SwiGLU reference (a 30.34% saving). A compatible SwiGLU point has only 6,314,496 parameters and better NLL (5.0196); thus neither a family-wide frontier advantage nor superiority at matched size follows. Perplexity and token accuracy are secondary metrics of models selected on NLL; five-seed means of perplexity are not reconstructed from mean NLL. The $K = 8$ rungs are now complete, from 5.2631 to 5.3213 NLL across the four capacity rungs (Table 39).

The single-block adverse results and the all-block favorable comparison to the larger named reference concern separate placements. Attention logits, residual adapters, and wider frozen heads remain

Table 37: FineWeb-Edu language model with the feedforward network of its second block replaced (Section E; data and protocol in Section G). Each group of rows shares one budget: a SwiGLU network rebuilt with the given number of site parameters, and PSNeuron sites of orders 2, 4, 6, and 8 built to within 1.3% of it, except order 8 at the budget of $24{,}576$, whose site has $25{,}024$ parameters. Entries are held out test NLL in nats, lower is better, as mean $\pm$ seed standard deviation over five seeds; the learning rate selected on validation is 4×10^{-4} for every row. The last column is the PSNeuron mean minus the SwiGLU mean of the same budget. The table lists every configuration with five completed seeds; orders 4, 6, and 8 at the smallest budget have none.

configuration	model params	site params	test NLL	difference
SwiGLU	4,573,056	98,304	5.04160 ± 0.00212	n/a
PSNeuron $K = 2$	4,573,056	98,304	5.06423 ± 0.00143	$+0.023$
PSNeuron $K = 4$	4,572,811	98,059	5.08950 ± 0.00612	$+0.048$
PSNeuron $K = 6$	4,572,974	98,222	5.10520 ± 0.00386	$+0.064$
PSNeuron $K = 8$	4,572,876	98,124	5.11333 ± 0.00897	$+0.072$
SwiGLU	4,523,904	49,152	5.04560 ± 0.00361	n/a
PSNeuron $K = 2$	4,523,904	49,152	5.06610 ± 0.00089	$+0.020$
PSNeuron $K = 4$	4,523,857	49,105	5.08902 ± 0.00462	$+0.043$
PSNeuron $K = 6$	4,523,650	48,898	5.09876 ± 0.00633	$+0.053$
PSNeuron $K = 8$	4,523,695	48,943	5.09924 ± 0.00414	$+0.054$
SwiGLU	4,499,328	24,576	5.04752 ± 0.00454	n/a
PSNeuron $K = 2$	4,499,328	24,576	5.07015 ± 0.00600	$+0.023$
PSNeuron $K = 4$	4,499,520	24,768	5.08294 ± 0.00510	$+0.035$
PSNeuron $K = 6$	4,499,648	24,896	5.08689 ± 0.00443	$+0.039$
PSNeuron $K = 8$	4,499,776	25,024	5.08647 ± 0.00155	$+0.039$
SwiGLU	4,487,040	12,288	5.05178 ± 0.00618	n/a
PSNeuron $K = 2$	4,487,040	12,288	5.06596 ± 0.00428	$+0.014$
PSNeuron $K = 4$	4,487,088	12,336	5.07433 ± 0.00525	$+0.023$
PSNeuron $K = 6$	4,487,120	12,368	5.07649 ± 0.00391	$+0.025$
PSNeuron $K = 8$	4,487,152	12,400	5.07470 ± 0.00387	$+0.023$
SwiGLU	4,480,896	6,144	5.05233 ± 0.00362	n/a
PSNeuron $K = 2$	4,480,896	6,144	5.06060 ± 0.00487	$+0.008$
PSNeuron $K = 4$	4,480,920	6,168	5.06499 ± 0.00418	$+0.013$
PSNeuron $K = 6$	4,480,936	6,184	5.06722 ± 0.00498	$+0.015$
PSNeuron $K = 8$	4,480,952	6,200	5.06457 ± 0.00360	$+0.012$
SwiGLU	4,478,094	3,342	5.05266 ± 0.00509	n/a
PSNeuron $K = 2$	4,478,080	3,328	5.05850 ± 0.00487	$+0.006$

Table 38: Residual adapters on frozen DINOv3 features. Each adapter sits after the trunk of a trainable MLP of width 128: the PSNeuron adapter is $h + U\,P(Dh)$, and the standard adapter has the same geometry with a GELU in place of P (Section E). Accuracy is held out test mean $\pm$ seed standard deviation over five seeds, higher is better. Counts are trainable parameters of the model, which exclude the frozen backbone, and of the adapter. Every PSNeuron adapter keeps the identity path and has 128 to 184 more parameters than the standard one, so no row is a size result.

task	adapter	model params	adapter params	accuracy, %
ImageNet100	standard	245,988	2,048	95.056 ± 0.118
	PSNeuron $K = 2$	246,116	2,176	94.996 ± 0.061
	PSNeuron $K = 4$	246,140	2,200	94.896 ± 0.080
	PSNeuron $K = 6$	246,156	2,216	94.900 ± 0.058
	PSNeuron $K = 8$	246,172	2,232	94.788 ± 0.119
EuroSAT	standard	234,378	2,048	97.048 ± 0.321
	PSNeuron $K = 2$	234,506	2,176	96.819 ± 0.167
	PSNeuron $K = 4$	234,530	2,200	96.626 ± 0.252
	PSNeuron $K = 6$	234,546	2,216	96.941 ± 0.222
	PSNeuron $K = 8$	234,562	2,232	96.715 ± 0.166

Table 39: FineWeb-Edu six-layer width-256 transformer host, feedforward networks replaced independently in all six blocks. All rows have five final seeds; P counts the summed replacement stack and whole-model trainable parameters, with no external frozen backbone. NLL is the primary held-out native metric (lower is better); perplexity (lower) and token accuracy (higher) are separately averaged secondary metrics where recorded, and – denotes an unavailable summary, not zero. The pe01/02/04/08 tags identify archived capacity rungs; comparisons use the same all-block placement. These final test means are descriptive across a tested capacity set.

method, rung	site P	model P	NLL	perplexity	token acc. %	seeds
PSNeuron $K = 2$, pe01	3,142,656	9,065,472	5.3046	201.29	23.44	5
PSNeuron $K = 2$, pe02	1,597,440	7,520,256	5.3472	–	22.76	5
PSNeuron $K = 4$, pe01	3,142,386	9,065,202	5.3352	–	22.89	5
PSNeuron $K = 4$, pe08	392,286	6,315,102	5.2631	193.09	23.25	5
PSNeuron $K = 6$, pe08	390,612	6,313,428	5.2709	194.62	23.19	5
PSNeuron $K = 8$, pe08	392,028	6,314,844	5.2631	193.11	23.20	5
PSNeuron $K = 8$, pe04	784,392	6,707,208	5.2856	197.47	23.14	5
PSNeuron $K = 8$, pe02	1,569,120	7,491,936	5.2910	198.56	23.15	5
PSNeuron $K = 8$, pe01	3,141,690	9,064,506	5.3213	204.67	22.95	5
SwiGLU, pe08	391,680	6,314,496	5.0196	–	–	5
SwiGLU, pe04	783,360	6,706,176	5.0284	–	–	5
SwiGLU, pe02	1,571,328	7,494,144	5.0761	–	–	5
SwiGLU, pe01	3,142,656	9,065,472	5.3081	202.10	22.18	5

adverse; on MQAR the host with the projection deleted comes within the registered allowance of one point of the cross-head site (Table 30).

Depth-wise order allocation on HIGGS Small. With two successive deep-block sites, a uniform $4 \to 4$ schedule allows orders up to 4 at both sites, while progressive $2 \to 4$ allocates order 2 to the first site and order 4 to the second. In the same three-seed study, the progressive configuration reaches 70.77751% with $4{,}188$ whole-model trainable parameters, versus 70.20428% at $4{,}212$ for uniform order, an observed gain of 0.57323 points with 24 fewer parameters (Table 40). Secondary means favor the progressive schedule in balanced accuracy (0.705734 versus 0.700588), macro F1 (0.705841 versus 0.700553), ROC AUC (0.782944 versus 0.776564), loss and NLL (0.559850 versus 0.566779), Brier (0.379814 versus 0.385333), and calibration error (0.016241 versus 0.020653); these are correlated secondary observations, not separate independently selected wins. The higher progressive $4 \to 6$ schedule reaches only 69.81789% at $4{,}228$ parameters. Tuning records for the first two schedules are incomplete and no eligible paired interval supports an inferential gain; the result identifies a favorable measured allocation, not a general rule for increasing order across depth.

Table 40: Depth-wise order allocation on HIGGS Small. A schedule a to b uses maximum order a in the first block and b in the second; both sites and the full trainable model are counted, with no frozen parameters. Accuracy is mean held-out test percent across the same three seed labels. The 2 to 4 and uniform 4 to 4 rows have archived incomplete tuning information and are descriptive; differences are not inferential.

order schedule	site P	model P	accuracy %	seeds	tier
progressive 2 to 4	2,074	4,188	70.78	3	descriptive
uniform 4 to 4	2,098	4,212	70.20	3	descriptive
progressive 4 to 6	2,114	4,228	69.82	3	eligible

J.3 Partition admissibility in the Boolean cube probes

The exact order probes on the Boolean cube (Section 5.2) ask whether matching K to the degree of a target is sufficient. We call a partition admissible for a target when it places the coordinates of the target's support in distinct blocks. In the two XOR probes on $d = 16$ coordinates, an inadmissible partition confounds that test, because the target then lies outside the span of the realized polynomial scores. Among seeds 0 to 4, the partition is inadmissible for every seed of the six coordinate task and for three seeds of the four coordinate task; only the two admissible four coordinate seeds and eight bit parity test optimization within the class.

A slot $q_{r,g}$ is a linear form in the coordinates of block g alone, so a product over distinct slots contributes at most one coordinate per block, and the target character $\prod_{i \in S} x_i$ is reachable only if the partition places the $|S|$ support coordinates in $|S|$ distinct blocks. With one partition shared by all

$R = 8$ bases, identity conditioning, and a linear core readout, as in these probes, a target support that contains two coordinates of one block has zero coefficient in every realized polynomial score at every available order. This obstructs exact polynomial representation by that predictor; it is not a theorem about an arbitrary nonlinear host or about its thresholded decisions (Section J.3).

Admissibility depends only on the partition and the target support, so the seeds were screened before training, and an exact enumeration confirms the screen: the full input cube ($65{,}536$ points at $d = 16$, 256 at $d = 8$), evaluated in double precision, was projected onto the target character for the initial draw and three further independent redraws of all trainable parameters. The predicate separates the seeds exactly and in both directions: seeds whose partition collides reach a maximum absolute coefficient of 2.65×10^{-17} on the six coordinate task and 3.38×10^{-17} on the four coordinate task, seeds that separate at least 1.10×10^{-2} and 7.43×10^{-2}. This gap of about fifteen orders of magnitude validates the tested cases, not a universal implication from sampled weights; necessity follows analytically from the support of disjoint blocks. For this unconstrained shared family, sufficiency for a target character of available degree $|S| \geq 2$ follows by setting one weight per separated target block to a nonzero value and all other weights to zero and using only the channel of degree $|S|$, whose fixed nonzero calibration and readout scale can be compensated. This constructive argument assumes those coefficient and weight choices are allowed and does not extend the sampled check to a constrained model.

On the first five admissible seeds of each task, chosen before training from the partition alone, the order matched PSNeuron and every PSNeuron with larger K solve both tasks on every seed (Table 41). In these 25 solved models and the five parity models, masking the channel of the target's order lowers accuracy from at least 97.1% to between 26.9 and 51.6% in all 30 models, while masking any other single order leaves accuracy exactly unchanged in 134 of 140 cases; the six exceptions come from one parity seed and change accuracy by at most 4.3 points. Eight bit parity is the positive control: there $d = G$, every block is a single bit, and the support is separated at every seed, as a scan of twenty thousand seeds confirms, so the $K = 8$ unit can represent the target, while units with $K < 8$ cannot reach its degree (Corollary 2). This settles representability before training; the high parity accuracy of the $K = 8$ unit is a separate observation about learning, and accuracy alone does not certify representability (Section J.3).

Explanatory counterexample: scores versus decisions. For independent uniform $x_i \in \{-1, +1\}$, take $t = x_1 x_2$ and

$$f(x) = -[(x_1 + x_2)(x_3 + x_4) + x_3 x_4] = -e_2(x_1 + x_2, x_3, x_4).$$

Every term of f has support different from $\{1, 2\}$, so $\mathbb{E}[tf] = 0$. Yet direct enumeration gives $\Pr\{\mathrm{sign}(f) = t\} = 10/16$, and f is never zero on this cube. Thresholding is nonlinear and need not preserve the Walsh support of the score, so on seeds 0 to 4 of the six coordinate task neither the test accuracy near chance nor the zero coefficient of the target score establishes a ceiling at chance for the thresholded classifier. The counterexample only shows that this implication fails; the near chance accuracies remain the empirical result, and the zero coefficient remains an exact property of the score.

J.4 Architecture and order ablations

Anatomy screen. In the anatomies of Table 42 and Figure 6, which are every tabular anatomy tested at fixed $K = 4$, only $x + \lambda B(P(Ax))$ has a residual path, and its scale λ is a learned scalar, the one extra parameter of that column. The selection rule was fixed before any test evaluation and left Helena out, and its score differs from the validation accuracy at the end of training in Section J.1. On test the selected $B(P(x))$ is the best anatomy on none of the three tasks: $P(x)$ leads on HIGGS Small and $B(P(Ax))$ on Jannis and Helena, by at most 0.27 points.

Exposed interaction order. Table 43 changes only the maximum order K of the standalone bare PSNeuron, in which P alone maps the input features to the logits without a host network, conditioner, or readout, unlike the $P(x)$ anatomy inside the tabular host. HIGGS Small, Jannis, and ImageNet100 have their highest means at $K = 4$ and EuroSAT at $K = 2$, so exposing more orders does not raise the mean monotonically on these representations. Above $K = 2$, however, the means of each task differ by at most 0.88 points, comparable to seed standard deviations of up to 1.29 points, and every selected rate lies at the grid edge, so the sweep does not resolve which of these orders is best.

Table 41: Exact order probes on the Boolean cube, five seeds each; test accuracy in percent as mean $\pm$ seed standard deviation, higher is better. *The order matched PSNeuron of each task. Parity fits the truth table of eight bits with $G = 8$, one bit per slot, on seeds 0 to 4. The XOR tasks use $d = 16$ coordinates, $G = 8$ blocks shared by all $R = 8$ bases, and held out draws with disjoint rows. Columns marked 0 to 4 use those seeds; admissible columns use the first five seeds whose partition separates the target coordinates: 2, 3, 5, 6, 7 for four coordinates and 5, 6, 7, 8, 12 for six. A partition that does not separate the target leaves it outside the span of the predictor's squarefree polynomials, which does not bound the accuracy of the thresholded classifier at chance (Section J.3). The additive control sums a GELU of each slot instead of multiplying the slots, with 16 bases on the same kind of partition, which is why its count is larger. The PSNeuron rows are the bare PSNeuron without a host; the additive control and the GELU MLP are baselines at their own geometry, so no parameter ratio is quoted.

	parameters		eight bit	four coordinate XOR		six coordinate XOR	
model	parity	XOR	parity	0 to 4	admissible	0 to 4	admissible
PSNeuron $K = 2$	80	144	51.19 ± 1.08	49.35 ± 1.91	49.59 ± 1.82	50.36 ± 1.08	49.30 ± 1.41
PSNeuron $K = 4$	104	168	50.01 ± 2.68	$69.57 \pm 27.78^*$	$100.00 \pm 0.00^*$	50.04 ± 1.86	49.64 ± 1.41
PSNeuron $K = 6$	120	184	51.58 ± 1.51	69.31 ± 28.02	100.00 ± 0.00	$49.89 \pm 2.09^*$	$100.00 \pm 0.00^*$
PSNeuron $K = 8$	136	200	$99.36 \pm 1.44^*$	69.75 ± 27.63	100.00 ± 0.00	50.32 ± 1.84	100.00 ± 0.00
additive control	160	288	52.08 ± 1.02	50.51 ± 1.46	49.90 ± 1.77	49.00 ± 1.34	48.92 ± 0.77
GELU MLP	178	610	76.46 ± 24.41	92.37 ± 8.72	95.58 ± 9.43	96.43 ± 4.91	100.00 ± 0.00

Table 42: Tabular anatomy screen at fixed $K = 4$, five seeds. Entries are mean test accuracy (%), higher is better, with the trainable parameter count of the whole model in parentheses. The last row is the selection score on validation, the best validation accuracy during training averaged over seeds and over HIGGS Small and Jannis; the highest score selected $B(P(x))$ as the common anatomy.

task	$P(x)$	$B(P(x))$	$P(Ax)$	$B(P(Ax))$	$x + \lambda B(P(Ax))$
HIGGS Small	71.18 (19,994)	70.91 (21,018)	70.53 (21,018)	71.02 (22,042)	71.01 (22,043)
Jannis	69.18 (21,788)	70.07 (22,812)	70.04 (22,812)	70.15 (23,836)	70.02 (23,837)
Helena	35.42 (26,300)	35.65 (27,324)	35.40 (27,324)	35.72 (28,348)	35.63 (28,349)
selection score	71.464	**71.649**	71.454	71.514	71.525

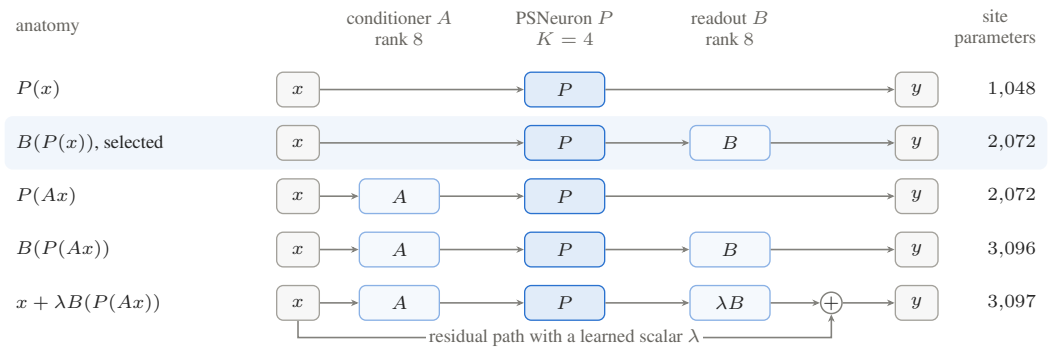

Figure 6: The five tabular anatomies of Table 42. All five share the PSNeuron block P with $K = 4$ and differ only in the host pieces around it, namely the rank 8 conditioner $A = A_{\text{host}} = I + U_A V_A$ and the rank 8 readout $B = B_{\text{host}} = U_B V_B$, each with 1,024 parameters, and a residual path with a learned scalar λ. The right column gives the trainable parameters of the replaced feedforward site, which are the same on all three tasks.

Table 43: Order sweep of the standalone bare PSNeuron, five seeds. Entries are mean test accuracy (%) ± seed standard deviation, higher is better, with the trainable parameter count of the whole model in parentheses; bold marks the best mean of each row. All 16 models selected 6.4×10^{-3}, the largest rate of the extended grid from 2×10^{-4} to 6.4×10^{-3}, so their tuning is not bracketed.

task	$K = 2$	$K = 4$	$K = 6$	$K = 8$
HIGGS Small	61.49 ± 0.29 (240)	**63.30** ± 0.98 (264)	62.68 ± 0.91 (280)	62.76 ± 0.81 (296)
Jannis	61.85 ± 0.20 (464)	**65.02** ± 0.49 (488)	64.96 ± 0.34 (504)	64.75 ± 0.51 (520)
ImageNet100 DINOv3	73.16 ± 0.95 (6,944)	**77.00** ± 0.89 (6,968)	76.20 ± 1.29 (6,984)	76.12 ± 1.06 (7,000)
EuroSAT DINOv3	**95.90** ± 0.20 (6,224)	95.66 ± 0.28 (6,248)	95.29 ± 0.33 (6,264)	95.30 ± 0.17 (6,280)

The $K = 2$ ImageNet100 entry and the $h = 8$ head of the frozen DINOv3 ladder (85.64%, 7,044 parameters; Section J.10) are different models, although both have $R = 8$, $G = 8$, and order 2, read the same cached features, and are scored on the same 5,000 held out images. Only the ladder head carries the class bias of length $C = 100$ that its dense control also carries, the whole difference in parameters. The sweep model trains for 2,000 steps at batch 256, about five passes over 100,000 training rows, and the ladder head for 12,400 steps at batch 512, about 52 passes over 121,689 rows, roughly twelve times as many examples; weight decay (0.01 against 10^{-4}), schedule (constant against cosine), learning rate grid (up to 6.4×10^{-3} against 3.2×10^{-3}), calibration rows (128 against 4,096), carved selection split (26,689 against 5,000 rows), and final seeds (five against three) also differ. So does the parameterization: the sweep model draws its bank from $\mathcal{N}(0, 1)$, multiplies the calibration rescale into it, and draws its partition from the seed of each run, whereas the ladder head starts its bank at the initialization scale of its GELU control, keeps the rescale as a fixed multiplier, and draws its partition from one fixed seed (Section G.1). The gap from 73.16 to 85.64% is attributed to no single change, and each model is compared only within its protocol: the sweep row with the other orders, the ladder head with the dense and additive heads of its width.

J.5 Placement inside and beside the attention block

Three sites of the vision and recall hosts lie in or next to attention, namely the point after the attention output projection, the projection itself, and the attention logits; the feedforward site of the same block serves as a reference.

What each site receives. In the ViT with six blocks of width 192, the feedforward and post attention sites both receive activations of shape $[B, 64, 192]$, and the same block with the same slot shape is placed at each, with 6,144, 6,168, 6,184, and 6,200 parameters at $K = 2$, 4, 6, and 8. The post attention site receives the output of the dense $192 \to 192$ map applied after the head outputs are merged, unchanged, as a direct comparison of the projection output with the site input confirms, so neither site sees head structure, and comparing them compares one channel mixer at two points of one block, which is no evidence about attention; at the post attention site the mixer acts in series on the attention output, which then reaches the residual stream only through a map whose outputs lie in a subspace of dimension 8. Only two placements see attention: the cross head site replaces that projection and receives the head outputs with the head axis intact, of shape $[B, T, H, D_{\text{head}}]$ (Table 29), and the logit correction adds a term to the query and key scores before the softmax.

Selected placement and adverse outcomes. On validation, with the test split held out, the same selected block reaches a mean over the five final seeds of 39.824, 39.776, 40.296, and 39.988% at the feedforward site and 39.872, 39.528, 32.256, and 35.664% at the post attention site for $K = 2$, 4, 6, and 8; these are validation values, not final test results. The prespecified rule compared the two sites only at $K = 4$, 39.776 against 39.528%, and returned the feedforward site, but the native host of 2,707,108 parameters, which keeps the GELU network that the block replaces and has no module at the post attention site, reaches 40.504% on validation, above all eight values, and at $K = 6$ and 8 the post attention site falls far below the feedforward site, although the rate selected at $K = 8$, 10^{-4}, is the lower edge of its extended grid. The comparison is not controlled: the block displaces a GELU network of 295,872 parameters at the feedforward site and nothing at the post attention site, so the two models differ by 295,872 parameters, 12.2% of the 2,417,404 of the feedforward model. The difference favors the site that lost, since the winner is the smaller model, so the ranking survives it. A second difference works against the post attention site: there the attention output of the last block passes through the rank 8 readout of the block, while at the feedforward site the block adds to

the stream. The post attention placement keeps every host parameter and inserts the block in series, 2,713,276 against 2,707,108 parameters, a ratio of 1.0023 that no width can move below one, and carries no size claim; the feedforward placement is a replacement with a ratio of 0.8930, and placing it in all six blocks gives Table 49. The logit correction is also an augmentation, and on MQAR at least three of five seeds collapse at every order and with either kernel (Table 34).

What is not tested. No placement acts on the value aggregation or the softmax, none replaces the query, key, and value projections, and the residual adapter sits outside every attention block. The attention results therefore concern what surrounds attention, replaces its output projection, or corrects its scores, never the value aggregation itself.

J.6 Statistics of the masks and component ablations

Populations. The 12 synthetic tasks are four and six coordinate XOR and eight bit parity (Table 41); the mixed order targets with equal shares through orders 6 and 8; the product targets $\prod_{i\in S}(1+\beta x_i)-1-\beta\sum_{i\in S}x_i$ at $\beta=0.7$ on the same supports through orders 6 and 8, where orders 7 and 8 carry only 0.3% of the variance of the order 8 target; and the coherent teachers of orders 2, 4, and 6 and incoherent teachers of orders 6 and 8, built like the targets of Section I.5 with one family of planted directions. The 14 settings with standard groups add the product target that keeps its linear term, $\prod_{i\in S}(1+\beta x_i)-1$, through orders 6 and 8. A group mask removes a contiguous block of orders: single orders at $K=4$; orders 2 and 3, 4, and 5 and 6 at $K=6$; and the standard groups, orders 2 to 4, 4 to 6, and 6 to 8, at $K=8$. Radial decrease, a diagnostic of training rather than a removal, is the median norm of the factor rows at initialization minus that at the end of training, in the five teachers. The controlled rows at $K=8$ are the 16 recovered runs among the 72 aligned order 8 runs of Section I.3.

Counts. The ablation analysis contains 2,207 paired comparisons, of which 1,811 are distinct measurements. The 425 measured conditions contribute 1,973 of them: 4 tabular component ablations (18 pairs), 32 tabular order masks (134), 192 synthetic order masks (960), 108 synthetic group masks (584), 5 diagnostics of radial decrease (25), 75 further masks on vision, attention score, and frozen head settings (225), and 9 separately trained structural comparisons (27). The 42 masks of the standard groups at $K=8$ over 14 synthetic settings contribute 234 pairs. These populations overlap: the standard group masks repeat the group masks at $K=8$ of the 12 synthetic tasks in 180 pairs, and on every synthetic task except eight bit parity the single order group masks at $K=4$ and the order 4 group mask at $K=6$ repeat order masks of the same models in 216 pairs, so the rows of Table 44 are not additive.

Evaluation split. Except for the controlled rows, which use the held out test split of the controlled teachers, every mask and component comparison in Tables 44 to 47 is scored on the validation split after selection, and none of these scores selected a model. The intact tabular model of the component comparisons, whose scores are the validation means of Section J.1, is however the anatomy that Table 42 selected on the same split and seeds by its best validation accuracy during training, so these comparisons may favor it; on test the anatomy screen likewise shows the readout helping on Jannis (70.07 against 69.18%) but not on HIGGS Small (70.91 against 71.18%). The CIFAR100 masks of Table 31 use the test split.

Paired effects and attainable significance. Each effect is in the native unit of its row, with accuracy and recall as fractions. The signed difference is intact minus ablated quality, reversed where lower is better, so positive differences indicate lost quality. Writing d_s for the signed difference of pair s, we report its median and the Hodges-Lehmann (HL) median of Walsh averages $(d_s+d_t)/2$ for $s\le t$, the counts of positive and negative pairs, and the matched pairs rank biserial correlation $(W_+-W_-)/(W_++W_-)$, where W_+ and W_- sum the absolute ranks of positive and negative paired differences, respectively. Exact two sided p values use the conditional sign distribution of the observed absolute ranks, with midranks for ties and the Pratt treatment of zeros. The smallest attainable two sided value is 2^{1-n_*} for n_* nonzero pairs, 0.0625 for five and 0.25 for three, so no single condition with five or three seeds can reach $p<0.05$, whatever its effect size. The two controlled mask rows of Table 45 reach $p=3.05\times10^{-5}$ only because they count 16 runs as independent; those runs come from three initialization seeds (9, 4, and 3 runs) and nine distinct pairs

of sign pattern and data stream (Section I.3), and grouped by initialization seed the three blocks give the floor $p = 0.25$ for either mask.

Bounds and the independent unit. Nominal 95% HL bounds use the ordered Walsh averages and the untied signed rank null: if c is the largest integer with $\Pr(W_+ \le c) \le 0.025$, their zero based indices are c and $M - 1 - c$, where n is the number of pairs in that condition and $M = n(n+1)/2$. These bounds are used in every table, but ties and dependent interventions preclude reading them as exact confidence intervals at the level of models; with fewer than six pairs no finite interval exists, and the tables mark the observed range with †. Several masks and configurations reuse training seeds, and the masks of one configuration usually reuse one trained model, so pooled bounds summarize dependent pairs and pooled p values assume independent signs that repeated masks do not provide. The seed check therefore averages all differences within each training seed and applies the same exact test to the seed means (floor 0.0020 for ten seeds), and the combined HIGGS Small and Jannis analysis keeps five seed blocks across both tasks rather than counting pairs of task and seed, or masks, as independent units.

Interpretation on real tasks. HIGGS Small and Jannis contribute 16 and 16 order mask conditions, one for each order 2 to K of the replacements with $K = 2$, 4, 6, and 8; 12 of the HIGGS Small and 7 of the Jannis conditions have five seeds, and the others were evaluated on seeds 0 to 2. The combined HL effect is 0.0021 with descriptive bounds $[0.0013, 0.0029]$, and its median over the five seed blocks is 0.0046, against 0.0019 for HIGGS Small and 0.0050 for Jannis alone, with block $p = 0.0625$. The component comparisons instead change the architecture and retrain it. Omitting the readout B_{host} retrains the $P(x)$ anatomy (19,994 parameters on HIGGS Small), and the native host, with its GELU network restored at the site (35,522 parameters), is retrained under the protocol of the replacement, so it is not the reference selected on validation in Table 28. Omitting the readout lowers Jannis accuracy in 5 of 5 pairs, whereas HIGGS Small shows no consistent benefit from the readout; the larger native host is level with the replacement on HIGGS Small and ahead of it on all three Jannis seeds, 0 to 2, the only seeds on which the Jannis native host was retrained. These comparisons support specific uses of the architecture without establishing that any component is necessary in general.

Represented order and functional load. The rows of Table 45 with equal shares and with parity use all five seeds and come from models other than the recovered controlled runs, so their mask effects, which differ from the controlled result, neither replicate nor contradict it. Synthetic pools aggregate different targets and describe the evaluated collection rather than a common effect on every task.

Further masks on real tasks. The attention score rows of Table 46 use the diagonal kernel selected on validation at its own size (Table 34). Averaged over seeds, their intact MQAR validation recall is 12.5 to 14.9% at $K = 2$, 4, 6, and 8, near the 12.5% of guessing among the eight stored values, so the MQAR masks act on models that have not solved the task. The frozen head rows mask direct replacement heads trained for these masks at $K = 2$, 4, 6, and 8, one mask for each order 2 to K (18,673 to 18,729 parameters on EuroSAT, 20,833 to 20,889 on ImageNet100), at learning rates of 1.6×10^{-3} to 3.2×10^{-3}. On ImageNet100 the masks at $K = 8$ come from two trainings: one at 2.4×10^{-3} for orders 2, 4, and 6, and one at 3.2×10^{-3} for the other orders. The heads at $K = 4$, which also enter Table 47, have 18,697 and 20,857 parameters, learning rates of 1.6×10^{-3} and 3.2×10^{-3}, and intact validation accuracy of 96.20% and 86.21%. They are separate three seed trainings, not the five seed runs at 6.4×10^{-3} of Table 35. Unlike the heads of the DINOv3 width ladder (Section J.10), they carry orders up to K, a conditioner, and a readout. Masking order 4 of the heads at $K = 4$ lowers validation accuracy in all three trained models of each task, by 41.67 to 48.91 points on EuroSAT and 0.51 to 2.09 points on ImageNet100, with HL estimates of 0.4463 and 0.0126 as accuracy fractions; with three pairs no two sided p below 0.25 is attainable.

J.7 Tabular budget comparisons

Section L.5 places every complete single site tabular replacement among the resized dense networks and gives the budget comparisons on each task. Table 48 adds all ten configurations that replace both feedforward networks of the tabular host at $K = 4$. They are much smaller models and less

Table 44: Pooled paired summaries for each comparison population, scored on validation (Section J.6) except radial decrease, which compares factor norms. Positive Δ favors the intact model; Δ is an accuracy fraction for the tabular tasks, XOR, and parity, R^2 for the regression targets and teachers, and a factor row norm for radial decrease. Several interventions reuse one trained model, so the HL bounds are descriptive summaries of the pairs, not confidence intervals over independent models, and pooled p values assume independent signs. The last column instead tests one mean difference per training seed: five seeds for the tabular rows except the Jannis native host, which has three, ten for the pools over all synthetic tasks and settings, and five for individual synthetic tasks and for radial decrease. The $+/-$ column counts pairs, not ranks, and Pratt zeros remain in the ranking. † marks an observed range where no finite two sided 95% interval exists. The signed rank statistic is $W = \min(W_+, W_-)$. Readout omitted retrains without B_{host}; native host retrains the host with its GELU network restored at the site.

comparison	conditions	pairs	zeros	med. Δ	HL [summary bounds]
HIGGS Small: readout omitted	1	5	0	-0.0001	-0.0001 $[-0.0064, 0.0032]^{\dagger}$
HIGGS Small: native host	1	5	0	-0.0004	-0.0004 $[-0.0049, 0.0037]^{\dagger}$
Jannis: readout omitted	1	5	0	0.0090	0.0082 $[0.0018, 0.0145]^{\dagger}$
Jannis: native host	1	3	0	-0.0011	-0.0018 $[-0.0048, -0.0003]^{\dagger}$
HIGGS Small: order mask	16	72	2	0.0005	0.0011 $[0.0005, 0.0019]$
Jannis: order mask	16	62	0	0.0028	0.0036 $[0.0020, 0.0058]$
both tabular tasks: order mask	32	134	2	0.0013	0.0021 $[0.0013, 0.0029]$
equal shares, $K = 8$: order mask	16	80	0	0.0586	0.0585 $[0.0476, 0.0692]$
eight bit parity: order mask	16	80	24	0.0034	0.0078 $[0.0037, 0.0139]$
coherent $K = 2$: order mask	16	80	0	0.0000	0.0000 $[0.0000, 0.0000]$
all 12 synthetic tasks: order mask	192	960	134	0.0073	0.0637 $[0.0577, 0.0707]$
all 12 synthetic tasks: group mask	108	584	49	0.0331	0.0949 $[0.0788, 0.1202]$
14 synthetic settings: standard groups	42	234	18	0.0231	0.1623 $[0.0961, 0.2095]$
five teachers: radial decrease	5	25	0	0.4354	0.4801 $[0.3082, 0.8066]$

comparison	$+/-$	r_{rb}	W	p	p floor	block p
HIGGS Small: readout omitted	2/3	-0.07	7.0	1.0000	0.0625	1.0000
HIGGS Small: native host	2/3	-0.07	7.0	1.0000	0.0625	1.0000
Jannis: readout omitted	5/0	1.00	0.0	0.0625	0.0625	0.0625
Jannis: native host	0/3	-1.00	0.0	0.2500	0.2500	0.2500
HIGGS Small: order mask	48/22	0.54	601.0	3.88×10^{-5}	1.69×10^{-21}	0.0625
Jannis: order mask	47/15	0.69	300.0	5.01×10^{-7}	4.34×10^{-19}	0.0625
both tabular tasks: order mask	95/37	0.62	1724.0	7.58×10^{-11}	3.67×10^{-40}	0.0625
equal shares, $K = 8$: order mask	64/16	0.85	236.0	1.07×10^{-13}	1.65×10^{-24}	0.0625
eight bit parity: order mask	45/11	0.61	575.5	6.44×10^{-6}	2.78×10^{-17}	0.0625
coherent $K = 2$: order mask	23/57	-0.24	1228.0	0.0602	1.65×10^{-24}	0.1250
all 12 synthetic tasks: order mask	522/304	0.57	97373.5	6.21×10^{-55}	4.47×10^{-249}	0.0020
all 12 synthetic tasks: group mask	364/171	0.66	28594.0	2.58×10^{-48}	1.78×10^{-161}	0.0020
14 synthetic settings: standard groups	133/83	0.56	5984.5	1.43×10^{-14}	1.90×10^{-65}	0.0020
five teachers: radial decrease	22/3	0.93	12.0	4.17×10^{-6}	5.96×10^{-8}	0.0625

accurate than the single site replacement (70.91% on HIGGS Small and 70.07% on Jannis), so they trade accuracy for size, and because no dense network with both feedforward networks resized to these budgets was trained, they enter no dense comparison (Table 62).

J.8 Replacement results on secondary placements

Beside the three bare configurations of Table 49, two bare CIFAR100 models apply $P(x)$ alone at the last feedforward site of the ViT: with $K = 2$ and $K = 6$ they have 2,414,308 and 2,414,348 parameters, 3,072 and 3,112 of them at the site, and reach $39.80 \pm 0.65\%$ and $39.54 \pm 0.78\%$ test accuracy over five seeds, both at a selected learning rate of 4×10^{-4}. The remaining three bare configurations, $K = 4$ at the last site and $K = 2$ and $K = 6$ in all six blocks, reach 39.41 ± 0.60,

Table 45: Masks of order groups at $K = 8$ in recovered models and over all seeds. The controlled rows use the 16 recovered runs among the 72 aligned order 8 runs of Section I.3, scored in held out test R^2; they come from three initialization seeds, so their p values treat dependent runs as independent (Section J.6). The rows over all seeds use all five seeds and are scored on validation. A controlled run counts as recovered when its held out R^2 reaches 0.9 (Section 4); on the intact validation score the same bar is reached by no run with equal shares (intact R^2 0.17 to 0.54) and by all five parity runs (intact accuracy 97.07% on one seed and 100% on four). Δ is in R^2 for the controlled and equal share rows and an accuracy fraction for the parity rows. Bounds use the nominal 95% signed rank Walsh construction; † marks the observed paired range at $n = 5$.

source	setting	comparison	n	zeros	med. Δ	HL [bounds/range]	$+/-$	r_{rb}	p	floor
controlled	recovered $K = 8$	mask 6 to 8	16	0	0.4286	0.4198 [0.3905, 0.4424]	16/0	1.00	3.05×10^{-5}	3.05×10^{-5}
controlled	recovered $K = 8$	mask 2 to 4	16	0	0.4346	0.4305 [0.4181, 0.4429]	16/0	1.00	3.05×10^{-5}	3.05×10^{-5}
controlled	recovered $K = 8$	high minus low	16	0	−0.0113	−0.0100 [−0.0248, 0.0026]	4/12	−0.43	0.1436	3.05×10^{-5}
all seeds	equal shares	mask 6 to 8	5	0	−0.0193	−0.0156 [−0.0281, 0.0822]†	1/4	−0.33	0.6250	0.0625
all seeds	equal shares	mask 2 to 4	5	0	0.2461	0.2473 [0.1970, 0.3620]†	5/0	1.00	0.0625	0.0625
all seeds	equal shares	high minus low	5	0	−0.2654	−0.2629 [−0.2798, −0.2251]†	0/5	−1.00	0.0625	0.0625
all seeds	parity	mask 6 to 8	5	0	0.4556	0.4563 [0.4258, 0.4966]†	5/0	1.00	0.0625	0.0625
all seeds	parity	mask 2 to 4	5	4	0.0000	0.0000 [−0.0142, 0.0000]†	0/1	−1.00	1.0000	1.0000
all seeds	parity	high minus low	5	0	0.4556	0.4563 [0.4399, 0.4966]†	5/0	1.00	0.0625	0.0625

Table 46: Further masks at evaluation on trained models for vision, attention score corrections, and heads on frozen features, three training seeds per condition, scored on validation; Δ is an accuracy or recall fraction. The CIFAR100 row pools the nine masks listed in Section J.2, which are disjoint from the seven of Table 31. A branch mask sets $\lambda = 0$ in the attention score correction. The frozen feature heads are the direct replacement heads at $K = 2$, 4, 6, and 8 described in Section J.6, not the heads of the DINOv3 width ladder. Pooled HL bounds and p values describe dependent pairs; the seed check keeps three independent seed blocks, with a smallest two sided p of 0.25. A positive effect shows reliance on the trained branch or order, not superiority over a separately trained reference.

comparison	conditions	pairs	zeros	med. Δ	HL [summary bounds]
CIFAR100: order mask	9	27	2	0.0014	0.0033 [0.0010, 0.0054]
MQAR scores: order mask	16	48	2	0.0005	0.0008 [0.0003, 0.0020]
MQAR scores: branch mask	1	3	0	0.0012	0.0008 [−0.0003, 0.0013]†
multi-needle scores: order mask	16	48	0	0.0751	0.1182 [0.0692, 0.1742]
multi-needle scores: branch mask	1	3	0	0.5364	0.5420 [0.4705, 0.6248]†
EuroSAT head: order mask	16	48	0	0.1077	0.1497 [0.0973, 0.2304]
ImageNet100 head: order mask	16	48	0	0.2651	0.3924 [0.2398, 0.4555]

comparison	$+/-$	r_{rb}	W	p	p floor	block p
CIFAR100: order mask	19/6	0.76	44.5	2.55×10^{-4}	5.96×10^{-8}	0.2500
MQAR scores: order mask	35/11	0.63	219.0	8.66×10^{-5}	2.84×10^{-14}	0.2500
MQAR scores: branch mask	2/1	0.67	1.0	0.5000	0.2500	0.5000
multi-needle scores: order mask	38/10	0.88	73.0	2.80×10^{-9}	7.11×10^{-15}	0.2500
multi-needle scores: branch mask	3/0	1.00	0.0	0.2500	0.2500	0.2500
EuroSAT head: order mask	48/0	1.00	0.0	7.11×10^{-15}	7.11×10^{-15}	0.2500
ImageNet100 head: order mask	48/0	1.00	0.0	7.11×10^{-15}	7.11×10^{-15}	0.2500

33.08 ± 0.75, and $31.14 \pm 0.53\%$ test accuracy over five seeds with 2,414,332, 950,308, and 950,548 parameters. The bare PSNeuron head with $K = 8$ on frozen DINOv3 EuroSAT features is the $K = 8$ EuroSAT model of Table 43, trained under the shorter protocol of the frozen feature heads without a dense head. The family of Table 49 that replaces the feedforward network of all six ViT blocks has $r = 0.3579$ at $K = 4$ against the native model, but its saving comes with a large loss of accuracy, so it is not compression that preserves quality.

J.9 Paired test gains and dominance of point estimates

Table 50 pairs by training seed four favorable final test comparisons from the budget studies, with the corresponding GELU comparisons beside them. Its intervals are the nominal paired intervals of Eq. (23), without adjustment for multiplicity; none keeps a positive lower bound after the joint Bonferroni correction of Section L, and five seeds cannot give an exact two sided signed rank p below

Table 47: Separately trained structural comparisons at $K = 4$, three seeds each, scored on validation; Δ is an accuracy or recall fraction, and positive effects favor the intact PSNeuron configuration. The intact score is the intact evaluation of the same seed and configuration at the same learning rate, data split, and optimization protocol, and every matching intact evaluation agrees exactly. Native model: the CIFAR100 host with its output projection restored in place of the cross head site, retrained at the learning rate of the cross head model, 2×10^{-4}. Full and low rank output kernel: the attention score correction with a full M or a factored M of rank 4 in place of the selected diagonal M. Conditioner omitted: the direct head without A_{host} (6,409 parameters on EuroSAT, 8,569 on ImageNet100). Readout omitted: the direct head without B_{host} (18,537 and 19,257). These are architecture comparisons, not masks on one model. Brackets give observed ranges (†), not 95% confidence intervals.

setting	comparator	n	zeros	med. Δ	HL [range†]	+/−	r_{rb}	p	floor
CIFAR100	native model	3	0	-0.0142	-0.0085 $[-0.0152, 0.0096]^{\dagger}$	1/2	-0.67	0.5000	0.2500
MQAR scores	full output kernel	3	0	0.0032	-0.0013 $[-0.0168, 0.0053]^{\dagger}$	2/1	0.00	1.0000	0.2500
MQAR scores	low rank output kernel	3	0	0.0033	0.0033 $[0.0016, 0.0049]^{\dagger}$	3/0	1.00	0.2500	0.2500
multi-needle scores	full output kernel	3	0	-0.0156	-0.0141 $[-0.0212, -0.0039]^{\dagger}$	0/3	-1.00	0.2500	0.2500
multi-needle scores	low rank output kernel	3	0	-0.0039	0.0048 $[-0.0134, 0.0405]^{\dagger}$	1/2	0.00	1.0000	0.2500
EuroSAT head	conditioner omitted	3	0	0.0300	0.0306 $[0.0270, 0.0354]^{\dagger}$	3/0	1.00	0.2500	0.2500
EuroSAT head	readout omitted	3	0	0.0009	0.0016 $[-0.0026, 0.0070]^{\dagger}$	2/1	0.33	0.7500	0.2500
ImageNet100 head	conditioner omitted	3	0	0.1936	0.1865 $[0.1592, 0.1996]^{\dagger}$	3/0	1.00	0.2500	0.2500
ImageNet100 head	readout omitted	3	0	-0.0041	-0.0042 $[-0.0072, -0.0013]^{\dagger}$	0/3	-1.00	0.2500	0.2500

Table 48: Replacing both feedforward networks of the tabular host with two blocks, at $K = 4$. All ten measured configurations use five final seeds; test accuracy is mean $\pm$ seed standard deviation. The site count includes both replaced networks, and † marks a selected learning rate at the upper edge of the evaluated grid.

task	configuration	model params	site params	test score (%)	learning rate
HIGGS Small	$P(x)$	4,466	2,096	70.20 ± 0.29	$0.0064^{\dagger}$
HIGGS Small	$B(P(x))$	6,514	4,144	70.16 ± 0.43	$0.0064^{\dagger}$
HIGGS Small	$P(Ax)$	6,514	4,144	70.44 ± 0.21	$0.0064^{\dagger}$
HIGGS Small	$B(P(Ax))$	8,562	6,192	70.14 ± 0.34	$0.0064^{\dagger}$
HIGGS Small	$x + \lambda B(P(Ax))$	8,564	6,194	70.27 ± 0.21	$0.0064^{\dagger}$
Jannis	$P(x)$	6,260	2,096	69.30 ± 0.16	$0.0064^{\dagger}$
Jannis	$B(P(x))$	8,308	4,144	68.98 ± 0.11	0.0032
Jannis	$P(Ax)$	8,308	4,144	69.45 ± 0.20	$0.0064^{\dagger}$
Jannis	$B(P(Ax))$	10,356	6,192	68.75 ± 0.31	0.0016
Jannis	$x + \lambda B(P(Ax))$	10,358	6,194	69.39 ± 0.21	$0.0064^{\dagger}$

Table 49: CIFAR100 ViT with the feedforward network of all six blocks replaced, and bare configurations. Accuracy is held out test mean $\pm$ seed standard deviation over five final seeds. The PSNeuron rows replace the feedforward network in all six blocks and are compared with the native model. The saving is the whole model parameter saving against the native model and does not imply preserved quality. The bare rows are a separate formulation and are not compared with a dense model.

family	configuration	model params	site params	test acc. (%)	saving (%)
PSNeuron, all blocks	$K = 2$	968,740	36,864	33.356 ± 0.499	64.21
PSNeuron, all blocks	$K = 4$	968,884	37,008	32.240 ± 1.049	64.21
PSNeuron, all blocks	$K = 6$	968,980	37,104	31.488 ± 0.532	64.21
PSNeuron, all blocks	$K = 8$	969,076	37,200	32.714 ± 0.442	64.20
native model	dense feedforward networks	2,707,108	1,775,232	40.618 ± 0.746	n/a
bare	$K = 8$, last block only	2,414,364	3,128	39.806 ± 0.825	n/a
bare	$K = 4$, all blocks	950,452	18,576	31.282 ± 0.997	n/a
bare	$K = 8$, all blocks	950,644	18,768	32.092 ± 0.746	n/a

0.0625 even when all five differences are positive. HIGGS Small with $K = 6$ and Jannis with $K = 2$ beat their SwiGLU controls under these intervals with site counts 0.38% and 1.15% larger, so they are gains in quality at comparable budgets, not parameter savings, and against GELU both intervals include zero. HIGGS Small with $K = 4$ uses 7 fewer site parameters than its SwiGLU control and has a higher test mean, and the CIFAR100 cross head with $K = 2$ has as many site parameters as the additive control and a higher mean; these two are dominance of point estimates only, without a frontier claim that accounts for uncertainty, and the CIFAR100 control is the additive one, not the larger native projection, whose mean remains higher. The pairs share task, data split, training seeds, host, optimizer, schedule, batch size, step budget, and evaluation split, but learning rates are chosen independently on validation, and the tabular PSNeuron selections stay at the upper edge of the grid. The fixed pairs are separate from the measured capacity envelopes and do not establish that one family dominates another.

Table 50: Test comparisons paired by seed at matched site budgets, with SwiGLU and GELU controls. Each row pairs final seeds 0 to 4; accuracy is test mean ± seed standard deviation. Bounds are nominal, unadjusted two sided Student t intervals on five seed differences. In the lower panel, P is PSNeuron and C the control, and the search column says whether each selected learning rate lies at the edge of the grid or inside it.

setting	control	PSNeuron accuracy	control accuracy	Δ pp [95% CI]
HIGGS Small, $K = 6$	SwiGLU	71.14 ± 0.26	70.66 ± 0.47	$+0.482$ [0.134, 0.829]
Jannis, $K = 2$	SwiGLU	69.99 ± 0.29	69.52 ± 0.26	$+0.464$ [0.076, 0.853]
HIGGS Small, $K = 4$	SwiGLU	71.03 ± 0.42	70.88 ± 0.17	$+0.148$ [-0.192, 0.489]
CIFAR100, $K = 2$	additive	38.29 ± 0.49	37.87 ± 0.85	$+0.418$ [-0.593, 1.429]
HIGGS Small, $K = 6$	GELU	71.14 ± 0.26	71.02 ± 0.43	$+0.116$ [-0.456, 0.687]
Jannis, $K = 2$	GELU	69.99 ± 0.29	69.83 ± 0.24	$+0.157$ [-0.298, 0.612]
HIGGS Small, $K = 4$	GELU	71.03 ± 0.42	71.06 ± 0.18	-0.032 [-0.577, 0.512]

setting	control	site P/C	LR P/C	search P/C	steps / batch
HIGGS Small, $K = 6$	SwiGLU	2,088/2,080	0.0064/0.0016	edge/interior	3000/256
Jannis, $K = 2$	SwiGLU	1,408/1,392	0.0064/0.0004	edge/interior	3000/256
HIGGS Small, $K = 4$	SwiGLU	2,739/2,746	0.0064/0.0032	edge/interior	3000/256
CIFAR100, $K = 2$	additive	7,936/7,936	0.0002/0.0002	interior/interior	8000/128
HIGGS Small, $K = 6$	GELU	2,088/2,064	0.0064/0.0064	edge/edge	3000/256
Jannis, $K = 2$	GELU	1,408/1,354	0.0064/0.0008	edge/interior	3000/256
HIGGS Small, $K = 4$	GELU	2,739/2,773	0.0064/0.0008	edge/interior	3000/256

J.10 Frozen DINOv3 width ladders and matched controls

This subsection gives the complete results behind the frozen feature heads of Section 5.2 and Figure 2(b): a width ladder with three seeds, a separate study with five seeds and a wider learning rate grid, its extension to wider heads with a linear probe, and a comparison with generic bilinear heads. At each width, $h \in \{1, 2, 3, 4, 8\}$ in the ladder and $h \in \{2, 3, 4, 8, 16, 32, 64, 128\}$ in the five seed study, three heads map the frozen feature $x \in \mathbb{R}^{768}$ to h values u through a bank $W \in \mathbb{R}^{h \times 768}$, and then to class scores $\ell = Bu + b$ with $B \in \mathbb{R}^{C \times h}$ and $b \in \mathbb{R}^C$ for C classes. They differ only in how u is formed. The GELU head, the dense control, uses $u = \mathrm{GELU}(Wx + b_1)$ with $b_1 \in \mathbb{R}^h$. The PSNeuron head with $K = 2$ splits the 768 coordinates of each row of W into $G = 8$ slots by a fixed balanced partition, one for each row, so that each row gives eight scaled partial inner products with x; it takes their elementary symmetric polynomial of order 2 times a frozen calibration constant, and the additive head sums a GELU of each of them instead (Section B); neither has a hidden bias. Both therefore have $768h + Ch + C$ parameters, exactly h fewer than the GELU head, a parameter ratio between 0.9987 and 0.9990 at every width, so the ladder compares nonlinearities at matched width and measures no parameter saving. The partition rule is given in Table 9, calibration in Sections G and G.1, and the learning rate grids and training budgets in Section G; in the ladder most heads, PSNeuron and controls alike, select the largest rate of a grid ending at 3.2×10^{-3}.

The features are the class token of DINOv3 ViT-B/16 (Siméoni et al., 2025) after its final LayerNorm, of width 768, with no learned pooling and no normalization after extraction. Images are resized to 256 with bicubic interpolation, cropped at the center to 224, and normalized with the ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. ImageNet100 is the fixed subset

of 100 ImageNet classes introduced for Contrastive Multiview Coding by Tian et al. (2020); the supplementary material identifies the checkpoint, dataset revision, and constant defining the class order.

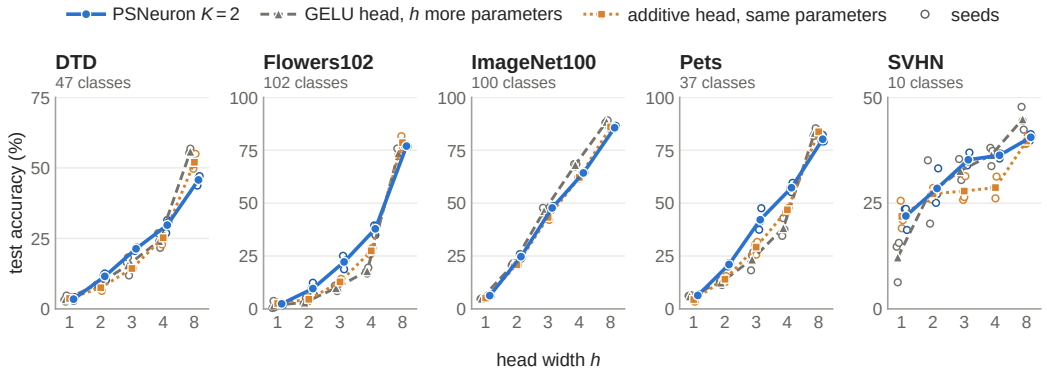

Figure 7: Frozen DINOv3 width ladders, one panel per dataset. Lines give the mean test accuracy over three seeds of the PSNeuron head with $K = 2$, the GELU head, which has h more parameters, and the additive head, which has exactly the parameters of PSNeuron; circles mark the three seeds. With five seeds and a wider grid the GELU head has the higher mean on ImageNet100 at $h = 2$, 3, and 4 and on DTD and Pets at $h = 4$ (Table 52). Table 51 lists every value.

Table 51: Frozen DINOv3 width ladders: mean test accuracy over three seeds, in percent, with the values of seeds 0, 1, and 2 in parentheses; higher is better, and the highest mean at each width is in bold. Parameters are those of the PSNeuron head and of the additive head, which have the same count; the GELU head has h more. *The selected learning rate is below the largest grid value, 3.2×10^{-3}, which every unmarked head selected.

h	parameters	GELU head	PSNeuron $K = 2$	additive head
DTD, 47 classes				
1	862	3.51* (3.19, 2.61, 4.73)	3.46* (3.35, 2.87, 4.15)	**3.69*** (3.35, 3.51, 4.20)
2	1,677	8.39 (8.09, 8.56, 8.51)	**11.54** (11.91, 12.55, 10.16)	7.52 (7.66, 8.51, 6.38)
3	2,492	15.51 (16.44, 18.24, 11.86)	**21.31** (20.59, 21.65, 21.70)	14.33 (14.10, 14.52, 14.36)
4	3,307	24.11 (26.81, 23.88, 21.65)	**29.72** (26.91, 31.49, 30.74)	25.25 (22.87, 27.02, 25.85)
8	6,567	**55.78** (55.27, 56.81, 55.27)	45.73 (43.72, 46.38, 47.07)	51.99 (49.68, 51.33, 54.95)
Flowers102, 102 classes				
1	972	1.65* (0.42, 3.76, 0.76)	2.41* (2.63, 2.85, 1.74)	**2.56** (2.54, 2.96, 2.20)
2	1,842	3.07* (2.36, 1.79, 5.07)	**9.62** (9.89, 12.36, 6.60)	4.54* (3.38, 5.53, 4.72)
3	2,712	9.90 (10.96, 10.29, 8.44)	**22.22** (25.09, 18.73, 22.83)	12.78 (11.77, 12.39, 14.18)
4	3,582	17.93 (17.58, 16.70, 19.52)	**37.84** (39.45, 34.92, 39.16)	27.44 (25.68, 27.89, 28.75)
8	7,062	73.66 (75.82, 72.69, 72.47)	76.97 (76.73, 77.38, 76.81)	**78.60** (81.66, 76.32, 77.83)
ImageNet100, 100 classes				
1	968	4.91 (4.60, 5.00, 5.12)	**6.28** (6.06, 6.64, 6.14)	5.14 (4.94, 5.34, 5.14)
2	1,836	20.78 (20.56, 21.44, 20.34)	**24.70** (25.86, 23.80, 24.44)	20.85 (21.54, 20.28, 20.74)
3	2,704	46.31 (47.74, 45.16, 46.02)	**47.73** (46.68, 47.84, 48.66)	43.28 (44.48, 43.28, 42.08)
4	3,572	**68.09** (68.46, 67.30, 68.52)	64.27 (64.12, 64.28, 64.42)	62.13 (62.30, 61.68, 62.40)
8	7,044	**88.49** (88.00, 88.34, 89.14)	85.64 (85.08, 85.32, 86.52)	86.00 (85.84, 86.04, 86.12)
Pets, 37 classes				
1	842	6.04 (6.13, 5.45, 6.54)	**6.37** (5.83, 6.41, 6.87)	4.42* (5.15, 4.69, 3.43)
2	1,647	12.53 (12.65, 13.63, 11.31)	**21.00** (20.09, 21.59, 21.31)	14.02 (13.93, 12.62, 15.51)
3	2,452	23.38 (18.23, 25.16, 26.74)	**42.14** (37.42, 41.40, 47.59)	29.22 (30.36, 25.62, 31.67)
4	3,257	38.28 (34.51, 42.87, 37.45)	**57.28** (57.10, 55.16, 59.58)	46.83 (45.71, 47.10, 47.70)
8	6,477	83.27 (81.77, 82.67, 85.39)	80.21 (79.23, 82.23, 79.18)	**83.72** (83.51, 84.14, 83.51)
SVHN, 10 classes				
1	788	12.20* (14.70, 6.28, 15.60)	**21.96** (23.59, 23.64, 18.66)	21.87 (25.57, 19.05, 20.97)
2	1,566	27.13 (35.11, 26.12, 20.16)	**28.47** (25.07, 27.14, 33.21)	27.21 (25.86, 28.57, 27.19)
3	2,344	32.64 (35.41, 32.06, 30.45)	**35.24** (33.92, 34.85, 36.95)	27.85* (25.76, 26.43, 31.37)
4	3,122	**36.38** (38.06, 33.74, 37.35)	36.31 (36.70, 35.53, 36.69)	28.67 (28.63, 26.13, 31.27)
8	6,234	**44.85** (47.75, 44.44, 42.36)	40.56 (39.70, 41.31, 40.66)	39.69* (39.01, 39.12, 40.94)

In the ladder (Figure 7 and Table 51), PSNeuron has a higher mean than the GELU head on 4, 5, 5, 3, and 1 of the five datasets at $h = 1$, 2, 3, 4, and 8, and than the additive head on 3, 5, 5, 5, and 1. At $h = 2$ and 3 it is ahead of the GELU head on all three seeds in 7 of the 10 combinations of dataset and width and on two of three seeds in the other 3 (SVHN at both widths and ImageNet100 at $h = 3$); against the additive head it wins 13, 15, and 15 of the 15 seed pairs at $h = 2$, 3, and 4. The lead does not survive widening: at $h = 8$ the GELU head is ahead on every dataset except Flowers102 and

the additive head on every dataset except SVHN, so the wide end of the ladder gives no evidence that the product helps. Under the protocol of the five seed study the reversal persists up to $h = 128$ (Section J.10), and wider replacement heads of a different design also stay below GELU heads of the same budget (Section J.2). A linear head on the ImageNet100 features reaches 95.44% under the shorter protocol of the frozen head study (Table 35); it is not a control for the ladder, but it shows that these narrow heads operate far below what the features support.

Five seeds and a wider learning rate grid. A separate study trains the PSNeuron, additive, and GELU heads at $h = 2$, 3, and 4, and the single order and mixture heads at $h = 3$ (lower block of Table 53), with the splits, calibration rows, optimizer, schedule, step counts, and selection rule of the ladder, final seeds 0 to 4, and eight learning rates, 2×10^{-4} to 2.56×10^{-2} in factors of two, shared by every head and never extended; its tuning runs evaluate only the validation split, whereas the tuning runs of the ladder also recorded test accuracy, which its selection never read. The study runs on CPU: the ImageNet100 features are identical to those of the ladder, while the VTAB-1k features are extracted again and differ numerically, so the two studies are reported separately and never pooled. PSNeuron has a higher mean than the additive head in 14 of the 15 combinations of dataset and width and ties on DTD at $h = 4$ (Table 52). It wins 25, 23, and 21 of the 25 seed pairs at $h = 2$, 3, and 4, 69 of 75 in all, by 7.34, 7.36, and 5.70 points on average; the nominal paired interval lies above zero in 11 combinations and below zero in none, and 8 stay above zero after a Bonferroni correction over the 15 combinations. The largest rate is selected in 32 of the 60 settings, by 8, 9, and 11 of the 15 PSNeuron, additive, and GELU heads; the three combinations with both heads inside the grid and a nominal interval above zero, Pets at $h = 4$ and SVHN at $h = 3$ and 4, do not keep it after the correction, so the comparison with the additive head remains descriptive. Against the GELU head, which has h more parameters, PSNeuron has the higher mean in 10 combinations and the lower mean in 5, ImageNet100 at every width and DTD and Pets at $h = 4$; after the correction 6 combinations favor PSNeuron and 3 the GELU head, ImageNet100 at $h = 3$ and 4 and DTD at $h = 4$. On ImageNet100, whose features the two studies share, PSNeuron is ahead of the GELU head on validation at $h = 3$ at 3.2×10^{-3}, the largest rate of the ladder (51.56 against 48.70%), but the GELU head keeps improving up to the largest rate of this study (58.57%), which suggests that the ordering of the ladder on this dataset reflects its shorter grid.

Table 52: Five seed study of the narrow heads. Accuracy columns give the mean test accuracy over seeds 0 to 4, in percent; higher is better; †the head selected the largest rate of the grid, 2.56×10^{-2}. Difference columns give PSNeuron minus the control, averaged over paired seeds, with a nominal two sided 95% Student t interval, not corrected for the 15 combinations, and the number of the five seeds on which PSNeuron is ahead. The additive head has exactly the parameters of PSNeuron, and the GELU head has h more.

h	dataset	GELU	PSNeuron	additive	minus additive	minus GELU
2	DTD	13.46†	17.93†	11.98†	5.95 [4.45, 7.45], 5	4.47 [2.37, 6.57], 5
2	Flowers102	7.71†	15.48†	4.53	10.95 [8.96, 12.94], 5	7.77 [5.41, 10.13], 5
2	ImageNet100	28.32†	27.48	23.22†	4.25 [2.64, 5.86], 5	−0.84 [−3.06, 1.37], 2
2	Pets	21.66†	33.16†	21.72†	11.44 [8.32, 14.56], 5	11.50 [7.55, 15.45], 5
2	SVHN	27.83	29.52	25.41	4.11 [−0.19, 8.41], 5	1.69 [−8.05, 11.43], 2
3	DTD	25.52†	26.80†	19.91†	6.88 [5.20, 8.57], 5	1.28 [−2.78, 5.34], 3
3	Flowers102	22.23†	32.79†	21.08†	11.71 [9.18, 14.23], 5	10.55 [6.35, 14.76], 5
3	ImageNet100	56.40†	48.27	46.67†	1.60 [−0.79, 3.99], 3	−8.14 [−10.57, −5.70], 0
3	Pets	45.04†	51.11†	41.55†	9.56 [6.46, 12.67], 5	6.07 [5.15, 6.99], 5
3	SVHN	30.39	33.82	26.75	7.06 [2.39, 11.74], 5	3.42 [−8.48, 15.33], 3
4	DTD	37.82†	29.97	29.97†	0.00 [−0.96, 0.96], 2	−7.85 [−8.95, −6.75], 0
4	Flowers102	42.14†	49.09†	36.49†	12.60 [10.28, 14.91], 5	6.95 [5.04, 8.86], 5
4	ImageNet100	72.37†	65.60†	64.26	1.34 [−0.23, 2.92], 4	−6.77 [−7.88, −5.65], 0
4	Pets	62.58	62.53	55.22	7.31 [1.58, 13.04], 5	−0.05 [−5.94, 5.83], 2
4	SVHN	29.52	37.63	30.38	7.25 [3.16, 11.35], 5	8.12 [5.18, 11.05], 5

Matched structured controls. Table 53 asks whether another structured head with the parameters of the PSNeuron head does as well. With $K = 2$ the PSNeuron head is itself the single order 2 head, so the controls are single order heads of orders 3 and 4, the fixed equal mixture of orders 2 to 4, and the additive head. At $h = 3$ with three seeds (Tables 51 and 53), PSNeuron has the highest mean of the six heads on all five datasets, and the single order 3 head, the strongest structured control, beats the GELU head on DTD, Flowers102, and Pets. At $h = 2$, 4, and 8 the single order 3 and 4 heads

and the mixture stay below PSNeuron on every dataset except the single order 3 head on SVHN at $h = 2$, so the reversal at $h = 8$ comes from the GELU and additive heads alone. At $h = 1$ each of these three heads is below PSNeuron on only two datasets, and on each of DTD, Flowers102, ImageNet100, and Pets the highest of the six means belongs to one of them. With five seeds at $h = 3$, PSNeuron again has the highest mean of the five heads of its size on every dataset and wins all 25 seed pairs against each of the single order 3 and 4 heads and the mixture, but the GELU head is highest on ImageNet100 and the single order 3 head beats the GELU head only on Flowers102, so these controls narrow the observation rather than identify a mechanism. The additive head asks what multiplication adds over a sum of the same slots; whether the grouped quadratic form does better than a generic product is the question of Section J.10.

Table 53: Matched structured controls: single order heads of orders 3 and 4 and the fixed equal mixture of orders 2 to 4, each with the parameters of the PSNeuron head, h fewer than the GELU head. Entries are mean test accuracy, in percent; higher is better. The upper block is the three seed ladder, whose GELU, PSNeuron, and additive heads are in Table 51; at $h = 3$ every head of the ladder selected the largest learning rate of the grid except the single order 4 and additive heads on SVHN. The lower block is the five seed study, whose other heads are in Table 52; †the head selected the largest rate, 2.56×10^{-2}.

h	head	DTD	Flowers102	ImageNet100	Pets	SVHN
three seeds						
1	single order 3	4.50	2.31	6.40	6.77	18.79
1	single order 4	3.05	3.34	6.35	6.93	20.78
1	mixture	4.47	2.35	6.31	7.77	18.78
2	single order 3	9.57	8.37	23.99	17.69	28.69
2	single order 4	9.10	5.18	21.95	16.09	23.51
2	mixture	9.66	7.24	23.88	17.85	24.49
3	single order 3	16.86	20.84	44.09	38.80	32.52
3	single order 4	12.59	18.23	37.82	29.49	31.38
3	mixture	12.34	16.82	37.71	29.14	30.29
4	single order 3	23.09	30.71	57.40	49.54	30.42
4	single order 4	16.52	26.60	51.03	38.54	22.54
4	mixture	19.41	24.36	49.88	40.96	29.37
8	single order 3	34.27	65.58	80.19	66.66	37.64
8	single order 4	25.80	53.90	74.00	56.09	29.41
8	mixture	29.04	54.02	72.67	62.80	33.09
five seeds, wider grid						
3	single order 3	19.47†	26.85†	44.21	40.87†	29.78
3	single order 4	12.82	20.52	40.17†	30.88	26.87
3	mixture	15.39	21.62	38.00	33.95	26.64

Wider heads. A registered extension of the five seed study trains the PSNeuron, additive, and GELU heads at $h = 8$, 16, 32, 64, and 128 with its protocol, grid, and seeds, and reruns $h = 4$ as a bridge that reproduces Table 52 bit for bit (Table 54). The lead of the narrow heads does not extend to these widths. PSNeuron minus GELU lies below zero in all 25 settings, by 0.51 to 14.48 points, and PSNeuron minus additive in 21, by 0.38 to 10.98 points; the other four, ImageNet100 and SVHN at $h = 8$ and 16, are unresolved, and no interval favors PSNeuron, which wins 14 of 125 seed pairs against the additive head and none against GELU. Only the comparison with the additive head at $h = 8$ is bracketed on every dataset, and there the additive head leads on DTD, Flowers102, and Pets. A linear probe trained under the same protocol, a scale reference rather than a control, reaches 78.20, 99.61, 95.58, 94.45, and 50.91% on DTD, Flowers102, ImageNet100, Pets, and SVHN, the ImageNet100 value at the lowest rate of its grid; its paired interval places every PSNeuron and additive head at $h \geq 8$ below it and the GELU heads below it in 19 of 25 settings; the other 6 are unresolved, and in two of them the GELU mean is at least that of the probe.

A generic product. A registered comparison asks whether the grouped quadratic form of the PSNeuron head does better than a generic product, with 405 runs in all: bilinear heads $u = (A_1 x) \odot (A_2 x)$ with $A_1, A_2 \in \mathbb{R}^{m \times 768}$, m product units, and class scores $\ell = Bu + b$ with $B \in \mathbb{R}^{C \times m}$, trained with the protocol, grid, and seeds of the five seed study and paired with its PSNeuron heads, whose values a rerun of seeds 0 and 4 of every head and width of that study reproduces bit for bit (Table 55). The bilinear heads have no biases inside the product and no calibration, and their maps start at the default initialization of linear layers, as in the GELU control, so their factors are not

Table 54: Wider heads on frozen DINOv3 features with $K = 2$, five seeds under the protocol of Table 52: mean test accuracy in percent, and PSNeuron minus control in points with a nominal paired 95% interval. The PSNeuron and additive heads have $h(768 + C) + C$ parameters with $C = 47$, 102, 100, 37, and 10 classes, the GELU head h more. †A learning rate selected at an edge of the grid. A linear probe under the same protocol, a scale reference and not a control, reaches 78.20, 99.61, 95.58†, 94.45, and 50.91% on DTD, Flowers102, ImageNet100, Pets, and SVHN.

h	dataset	PSNeuron	additive	GELU	minus additive	minus GELU
8	DTD	45.97	53.37	60.45	−7.40 [−10.78, −4.03]	−14.48 [−15.70, −13.26]
	Flowers102	75.42	84.18	89.08	−8.76 [−12.98, −4.53]	−13.66 [−17.20, −10.12]
	ImageNet100	86.37	86.20	89.06†	+0.16 [−0.70, +1.02]	−2.69 [−3.23, −2.15]
	Pets	78.84	84.67	87.82	−5.82 [−8.86, −2.78]	−8.98 [−12.31, −5.65]
	SVHN	41.35	38.48	45.74	+2.87 [−1.52, +7.25]	−4.39 [−7.66, −1.12]
16	DTD	56.77	67.74	71.01	−10.98 [−13.00, −8.95]	−14.24 [−15.23, −13.26]
	Flowers102	91.27	96.29	98.00	−5.02 [−6.60, −3.44]	−6.72 [−8.46, −4.99]
	ImageNet100	90.88†	90.71	92.56	+0.17 [−0.33, +0.67]	−1.68 [−2.17, −1.19]
	Pets	87.41	90.86	92.52	−3.45 [−4.22, −2.68]	−5.11 [−6.18, −4.05]
	SVHN	44.98	44.95	49.20	+0.04 [−1.55, +1.62]	−4.21 [−6.09, −2.33]
32	DTD	65.40	73.15	74.90	−7.74 [−8.74, −6.75]	−9.50 [−10.81, −8.19]
	Flowers102	96.22	98.90	99.29	−2.68 [−3.15, −2.21]	−3.07 [−3.42, −2.73]
	ImageNet100	92.16†	93.25†	94.55†	−1.09 [−1.86, −0.32]	−2.40 [−2.87, −1.92]
	Pets	90.89	92.81	93.76	−1.92 [−2.69, −1.16]	−2.88 [−3.69, −2.06]
	SVHN	45.21	47.65	50.20	−2.44 [−3.63, −1.25]	−4.99 [−5.55, −4.43]
64	DTD	70.52	74.70	75.65	−4.18 [−5.81, −2.55]	−5.13 [−6.61, −3.64]
	Flowers102	98.47	99.17	99.40	−0.70 [−0.97, −0.44]	−0.93 [−1.02, −0.84]
	ImageNet100	93.08†	94.17	95.27†	−1.09 [−1.33, −0.85]	−2.19 [−2.50, −1.89]
	Pets	92.07	93.46	94.33	−1.39 [−1.88, −0.90]	−2.26 [−2.48, −2.04]
	SVHN	47.07	48.38	50.91	−1.30 [−2.27, −0.34]	−3.83 [−4.70, −2.97]
128	DTD	72.88	74.88	77.47	−2.00 [−2.73, −1.27]	−4.59 [−5.81, −3.36]
	Flowers102	99.07	99.45	99.58	−0.38 [−0.54, −0.22]	−0.51 [−0.57, −0.46]
	ImageNet100	94.30†	95.07†	95.41†	−0.77 [−1.33, −0.22]	−1.11 [−1.54, −0.69]
	Pets	93.20	93.97	94.47	−0.76 [−1.06, −0.46]	−1.27 [−1.69, −0.85]
	SVHN	47.51	48.54	50.37	−1.03 [−1.81, −0.25]	−2.86 [−3.52, −2.20]

rescaled to unit root mean square as the PSNeuron slots are. No bilinear width matches a PSNeuron head in parameters, so each PSNeuron head is compared with the largest bilinear head of no more parameters, which has fewer product units, and with the smallest of more parameters, 25 to 98% larger. PSNeuron has the higher mean against the smaller bilinear heads in all 15 settings and a nominal interval above zero in 13, 12 after a Bonferroni correction, a lead that their lower readout rank alone can explain; on Flowers102 the bilinear head with $m = 1$ stays near chance. Against the larger heads the intervals favor PSNeuron in 6 settings and the bilinear head in 6, with 3 unresolved (3, 2, and 10 after correction); the Pets interval at $h = 2$ misses favoring the bilinear head by 0.007 points. With as many product units as PSNeuron, $m = h$ at $h = 2$ and 3 with 83 to 98% more parameters, the bilinear head has the higher mean in all 10 settings and an interval above PSNeuron in 8, 3 after correction. Nine of the fifteen bilinear heads select the upper edge of the grid, among them the head with $m = 2$ on all five datasets, and on DTD and Pets the heads with $m = 3$ tie exactly with that edge on validation and take the smaller rate. The PSNeuron head lies at an edge in 8 of the 15 settings; no setting reaches the registered label of either a grouped structure advantage or a generic product advantage, so the registered reading reaches no conclusion for the family and is inconclusive. A generic bilinear head therefore neither isolates nor refutes the grouped structure of the PSNeuron head; the comparison with the additive head, which this study leaves unchanged, does not by itself show that the grouping is what helps.

K Matched Controls and Preprocessing Sensitivity

These studies use the same hosts, splits, final test evaluation rule, and seed identities as their stated comparators, except the fresh seed holdout of Section K.2, and each fixes its learning rates on validation before the final test evaluations. They are separate cohorts: the original results remain in Tables 1 and 32, and outcomes of these studies are never substituted into those tables. Complete result records, paired seed measurements, and acceptance audits are identified in the evidence manifest. The vendored records carry the source commit and each study's execution commit and result path, and the accepted audits recomputed every reported test quantity and validation selection from them. These cohorts ran in dedicated GPU jobs rather than the planned overlapping steps in existing allocations.

Table 55: PSNeuron heads with $K = 2$ against generic bilinear heads $u = (A_1x) \odot (A_2x)$, class scores $Bu + b$, with m product units and $2 \cdot 768m + (m + 1)C$ parameters, against $h(768 + C) + C$ for PSNeuron; frozen DINOv3 features, five seeds, protocol of Table 52. Upper block: mean test accuracy in percent; †a learning rate selected at an edge of the grid. Lower block: PSNeuron minus bilinear head in points with a nominal paired 95% interval; *the interval also excludes zero after a Bonferroni correction. No bilinear width matches a PSNeuron head in parameters; the head with fewer has $m = 1$ at $h = 2$ and 3 and $m = 2$ at $h = 4$, the head with more has $m = 2$ at $h = 2$ and 3 and $m = 3$ at $h = 4$, and the last column compares $h = 3$ with $m = 3$, as many product units.

head	width	DTD	Flowers102	ImageNet100	Pets	SVHN
PSNeuron	$h = 2$	17.93†	15.48†	27.48	33.16†	29.52
PSNeuron	$h = 3$	26.80†	32.79†	48.27	51.11†	33.82
PSNeuron	$h = 4$	29.97	49.09†	65.60†	62.53	37.63
bilinear	$m = 1$	7.68	1.54	10.54†	12.51†	27.90
bilinear	$m = 2$	22.19†	17.39†	38.96†	39.61†	35.58†
bilinear	$m = 3$	34.05	37.58	68.23†	56.60	41.06†

dataset	h	fewer parameters	more parameters	$m = 3$
DTD	2	$+10.24$ $[+8.57, +11.92]^{*}$	-4.27 $[-6.87, -1.66]$	
	3	$+19.12$ $[+17.17, +21.07]^{*}$	$+4.61$ $[+1.97, +7.25]$	-7.26 $[-9.05, -5.46]^{*}$
	4	$+7.78$ $[+4.72, +10.84]^{*}$	-4.09 $[-5.74, -2.43]^{*}$	
Flowers102	2	$+13.94$ $[+10.60, +17.28]^{*}$	-1.91 $[-4.02, +0.21]$	
	3	$+31.25$ $[+28.40, +34.10]^{*}$	$+15.40$ $[+12.20, +18.60]^{*}$	-4.79 $[-8.34, -1.24]$
	4	$+31.70$ $[+29.10, +34.30]^{*}$	$+11.51$ $[+8.19, +14.84]^{*}$	
ImageNet100	2	$+16.93$ $[+14.91, +18.95]^{*}$	-11.49 $[-15.25, -7.73]^{*}$	
	3	$+37.72$ $[+35.84, +39.60]^{*}$	$+9.30$ $[+5.46, +13.14]^{*}$	-19.96 $[-22.99, -16.93]^{*}$
	4	$+26.64$ $[+23.98, +29.30]^{*}$	-2.62 $[-4.55, -0.70]$	
Pets	2	$+20.65$ $[+18.92, +22.39]^{*}$	-6.44 $[-12.89, +0.01]$	
	3	$+38.60$ $[+36.26, +40.95]^{*}$	$+11.51$ $[+5.65, +17.36]$	-5.49 $[-8.75, -2.23]$
	4	$+22.92$ $[+17.93, +27.91]^{*}$	$+5.93$ $[+2.74, +9.11]$	
SVHN	2	$+1.62$ $[-1.21, +4.46]$	-6.07 $[-9.87, -2.26]$	
	3	$+5.92$ $[+3.20, +8.64]$	-1.77 $[-4.79, +1.25]$	-7.24 $[-12.04, -2.45]$
	4	$+2.05$ $[-2.54, +6.64]$	-3.43 $[-6.42, -0.43]$	

Some executions used a per job GPU sharing daemon, and the dense mixer study combined time sliced and shared executions, so trainer wall times are not compared. Interrupted tuning starts on HIGGS Small and on the SwiGLU mixer were restarted before a result existed.

K.1 Dense mixers with the same recurrent inputs

The GELU and SwiGLU controls and the conditional GELU structure control receive the same state readout r_t and normalized input u_t as the joint $K = 4$ PSNeuron mixer of the selected layout, with 4,400 site parameters (Section G), in the same four layer DeltaNet. The GELU control is the network $256 \to 11 \to 128$ with biases of Section E (4,363 parameters), and the SwiGLU control is bias free with hidden width 7 and no input projection (4,480 parameters), unlike the SwiGLU mixers described there. Every model trains for 4,000 steps at batch 64 and selects its learning rate on validation with the tuning seeds over seven rates from 10^{-4} to 6.4×10^{-3}, with ties going to the smaller rate. For the structure control, $z_t = [A_s r_t; A_x u_t]$ and $C(z_t) = W_2\,\mathrm{GELU}(W_1 z_t)$, with bias free $A_s, A_x \in \mathbb{R}^{8\times128}$, $W_1 \in \mathbb{R}^{16\times16}$, and $W_2 \in \mathbb{R}^{128\times16}$, so its 4,352 parameters split into 2,048 conditioner, 256 hidden map, and 2,048 output map parameters. Its two conditioners are initialized from copies of the PSNeuron conditioners at the same seed and then trained independently. This matches state and input access, conditioner capacity, and initialization while changing the aggregation. Every result in Table 56 comes from the common chunked kernel.

In this earlier cohort of seeds 0 to 4, the product has higher MQAR point estimates than GELU and SwiGLU, but neither interval excludes zero. The PSNeuron and GELU runs on MQAR exactly reproduce archived runs, so they are not independent replications, and the PSNeuron layout was chosen on validation of the same seed identities (Section G). The nominal interval against conditional GELU excludes zero, but that control learns MQAR on only one of five seeds, the outcomes are bimodal, a Bonferroni correction over the three MQAR controls widens the interval to $[-10.8, 114.0]$ points, and an additional distribution free sign flip check, which was not the registered decision rule, gives $p = 0.125$. On multi-needle retrieval the SwiGLU and conditional GELU controls lie within

one point of the product with intervals that contain zero, the conditional GELU control higher on all five seeds, and the GELU control has two partially failed seeds and an unresolved interval. On state tracking every interval is unresolved, so the higher conditional GELU mean does not establish a control advantage; that control selected the lowest rate tried, and smaller rates were not tested. The chunked kernel changes the PSNeuron selection and its recall relative to the sequential kernel result of Table 32, from $80.18 \pm 19.60\%$ to $65.06 \pm 39.47\%$ on state tracking and from $99.61 \pm 0.17\%$ to $99.16 \pm 0.92\%$ on multi-needle retrieval, so the kernel cohorts must not be pooled. Two PSNeuron tuning runs on MQAR at the highest rate, 6.4×10^{-3}, ended with nonfinite values; they remain in the records and are ineligible for selection. On MQAR the fresh seed holdout below replaces this five seed cohort as the main evidence and is never pooled with it.

Table 56: Same input DeltaNet controls. Test recall in percent, mean and sample standard deviation over five seeds; difference is PSNeuron minus control with a nominal paired 95% interval. All mixers receive the state readout and input.

task	control (site parameters)	PSNeuron	control	difference [interval]
MQAR	GELU 4,363	77.98 ± 23.87	42.38 ± 47.87	$+35.60$ $[-38.40,\ +109.60]$
MQAR	SwiGLU 4,480	77.98 ± 23.87	46.54 ± 46.92	$+31.44$ $[-31.83,\ +94.71]$
MQAR	conditional GELU 4,352	77.98 ± 23.87	26.39 ± 39.88	$+51.60$ $[+7.86,\ +95.33]$
Multi-needle	GELU 4,363	99.16 ± 0.92	89.42 ± 14.24	$+9.73$ $[-8.20,\ +27.66]$
Multi-needle	SwiGLU 4,480	99.16 ± 0.92	99.46 ± 0.31	-0.30 $[-1.17,\ +0.56]$
Multi-needle	conditional GELU 4,352	99.16 ± 0.92	99.78 ± 0.15	-0.62 $[-1.58,\ +0.33]$
State tracking	GELU 4,363	65.06 ± 39.47	66.76 ± 39.02	-1.70 $[-88.91,\ +85.51]$
State tracking	SwiGLU 4,480	65.06 ± 39.47	73.05 ± 23.41	-7.99 $[-60.63,\ +44.65]$
State tracking	conditional GELU 4,352	65.06 ± 39.47	92.56 ± 3.31	-27.50 $[-74.98,\ +19.98]$

K.2 Fresh seed holdout of the DeltaNet mixer

The comparisons above reuse the seed identities on which the joint layout was chosen (Section G). A registered holdout therefore trains every mixer on 20 fresh seeds, 1000 to 1019, disjoint from every earlier DeltaNet record, in the same four layer DeltaNet with $d_{\text{model}} = 128$, on token MQAR with the archived split of 8,192, 1,024, and 1,024 sequences (Table 57). Every mixer occupies the channel mixer of the last block, and the 739,780 host parameters outside it are identical bit for bit across mixers at each seed. The architectures and learning rates are frozen from the earlier validation selection. One control is new: an exact additive twin that copies the conditioners, factor bank, slot partition, bank rescale, and readout of the PSNeuron mixer at the same seed and replaces the product by $s_r = \sum_g \mathrm{GELU}(q_{r,g})$ over the same slots, so that its parameters differ only by the 48 order coefficients; the twin and the conditional GELU control therefore have 48 fewer parameters than PSNeuron, 1.09% of the site. The twin selected its rate, 1.6×10^{-3}, on validation with the tuning seeds 101 and 102, inside its grid. The registered primary comparisons are PSNeuron against the twin and against the conditional GELU control, decided by an exhaustive paired sign flip test over all 2^{20} sign assignments with a Holm correction over the two; GELU, SwiGLU, and the cross term layout are secondary and uncorrected, and the 4,608 parameter additive control of Section J.2 is context. Reruns of 30 archived runs reproduced their values bit for bit.

The joint $K = 4$ mixer reaches $68.22 \pm 31.84\%$ recall, below the $77.98 \pm 23.87\%$ of the archived seeds 0 to 4, and exceeds the twin, at $24.15 \pm 31.96\%$, by 44.08 points (nominal interval $[22.39, 65.77]$, ahead on 16 of 20 seeds, raw $p = 9.86 \times 10^{-4}$, Holm adjusted $p = 0.0020$), and the conditional GELU control, at $30.84 \pm 37.39\%$, by 37.38 points ($[15.69, 59.08]$, 14 of 20, Holm adjusted $p = 0.0027$). The GELU comparison is positive only at the nominal level, as is the lead of 51.06 points over the 4,608 parameter additive control, and the SwiGLU comparison is unresolved ($+16.75$ points, $[-5.62, 39.12]$). The joint layout does not lead the cross term layout on these seeds (-1.14 points, $[-22.15, 19.86]$), whereas the archived seeds 0 to 4 gave $+47.87$, $[8.96, 86.77]$; the earlier layout advantage therefore does not replicate, and the two cohorts are reported side by side and never pooled. Outcomes remain bimodal: 8 PSNeuron seeds solve the task and 6 stall, against 3 and 17 for the twin and 5 and 15 for the conditional GELU control, and two PSNeuron seeds end near chance. The conditional GELU comparison rests on the size of the gains rather than on how often the task is solved: an exact sign test gives $p = 0.115$, and the seeds solved by only one of the two number 7 against 4. The rates of PSNeuron, the twin, the conditional GELU control, and the GELU control each rest on one of the two tuning seeds solving the task; the SwiGLU and additive rates were chosen among stalled tuning runs, and only the cross term rate is supported by both seeds. The holdout is an

attribution result at nearly equal parameters and supports no size comparison; it establishes neither an advantage over SwiGLU, nor reliable solving of MQAR, nor an advantage of the joint layout.

Table 57: Fresh seed MQAR holdout of the DeltaNet mixer: 20 training seeds, 1000 to 1019, disjoint from every earlier DeltaNet record, with the architecture and learning rates frozen from the earlier selection; the additive twin, new here, selected its rate on validation with the tuning seeds. Recall: test recall@1 in percent, mean $\pm$ seed standard deviation over the 20 seeds, higher is better; solved/stalled: seeds classified as solving and as stalled by the registered analysis, the rest lying between. Difference: PSNeuron minus the row, mean over paired seeds with its nominal 95% t interval, in points; seeds: PSNeuron ahead and behind. The registered test is an exhaustive paired sign flip test over all 2^{20} sign assignments for the twin and the conditional GELU control, with raw $p = 9.86 \times 10^{-4}$ and 2.69×10^{-3}, Holm corrected over the two; the other rows are uncorrected and read as nominal or unresolved. This cohort is never pooled with the earlier five seed cohort of Tables 32 and 56.

mixer	site params	recall	solved/stalled	difference	seeds	Holm p
PSNeuron, joint $K = 4$	4,400	68.22 ± 31.84	8/6			
additive twin on the same slots	4,352	24.15 ± 31.96	3/17	$+44.08$ $[+22.39, +65.77]$	16/4	0.0020
conditional GELU	4,352	30.84 ± 37.39	5/15	$+37.38$ $[+15.69, +59.08]$	14/6	0.0027
GELU	4,363	25.72 ± 37.08	4/16	$+42.50$ $[+23.05, +61.95]$	14/6	nominal
SwiGLU	4,480	51.48 ± 42.47	8/10	$+16.75$ $[-5.62, +39.12]$	11/9	unresolved
PSNeuron, cross terms, $K = 4$	4,400	69.37 ± 37.66	11/6	-1.14 $[-22.15, +19.86]$	8/12	unresolved
additive control of Section J.2	4,608	17.16 ± 27.99	2/18	$+51.06$ $[+32.76, +69.37]$	16/4	nominal

K.3 The common tabular replacement on a bracketed grid

On each of HIGGS Small, Jannis, and Helena, PSNeuron, same budget GELU, the native GELU host, the validation selected dense reference (the native host on Helena), and a smaller SwiGLU network share a fixed host and eight learning rates from 2×10^{-4} to 2.56×10^{-2}, and every selected rate is interior on this grid. HIGGS Small and Helena PSNeuron final results, and the native hosts on all three tasks, reproduce the original cohort exactly, while Jannis chooses another rate and obtains $70.23 \pm 0.21\%$ here rather than $70.07 \pm 0.27\%$ there (Table 1).

On HIGGS Small and Jannis the same budget GELU difference is unresolved; on Helena GELU is higher by 0.30 points, with PSNeuron minus GELU interval $[-0.39, -0.22]$. On Jannis, PSNeuron is smaller than the native GELU host and more accurate (Table 58), but the same budget GELU is also more accurate than that host, by 0.66 points with interval $[0.43, 0.90]$. The PSNeuron interval against the Jannis host does not survive a Bonferroni correction over the four controls of the task, which widens it to $[-0.01, 1.11]$ points. This is nominal evidence for a compact replacement on Jannis under the transductive preprocessing, without attributing the host gain to the grouped product; its interval against the host becomes unresolved when the preprocessing is fitted on the training split (Table 59). The Jannis PSNeuron learning rate beats the grid edge by only 7 correct validation predictions out of $20{,}000$, summed over the two tuning seeds. The host selects 4×10^{-4}, one rate above the lower edge of the grid, and its validation profile is bimodal: the upper edge 2.56×10^{-2} is only 36 predictions behind. The Jannis SwiGLU reference ties exactly on validation, with $14{,}002$ of $20{,}000$ correct predictions at both 1.28×10^{-2} and the grid edge 2.56×10^{-2}, and takes the smaller rate under a tie rule fixed in advance. None of these fragile selections is hidden by the bracketed label.

K.4 Preprocessing fitted on the training split

This sensitivity study refits the categorical vocabulary, imputation, and standardization of Section G on the training split only, keeps the split indices and the eight rate grid, and retunes on validation (Table 59). The same budget GELU reading does not change: unresolved on HIGGS Small and Jannis and favoring GELU on Helena, so the Helena finding persists. On HIGGS Small the comparison with the native host changes from unresolved to a positive interval; on Jannis the positive host interval becomes unresolved and the SwiGLU reference has a negative interval. These changed comparisons rely on fragile selections, including a grid edge for Jannis PSNeuron. Because the comparison includes retuning and numerical variation, the changes cannot be assigned solely to preprocessing; they do not establish a new HIGGS Small parameter efficiency result.

Table 58: Bracketed $K = 4$ tabular study. Test accuracy in percent, mean and sample standard deviation over five seeds; paired difference is PSNeuron minus control with a nominal 95% interval. The site has $2{,}072$ parameters.

task	control (site parameters)	PSNeuron	control	difference [interval]
HIGGS Small	GELU at site budget $2{,}064$	70.91 ± 0.62	71.02 ± 0.43	-0.11 $[-0.46, +0.24]$
HIGGS Small	native GELU $16{,}576$	70.91 ± 0.62	71.24 ± 0.37	-0.33 $[-0.93, +0.27]$
HIGGS Small	SwiGLU reference $24{,}576$	70.91 ± 0.62	71.27 ± 0.34	-0.36 $[-1.04, +0.33]$
HIGGS Small	small SwiGLU 544	70.91 ± 0.62	70.84 ± 0.41	$+0.08$ $[-0.43, +0.58]$
Jannis	GELU at site budget $2{,}064$	70.23 ± 0.21	70.34 ± 0.22	-0.11 $[-0.41, +0.18]$
Jannis	native GELU $16{,}576$	70.23 ± 0.21	69.68 ± 0.19	$+0.55$ $[+0.19, +0.91]$
Jannis	SwiGLU reference $24{,}576$	70.23 ± 0.21	70.51 ± 0.29	-0.28 $[-0.63, +0.06]$
Jannis	small SwiGLU 544	70.23 ± 0.21	70.20 ± 0.22	$+0.03$ $[-0.31, +0.37]$
Helena	GELU at site budget $2{,}064$	35.65 ± 0.33	35.95 ± 0.32	-0.30 $[-0.39, -0.22]$
Helena	native GELU $16{,}576$	35.65 ± 0.33	36.09 ± 0.31	-0.44 $[-0.84, -0.03]$
Helena	small SwiGLU 544	35.65 ± 0.33	35.88 ± 0.19	-0.23 $[-0.51, +0.05]$

Table 59: Preprocessing sensitivity. PSNeuron minus control test accuracy in percentage points, with nominal paired 95% intervals over five seeds. Both regimes retune learning rates on validation.

task	control	full feature pool	training split only
HIGGS Small	same budget GELU	-0.11 $[-0.46, +0.24]$	$+0.25$ $[-0.30, +0.81]$
HIGGS Small	native GELU	-0.33 $[-0.93, +0.27]$	$+0.32$ $[+0.03, +0.61]$
HIGGS Small	SwiGLU reference	-0.36 $[-1.04, +0.33]$	-0.09 $[-0.57, +0.40]$
Jannis	same budget GELU	-0.11 $[-0.41, +0.18]$	$+0.11$ $[-0.67, +0.89]$
Jannis	native GELU	$+0.55$ $[+0.19, +0.91]$	$+0.24$ $[-0.37, +0.85]$
Jannis	SwiGLU reference	-0.28 $[-0.63, +0.06]$	-0.55 $[-1.07, -0.02]$
Helena	same budget GELU	-0.30 $[-0.39, -0.22]$	-0.32 $[-0.35, -0.29]$
Helena	native GELU	-0.44 $[-0.84, -0.03]$	-0.44 $[-0.81, -0.06]$

K.5 Shared degree singleton geometry

A vector output shared degree HOFM style control applies orders 2 through 4 to the 64 individual input coordinates, with a shared factor bank and per basis order coefficients, where PSNeuron uses 8 grouped slots on the same host. Both sites have exactly $2{,}072$ parameters, both whole models have identical parameter counts, and both use the same initialization, calibration policy, validation grid, and paired final seeds. This isolates the slot partition at this $K = 4$ operating point; the control is a neural site construction rather than a standard HOFM implementation with independent per degree factors. The comparison leaves the other claims of Sections 3 and 5.3 untested: replacement of a dense module at a site, nesting of the fixed geometry across orders, order interventions within a trained model, and the separate representation, recovery, and functional use claims.

Table 60 shows no resolved accuracy difference on any of the three tasks at $n = 5$; these intervals establish neither equivalence nor superiority for either geometry. The PSNeuron measurements exactly repeat those of the bracketed tabular study and are not additional independent seeds. As implemented, the singleton recurrence dispatches more operations and takes 3.8 to 4.0 times as long for an entire training step at batch 256; this is a local implementation timing under GPU sharing, not a general HOFM cost comparison.

Table 60: Grouped versus singleton geometry at exact parameter parity. Test accuracy in percent, mean and sample standard deviation over five seeds. The difference is grouped minus singleton with a nominal paired 95% interval.

task	grouped PSNeuron	singleton control	difference [interval]
HIGGS Small	70.91 ± 0.62	71.00 ± 0.43	-0.09 $[-0.48, +0.30]$
Jannis	70.23 ± 0.21	70.30 ± 0.15	-0.07 $[-0.50, +0.36]$
Helena	35.65 ± 0.33	35.60 ± 0.42	$+0.05$ $[-0.25, +0.34]$

L Parameter Efficiency and Pareto Comparisons

The definitions behind Section 5.1 separate the parameter saving of a named comparison, the parameters needed to reach a common quality, and Pareto relations, all of which summarize the measured configurations as trained and selected under Section G. An empirical envelope describes

the evaluated configurations, not the smallest model that an unrestricted architecture family could reach.

L.1 Definitions and inference rules

Comparable observations and parameter scope. Two configurations are compared only when task, split, preprocessing, host and placement, training budget, evaluation rule, and checkpoint selection rule agree; such configurations form a stratum of the descriptive tier, which keeps every configuration whose final seeds are all complete, low scores and collapsed runs included. The strict tier keeps only the configurations whose learning rate search stayed inside its grid and whose tuning records agree across seeds (Section 4), and divides the strata further so that each also shares one learning rate grid. It constrains tuning and does not mean that a layout was chosen on independent seeds: 65 of the 66 strict DeltaNet pairs with a nominal positive interval use the joint layout selected on validation of the same seeds (Section G). A family fixes the operator form and K and varies only the measured capacity; the bare EuroSAT heads, which apply the PSNeuron directly to the frozen feature, and the latent heads form separate families, and replacing both tabular feedforward networks is a separate placement. For configuration i, $\bar{m}_i$ is the arithmetic mean of the task metric over the final seeds in its own units, $\bar{q}_i = \sigma \bar{m}_i$ with $\sigma = +1$ for metrics to maximize and -1 for metrics to minimize, and P_i is the exact trainable parameter count. Site, trainable model, and model with its frozen parameters are separate scopes; every learned projection, bias, readout, and replacement site is counted, shared parameters once, and counts of heads on frozen features exclude the backbone, which a smaller head does not shrink. Arithmetic, latency, and memory are separate resources (Section H). The dense set of a comparison holds every complete control of the same task and host: the native model, the references of Table 28, the GELU and SwiGLU networks resized at the same site, the additive controls that sum a GELU of each slot, the low rank score corrections, the LoRA and standard adapters, and, for heads on frozen features, the linear, GELU, and SwiGLU heads. The additive controls keep the slot partition, and most strict smaller and better pairs with a nominal interval above zero are against them.

Savings at a named comparison. For a PSNeuron configuration O and a stated dense model D in the same scope, $r = P_O/P_D$, $C = 1/r$, and $S = 100(1 - r)\%$ (Section 4), computed before rounding; Table 1(a) gives C at site scope and r at whole model scope, so there C differs from $1/r$. Placements that keep the dense path and add a branch beside it have $r \geq 1$ by construction and make no size claim. A saving is always reported with the quality difference and on its own shows neither equal quality nor an advantage over the resized dense networks.

Common quality and common budget. The measured threshold $P_A^{\mathrm{obs}}(q)$, the empirical frontier $F_A^{\mathrm{obs}}(p)$, the efficiency $\eta(q)$ with its saving S_Q, and the budget difference $\Delta_F(p)$ are those of Eq. (8), over the complete configurations of each family, or in Table 62 of every PSNeuron configuration of a stratum, against the dense configurations of the same stratum, without interpolation. Both sides use the same threshold, metric, and scope, and the achieved means are reported with the threshold, since reaching one threshold does not make the true means equal. A threshold that no measured configuration reaches is reported as not reached, without a ratio or extrapolation; when the cheapest measured model already reaches it, the true crossing may lie lower, and sparse grids cannot establish the minimum size of an architecture. Both measured thresholds $P_{\mathrm{PS}}^{\mathrm{obs}}$ and $P_{\mathrm{D}}^{\mathrm{obs}}$ are step functions of q, so η changes in jumps and can cross one when the threshold moves by a small fraction of a seed standard deviation, and a negative Δ_F at one budget does not contradict $\eta > 1$ at another threshold.

Pareto relations. Point estimate dominance and nondominated status against a named set B are

$$\begin{aligned} i \succ j \iff & P_i \leq P_j \ \wedge \ \bar{q}_i \geq \bar{q}_j \ \wedge \ (P_i < P_j \ \vee \ \bar{q}_i > \bar{q}_j), \\ \mathrm{ND}(i; B) = & \mathbf{1}\{\nexists j \in B : j \succ i\}. \end{aligned} \tag{22}$$

Here $\mathbf{1}\{\cdot\}$ is the indicator of a true condition. Exact ties do not dominate. A configuration that no member of B dominates is not thereby better than a reference, and a smaller configuration of lower quality is a tradeoff rather than a dominated point. These relations are computed from seed means and exact parameter counts and are separate from quality comparisons at a different budget and from statistical evidence.

Paired intervals and multiplicity. For a fixed comparison paired by training seed over n seeds, let $q_{O,s}$ and $q_{D,s}$ be the oriented metrics of the two models at seed s, let $d_s = q_{O,s} - q_{D,s}$, and let s_d be the sample standard deviation of the differences; $\bar{d} = n^{-1} \sum_s d_s$. With $t_{\alpha,\nu}$ the α quantile of Student's t distribution with ν degrees of freedom, the nominal two sided interval is

$$[L, U] = \bar{d} \pm t_{0.975,n-1} \frac{s_d}{\sqrt{n}}, \qquad s_d^2 = \frac{1}{n-1} \sum_s (d_s - \bar{d})^2. \tag{23}$$

It describes seed variation for the fixed data split. $L > 0$ supports a nominal quality gain for that pair, and together with $P_O < P_D$ a smaller and better comparison; an interval that contains zero is not evidence of equivalence, and comparisons without seed pairing remain descriptive. The inventories attach intervals only to strict pairs whose seeds are verifiably paired, and the counts in parentheses in Tables 2 and 62 are the pairs whose interval stays above zero after a Bonferroni correction over the comparisons of their stratum. Noninferiority would need an allowable loss ϵ fixed in advance and $L > -\epsilon$; none was fixed, so the paper makes no noninferiority claim. A fixed pair interval is neither a simultaneous band over an envelope nor an interval for η, a ratio of minima selected on test data, and multiplicity correction does not remove that selection.

As a sensitivity check, one family is corrected jointly: every pair of a PSNeuron configuration and a control matched to it in size in advance, with all five final seeds complete on both sides, compared on accuracy, so that membership does not depend on the observed quality. The family holds 315 pairs: 108 pairs of heads on frozen ImageNet100 features and 96 on EuroSAT, 34 pairs at the feedforward site of each of HIGGS Small, Jannis, and Helena, and 9 at the CIFAR100 cross head site. It was fixed at an earlier evidence cutoff and is not extended: configurations completed later, among them the recurrent capacity rungs of Section J.2, lie outside it and enter only the corrections within their strata. The Bonferroni correction at familywise level 0.05 replaces $t_{0.975,4}$ by $t_{1-0.05/630,4}$. Three pairs have a nominal interval above zero, HIGGS Small at $K = 6$ and two Jannis configurations at $K = 2$, each against a SwiGLU network of matched size, and none keeps a positive lower bound after correction; the nominal HIGGS Small interval of 0.13 to 0.83 points, for example, becomes -1.25 to 2.21. Of the 233 pairs with a nominal interval below zero, 113 stay below zero after correction. Other intervals in the appendix are nominal unless a correction is stated with them.

Metrics and reference normalization. The analyzed metrics are accuracy, balanced accuracy, macro F1, ROC AUC, recall at one, token accuracy, negative log likelihood, loss, perplexity, expected calibration error, Brier score, R^2, mean absolute error, and root mean squared error. Each keeps its own direction and units and is averaged over independent training seeds for the fixed split; Pareto relations, thresholds, and budget summaries are computed separately for each, and missing or failed measurements stay visible. Secondary metrics are read with the model and learning rate selected on the primary metric, keep their own comparator, and never count as additional primary wins; loss and NLL, which coincide, count once. Averaging after a transform would change the estimand: for perplexities $(1, 10)$ against $(4, 4)$ the arithmetic means favor the second model and the means of log perplexity the first, so mean perplexity is never replaced by the exponential of mean NLL or by mean log perplexity. A justified normalization acts on these means with the same transform T for both models and the reference:

$$\begin{aligned} z_i = T(\bar{m}_i), \quad G_D = z_D - T(m_{\mathrm{ref}}) > 0, \\ g_N = \frac{z_O - z_D}{G_D}, \quad R_N = 1 + g_N. \end{aligned} \tag{24}$$

R_N is the share of the dense model's gain over the reference that PSNeuron retains; it may exceed one or fall below zero. For the accuracies of Table 61, T is the identity and $m_{\mathrm{ref}} = 1/n_c$ is the expected accuracy of a uniform random classifier over the task's n_c classes ($n_c = 2$, 4, 100, 10, and 100 for HIGGS Small, Jannis, Helena, EuroSAT, and CIFAR100), a reference that follows from the number of classes alone, depends on no model, and does not assume balanced classes. Only accuracies are normalized; every other metric keeps its native units. A ratio is reported only when the dense model's gain over the reference is positive and resolved, that is, when the lower end of the dense model's nominal 95% seed interval lies above the reference, and no constant, absolute value, or change of metric is substituted otherwise. Normalized values carry no confidence interval.

L.2 MEASURED THRESHOLDS, FRONTIERS, AND SELECTED REPLACEMENTS

The thresholds of Table 1(b), whose models Table 61 lists, were chosen after the test means were known, each at or just below the mean of a PSNeuron configuration, so they illustrate the envelopes rather than test targets fixed in advance; learning rates and architectures remain selected on validation. On HIGGS Small at $K = 4$, the saving is 43.65% for thresholds above 71.165 and up to 71.178, 26.54% from above 71.063 to 71.165, and -0.10% from above 71.045 to 71.063; at or below 71.045 the resized SwiGLU network with $19{,}490$ parameters is smaller than every PSNeuron configuration of that order ($\eta = 0.975$). Each HIGGS Small saving holds against every resized GELU and SwiGLU network but only in such a narrow band, and every HIGGS Small model behind the thresholds of Table 1(b), PSNeuron and dense, selected the largest learning rate of its grid, whereas the dense networks with $19{,}962$ and $19{,}490$ parameters behind the lower bands selected interior rates. On Jannis and Helena a resized dense network reaches the threshold with fewer parameters, and at the CIFAR100 feedforward site the smallest dense network is smaller by 420 of 2.4 million parameters and slightly more accurate, a tie in size. The latent EuroSAT head reaches its threshold with fewer trainable parameters than the linear head, which is more accurate. Replacing all six feedforward networks trades accuracy for size (Section 5.4), and the bare EuroSAT head of Table 43 was trained in a separate experiment without a dense head; neither enters a threshold comparison.

Table 61: Models behind the thresholds of Table 1(b): the smallest PSNeuron configuration of the stated order and the smallest dense network whose mean test accuracy over five seeds reaches the threshold q, each with its seed standard deviation. Counts are trainable parameters of the whole model, or of the head for EuroSAT. R_N (Eq. (24), in percent) is the share of the dense network's gain over class chance that the PSNeuron configuration retains. †The selected learning rate is the largest of its grid. The EuroSAT heads are described in Table 35; every other model is a point of Figure 8, and its selected learning rate is in the machine readable records of the supplementary material.

		smallest PSNeuron		smallest dense network			
task, order	q (%)	parameters	accuracy (%)	network	parameters	accuracy (%)	R_N (%)
HIGGS Small, $K = 2$	71.10	19,970†	71.116 ± 0.485	resized SwiGLU	27,202†	71.165 ± 0.237	99.771
HIGGS Small, $K = 4$	71.17	19,982†	71.178 ± 0.447	resized SwiGLU	35,458†	71.241 ± 0.286	99.705
HIGGS Small, $K = 2$	71.20	21,762†	71.230 ± 0.552	resized SwiGLU	35,458†	71.241 ± 0.286	99.951
Jannis, $K = 8$	70.00	23,547†	70.059 ± 0.228	resized SwiGLU	21,578†	70.139 ± 0.176	99.825
Helena, $K = 2$	35.50	26,276	35.637 ± 0.328	resized SwiGLU	25,796	35.882 ± 0.188	99.300
EuroSAT latent head, $K = 4$	95.50	6,376†	95.878 ± 0.228	linear head	7,690	97.207 ± 0.048	98.475
CIFAR100 ViT, $K = 2$	39.90	2,416,612	39.998 ± 0.952	resized GELU	2,416,192	40.146 ± 1.067	99.622

Replacements selected on validation. The common replacement fixes $K = 4$ in advance; ranking all complete replacements of each tabular task by validation accuracy instead selects $K = 6$ on HIGGS Small ($20{,}635$ parameters, $r = 0.581$), $K = 2$ on Jannis ($22{,}788$, $r = 0.611$), and $K = 2$ on Helena ($26{,}276$, $r = 0.628$). Paired by seed with the host, their test accuracy changes by -0.23 points (95% interval -0.83 to 0.37), $+0.44$ points (0.25 to 0.63, higher on all five seeds), and -0.45 points (-0.67 to -0.23); the first two selected the largest learning rate of their grid. Over all replacements, the paired interval against the host lies above zero for 0 of 33 on HIGGS Small, 7 of 32 on Jannis, and 0 of 32 on Helena, and below zero for 7, 8, and 27. The Jannis interval of the selected replacement does not survive a Bonferroni correction over its 32 configurations, and a resized SwiGLU network with $21{,}578$ parameters reaches 70.14%, level with it. These comparisons with the host are nominal and include configurations of the descriptive tier; the inventory attaches intervals only to strict pairs, and none of its tabular pairs has a nominal interval above zero (Table 62).

Secondary metrics. Secondary metrics were recorded for every task: balanced accuracy, Brier score, calibration error, NLL, and macro F1 for HIGGS Small, Jannis, Helena, and the heads and adapters on frozen features, with ROC AUC also for HIGGS Small; Brier score, calibration error, NLL, and macro F1 for CIFAR100; Brier score, calibration error, and NLL for the Boolean cube probes; NLL for the sequence tasks, reported for recurrent MQAR also where recall is worse; perplexity and token accuracy for FineWeb-Edu; and mean absolute and root mean squared error for the teachers and mixed order targets. Three favorable examples show why they are kept separate. The CIFAR100 cross head configuration with $K = 4$ and 995 site parameters has NLL 3.44910 against 3.53895 and calibration error 0.33820 against 0.34790 for the native model, while its accuracy is lower. Helena at $K = 4$ with $1{,}036$ site parameters has calibration error 0.014676 against 0.036275

for the GELU host without winning accuracy, and Jannis at $K = 2$ with $1{,}664$ site parameters has balanced accuracy 0.543839 against 0.537843 for the GELU host. Secondary metrics can also go against PSNeuron: the Jannis configuration at $K = 2$ with $37{,}636$ parameters has a higher calibration error (0.037869 against 0.022711) and NLL (0.748128 against 0.740201) than a resized SwiGLU network with $37{,}252$ parameters (both in Figure 8(b)).

L.3 Pareto inventory

Table 62 breaks Table 2 down by task and tier. On MQAR every run, the native model included, evaluates the delta rule chunkwise, and all share one descriptive stratum. On multi-needle retrieval and state tracking, the runs marked in Table 32 as evaluated chunkwise come from a later code revision than the sequentially evaluated native model and form strata of their own. The MQAR stratum and the chunkwise multi-needle stratum also hold the GELU, SwiGLU, and additive mixers that read the state (Section J.2): the inventory counts 1079 smaller and better pairs among 3720 comparisons of a mixer with a control of its stratum, most at nearly equal size. 66 strict pairs, all on MQAR and multi-needle retrieval, have a nominal interval above zero, among them the joint $K = 4$ mixer against the additive mixer on MQAR, and none stays above zero after a Bonferroni correction within its stratum, which holds 504 comparisons on MQAR; a capacity rung that repeats this mixer exactly counts as a second pair. A PSNeuron rung forms strict pairs only with controls tuned on its own learning rate grid, on MQAR mostly additive mixers, which is why most MQAR rungs are non dominated in the strict tier but only 10 of 56 against every configuration of the descriptive tier. A separate study with the same state and input access finds a nominal positive MQAR paired interval against conditional GELU that does not survive a Bonferroni correction over its three MQAR controls, while its GELU and SwiGLU comparisons on MQAR and its dense comparisons on state tracking remain unresolved (Section K); a fresh seed holdout finds Holm corrected gains over conditional GELU and an exact additive twin, a gain over GELU only at the nominal level, and an unresolved SwiGLU comparison, and supports no size comparison (Section K.2). Both cohorts are distinct and do not enter this inventory.

On the synthetic teachers every strict smaller and better pair is against the additive control (528 parameters). The other dense control, a GELU network with $2{,}177$ parameters, reaches held out $R^2 \geq 0.965$ on the 9 targets whose orders share directions, and no PSNeuron unit is smaller and better than it there; on the incoherent teachers of orders 6 and 8 it reaches only 0.39 and 0.14, and 3 PSNeuron units with 288 to 320 parameters exceed it on the teacher of order 8, the best at $R^2 = 0.28$, all in the descriptive tier.

L.4 Additional measured configurations

Jannis capacity points. Table 63 reports all twelve completed five-seed PSNeuron capacity configurations from a separate sweep alongside the compatible 665-site-parameter GELU control. At $K = 6$, a 527-site-parameter PSNeuron has $21{,}267$ model parameters and 70.03245% accuracy, versus $21{,}405$ and 69.85969% for that control; at $K = 8$, the 533-site-parameter point has $21{,}273$ and 70.00969%. The corresponding named-reference savings are 0.6447% and 0.6167%, both descriptive point-estimate relations with incomplete tuning metadata. A $K = 6$ 660-site-parameter point is also five parameters smaller and scores higher than the same control, but the saving is negligible. The $K = 6$ 793-site-parameter point has $21{,}533$ parameters and 68.61838%; its eligible paired interval against that control is wholly unfavorable. These local points do not replace the earlier thresholds, larger dense frontier, or strict-tier conclusions above.

Other complete and incomplete arrivals. Table 64 retains the completed CIFAR100 conditioners, cross-head controls, HIGGS Small feedforward budgets, and recurrent capacity observations. The three CIFAR100 conditioner points lack a compatible resized dense comparator in their placement, so their accuracies support no parameter-efficiency comparison; the cross-head rows remain adverse. The recurrent rows provide additional seed means, not a new controlled mixer advantage. Two configurations remain incomplete, both HIGGS Small feedforward budgets (Table 65); neither contributes a completed mean. The FineWeb-Edu $K = 8$ all-block rungs, previously incomplete here, are now complete (Table 39). A newly arrived CIFAR100 cross-head $K = 2$ local comparison is in Table 66. The HIGGS depth schedule and the all-block language-model placement are shown separately in Tables 39 and 40. These tables add affected configuration views to the frozen earlier full

Table 62: Pareto inventory by task, primary metric, whole model scope (head scope on frozen features), for replacements and, below them, augmentations, which keep a dense carrier and carry no size claim; see also Table 2. Tiers as in Section 4, with all for the descriptive tier. O / D counts PSNeuron and dense configurations, and a measurement that is both a capacity rung and the host or the common replacement counts once in each role; non dominated counts PSNeuron configurations that no dense configuration of the stratum dominates, among those whose stratum has one; $L > 0$ counts smaller and better pairs whose nominal paired 95% interval lies above zero (Bonferroni survivors in parentheses); savings are those of the smaller and better pairs; $\Delta_F > 0$ counts, summed over the strata of the tier, the common budgets at which the best PSNeuron configuration of a stratum, of any order, is better than the best dense configuration of the same stratum, over all budgets where both have a configuration. [§]Because the identity carrier alone reaches 99.96% with no site parameters (Table 30), the MQAR attention output rows are not read as PSNeuron parameter efficiency evidence; their inventory, fixed at an earlier cutoff, is kept for provenance.

host	task	tier	O / D	non dominated	smaller, better	$L > 0$	saving (%)	$\Delta_F > 0$
language model	FineWeb-Edu	strict	21 / 6	1 of 21	0			0 of 21
language model	FineWeb-Edu	all	21 / 6	1 of 21	0			0 of 21
unit	XOR, 4 coordinates	all	8 / 4	8 of 8	9		30.6 to 72.5	3 of 4
unit	XOR, 6 coordinates	strict	8 / 0	no dense	0			
unit	XOR, 6 coordinates	all	8 / 4	8 of 8	8		30.6 to 50.0	2 of 4
unit	mixed equal, $K = 6$	strict	4 / 2	4 of 4	4	4 (2)	39.4 to 50.0	1 of 2
unit	mixed equal, $K = 6$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	mixed equal, $K = 8$	strict	2 / 1	2 of 2	2	2 (2)	39.4 to 50.0	1 of 1
unit	mixed equal, $K = 8$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	mixed, no order 1, $K = 6$	strict	4 / 2	4 of 4	4	4 (4)	39.4 to 50.0	1 of 2
unit	mixed, no order 1, $K = 6$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	mixed, no order 1, $K = 8$	strict	3 / 1	3 of 3	3	3 (3)	39.4 to 45.5	1 of 1
unit	mixed, no order 1, $K = 8$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	mixed with order 1, $K = 6$	strict	4 / 2	4 of 4	3	2 (0)	39.4 to 45.5	1 of 2
unit	mixed with order 1, $K = 6$	all	4 / 2	4 of 4	3		39.4 to 45.5	1 of 2
unit	mixed with order 1, $K = 8$	strict	3 / 1	3 of 3	3	2 (1)	39.4 to 45.5	1 of 1
unit	mixed with order 1, $K = 8$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	parity, 8 bits	all	4 / 2	4 of 4	2		15.0 to 23.6	2 of 2
unit	coherent, $K = 2$	strict	4 / 2	4 of 4	4	4 (0)	39.4 to 50.0	1 of 2
unit	coherent, $K = 2$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	coherent, $K = 4$	strict	4 / 2	4 of 4	4	4 (1)	39.4 to 50.0	1 of 2
unit	coherent, $K = 4$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	coherent, $K = 6$	strict	4 / 2	4 of 4	4	4 (2)	39.4 to 50.0	1 of 2
unit	coherent, $K = 6$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	incoherent, $K = 6$	strict	2 / 2	2 of 2	2	2 (2)	45.5 to 50.0	1 of 2
unit	incoherent, $K = 6$	all	4 / 2	4 of 4	4		39.4 to 50.0	1 of 2
unit	incoherent, $K = 8$	strict	1 / 2	1 of 1	1	1 (1)	50.0	1 of 2
unit	incoherent, $K = 8$	all	4 / 2	4 of 4	7		39.4 to 86.8	2 of 2
MLP	Helena	strict	32 / 22	0 of 31	2	0 (0)	2.6 to 3.9	0 of 45
MLP	Helena	all	32 / 22	0 of 32	2		2.6 to 3.9	0 of 45
MLP	HIGGS Small	strict	6 / 15	0 of 6	0			0 of 13
MLP	HIGGS Small	all	33 / 22	10 of 33	103		0.0 to 26.5	41 of 45
MLP, all blocks	HIGGS Small	all	4 / 0	no dense	0			
MLP	Jannis	strict	10 / 20	4 of 10	6	0 (0)	0.1 to 41.7	1 of 22
MLP	Jannis	all	33 / 22	1 of 33	206		0.0 to 51.9	20 of 45
MLP, all blocks	Jannis	strict	2 / 0	no dense	0			
MLP, all blocks	Jannis	all	4 / 0	no dense	0			
MLP	mixed equal, $K = 6$	strict	2 / 2	2 of 2	0			0 of 2
MLP	mixed equal, $K = 6$	all	4 / 2	4 of 4	2		43.5 to 53.8	2 of 2
DeltaNet	MQAR	strict	56 / 9	43 of 56	212	58 (0)	0.0 to 14.5	53 of 53
DeltaNet	MQAR	all	56 / 35	10 of 56	515		0.0 to 14.5	78 of 80

inventory; no newly completed point-estimate Pareto relation is treated as an uncertainty-qualified family win.

L.5 Measured configurations against resized dense networks

Table 1(a) compares each replacement with the dense module it removes. Figure 8 makes the stricter comparison with dense networks resized to the same budgets: it shows every complete single site replacement of the tabular study and, for CIFAR100, those complete at an earlier evidence cutoff, which lack five later PSNeuron configurations of Table 62 that change none of the statements below. The tabular anatomy that keeps a carrier, which the figure omits, is in Table 42, and the mean, seed standard deviation, and selected learning rate of every point are in the machine readable records of

Table 62: Pareto inventory by task (continued).

host	task	tier	O / D	non dominated	smaller, better	$L > 0$	saving (%)	$\Delta_F > 0$
DeltaNet	multi-needle	strict	55 / 33	12 of 49	256	8 (0)	0.0 to 14.6	34 of 76
DeltaNet	multi-needle	all	56 / 36	11 of 56	561		0.0 to 14.6	24 of 81
DeltaNet	state tracking	strict	8 / 0	none in tier	0			
DeltaNet	state tracking	all	8 / 1	4 of 4	3		14.7	1 of 1
cross head	CIFAR100	strict	10 / 7	1 of 10	4	0 (0)	0.1 to 0.3	0 of 11
cross head	CIFAR100	all	16 / 7	0 of 16	8		0.1 to 0.3	0 of 18
cross head	MQAR[§]	strict	1 / 0	not read	–	–	no size claim	–
cross head	MQAR[§]	all	4 / 1	not read	–	–	no size claim	–
ViT	CIFAR100	strict	48 / 13	0 of 41	18	0 (0)	0.0 to 2.2	0 of 53
ViT	CIFAR100	all	48 / 13	0 of 44	18		0.0 to 2.2	0 of 56
ViT, all blocks	CIFAR100	strict	4 / 1	4 of 4	0			0 of 1
ViT, all blocks	CIFAR100	all	4 / 1	4 of 4	0			0 of 1
head	EuroSAT	strict	43 / 14	3 of 43	0			0 of 53
head	EuroSAT	all	49 / 28	1 of 49	0			0 of 71
head	ImageNet100	strict	37 / 20	0 of 16	0			0 of 28
head	ImageNet100	all	50 / 30	2 of 50	0			0 of 77
augmentations, which keep a dense carrier								
MLP	Helena	strict	1 / 22	0 of 1	0			0 of 8
MLP	Helena	all	1 / 22	0 of 1	0			0 of 8
MLP	HIGGS Small	all	1 / 22	0 of 1	2		no size claim	0 of 8
MLP, all blocks	HIGGS Small	all	1 / 0	no dense	0			
MLP	Jannis	all	1 / 22	0 of 1	8		no size claim	0 of 8
MLP, all blocks	Jannis	all	1 / 0	no dense	0			
MLP	mixed equal, $K = 6$	strict	1 / 2	1 of 1	0			0 of 2
MLP	mixed equal, $K = 6$	all	1 / 2	1 of 1	0			0 of 2
score correction	MQAR	strict	8 / 1	0 of 4	0			0 of 4
score correction	MQAR	all	8 / 1	0 of 4	0			0 of 4
score correction	multi-needle	all	8 / 1	0 of 4	0			0 of 4
ViT, post attention	CIFAR100	strict	3 / 1	0 of 1	0			0 of 1
ViT, post attention	CIFAR100	all	4 / 1	0 of 4	0			0 of 4
adapter	EuroSAT	strict	9 / 3	1 of 6	1	0 (0)	no size claim	0 of 7
adapter	EuroSAT	all	11 / 3	3 of 11	3		no size claim	2 of 12
adapter	ImageNet100	strict	1 / 0	no dense	0			
adapter	ImageNet100	all	11 / 3	1 of 11	3		no size claim	12 of 12
adapter	IMDb	strict	4 / 1	2 of 3	0			2 of 3
adapter	IMDb	all	4 / 1	3 of 4	0			3 of 4
head	EuroSAT	strict	1 / 14	0 of 1	0			0 of 1
head	EuroSAT	all	1 / 28	0 of 1	0			0 of 26
head	ImageNet100	strict	1 / 20	0 of 1	0			0 of 2
head	ImageNet100	all	1 / 30	0 of 1	11		no size claim	0 of 19

Table 63: Additional Jannis feedforward capacity points and compatible resized GELU reference at 665 site parameters. All accuracies are five-seed final test means; Δ is the accuracy difference in percentage points against that named reference. Strict and descriptive distinguish complete tuning metadata from incomplete tuning metadata, not statistical significance. The $K = 6$, 793-site-parameter row (†) has a seed-paired nominal 95% interval entirely below zero against the 665-site-parameter GELU, $[-1.62, -0.86]$ points. The favorable 527- and 533-site-parameter rows and the five-parameter 660-site-parameter relation are observed point comparisons only.

K	site P	model P	accuracy %	Δ pp	seeds	tier
2	512	21,252	69.64	-0.21	5	strict
2	640	21,380	69.79	-0.07	5	strict
2	768	21,508	70.15	+0.29	5	descriptive
4	652	21,392	69.32	-0.54	5	strict
6	527	21,267	70.03	+0.17	5	descriptive
6	660	21,400	70.06	+0.20	5	descriptive
6	793	21,533	68.62†	-1.24	5	strict
6	1,325	22,065	69.83	-0.03	5	descriptive
8	533	21,273	70.01	+0.15	5	descriptive
8	668	21,408	69.93	+0.07	5	descriptive
8	803	21,543	70.04	+0.18	5	descriptive
8	1,343	22,083	69.96	+0.10	5	descriptive
GELU	665	21,405	69.86	0.00	5	strict

Table 64: Additional complete primary-metric configurations: CIFAR100 conditioner and cross-head arms, HIGGS Small resized feedforward sites, and recurrent MQAR and multi-needle capacity points. CIFAR100 and HIGGS Small report test accuracy; MQAR and multi-needle report recall at one, each as a percent. Site and whole-model parameters are trainable counts. A completed row lacking a compatible dense comparator is a measurement, not a frontier claim. Tier refers to archived tuning coverage. The broader full inventory is in Table 62.

task	site/form	K	site P	model P	metric %	seeds	tier
CIFAR100	additive nearest build 1216p	–	1,216	2,671,268	38.77	5	strict
CIFAR100	cross head nearest build 1888p	2	1,888	2,671,940	37.81	5	strict
CIFAR100	cross head nearest build 1222p	4	1,222	2,671,274	33.17	5	strict
CIFAR100	conditioner lowrank	4	9,241	2,420,477	40.11	3	strict
CIFAR100	conditioner raw	4	6,169	2,417,405	40.57	3	strict
CIFAR100	conditioner lowrank	8	9,273	2,420,509	39.43	3	strict
HIGGS Small	nearest build 512p	2	512	19,458	71.01	5	strict
HIGGS Small	nearest build 768p	2	768	19,714	70.92	5	strict
HIGGS Small	nearest build 652p	4	652	19,598	70.84	5	strict
HIGGS Small	nearest build 783p	4	783	19,729	71.02	5	descriptive
HIGGS Small	nearest build 1325p	6	1,325	20,271	70.84	5	descriptive
HIGGS Small	nearest build 527p	6	527	19,473	70.69	5	strict
HIGGS Small	nearest build 660p	6	660	19,606	70.92	5	descriptive
HIGGS Small	nearest build 793p	6	793	19,739	71.04	5	descriptive
HIGGS Small	nearest build 668p	8	668	19,614	70.92	5	descriptive
MQAR	nearest build 4224p	–	4,224	744,004	9.13	5	strict
multi-needle	nearest build 4224p	–	4,224	744,004	99.12	5	strict

Table 65: Additional incomplete configurations. The seed lists are actual observations, not completed means; none of these rows enters an empirical frontier or a five-seed summary. The FineWeb-Edu $K = 8$ all-block rungs previously listed here are now complete (Table 39).

host/placement	K, rung	observed seeds	expected n	model P
HIGGS Small, feedforward	dense nearest build 665p	[0]	5	19,611
HIGGS Small, feedforward	$K = 4$, nearest build 1307p	[0,1,2,3]	5	20,253

Table 66: CIFAR100 cross-head order-2 local comparison (Sep-26 arrival), five seeds each, whole-model counts. PSNeuron leads the larger additive control by $+0.498$ points (nominal paired 95% CI $[+0.035, +0.961]$); the exact-size additive control, at the same 2,671,268 parameters as the PSNeuron point, reaches 38.772% and is not beaten (-0.430 points). This is a local larger-control Pareto observation, not a general product win.

configuration	model P	accuracy %	seeds
PSNeuron, $K = 2$	2,671,268	38.342	5
additive, exact size	2,671,268	38.772	5
additive, larger	2,671,940	37.844	5

the supplementary material. The tabular grid covers $K = 2$, 4, 6, and 8 with five final seeds each; it includes compact Jannis configurations near the quality of the selected reference, but it does not establish an optimal interaction order across tasks. On HIGGS Small the common $K = 4$ model, with $21{,}018$ model and $2{,}072$ site parameters, reaches $70.912 \pm 0.622\%$ at the largest learning rate of its grid, and the ladder model with the same counts selected 8×10^{-4} and reaches $70.165 \pm 0.328\%$; the two were selected and trained separately, so neither replaces or is averaged with the other.

In the descriptive tier, which compares every configuration of a task with every dense network of that task, the PSNeuron configurations that no dense network dominates share one caveat: all 10 of them on HIGGS Small and the single one on Jannis selected the largest learning rate of the extended grid, so their tuning is not bracketed. The strict tier of Tables 2 and 62 leaves such configurations out and compares each remaining one only with the dense networks of its own stratum, tuned on the same learning rate grid (Section L); in that tier no HIGGS Small configuration and four Jannis configurations are non dominated. On Jannis only two PSNeuron configurations whose selected learning rate lies inside its grid are smaller than the GELU host with $37{,}316$ parameters and have a higher mean than its 69.68%, the $K = 2$ configurations with $1{,}664$ and $1{,}024$ site parameters, which save 39.96% and 41.68%; the SwiGLU network with a site of 544 parameters is smaller and more accurate than both. On Helena the SwiGLU network with a site of 544 parameters is smaller and more accurate than every PSNeuron configuration. At the CIFAR100 site the PSNeuron envelope stays below the dense envelope: at every budget a resized GELU network is at least as accurate as the best

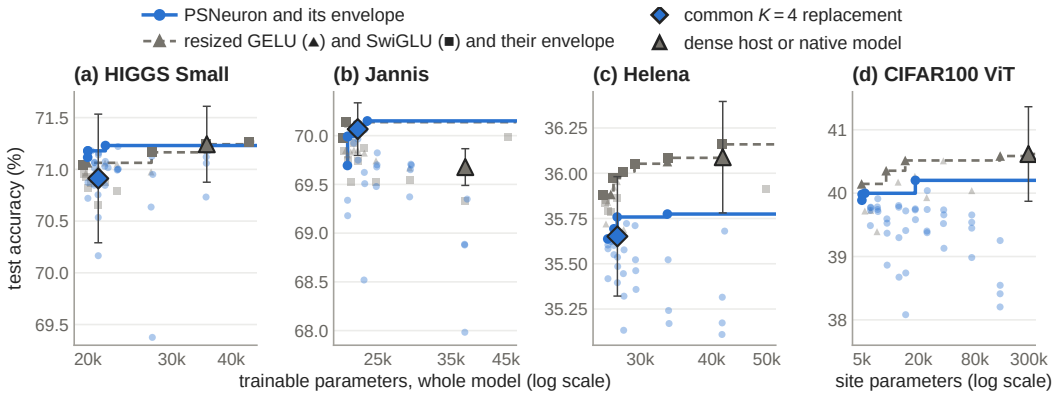

Figure 8: Measured configurations against resized dense networks: tabular single-site replacements and CIFAR100 configurations complete at an earlier evidence cutoff. (a) to (c) Tabular feedforward site: mean test accuracy over five seeds against the trainable parameters of the whole model. PSNeuron capacity ladders ($K \in \{2, 4, 6, 8\}$) and anatomies form one group, resized GELU and SwiGLU networks another; each envelope gives the best mean at or below a budget, faded points are dominated within their group, error bars are seed standard deviations, and the anatomy that keeps a carrier is omitted. (d) CIFAR100 ViT, last feedforward site, against the parameters of the site, with the resized GELU networks and the native network.

PSNeuron configuration of equal or smaller size, and the highest PSNeuron mean, $40.202 \pm 0.304\%$ for $K = 6$ with $18{,}512$ site parameters ($2{,}429{,}748$ in the model), is still below the $40.618 \pm 0.746\%$ of the native model.