The ARC White-Box Estimation Challenge
Phase 1: Two Final Submissions

ABSTRACT

We describe Team Puffi's two final Phase 1 submissions, #327950 and #328046, for estimating mean post-ReLU activations in a known deep network. Submission #327950 achieved the stronger public platform score of the two, $9.0914304340 \times 10^{-8}$; lower is better. Submission #328046 used 10.47% less mean effective compute and scored $9.7126285217 \times 10^{-8}$. We include it as a lower-compute alternative, not as evidence of better unseen performance. Both submissions combine structured deterministic integration, a shallow analytical reference, a fitted inter-layer control variate, and compute-aware execution. Technical details and evidence limits are in the Appendix.

Statement on the use of AI tools. Large-language-model assistants were used under supervision during research and manuscript preparation. The authors reviewed the archived source, receipts, equations, numerical claims, and final text, and remain responsible for the report's contents.

1. INTRODUCTION

We discuss Team Puffi's two final Phase 1 submissions. The task is to estimate mean post-ReLU activations in a known 32-layer, width-256 network under standard Gaussian input. The per-network compute budget is $B = 2.72 \times 10^{11}$ floating-point-operation equivalents. Submission #327950 achieved the stronger public score; #328046 is a lower-compute second configuration.

Main contributions.—

- A deterministic hybrid estimator for deep activation means.
- An improved shallow analytical reference, denoted A716, through layer 9.
- A fitted inter-layer control variate that transports the layer-9 sampling discrepancy to correct the final-layer estimate.
- A compute-aware implementation for the competition budget.
- Two operating points: the stronger public result and a lower-compute alternative.

Main ingredients.—

- Exact integration over the Gaussian input radius.
- Structured deterministic directions, evaluated in opposite pairs.
- A small full-width diagnostic sample, called the probe, for pruning and filling omitted coordinates.
- A fitted discrepancy transport from post-ReLU layer 9 to the final layer.
- Exact skipping of trailing zeros and fast matrix multiplication.

Section 2 gives the measured results. The Appendix gives the construction and statistical qualifications.

2. MAIN RESULTS

Table 1 reports the two preserved public platform aggregates. Mean-squared error (MSE) measures prediction accuracy. The adjusted score combines final-layer MSE with effective compute; lower values are better. Appendix B gives the exact definition and evidence scope.

Table 1. Public platform aggregates for the two final submissions. Values are copied from the archived dashboard receipts, not reconstructed from local runs.

quantity	#327950	#328046
adjusted final-layer score	$9.0914304340 \times 10^{-8}$	$9.7126285217 \times 10^{-8}$
raw final-layer MSE	$1.8944517805 \times 10^{-7}$	$2.2586823746 \times 10^{-7}$
all-layers MSE	$1.6073822226 \times 10^{-5}$	$1.6756143741 \times 10^{-5}$
mean effective compute	$1.3136782035 \times 10^{11}$	$1.1761760108 \times 10^{11}$

Shared approach.—Both submissions use two main ideas. First, structured orthonormal bases, called frames below, and opposite pairs provide deterministic angular quadrature. Second, the shallow analytical reference A716 supplies a layer-9 discrepancy. A fitted inter-layer control variate transports that discrepancy to correct the final-layer sample mean. Compute-saving execution makes this hybrid estimator fit the budget.

Inter-layer response.—For a fixed network the exact layer means are deterministic. The relevant fluctuations are instead the errors induced by applying the same finite quadrature bank at successive layers. Let $\delta\boldsymbol{x}^{(L)}$ denote a small perturbation of the post-ReLU mean at zero-based layer L. A local linear response model writes

$$\delta\boldsymbol{x}^{(L')} \simeq \mathbf{R}_{L\to L'}^{\top}\delta\boldsymbol{x}^{(L)}, \qquad L' > L, \tag{1}$$

where $\mathbf{R}_{L\to L'}$ is a matrix acting on the whole perturbation vector; there is no implicit summation over layer labels.

Our implementation uses $L = 8$ and $L' = 31$, corresponding to post-ReLU layers 9 and 32 in the figures. Let $\mathcal{I}_8$ be the retained layer-8 coordinates, $\mathcal{J}$ the directly propagated final coordinates, $\widehat{\boldsymbol{\mu}}_{\mathrm{samp}}^{(L)}$ the finite-design mean, and $\boldsymbol{\mu}^{A,(8)}$ the A716 reference. With a fitted response $\widehat{\mathbf{R}}_{\mathcal{J}}$ and shrinkage $\rho = 0.75$, the main correction has the form

$$\left(\widehat{\boldsymbol{\mu}}_{\mathrm{CV}}^{(31)}\right)_{\mathcal{J}} = \left(\widehat{\boldsymbol{\mu}}_{\mathrm{samp}}^{(31)}\right)_{\mathcal{J}} - \rho\,\widehat{\mathbf{R}}_{\mathcal{J}}^{\top}\left[\left(\widehat{\boldsymbol{\mu}}_{\mathrm{samp}}^{(8)}\right)_{\mathcal{I}_8} - \left(\boldsymbol{\mu}^{A,(8)}\right)_{\mathcal{I}_8}\right]. \tag{2}$$

The bracket estimates the layer-9 quadrature discrepancy; the fitted response predicts the component of that discrepancy that reaches the final layer. Appendix A.8 gives the ridge fit, the submitted terminal-bypass branch, and the limitations of this model-assisted control variate.

Different operating points.—Submission #327950 uses a larger, unrotated direction bank. Submission #328046 uses a smaller, rotated bank. On the public aggregate, #328046 used 10.4670% less mean effective compute but had 19.2262% higher raw MSE and 6.8328% higher adjusted score. We therefore describe it as a lower-compute alternative. Appendix A.3 and Table 2 give the exact differences.

Across the 50 matched networks, the correction lowered final-layer mean MSE from 3.1120×10^{-7} to 2.5644×10^{-7}, a 17.60% reduction relative to the uncorrected estimate. The paired-bootstrap 95% interval is $[7.84\%, 26.33\%]$. Appendix B.1 gives the protocol and qualifications.

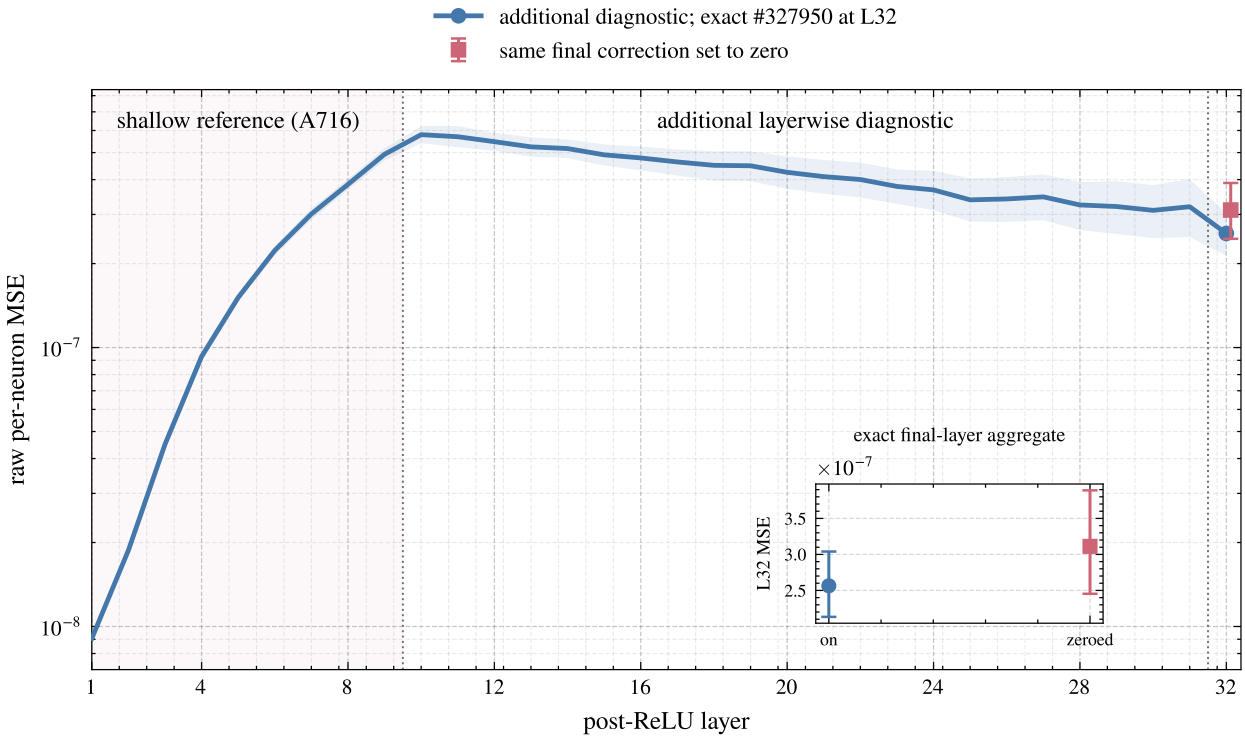

Figure 1. Raw per-neuron MSE by post-ReLU layer on a reused, matched 50-network panel from a locally generated evaluation set; these networks were not supplied by the organizers. Layers 1–9 reproduce the shallow analytical reference in #327950. Layers 10–31 are an additional diagnostic: retained coordinates use the complete submitted direction bank, while omitted coordinates keep the three-frame probe mean. These added diagnostic averages do not affect the submission. Layer 32 is the exact #327950 output; the square shows the same final calculation with its fitted correction set to zero. Lines are network means and bands are pointwise network-bootstrap 95% intervals. This exposed-panel diagnostic is not a platform score or fresh confirmation.

3. CONCLUSION AND FUTURE WORK

The measured evidence favors #327950 as the stronger of our two final configurations. Submission #328046 is a lower-compute alternative with a different angular design; we do not claim that it is more accurate on unseen networks.

Our main open problem is how to choose the angular quadrature rule: the finite set of directions and weights used to approximate the Gaussian expectation. Retrospective, target-using diagnostics found lower error for some network-specific selections of a few dozen complete frames. These results show capacity, not a deployable selection rule or evidence of generalization.

We also tested herding, a greedy procedure that does not use the exact activation means and adds the frame that most reduces the difference between the selected feature average and a reference average. Herding reduced the discrepancy it was designed to measure, but it did not consistently reduce final-layer MSE. Future work should seek a mechanistic selection criterion that predicts how each frame's error propagates through the later layers. The rule should be fixed before evaluation and tested on untouched networks. Appendix B.2 gives the relevant qualifications.

APPENDIX

A. DETAILS ABOUT OUR ESTIMATORS

This Appendix describes the shared numerical estimator and the configuration differences between #327950 and #328046. The description follows the archived source used for the two submissions.

A.1. *Problem and notation*

Vectors are bold italic lowercase, matrices are upright bold uppercase, and activations are represented as rows. The network has width $n = 256$, depth $D = 32$, and per-network budget $B = 2.72 \times 10^{11}$ floating-point-operation (FLOP) equivalents. We use zero-based layer indices $L = 0, \dots, D-1$. Let $\mathbf{I}_n$ be the $n \times n$ identity matrix, $[\cdot]_+$ the componentwise ReLU, and $\{\mathbf{W}^{(L)}\}_{L=0}^{D-1}$ the fixed network weights. We write $\boldsymbol{z}^{(L)}$ and $\boldsymbol{a}^{(L)}$ for the preactivation and post-ReLU activation rows. With a standard Gaussian input, the network is

$$\boldsymbol{a}^{(-1)} = \boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n), \qquad \boldsymbol{z}^{(L)} = \boldsymbol{a}^{(L-1)}\mathbf{W}^{(L)}, \qquad \boldsymbol{a}^{(L)} = [\boldsymbol{z}^{(L)}]_+, \quad L = 0, \dots, D-1. \tag{A1}$$

For one fixed realization of all weights, the requested per-neuron mean is

$$\boldsymbol{\mu}^{(L)} = \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{a}^{(L)}\right]^\top, \qquad \mu_i^{(L)} = \mathbb{E}_{\boldsymbol{x}}\left[a_i^{(L)}\right], \quad L = 0, \dots, D-1. \tag{A2}$$

All expectations below are over the Gaussian input for one fixed realization of the weights.

A.2. *Exact radial integration*

Every layer in Equation (A1) is positively homogeneous. For a direction $\boldsymbol{u} \in \mathbb{R}^n$ and scale $r \geq 0$,

$$\boldsymbol{a}^{(L)}(r\boldsymbol{u}) = r\,\boldsymbol{a}^{(L)}(\boldsymbol{u}). \tag{A3}$$

Write the Gaussian input as $\boldsymbol{x} = r\boldsymbol{u}$, where $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ and $r \sim \chi_n$ are independent. Then

$$\boldsymbol{\mu}^{(L)} = \bar{r}_n \mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{a}^{(L)}(\boldsymbol{u})\right]^\top, \qquad \bar{r}_n = \sqrt{2}\,\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}. \tag{A4}$$

The code evaluates all directions at the single radius $R_\star = \bar{r}_{256}$. Thus the Gaussian radial expectation is exact; only the angular integral is approximated.

A.3. *Structured angular designs and the rotation used by #328046*

Both submissions use a seeded real Kerdock–Walsh family. Each basis contributes 256 directions, and every direction and its antipode are propagated. Submission #327950 uses all 128 Kerdock bases and appends the identity basis. The identity contributes another 256 positive-half rows and is especially cheap at layer 0 because its preactivations are $R_\star\mathbf{W}^{(0)}$ up to row order. Submission #328046 uses the first 114 Kerdock bases and does not append the identity. Thus

$$N_{\mathrm{ang}} = \begin{cases} 2(128 \times 256 + 256) = 66{,}048, & \#327950, \\ 2(114 \times 256) = 58{,}368, & \#328046. \end{cases} \tag{A5}$$

Both estimators are deterministic conditional on the setup seed.

To state the underlying equal-weight angular rule explicitly, let M be the number of bases used by one submission, let $\mathbf{U}_b \in \mathbb{R}^{n \times n}$ be basis b, and let $\boldsymbol{u}_{bj}$ be row j of $\mathbf{U}_b$. Thus $M = 129$ for #327950 and $M = 114$ for #328046. Before pruning and mean fill, the full-bank estimate is

$$\widehat{\boldsymbol{\mu}}_{\mathrm{ang}}^{(L)} = \frac{1}{2Mn}\sum_{b=1}^{M}\sum_{j=1}^{n}\left[\boldsymbol{a}^{(L)}(R_\star\boldsymbol{u}_{bj}) + \boldsymbol{a}^{(L)}(-R_\star\boldsymbol{u}_{bj})\right]^\top. \tag{A6}$$

Equation (A6) isolates the deterministic quadrature rule. The deployed estimator then applies the adaptive restrictions and omitted-mean approximation described in Section A.4.

Submission #328046 changes the orientation of its finite design. From the setup seed it constructs a diagonal sign matrix $\mathbf{D}_s$ and a Walsh–Hadamard matrix $\mathbf{H}$, then sets

$$\mathbf{Q}_s = \frac{1}{\sqrt{n}}\mathbf{D}_s\mathbf{H}, \qquad \widetilde{\mathbf{W}}^{(0)} = \mathbf{Q}_s\mathbf{W}^{(0)}. \tag{A7}$$

Write $\overset{d}{=}$ for equality in distribution. In exact real arithmetic, because $\mathbf{Q}_s\mathbf{Q}_s^\top = \mathbf{I}_n$ and $\boldsymbol{x}\mathbf{Q}_s \overset{d}{=} \boldsymbol{x}$ for $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, this precomposition leaves the exact target unchanged. It does change which directions the finite deterministic bank places against a fixed network. The analytical reference still receives the original weights; this is consistent because it approximates the same rotation-invariant Gaussian expectation. The implementation materializes $\mathbf{Q}_s\mathbf{W}^{(0)}$ in float32, so bitwise invariance is not asserted.

For a Kerdock frame, the signed-Walsh representation evaluates layer 0 with a fast Walsh–Hadamard transform rather than a dense design-by-weight product. Antipodal structure is also reused at layer 1. If $\boldsymbol{z} = \boldsymbol{x}\mathbf{W}^{(0)}$, then

$$[\boldsymbol{z}]_+\mathbf{W}^{(1)} + [-\boldsymbol{z}]_+\mathbf{W}^{(1)} = |\boldsymbol{z}|\mathbf{W}^{(1)}, \tag{A8}$$

$$[\boldsymbol{z}]_+\mathbf{W}^{(1)} - [-\boldsymbol{z}]_+\mathbf{W}^{(1)} = \boldsymbol{z}\mathbf{W}^{(1)}. \tag{A9}$$

The two layer-1 preactivations are recovered from half the sum and difference; the linear difference uses a transformed composition $\mathbf{W}^{(0)}\mathbf{W}^{(1)}$.

The 128 Kerdock bases together with the identity form a complete real set of 129 mutually unbiased bases, abbreviated MUB-129. Vectors from two different bases have absolute inner product $1/\sqrt{256}$. This fixed cross-basis angle is the angular-coverage property that motivated the design. Figure 2 compares the corresponding uncorrected quadrature rules on development data.

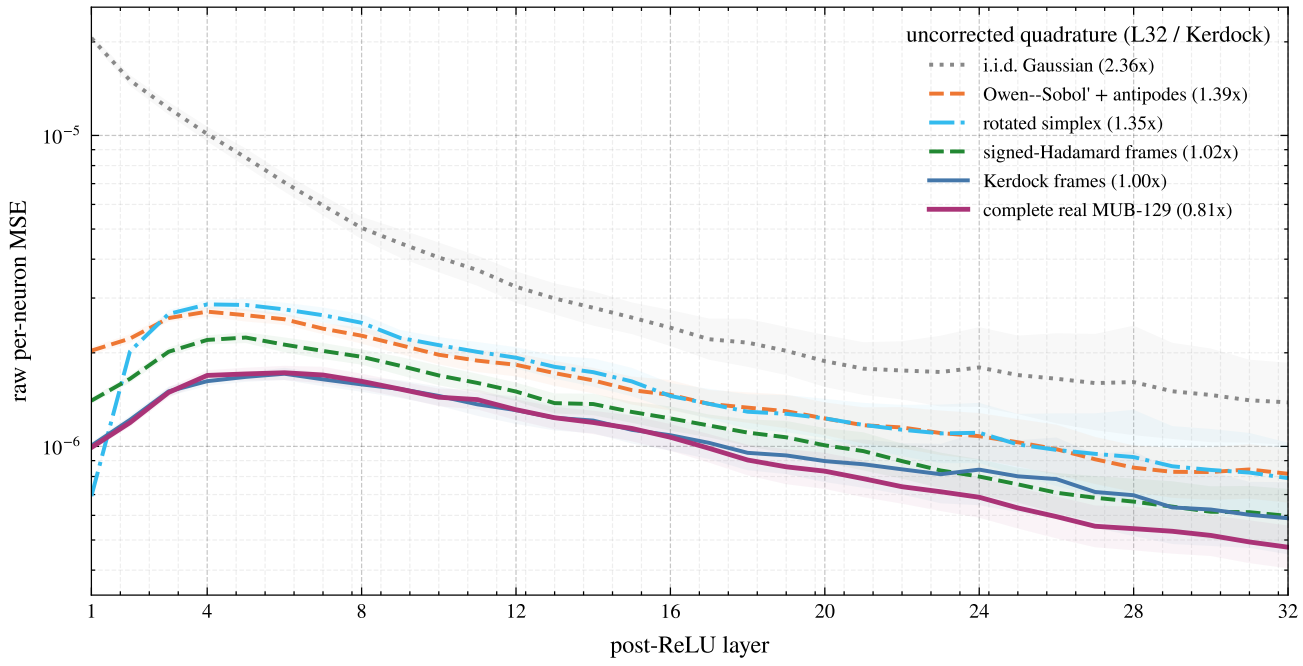

Figure 2. Development-set comparison of uncorrected quadrature rules on one previously used 50-network Phase 1 panel. Every curve is an equal-weight sample mean at matched nominal forward compute, without the analytical reference, control variate, pruning, or final readout used by the submissions. Lines are network means; bands are pointwise network-bootstrap 95% intervals. Parenthetical values are final-layer MSE relative to Kerdock. The displayed alternatives and their ordering were examined on the same panel, so this is exploratory design evidence rather than a submitted-score comparison.

A.4. *Three-frame probe, pruning, and omitted-mean fill*

We call the small full-width diagnostic sample the *probe*. It consists of the first three Kerdock bases propagated through all 32 layers. With antipodes, it has $P = 2 \times 3 \times 256 = 1536$ rows. Let

$$s_i^{(L)} = \sum_{r=1}^{P} a_{r,i}^{(L)}, \qquad \bar{a}_{P,i}^{(L)} = s_i^{(L)}/P. \tag{A10}$$

Let $\bar{\boldsymbol{a}}_P^{(L)}$ denote the row vector with entries $\bar{a}_{P,i}^{(L)}$. The retained set is

$$\mathcal{I}_L = \left\{ i : s_i^{(L)} > \epsilon_L P \right\} \cup \left\{ i : s_i^{(L)} = \max_j s_j^{(L)} \right\}, \quad \epsilon_L = \begin{cases} 3 \times 10^{-4}, & L < 16, \\ 1 \times 10^{-4}, & L \geq 16. \end{cases} \tag{A11}$$

Keeping a maximum-mass coordinate guarantees a nonempty layer. These are frozen empirical thresholds, not confidence bounds. The probe both selects the mask and contributes to the estimate, so the construction is adaptive quadrature rather than sample splitting.

Dropping inactive predecessors would bias the next preactivation. The restricted forward instead replaces each omitted coordinate with its probe mean. For $L \geq 1$,

$$\boldsymbol{z}_{\mathcal{I}_L}^{(L)} \approx \boldsymbol{a}_{\mathcal{I}_{L-1}}^{(L-1)} \mathbf{W}^{(L)}[\mathcal{I}_{L-1}, \mathcal{I}_L] + \bar{\boldsymbol{a}}_{P,\mathcal{I}_{L-1}^c}^{(L-1)} \mathbf{W}^{(L)}[\mathcal{I}_{L-1}^c, \mathcal{I}_L]. \tag{A12}$$

The second term is a row-independent bias; all 31 downstream biases are formed in one batched contraction. Mean fill inserts a probe estimate of the omitted first moment and discards omitted rowwise variation and dependence, so it is an explicit approximation.

A.5. *Prefix-based sparse execution*

Both submissions reduce the cost of restricted matrix products from zero-based layer 2 onward by exploiting trailing zeros. Probe firing counts first order the retained input coordinates from frequently to infrequently active. For each nonzero sample row, the implementation locates the final nonzero activation in that order and calls its position the required prefix length p. An all-zero row is conservatively assigned the full input width. Because ReLU activations are nonnegative and the excluded suffix is exactly zero, replacing

$$\boldsymbol{a}\mathbf{W} \quad \text{by} \quad \boldsymbol{a}_{1:p}\mathbf{W}_{1:p,:} \tag{A13}$$

is algebraically exact for that row.

Rows are stably sorted by required prefix. Adjacent rows with similar prefix lengths are grouped, and longer groups add only the exact tail products that they require. The grouping depends only on matrix geometry; it does not use targets, scores, network identifiers, or evaluation results.

The row permutation is carried explicitly and canonical order is restored before the layer-8, layer-30, and final-layer rows are captured. This preserves the rowwise alignment required by the control variate. The prefix restriction is exact over real arithmetic, although regrouping float32 sums can change their last bits.

A.6. *Positive-stable terminal bypass*

The estimator avoids some last-layer rowwise products when the full-width probe indicates a preactivation is safely positive. From probe preactivations $z_{r,i}^{(31)}$, define

$$\bar{z}_{P,i} = P^{-1} \sum_r z_{r,i}^{(31)}, \tag{A14}$$

$$\widehat{\sigma}_{P,i}^2 = \max\left(P^{-1} \sum_r (z_{r,i}^{(31)})^2 - \bar{z}_{P,i}^2, 10^{-12} \right), \tag{A15}$$

$$\alpha_{P,i} = \bar{z}_{P,i}/\widehat{\sigma}_{P,i}, \tag{A16}$$

$$h_{P,i} = \frac{\sum_r [-z_{r,i}^{(31)}]_+}{\max(\sum_r [z_{r,i}^{(31)}]_+, 10^{-12})}. \tag{A17}$$

A logically active coordinate is bypassed when $\alpha_{P,i} \geq 2.5$, $h_{P,i} \leq 2.5 \times 10^{-4}$, and it is not the maximum-mass safety coordinate.

The reconstruction uses the exact scalar identity $[z]_+ = z + [-z]_+$. The linear term comes from the all-design layer-30 mean, the exact final weight column, and omitted-mean bias. Denote that last term, defined by the second term of Equation (A12), by $b^{(31)}_{\mathrm{fill},i}$. Only the small negative hinge is estimated from the probe:

$$\widehat{\mu}^{(31)}_{i,\mathrm{byp}} = \left(\widehat{\boldsymbol{\mu}}^{(30)}_{\mathcal{I}_{30}}\right)^{\top} \mathbf{W}^{(31)}[\mathcal{I}_{30}, i] + b^{(31)}_{\mathrm{fill},i} + P^{-1} \sum_r [-z^{(31)}_{r,i}]_+. \tag{A18}$$

The identity is exact; deciding that the probe hinge is adequate remains an approximation.

A.7. *Shallow analytical reference (A716) through post-ReLU layer 9*

We use *A716* as a short label for our shallow analytical reference. At evaluation time it approximates the early activation means from the known weights without using the directional sample bank. Its fitted coefficients were selected earlier on labeled development networks, so "weights-only" describes evaluation-time inputs rather than how the method was developed. The reference stops at zero-based layer 8, which is post-ReLU layer 9 in the figures.

All vectors in this subsection are columns. The symbols $\odot$, $\oslash$, and $(\cdot)^{\odot k}$ denote elementwise multiplication, division, and powers; scalar functions of vectors also act elementwise. A716 carries a post-ReLU mean $\boldsymbol{m}^{(L)}$, a full covariance $\mathbf{C}^{(L)}$, marginal third- and fourth-cumulant information, a paired $(2,2)$ fourth-cumulant matrix, and at most eight generations of a mixed $(2,1)$ third-cumulant history. The initial Gaussian input state is $\boldsymbol{m}^{(-1)} = \mathbf{0}$, $\mathbf{C}^{(-1)} = \mathbf{I}_n$, with all higher cumulants zero.

Gaussian push-forward and response table.—At layer L, write $\mathbf{W} = \mathbf{W}^{(L)}$. The linear push-forward of the carried state is

$$\begin{aligned} \boldsymbol{m}_z &= \mathbf{W}^{\top}\boldsymbol{m}^{(L-1)}, & \boldsymbol{\Lambda} &= \mathbf{W}^{\top}\mathbf{C}^{(L-1)}\mathbf{W}, & & \\ \boldsymbol{v} &= \max\left(\operatorname{diag}\boldsymbol{\Lambda}, \sigma^2_{\min}\mathbf{1}\right), & \boldsymbol{\sigma} &= \sqrt{\boldsymbol{v}}, & \boldsymbol{r} &= \boldsymbol{m}_z \oslash \boldsymbol{\sigma}, \end{aligned} \tag{A19}$$

with $\sigma_{\min} = 10^{-6}$. For a Gaussian variable with mean m and variance $v = \sigma^2$, define its positive-part raw moments $M_k = \mathbb{E}[Z_+^k]$. Applied entrywise, the exact marginal recurrence is

$$\begin{aligned} &\boldsymbol{M}_0 = \Phi(\boldsymbol{r}), \qquad \boldsymbol{M}_1 = \boldsymbol{m}_z \odot \boldsymbol{M}_0 + \boldsymbol{\sigma} \odot \phi(\boldsymbol{r}), \\ &\boldsymbol{M}_k = \boldsymbol{m}_z \odot \boldsymbol{M}_{k-1} + (k-1)\boldsymbol{v} \odot \boldsymbol{M}_{k-2}, \qquad k = 2,3,4. \end{aligned} \tag{A20}$$

Thus Equation (A20) includes the familiar ReLU mean $m\Phi(m/\sigma) + \sigma\phi(m/\sigma)$, but also supplies the second through fourth raw moments needed by the next layer.

The approximation uses the needed entries of a response table

$$\boldsymbol{F}_{k,q} = \frac{\partial^q \boldsymbol{M}_k}{\partial \boldsymbol{m}_z^q} \quad \text{at fixed } \boldsymbol{v}. \tag{A21}$$

For the mean channel, the four responses used later are

$$\boldsymbol{F}_{1,1} = \Phi(\boldsymbol{r}), \quad \boldsymbol{F}_{1,2} = \phi(\boldsymbol{r}) \oslash \boldsymbol{\sigma}, \quad \boldsymbol{F}_{1,3} = -\boldsymbol{r} \odot \phi(\boldsymbol{r}) \oslash \boldsymbol{v}, \quad \boldsymbol{F}_{1,4} = (\boldsymbol{r}^{\odot 2} - \mathbf{1}) \odot \phi(\boldsymbol{r}) \oslash (\boldsymbol{v} \odot \boldsymbol{\sigma}). \tag{A22}$$

The other entries follow by differentiating the recurrence in Equation (A20); for example, $\boldsymbol{F}_{2,1} = 2\boldsymbol{M}_1$, $\boldsymbol{F}_{2,2} = 2\boldsymbol{M}_0$, $\boldsymbol{F}_{3,1} = 3\boldsymbol{M}_2$, and $\boldsymbol{F}_{4,4} = 24\boldsymbol{M}_0$.

Marginal non-Gaussian correction.—Let $\boldsymbol{d}_3$ and $\boldsymbol{d}_4$ be the transported marginal preactivation cumulants at this layer. For each $k = 1, \ldots, 4$, A716 forms

$$\begin{aligned} \Delta\boldsymbol{M}_k &= \operatorname{clip}_{0.25\,\boldsymbol{\sigma}^{\odot k}}\left(\tfrac{1}{6}\boldsymbol{F}_{k,3} \odot \boldsymbol{d}_3 + \tfrac{1}{24}\boldsymbol{F}_{k,4} \odot \boldsymbol{d}_4\right), \\ \widetilde{\boldsymbol{M}}_k &= \boldsymbol{M}_k + \Delta\boldsymbol{M}_k, \end{aligned} \tag{A23}$$

where $\mathrm{clip}_{\boldsymbol{c}}(\boldsymbol{y})$ clips component i to $[-c_i, c_i]$. The 25% bound is empirical. The corrected post-ReLU mean and marginal cumulants are then

$$\begin{aligned}
\boldsymbol{m}_{\mathrm{b}} &= \widetilde{\boldsymbol{M}}_1, & \boldsymbol{v}_{\mathrm{a}} &= \max\Big(\widetilde{\boldsymbol{M}}_2 - \boldsymbol{m}_{\mathrm{b}}^{\odot 2}, 10^{-12}\mathbf{1}\Big),\\
\boldsymbol{\kappa}_{3,\mathrm{a}} &= \widetilde{\boldsymbol{M}}_3 - 3\boldsymbol{m}_{\mathrm{b}} \odot \widetilde{\boldsymbol{M}}_2 + 2\boldsymbol{m}_{\mathrm{b}}^{\odot 3},\\
\boldsymbol{\kappa}_{4,\mathrm{a}} &= \widetilde{\boldsymbol{M}}_4 - 4\boldsymbol{m}_{\mathrm{b}} \odot \widetilde{\boldsymbol{M}}_3 - 3\widetilde{\boldsymbol{M}}_2^{\odot 2} + 12\boldsymbol{m}_{\mathrm{b}}^{\odot 2} \odot \widetilde{\boldsymbol{M}}_2 - 6\boldsymbol{m}_{\mathrm{b}}^{\odot 4}.
\end{aligned} \tag{A24}$$

Cross-neuron closure.—The covariance is not factorized. To expose the actual response contraction, let $\mathbf{P}$ denote the finite-window transported $(2,1)$ third-cumulant state, let $\mathbf{D}_4$ denote the paired fourth-order surrogate, and let $\alpha_{\mathrm{p}} = 0.8$ be the third-cumulant propagation scale. Define

$$\begin{aligned}
(\mathbf{Q}_1, \ldots, \mathbf{Q}_9) = \big(&\boldsymbol{\Lambda}, \tfrac{1}{2}\boldsymbol{\Lambda}^{\odot 2} + \tfrac{1}{4}\mathbf{D}_4, \tfrac{1}{2}\mathbf{P}, \tfrac{1}{2}\mathbf{P}^\top, \tfrac{1}{2}\mathbf{P} \odot \boldsymbol{\Lambda}, \tfrac{1}{2}(\mathbf{P} \odot \boldsymbol{\Lambda})^\top,\\
&\tfrac{1}{4}\mathbf{P} \odot \mathbf{P}^\top, \tfrac{\alpha_{\mathrm{p}}}{6}(\boldsymbol{d}_3\mathbf{1}^\top) \odot \boldsymbol{\Lambda}, \tfrac{\alpha_{\mathrm{p}}}{6}\big[(\boldsymbol{d}_3\mathbf{1}^\top) \odot \boldsymbol{\Lambda}\big]^\top\big).
\end{aligned} \tag{A25}$$

For $\boldsymbol{a} = (1,2,2,1,3,2,3,4,1)$ and $\boldsymbol{b} = (1,2,1,2,2,3,3,1,4)$, the off-diagonal covariance update is

$$\mathbf{C}_{\mathrm{off}}^{(L)} = \sum_{s=1}^{9} \mathrm{diag}(\boldsymbol{F}_{1,a_s})\, \mathbf{Q}_s\, \mathrm{diag}(\boldsymbol{F}_{1,b_s}). \tag{A26}$$

Its diagonal is replaced by $\boldsymbol{v}_{\mathrm{a}}$ from Equation (A24). The paired fourth-order state uses the exact diagonal contraction of its carried surrogate and a symmetric row-mean approximation off the diagonal; the mixed third-order history is truncated to the latest eight generations. Equations (A25)– (A26) therefore describe a finite closure, not exact non-Gaussian moment propagation.

Fitted correction transported across layers.—The base mean $\boldsymbol{m}_{\mathrm{b}}$ is followed by the part that distinguishes A716 from a direct cumulant truncation. Define the standardized cumulants $\boldsymbol{\gamma}_3 = \boldsymbol{d}_3 \oslash (\boldsymbol{v} \odot \boldsymbol{\sigma})$ and $\boldsymbol{\gamma}_4 = \boldsymbol{d}_4 \oslash \boldsymbol{v}^{\odot 2}$, the scale $\boldsymbol{\chi} = \boldsymbol{\sigma} \odot \phi(\boldsymbol{r})$, the polynomials $H_2(\boldsymbol{r}) = \boldsymbol{r}^{\odot 2} - \mathbf{1}$ and $H_4(\boldsymbol{r}) = \boldsymbol{r}^{\odot 4} - 6\boldsymbol{r}^{\odot 2} + 3\mathbf{1}$, and let $\boldsymbol{\tau}$ be the row-mean paired-fourth surrogate. The ten n-vectors used at layer L are

$$\begin{aligned}
\boldsymbol{h}_1 &= \Delta\boldsymbol{M}_1, & \boldsymbol{h}_2 &= \tfrac{1}{6}\boldsymbol{F}_{1,3} \odot \boldsymbol{d}_3, & \boldsymbol{h}_3 &= \tfrac{g_{4,L}}{24}\, \boldsymbol{\chi} \odot H_2(\boldsymbol{r}),\\
\boldsymbol{h}_4 &= \tfrac{1}{24}\boldsymbol{F}_{1,4} \odot \boldsymbol{d}_4, & \boldsymbol{h}_5 &= \mathbf{0}, & \boldsymbol{h}_6 &= \boldsymbol{\chi},\\
\boldsymbol{h}_7 &= \boldsymbol{\chi} \odot \boldsymbol{\tau} \oslash \boldsymbol{v}^{\odot 2}, & \boldsymbol{h}_8 &= \boldsymbol{\chi} \odot H_2(\boldsymbol{r}) \odot \boldsymbol{\gamma}_3^{\odot 2}, & \boldsymbol{h}_9 &= \boldsymbol{\chi} \odot (3\boldsymbol{r} - \boldsymbol{r}^{\odot 3}) \odot \boldsymbol{\gamma}_3 \odot \boldsymbol{\gamma}_4,\\
\boldsymbol{h}_{10} &= \boldsymbol{\chi} \odot H_4(\boldsymbol{r}) \odot \boldsymbol{\gamma}_3^{\odot 2}.
\end{aligned} \tag{A27}$$

Here $g_{4,L}$ is a fixed layer-dependent radial coefficient, and the fifth channel is zero in the submitted estimator. With $\mathbf{H}^{(L)} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{10}]$ and a frozen coefficient vector $\boldsymbol{\beta}^{(L)} \in \mathbb{R}^{10}$, the transported correction and A716 row obey

$$\begin{aligned}
\boldsymbol{s}^{(-1)} &= \mathbf{0},\\
\boldsymbol{s}^{(L)} &= \Phi(\boldsymbol{r}^{(L)}) \odot \Big[(\mathbf{W}^{(L)})^\top \boldsymbol{s}^{(L-1)}\Big] + \mathbf{H}^{(L)}\boldsymbol{\beta}^{(L)},\\
\boldsymbol{\mu}^{A,(L)} &= \boldsymbol{m}_{\mathrm{b}}^{(L)} - \boldsymbol{s}^{(L)}.
\end{aligned} \tag{A28}$$

There are ten coefficients at each of nine layers, hence 90 frozen coefficients, fitted jointly against the transported zero-based layer-8 residual. The mean readout uses the full $\boldsymbol{d}_3$ term in Equation (A23); propagation uses $\alpha_{\mathrm{p}} = 0.8$. The final reported A716 rows are clamped nonnegative as shown later in Equation (A39).

Small standardization variances are additionally floored at 10^{-6} before the last four channels are formed. A716 remains an empirical finite-order closure rather than an exact analytical solution.

Figure 3 compares A716 as implemented in the archived submission with the factorized Wu $K = 3$ analytical method. The latter is a weights-only cumulant-propagation baseline; $K = 3$ denotes its third-order truncation. Both curves use the same 50 previously used official Phase 1 development networks. A716 is shown only through post-ReLU layer 9, where it is used by #327950; this is an accuracy comparison, not a common-compute or adjusted-score comparison. On this exposed panel A716 has higher error at layer 1, lower mean MSE at layers 2–9, and a layer-9 ratio of 0.511 relative to Wu $K = 3$.

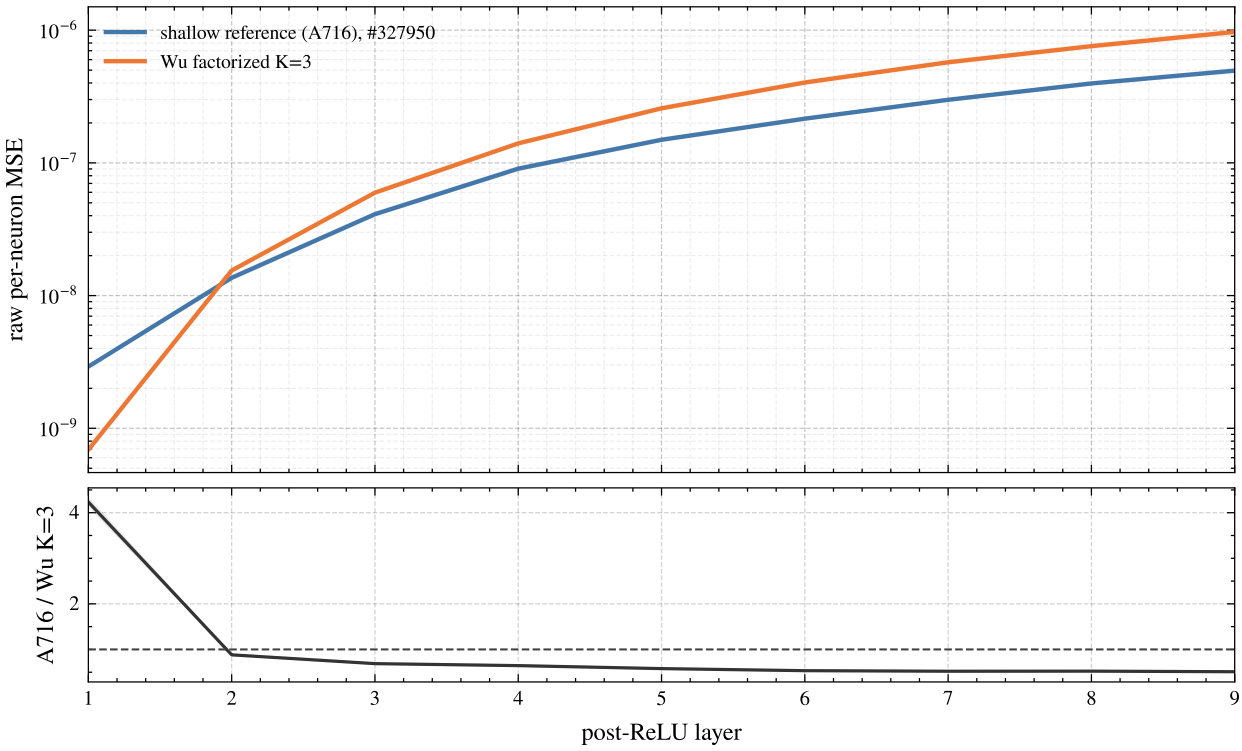

Figure 3. The shallow analytical reference A716 embedded in #327950 versus factorized Wu $K = 3$ on the same exposed 50-network panel. Curves are mean raw per-neuron MSE; bands are pointwise paired-network-bootstrap 95% intervals. The lower panel gives the ratio of cohort means, A716 over Wu $K = 3$. The A716 extraction was checked against the archived #327950 output at layers 1–9. The methods use different compute accounting, so the figure does not rank their adjusted scores.

A.8. *Inter-layer control variate: post-ReLU layer 9 to the final layer*

Here "control variate" (CV) refers to the fitted correction described below; it is not an independent random sample. The reason an inter-layer correction is possible is that the same deterministic directions are propagated through the network. For $L' > L$, if $\Delta \boldsymbol{a}_r^{(L)}$ is the centered activation row of direction r at layer L, the response approximation is

$$\Delta \boldsymbol{a}_r^{(L')} \simeq \Delta \boldsymbol{a}_r^{(L)} \mathbf{R}_{L \to L'}. \tag{A29}$$

The map is fitted within each network from aligned direction rows. It is not an assertion that the nonlinear network has one global Jacobian.

Let $p = |\mathcal{I}_8|$. From the directions outside the probe, each submission captures the first F Kerdock frames from both antipodal halves, excluding identity rows. The setting is $F = 36$ for #327950 and $F = 16$ for #328046, so

$$N_{\text{fit}} = 2F \times 256 = \begin{cases} 18{,}432, & \#327950, \\ 8{,}192, & \#328046. \end{cases} \tag{A30}$$

Let $\mathbf{X} \in \mathbb{R}^{N_{\text{fit}} \times p}$ contain the retained layer-8 activations. Let $\mathcal{J}$ be the explicitly propagated terminal coordinates, $\mathcal{H}$ the bypassed coordinates, $\mathbf{Y}_{\mathcal{J}}$ the aligned post-ReLU terminal activations on $\mathcal{J}$, and $\mathbf{X}_{30}$ the aligned retained layer-30 activations. A subscript c denotes subtraction of the corresponding column mean. After centering,

$$\mathbf{G} = \mathbf{X}_c^\top \mathbf{X}_c, \qquad \lambda = 0.25 \frac{\operatorname{tr}(\mathbf{G})}{p}. \tag{A31}$$

For the directly propagated route, ridge regression gives the empirical response operator

$$\widehat{\mathbf{R}}_{\mathcal{J}} = \underset{\mathbf{R}}{\operatorname{argmin}} \, \|\mathbf{Y}_{\mathcal{J},c} - \mathbf{X}_c \mathbf{R}\|_F^2 + \lambda \|\mathbf{R}\|_F^2 = (\mathbf{G} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_c^\top \mathbf{Y}_{\mathcal{J},c}. \tag{A32}$$

For bypassed terminal coordinates the code does not materialize their post-ReLU trajectories. With $\mathbf{W}_b = \mathbf{W}^{(31)}[\mathcal{I}_{30}, \mathcal{H}]$, the corresponding linear response is

$$\widehat{\mathbf{R}}_{\mathcal{H}} = (\mathbf{G} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_c^\top \mathbf{X}_{30,c} \mathbf{W}_b. \tag{A33}$$

Let $\widehat{\boldsymbol{\mu}}^{(8)}_{\text{samp}}$ be the all-design sampled layer-8 mean. The discrepancy from the A716 mean on retained coordinates is

$$\boldsymbol{\delta} = \left(\widehat{\boldsymbol{\mu}}^{(8)}_{\text{samp}} - \boldsymbol{\mu}^{A,(8)} \right)_{\mathcal{I}_8}. \tag{A34}$$

If A716 were exact, $\boldsymbol{\delta}$ would be the finite-design error at layer 8. In practice it also contains the negative of the A716 approximation error. Writing $\rho = 0.75$, the transparent matrix form of the two corrections is

$$\begin{aligned} \widehat{\boldsymbol{\mu}}^{(31)}_{\text{CV},\mathcal{J}} &= \widehat{\boldsymbol{\mu}}^{(31)}_{\text{samp},\mathcal{J}} - \rho\, \widehat{\mathbf{R}}_{\mathcal{J}}^\top \boldsymbol{\delta}, \\ \widehat{\boldsymbol{\mu}}^{(31)}_{\text{CV},\mathcal{H}} &= \widehat{\boldsymbol{\mu}}^{(31)}_{\text{byp},\mathcal{H}} - \rho\, \widehat{\mathbf{R}}_{\mathcal{H}}^\top \boldsymbol{\delta}. \end{aligned} \tag{A35}$$

The first line is Equation (2); the second transports the centered layer-30 linear preactivation through the final weight. Per-row negative-hinge variation is not fitted on $\mathcal{H}$, because the probe hinge enters only the reconstructed mean in Equation (A18). This is an additional bypass approximation.

The submitted implementation avoids materializing either response matrix. It solves

$$(\mathbf{G} + \lambda \mathbf{I})^\top \boldsymbol{v} = \boldsymbol{\delta} \tag{A36}$$

and evaluates

$$\widehat{\mathbf{R}}_{\mathcal{J}}^\top \boldsymbol{\delta} = \mathbf{Y}_{\mathcal{J},c}^\top \mathbf{X}_c \boldsymbol{v}, \qquad \widehat{\mathbf{R}}_{\mathcal{H}}^\top \boldsymbol{\delta} = \mathbf{W}_b^\top \mathbf{X}_{30,c}^\top \mathbf{X}_c \boldsymbol{v}. \tag{A37}$$

If the solve raises a linear algebra error, the code returns the uncorrected final estimate.

The role of layer correlation can be made explicit without a probabilistic claim. Let $\boldsymbol{e}_8$ and $\boldsymbol{e}_{31}$ be the finite-design errors of the layer-8 and directly propagated final means, and let $\boldsymbol{e}_A = (\boldsymbol{\mu}^{A,(8)} - \boldsymbol{\mu}^{(8)})_{\mathcal{I}_8}$ be the A716 error. Then $\boldsymbol{\delta} = \boldsymbol{e}_8 - \boldsymbol{e}_A$ and, for one fixed network,

$$\begin{aligned} \boldsymbol{e}_{\text{CV}} &= \boldsymbol{e}_{31} - \rho \widehat{\mathbf{R}}_{\mathcal{J}}^\top (\boldsymbol{e}_8 - \boldsymbol{e}_A), \\ \|\boldsymbol{e}_{\text{CV}}\|_2^2 &= \|\boldsymbol{e}_{31}\|_2^2 - 2\rho\, \boldsymbol{e}_{31}^\top \widehat{\mathbf{R}}_{\mathcal{J}}^\top \boldsymbol{\delta} + \rho^2 \|\widehat{\mathbf{R}}_{\mathcal{J}}^\top \boldsymbol{\delta}\|_2^2. \end{aligned} \tag{A38}$$

The correction helps only when the middle alignment term exceeds the added quadratic term. This identity explains the mechanism but does not guarantee improvement on a new network.

This is model-assisted rather than a textbook unbiased control variate. A716 is approximate, the same deterministic design contributes to the fit and discrepancy, pruning changes the forward, and both ridge and shrinkage are empirical. On the exposed Figure 1 panel, setting only this correction to zero increased mean final-layer MSE from 2.5644×10^{-7} to 3.1120×10^{-7}; Section B.1 gives the paired interval and protocol.

Figure 4 summarizes how the two paths meet in #327950. Layer numbers in that figure use the one-based convention of the plots.

A.9. *Returned rows and fast kernels*

The returned array has shape 32×256:

$$\widehat{\boldsymbol{\mu}}^{(L)} = \begin{cases} [\boldsymbol{\mu}^{A,(L)}]_+, & 0 \le L \le 8, \\ \bar{\boldsymbol{a}}_P^{(L)}, & 9 \le L \le 30, \\ [\widehat{\boldsymbol{\mu}}^{(31)}_{\text{CV}}]_+, & L = 31. \end{cases} \tag{A39}$$

The final clamp enforces the range of a post-ReLU mean. The large restricted trajectories exist to estimate the ranked final row; intermediate rows 9–30 remain inexpensive probe estimates.

The implementation combines this prefix restriction with fast dense matrix multiplication and bounded reusable workspaces. These choices reduce the instrumented arithmetic and memory-allocation overhead but do not change the estimator equations. Both submissions perform the sampled propagation in float32 arithmetic.

In Algorithm 1, K is the number of Kerdock bases, I is the identity-basis indicator, F is the number of fit frames per antipodal half, and Q is the input orientation.

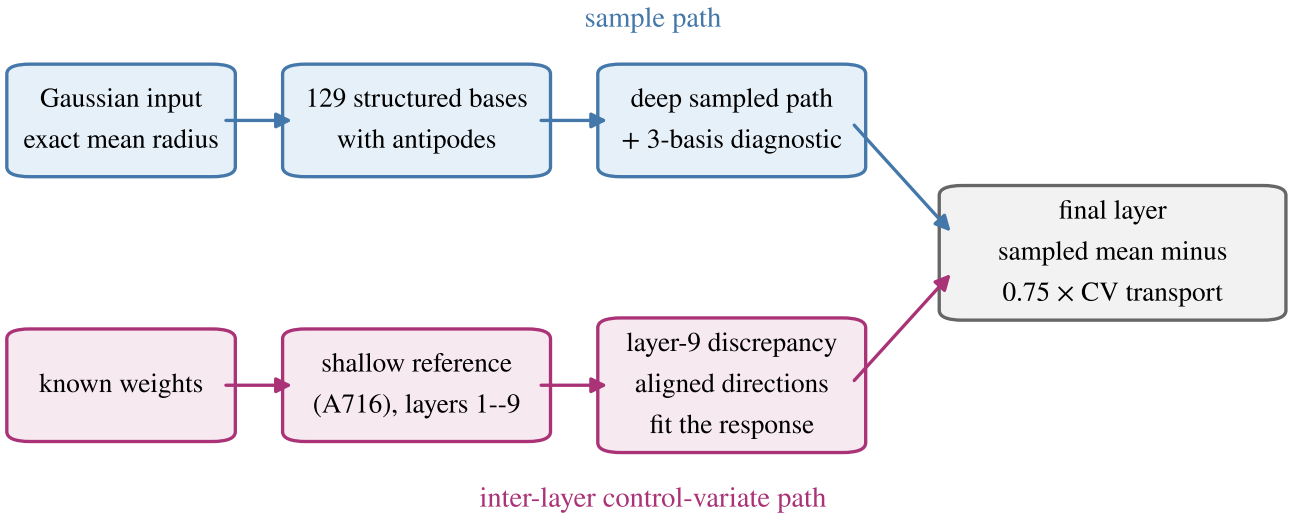

Figure 4. Executable structure of submission #327950. A716 is the weights-only analytical approximation defined in Section A.7. Its means are returned at post-ReLU layers 1–9 and its layer-9 discrepancy supplies the fitted inter-layer control-variate transport. The sampled path supplies the later probe rows and the uncorrected final mean. The fitted correction changes only layer 32. Box size does not represent compute.

Algorithm 1: Shared conceptual path for the two final submissions

Input: Weights $\{\mathbf{W}^{(L)}\}_{L=0}^{31}$, budget B, setup seed, variant (K, I, F, Q)
Output: Estimated means $\{\widehat{\boldsymbol{\mu}}^{(L)}\}_{L=0}^{31}$
Use $(K, I, F, Q) = (128, 1, 36, \mathbf{I})$ for #327950 and $(114, 0, 16, \mathbf{Q}_s)$ for #328046; apply Q only to the sampled first-layer weight
Build the first K Kerdock–Walsh bases and append the identity basis if $I = 1$; scale by $\bar{r}_{256}$ and add antipodes
Propagate the first three Kerdock bases at full width; collect masses, firing counts, and terminal preactivation diagnostics
Construct active sets, order retained coordinates by firing count, and batch all omitted-mean biases
Choose positive-stable final coordinates for terminal bypass
Propagate the remaining $K - 3$ Kerdock bases and the optional identity basis through restricted weights; from layer 2 use exact row prefixes and restore the aligned row order used to fit the response
Reconstruct bypassed terminal means using Equation (A18)
Compute A716 through layer 8 and form the discrepancy in Equation (A34)
Fit the $2F \times 256$-row ridge transport and apply the 0.75-shrunk correction
Assemble Equation (A39) and clamp the final row nonnegative
return $\{\widehat{\boldsymbol{\mu}}^{(L)}\}_{L=0}^{31}$

A.10. *Principal configurations*

Table 2. Principal settings in the two archived submissions.

component	#327950	#328046
challenge geometry	width 256, depth 32	width 256, depth 32
angular design	128 Kerdock + identity	114 Kerdock, no identity
angular evaluations	66,048	58,368
sample orientation	original first weight	signed-Hadamard-rotated first weight
probe	first 3 bases; 1,536 rows	first 3 bases; 1,536 rows
pruning threshold	3×10^{-4}; 10^{-4} from layer 16	same
prefix-based multiplication	from zero-based layer 2	same
analytical-reference layer / history window	layer 8 / 8 generations	same
reference scales / clipping	1.0 readout; 0.8 propagation / 0.25	same
response-fit sample	36 frames per half; 18,432 rows	16 frames per half; 8,192 rows
CV ridge / shrinkage	0.25 / 0.75	0.25 / 0.75
terminal bypass	$\alpha_P \geq 2.5$; hinge $\leq 2.5 \times 10^{-4}$	same
sampling arithmetic	float32	float32

B. STATISTICAL EVIDENCE AND MEASURED TRADEOFFS

B.1. *Layerwise diagnostic for submission #327950*

Figure 1 uses the archived source for #327950. The matched panel contains 50 networks from a locally generated evaluation set, not networks supplied by the organizers. Because this panel had already been examined, it is reused development evidence rather than fresh confirmation.

The submitted coefficient multiplying the final correction is 0.75. For the ablation, the uncorrected final row was captured immediately before that correction, so the comparison changes only the correction term. The submitted source returns A716 at plot layers 1–9, the three-frame probe at layers 10–31, and the full-bank corrected estimate at layer 32.

For Figure 1, an analysis-only copy adds column reductions to the same forward pass. At layers 10–31 it reports the complete-bank mean on retained coordinates and the unchanged probe mean on omitted coordinates. These rows do not set masks, enter the control variate, or modify the final calculation. The plotted layer-32 mean and bootstrap bounds come directly from the separately validated #327950 aggregate. The intermediate curve is therefore a mechanism diagnostic, while its layer-32 endpoint is the submission result. The additional reductions are not included in the submission's compute claim.

The correction-on and correction-zeroed layer-32 means are 2.5643762×10^{-7} and 3.1120390×10^{-7}. Thus the correction reduces mean MSE by 17.5982% relative to the zeroed estimate and wins on 34 of 50 networks. The paired network-bootstrap 95% interval for that reduction is $[7.8377\%, 26.3327\%]$, using 4,000 paired resamples. The layerwise bands use the same resampling unit and are pointwise intervals. Only aggregates are included in the report package; it contains no network names, generation seeds, weights, targets, identifiers, or per-network errors.

B.2. *Retrospective frame selection and herding*

Here a complete frame is one antithetic orthonormal block of angular directions. A retrospective capacity check compared equal-weight subsets of 38 frames from the 128-frame Kerdock family. A fixed 38-frame prefix had uncorrected final-layer MSE 1.00116×10^{-6}, whereas a per-network greedy subset chosen using the exact final-layer target had MSE 5.46815×10^{-8}. The best of 512 random subsets, also selected using the target separately for each network, had MSE 2.39144×10^{-7}. These are a posteriori oracle results on exposed data. They establish that useful small subsets can exist, but not that they can be identified at inference time.

The target-free herding tests, which do not use exact activation means, assigned each candidate frame an observable feature vector. Starting from an empty subset, they repeatedly added the frame that most reduced the squared difference between the selected-feature average and a full-bank or analytical reference average. The tested features included shallow mean discrepancies and their estimated downstream response. Across the tested frame counts, better feature matching did not reliably imply lower downstream MSE; in one 30-frame sweep the uncorrected MSE was about 1.4×10^{-6}. The later nonlinear layers can rotate or regenerate error that was reduced at the selection layer.

We therefore report the oracle result only as evidence that small useful subsets can exist and treat reliable target-free selection as future work. Figure 5 shows the gap directly.

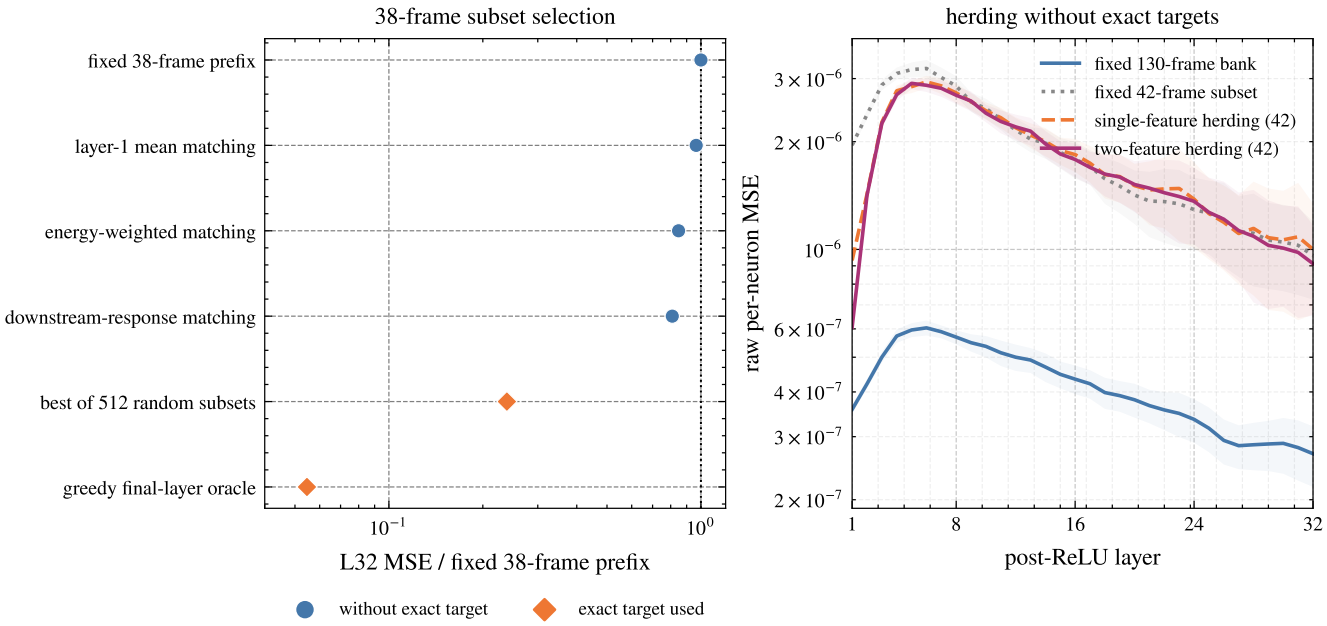

Figure 5. Two retrospective development diagnostics; the panels use different network sets and are not compared vertically. Left: final-layer MSE for 38-of-128 Kerdock subsets on 40 networks, relative to a fixed 38-frame prefix. Orange diamonds use the final-layer target to choose each subset and are retrospective oracles, not deployable rules. Blue circles are target-free. Right: uncorrected layerwise means on a separate 50-network panel for a fixed 130-frame bank, a fixed 42-frame subset, and two 42-frame herding rules; bands show mean ±1.96 standard errors. Herding greedily adds the frame that best matches an observable feature average. These studies use uncorrected sample means and do not measure the submitted estimator.

B.3. *Official score and paired public comparison*

For M valid networks indexed by $m = 1, \dots, M$, let e_m be final-layer per-neuron MSE, F_m the instrumented arithmetic, and R_m the billed residual wall-time term. The effective compute and multiplier are

$$C_m = F_m + 10^{11} R_m, \qquad q_m = \max\left(0.1, \frac{C_m}{B}\right). \tag{B40}$$

The ranked metric is the mean of the per-network products,

$$S = \frac{1}{M} \sum_{m=1}^{M} e_m q_m. \tag{B41}$$

With the $1/M$ empirical convention and $\bar{e}, \bar{q}$ denoting sample means, $S = \bar{e}\,\bar{q} + \mathrm{Cov}_M(e, q)$. For #327950, $S/\bar{e} = 0.479898$ while $\bar{q} = 0.482970$; for #328046, the corresponding values are 0.430013 and 0.432418. The difference in each pair is the normalized empirical covariance $\mathrm{Cov}_M(e, q)/\bar{e}$, not a rounding discrepancy.

Table 3. Descriptive paired comparison on the 50 aligned public score rows. Ratios are #328046 over #327950; values below one favor #328046. Intervals are 200,000-resample paired percentile bootstraps, computed after both receipts were visible.

metric	ratio of means	paired 95% interval	#328046 wins
adjusted final-layer score	1.068328	[0.908484, 1.254405]	27/50
raw final-layer MSE	1.192262	[1.014724, 1.399286]	19/50

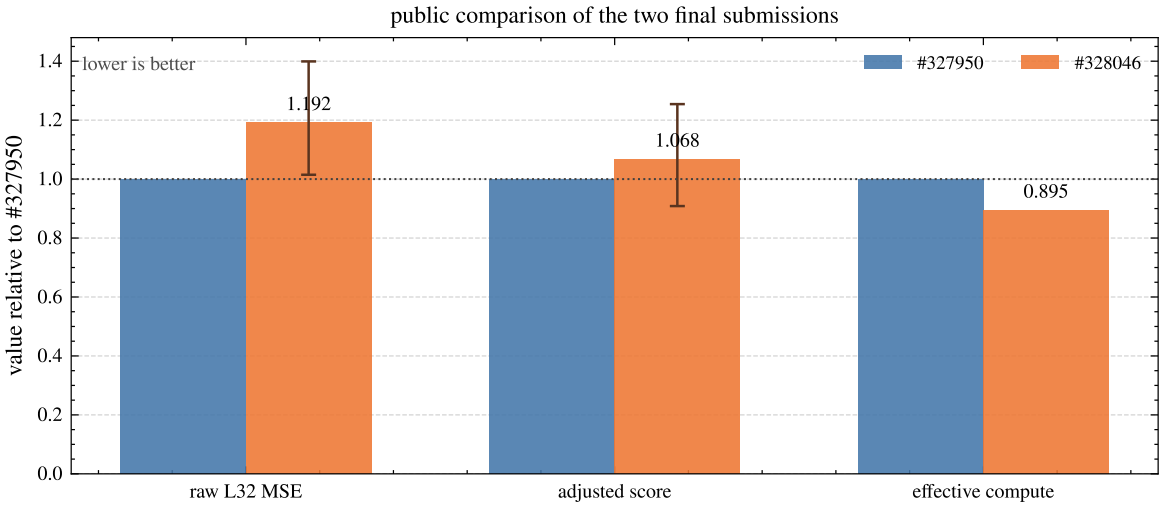

Figure 6. Direct public comparison of the two final submissions, normalized to #327950. Lower is better for all three quantities. Error bars are paired-bootstrap 95% intervals for raw final-layer MSE and adjusted score; the effective-compute bar is the ratio of the two platform aggregates and has no interval. These repeatedly exposed public rows provide descriptive, not fresh, evidence.