Separate-Universe Response Closure for Deep ReLU Networks: submission #327725

Cipo[1]

[1] *On behalf of Team Puffi*

ABSTRACT

We describe an estimator for the mean activations of a fixed-weight deep ReLU network under Gaussian inputs that receives the network weights and propagates no sampled inputs at inference. The Alignment Research Center (ARC) White-Box Estimation Challenge asks for these per-neuron means subject to a competition budget $B = 2.72 \times 10^{11}$ FLOPs, with $L = 32$ layers and constant width $n = 256$. ARC's cumulant propagation algorithm is one such method. It belongs to the general class of deduction–projection estimators: in the idealized decomposition, an exact deduction step advances the computational state and a projection step compresses it. Here, we modify that projection. Like the $K = 3$ cumulant propagation algorithm, we carry a mean, a covariance, and a structured third cumulant; in addition, we regenerate the fourth cumulant from at most four covariance-response modes, with a small fitted coupling matrix. We call this the SSC response closure. The motivating toy model is a Gaussian whose covariance is modulated by background fluctuations, following the idea of super-sample covariance in cosmology. The submitted estimator also includes corrections for third-order dependencies, a one-layer memory of the previous response, and accumulated mean drift. On a matched panel of 30 independently generated Phase-1 networks, SSC reduces layer-32 MSE by 77.2% relative to ARC's factorized $K = 3$ algorithm and by 48.2% relative to factorized $K = 4$; paired-network bootstrap 95% intervals are 68.9–83.2% and 30.3–59.4%, respectively. The displayed normalized operation ratios are $0.501B$ for SSC and $0.373B$ for a parity-checked factorized $K = 3$ port, using FlopScope 0.10 named-operation accounting for both. On this common ledger, SSC uses 34.3% more named operations than $K = 3$. Full factorized $K = 4$ has a separate analytical estimate on the order of $10^2 B$. The corresponding archived submission, #327725, achieved public raw final-layer MSE 2.6597×10^{-6} and adjusted score 1.6533×10^{-6} at public mean effective compute $0.622B$. On the same matched panel, SSC's raw layer-32 MSE is 7.94 times that of full-budget Kerdock QMC. At layer 32, the submitted method therefore remains less accurate than the sampling references used here.

Statement on the use of AI tools. Large-language-model assistants were used under supervision during research, derivation, auditing, and manuscript preparation. The authors reviewed the text and checked the numerical claims against retained experiment and platform receipts.

1. INTRODUCTION

Phase 1 of the ARC White-Box Estimation Challenge (Hilton et al. 2026) provides the weights of one realized network: a bias-free ReLU MLP of width $n = 256$ and depth $L = 32$. With $\boldsymbol{h}^{(0)} \sim \mathcal{N}(0, I)$ and $\boldsymbol{h}^{(\ell)} = \mathrm{ReLU}(\mathbf{W}^{(\ell)}\boldsymbol{h}^{(\ell-1)})$, the task is to estimate

$$\mathbb{E}[\boldsymbol{h}^{(L)} \mid \{\mathbf{W}^{(r)}\}_{r=1}^{L}] \tag{1}$$

within $B = 2.72 \times 10^{11}$ floating-point operations. This report studies an estimator that uses the supplied weights directly rather than relying only on sampled forward passes.

1.1. *Why compare with sampling?*

For a standard sample mean, uncertainty generally decreases as $N^{-1/2}$ (Neyman et al. 2025). The challenge asks whether, under a fixed compute budget, the known weights and input distribution can yield a lower-error estimate than repeated forward passes. This report concerns only mean activations in randomly initialized ReLU networks. Its empirical comparisons are restricted to the named panels and platform receipts; they do not establish results for trained systems, other behavioral properties, or asymptotic performance.

1.2. *Mechanistic estimation*

Analytical moment propagation offers a different route. Precedents include Gaussian assumed-density and covariance-propagation methods (Frey & Hinton 1999; Wright et al. 2024), but every ReLU creates non-Gaussian

dependence. Wu et al. (2026) address this with cumulants and Hermite expansions. Linear layers transport the represented cumulants exactly; nonlinear updates create new terms; and a structured projection keeps the state tractable. This is a *deduction–projection* estimator: advance the calculation, then compress what must be carried onward. In code, deduction and projection can be interleaved in one approximate update.

The cumulant estimator of Wu et al. (2026) degrades as network depth increases. Each ReLU reopens the cumulant hierarchy, and each projection discards some non-Gaussian state. The discarded state can bias the next update of the mean and other cumulants, after which the remaining layers transport and compound the local errors. Wu et al. (2026) conjecture error of order $c_K(L/n)^K$. Their Appendix D tests the depth dependence by overlaying predicted L^K slopes anchored at the rightmost point. This is a finite-range diagnostic, not a free fit of the exponent or a theorem for this finite-width regime. On the challenge networks, MSE climbs layer after layer, and by layer 32 factorized $K = 3$ trails plain Monte Carlo by more than an order of magnitude. A dense fourth cumulant has $O(n^4)$ entries, while factorized $K = 4$ still costs on the order of 10^2B. These observations motivate investigating a more selective projection.

In this submission, we retain the cumulant-propagation backbone of Wu et al. (2026) but change what is kept at fourth order. We are inspired by the separate-universe picture in cosmology: a long background mode changes the covariance measured inside a survey, and averaging over that mode produces a factorized connected four-point function (Takada & Hu 2013; Li et al. 2014; Barreira et al. 2018). SSC, short for super-sample-covariance response closure, reuses that algebra. After each ReLU our algorithm regenerates a fourth-cumulant approximation from at most four covariance-response matrices computed from the live state. Exactness applies only to the cosmological toy model; Appendix B explains it without assuming cosmology background.

Contributions.

- **A covariance-response closure for the fourth cumulant** (Section 3, Appendices B–C): $\boldsymbol{\kappa}_4$ is regenerated each layer from $p \le 4$ live modes coupled by a small matrix $\boldsymbol{\Lambda}$. The closure is exact in a conditional-Gaussian toy model, is deployed as a signed ansatz with indefinite $\boldsymbol{\Lambda}$, and keeps its fourth-order work in matrix form.
- **A documented estimator and fitting boundary** (Section 3): we distinguish Wu's algebraic factorization from SSC's statistical closure, give the layer update, and state exactly which quantities were fitted offline and frozen.
- **A matched empirical comparison with factorized cumulant propagation** (Section 2): on a matched 30-network replication panel at fixed width $n = 256$, the submitted SSC stack has a lower fitted variance-normalized-MSE slope over $L = 8$–32. Its fitted log–log slope is 2.21, versus 2.69 for factorized $K = 3$ and 4.13 for factorized $K = 4$. At layer 32, SSC lowers raw MSE by 77.2% relative to $K = 3$ and by 48.2% relative to full $K = 4$. The common FlopScope named-operation ratios are $0.501B$ for SSC and $0.373B$ for the parity-checked $K = 3$ port; full $K = 4$ has a separate analytical estimate on the order of 10^2B. Deployed submission #327725 scores raw 2.6597×10^{-6} and adjusted 1.6533×10^{-6} at mean effective compute $0.622B$, without propagating sampled inputs.

Section 2 begins with the empirical result, Section 3 explains the estimator and its fitting boundary, and Section 4 concludes with future work. The appendices contain notation, the separate-universe toy model, derivations, implementation details, and the final-stack ablation.

2. MAIN RESULTS

Figure 1 compares all methods on the same independently generated panel of 30 Phase-1 networks. Among the analytic methods, factorized $K = 4$ has the lowest MSE at shallow depth, while SSC has lower MSE in the late-middle layers; the factorized $K = 3$ curve rises more steeply. Both samplers have lower MSE than the analytic methods at the deepest layers, including layer 32, where the platform score is computed.

At layer 32, SSC's MSE is 77.2% lower than factorized $K = 3$'s and 48.2% lower than full factorized $K = 4$'s, with paired-network bootstrap 95% intervals of 68.9–83.2% and 30.3–59.4%; ARC's analytical estimate for $K = 4$ is on the order of 10^2B, far outside the competition budget. The samplers are each run at the full $1B$ budget. At layer 32, SSC's MSE is 7.94 times the Kerdock-frame QMC value (Calderbank et al. 1997) and 2.94 times the i.i.d.-Monte-Carlo value.

Here "parity-checked" means that the port uses the pinned ARC recurrence and agrees with its reference predictions to a worst-layer relative error of 4.1×10^{-6} on the retained parity check. On the common ledger, SSC uses 34.3% more

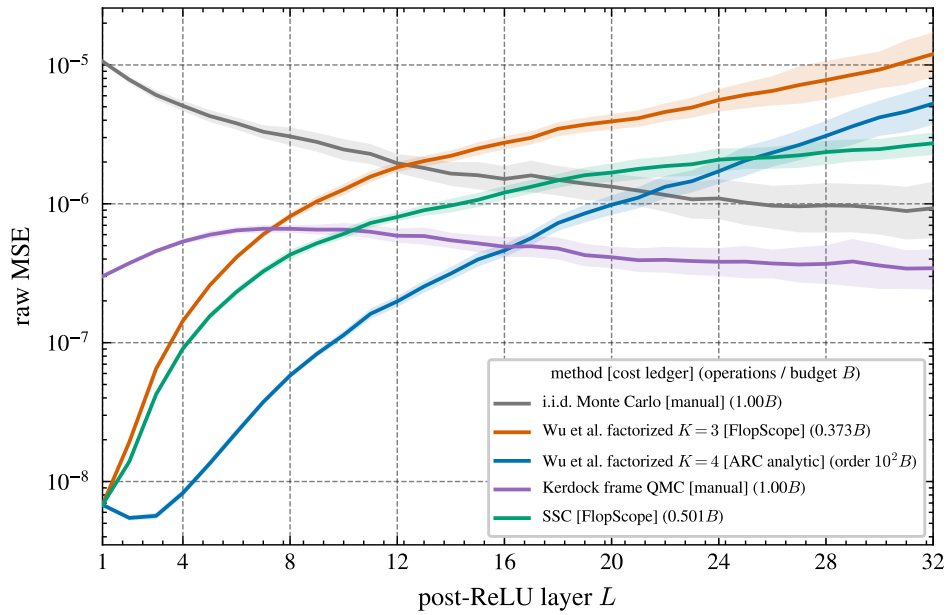

Figure 1. Raw MSE versus depth on matched networks. Curves are network-mean MSE after each ReLU; bands are pointwise network-bootstrap 95% confidence intervals. All methods use the same fixed 30-network panel, with reference means estimated from 10^8 Gaussian inputs per network. The analytic methods' common layer-1 MSE, 6.80×10^{-9}, agrees with the panel-mean reference-noise estimate $\overline{\mathrm{Var}(h_i^{(1)})}/10^8 = 6.82 \times 10^{-9}$, so shallow analytic values near this scale are reference-noise limited. Legend brackets identify the cost ledger: FlopScope 0.10 named operations for the exact SSC source and parity-checked $K = 3$ port, a manual dense-forward ledger for the 64,000-point samplers, and ARC's analytical counter for unrestricted $K = 4$. SSC's $0.501B$ label is F/B; its public $C/B = 0.622$ also includes the residual-time charge λR defined in Appendix A.

named operations than the $K = 3$ port; this compares complete implementations and does not assign the difference to any SSC component.

The platform executed 100 MLPs with zero failures. Fifty formed the public scoring panel; the other 50 were gated evaluation cases not included in the public score. On the public 50, submission #327725 scored raw final-layer MSE 2.6597×10^{-6} and adjusted 1.6533×10^{-6} at mean effective compute $0.622B$.

Submission #327725 is not our team's best-scoring platform submission; #327950 scored better. We nevertheless nominate #327725 because SSC provides a concrete starting point for improving the projection in cumulant propagation. We intend to continue developing deduction–projection estimators beyond this particular closure.

3. THE SSC ESTIMATOR

This section briefly describes the SSC estimator.

3.1. *What changes relative to factorized cumulant propagation?*

A cumulant records the part of a joint moment not explained by lower-order moments. In the state used here, $\boldsymbol{m} = \boldsymbol{\kappa}_1$ is the mean, $\mathbf{C} = \boldsymbol{\kappa}_2$ is the covariance, $\boldsymbol{\kappa}_3$ carries skew and cross-neuron third-order dependence, and $\boldsymbol{\kappa}_4$ is the connected fourth-order part. Retaining more orders preserves additional terms in the nonlinear update, but a general order-r tensor has n^r entries.

The word *factorized* has a specific meaning in Wu et al. (2026): their Appendix S.4.3 rewrites the BASIC truncation so its selected tensors can be contracted through factors rather than first materialized densely. It is an algebraic implementation of the same truncated estimator, not a fitted low-rank statistical model. Their factorized $K = 3$ method carries the state through third order, and factorized $K = 4$ carries the selected fourth-order state systematically at far higher cost, on the order of 10^2B in Figure 1.

SSC shares the lower-order backbone but makes a different projection at fourth order. Instead of retaining the full $K = 4$ state, it rebuilds an approximation after every ReLU from $p \le 4$ live response matrices:

$$\widehat{\kappa}_{4,ijkl} = \sum_{a,b=1}^{p} \Lambda_{ab} \left(B_{a,ij} B_{b,kl} + B_{a,ik} B_{b,jl} + B_{a,il} B_{b,jk} \right). \tag{2}$$

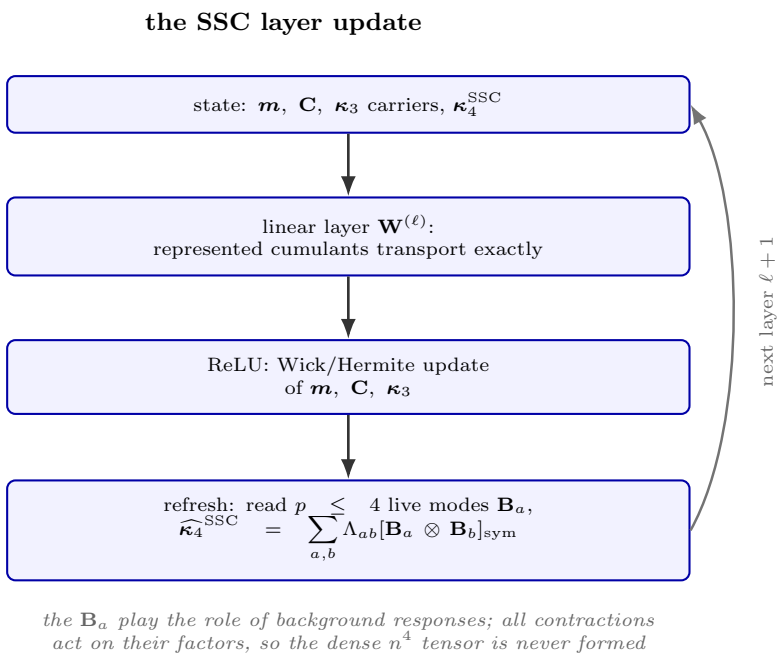

Figure 2. The recurring SSC update. Linear transport is exact for the represented carrier state. The nonlinear Wick/Hermite update and the subsequent projection are approximate. Response factors are refreshed from the live state, which keeps fourth-order work in matrix form. The smaller deployed repairs are described in the text and Appendix D.

Each symmetric $\mathbf{B}_a$ describes a direction in which the current covariance could change, and the small matrix $\boldsymbol{\Lambda}$ couples those directions; the conditional-Gaussian model in Appendix B makes this shape exact. In the network estimator, $\boldsymbol{\Lambda}$ may be indefinite, so Equation (2) is a signed ansatz rather than a literal mixture of Gaussian populations. Every contraction acts on the $\mathbf{B}_a$ factors; an n^4 array is never formed.

3.2. *One layer of SSC*

Figure 2 gives the recurring calculation. More explicitly, one layer performs four operations:

1. **Read the state.** The input consists of $\boldsymbol{m}$, $\mathbf{C}$, structured carriers for $\boldsymbol{\kappa}_3$, and the factored SSC response from the preceding layer.

2. **Apply the weight matrix.** Multiplication by $\mathbf{W}^{(\ell)}$ transports every represented cumulant exactly. Here "exactly" applies to the tensor encoded by the carriers; anything already projected away stays lost.

3. **Apply ReLU.** A Wick/Hermite expansion uses the incoming state to update the mean, covariance, and third-order carriers. This is the approximate deduction–projection step where non-Gaussian information is consumed and new information is created.

4. **Repair and refresh.** The deployed code adds a small set of fresh cross-family $\boldsymbol{\kappa}_3$ atoms, applies its frozen late-layer drift corrector, constructs up to four new $\mathbf{B}_a$ matrices from the live state, and reinstalls Equation (2) for the next layer. It also retains one signed response-memory term from the preceding update.

The third-order repair can add up to 64 fresh cross-family atoms, then caps that cross-family block to the four highest-scoring atoms before the next linear transport. The drift corrector maps 31 deterministic summaries of $\boldsymbol{m}$, $\mathbf{C}$, and $\boldsymbol{\kappa}_3$ through a small $31 \to 32 \to 32 \to 1$ network. These additions, together with response memory, pair-series terms, and numerical changes, all lie between the clean response closure and submission #327725, so the endpoint score measures the deployed system as a whole; Appendix D isolates the response branch operationally.

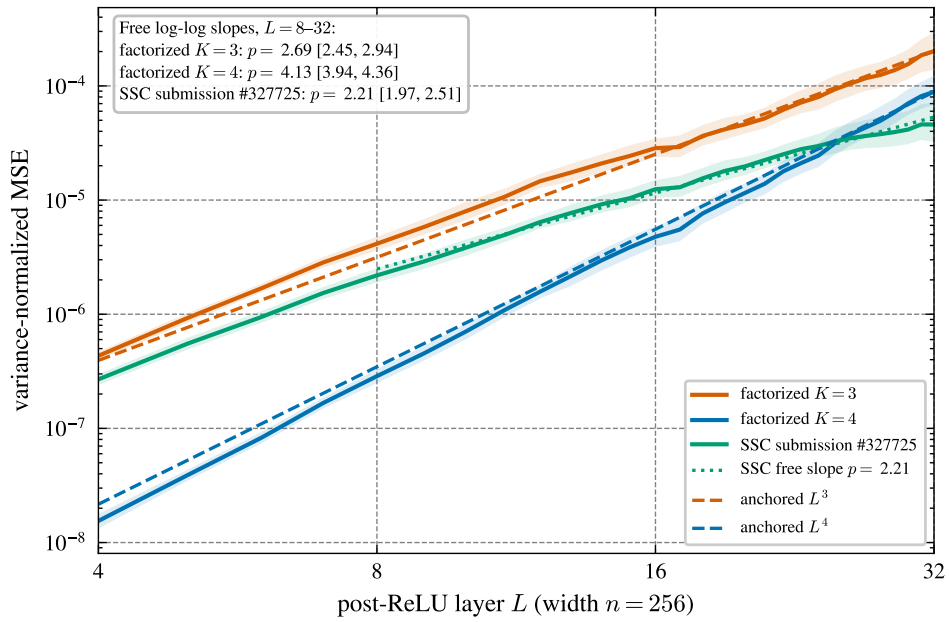

Figure 3. Depth trend at fixed width. Following the diagnostic of Wu et al. (2026), curves show variance-normalized MSE against depth for the same 30 networks as Figure 1. The matched 64,000-point i.i.d. Monte Carlo arm estimates the single-input variance at each layer, and the noise contribution of the 10^8-sample reference is removed. Bands are paired network-bootstrap 95% intervals. Dashed curves have the conjectured L^3 and L^4 slopes anchored at layer 32, as in Wu et al.; the dotted curve is a free SSC fit over layers 8–32. The fitted SSC exponent is 2.21 (paired-bootstrap 95% interval 1.97–2.51). Because $n = 256$ throughout, this figure tests depth dependence only, not joint L/n scaling.

The complete submitted stack also has a different observed depth-error trend. Figure 3 repeats the diagnostic of Appendix D in Wu et al. (2026) on our matched panel. We first convert raw MSE to variance-normalized MSE, using the matched i.i.d. Monte Carlo curve to estimate the single-input variance at each layer. Since every network has $n = 256$, $\log(L/n)$ and $\log L$ differ only by a constant. A free log–log fit over $L = 8$–32 gives exponent $p = 2.21$ for SSC, compared with 2.69 for factorized $K = 3$ and 4.13 for factorized $K = 4$. Over the later range $L = 16$–32, the SSC exponent is 1.97. On this panel and fit window, SSC therefore has the flatter observed depth-error trend at this width, as well as the lower layer-32 error. We do not identify its fitted exponent with a cumulant order: one fixed width cannot establish a joint $(L/n)^p$ law or asymptotic scaling.

3.3. *Fitting used in submission #327725*

Using fitted coefficients does not by itself make a model nonmechanistic. Effective field theory provides a familiar physics analogy: symmetries and scale separation determine the allowed terms, while matching or calibration sets coefficients that summarize unresolved small-scale effects (Burgess 2007). SSC uses fitting in this limited sense. Unlike a controlled effective field theory, however, SSC has no identified small expansion parameter and no corresponding bound on its truncation error. The analogy concerns the separation between structural form and matched coefficients, not control of the omitted terms. Equation (2) and the layerwise recurrence specify the structural form; offline fitting sets a small number of coefficients for higher-order effects that the compact state does not resolve explicitly. The analogy motivates this modeling choice but does not validate it by itself, so all coefficients are frozen before evaluation and judged on held-out networks.

At inference, SSC receives the supplied weight matrices and uses its frozen tables; it fits nothing to the test network, uses no reference activations or private targets, and propagates no sampled inputs. The same response tables and correction weights are frozen for every evaluated network.

Those constants were learned offline. First, the six coefficients defining $\mathbf{Q}$ in Equation (C3) were fitted layer by layer by ridge least squares to the measured off-diagonal target $\kappa_{iiij} + \kappa_{jjji}$. For the coupling fit, let $\boldsymbol{w}_j$ be the incoming weight vector of next-layer neuron j and define

$$q_{aj} = \boldsymbol{w}_j^\top \mathbf{B}_a \boldsymbol{w}_j, \qquad \widehat{\kappa}_{4,j} = 3\, \boldsymbol{q}_j^\top \boldsymbol{\Lambda} \boldsymbol{q}_j. \tag{3}$$

Each symmetric $\mathbf{\Lambda}$ was fitted by ridge least squares to the measured marginal fourth cumulant of the next preactivation. The ridge value was selected on separate development networks by the MSE of the resulting fourth-order Edgeworth contribution to the next ReLU mean. The tables were then frozen. At inference, the response update applies the four raw modes directly and diagonalizes the resulting 4×4 coupling matrix at each layer. The one-layer response-memory coefficient was selected in a separate development study and frozen at -0.3.

The recurrent drift corrector was fitted separately on 160 networks from the public Phase-1 corpus and checked on 16 held-out networks. The 30-network replication panel comes from a separate evaluation corpus and contains none of these 176 networks. The corrector's teacher-forced target was the predictable one-step mean drift over its fixed late-layer range; it was not trained through the final competition score. The third-order repair, pair-series degree, layer ranges, and all numerical settings were likewise fixed before evaluation.

4. CONCLUSION AND FUTURE WORK

SSC turns a dense fourth-order hierarchy into a compact covariance-response closure. The separate-universe toy model supplies algebra that is exact in a controlled setting; deployed in a network, the same shape becomes a fitted, signed estimator with frozen coefficients. Submission #327725 provides one tested mechanistic baseline, and its measured gap to sampling leaves the projection as a target for further study.

While finishing this work, we did a separate controlled audit using the same four underlying response modes, orthogonalized their readout for conditioning, but changed the regression target: instead of fitting $\mathbf{\Lambda}$ to marginal fourth-cumulant quantities, it fitted the response to the mean correction consumed after the next weight and ReLU update. On that audit's predecessor closure baseline, the refitted table reduced layer-32 MSE by 13.6% and had lower MSE on 26 of 30 networks. The measured reduction came from changing the target, not from adding modes or from orthogonalization alone. Those coefficients were not embedded in submission #327725, and they have not been evaluated inside its complete stack. Integrating them into the archived estimator and testing the result without retuning is therefore future work. Transfer to trained weights, other widths, biases, and residual architectures also remains open.

In Phase 2, we plan to investigate other deduction–projection estimators and mechanistic approaches to final-layer mean estimation.

APPENDIX

A. PROBLEM AND NOTATION

The challenge networks are width $n = 256$, depth $L = 32$, bias-free ReLU MLPs with He-scaled weights. With $\theta = \{\mathbf{W}^{(\ell)}\}_{\ell=1}^{L}$,

$$\boldsymbol{z}^{(\ell)} = \mathbf{W}^{(\ell)}\boldsymbol{h}^{(\ell-1)}, \qquad \boldsymbol{h}^{(\ell)} = \mathrm{ReLU}(\boldsymbol{z}^{(\ell)}), \qquad \boldsymbol{m}^{(\ell)}(\theta) = \mathbb{E}[\boldsymbol{h}^{(\ell)} \mid \theta]. \tag{A1}$$

Raw error is the coordinate-mean squared error of the final predicted mean. For each successful network m, the platform uses multiplier $\max(0.1, C_m/B)$, where $C_m = F_m + \lambda R_m$, F_m is the FlopScope named-operation count, R_m is residual wall time, and $\lambda = 10^{11}$ FLOP-equivalents per second. It averages the per-network products, so the published adjusted score need not equal mean raw MSE times mean utilization. We distinguish this effective compute from the research operation ledger.

Our notation is: $\boldsymbol{\kappa}_r$ for the order-r activation cumulant, with $\boldsymbol{\kappa}_1 = \boldsymbol{m}$ and $\boldsymbol{\kappa}_2 = \mathbf{C}$; $\mathbf{B}_a$ for a symmetric live covariance-response mode; $\mathbf{\Lambda}$ for the small symmetric matrix coupling those modes; $\mathrm{Sym}(\boldsymbol{v}, \mathbf{M})_{ijk} = v_i M_{jk} + v_j M_{ik} + v_k M_{ij}$; $\boldsymbol{s}^{(\ell)}$ for local mean-source error; and $\mathbf{J}_\ell$ for its linearized transport through layer ℓ.

For a scalar Gaussian, the leading ReLU moment is

$$\mathbb{E}[\mathrm{ReLU}(Z)] = \sigma\phi(\mu/\sigma) + \mu\Phi(\mu/\sigma). \tag{A2}$$

Higher corrections are assembled from the Wick/Hermite coefficients and stored cumulants as in Wu et al. (2026). Gaussian and assumed-density propagation are established ideas (Frey & Hinton 1999; Wright et al. 2024). The contribution described here is the particular response closure, the carrier reorganization, and their deployment in submission #327725; moment propagation itself is prior work.

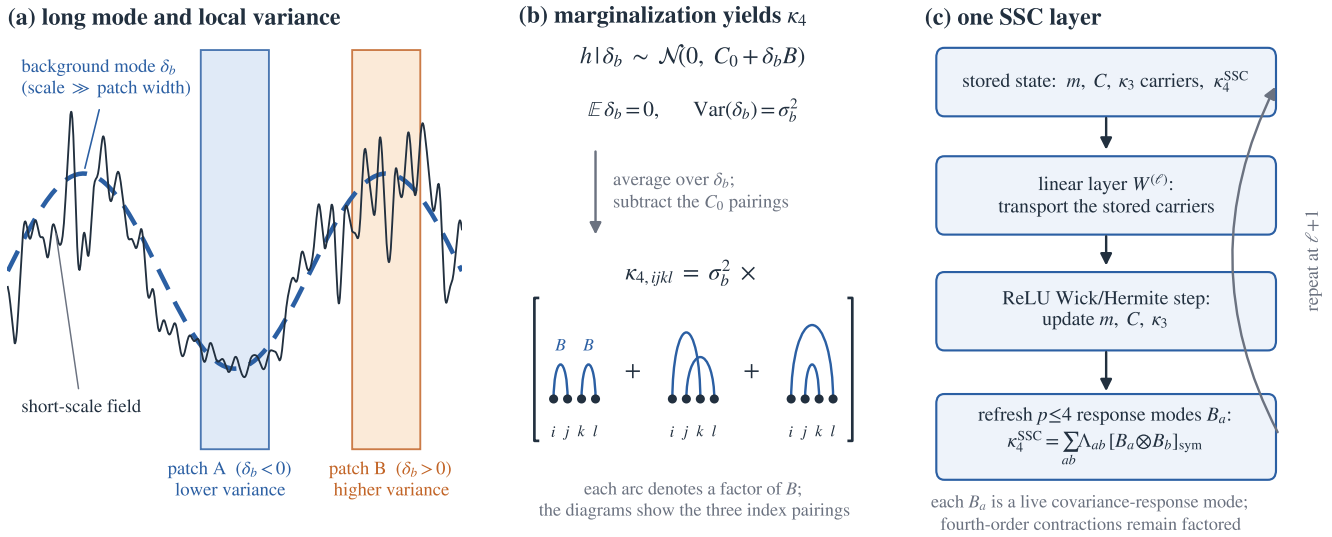

Figure B1. From covariance response to the SSC recurrence. In the exact toy model, a background variable changes the covariance of the short-scale variables, and averaging over that background gives Equation (B2). In SSC, up to four matrices derived from the current neural state are coupled by a frozen $\mathbf{\Lambda}$ and refreshed after each ReLU. The shared algebra motivates the closure; it does not assert a literal Gaussian-mixture model for neural activations.

B. A HALF-PAGE TOY UNIVERSE

In cosmology, a fluctuation with wavelength longer than the observed survey can look locally like a change in the background universe. That long mode can change the covariance of shorter modes inside the survey. The following minimal model isolates the algebra SSC borrows.

Let $\boldsymbol{h} \in \mathbb{R}^n$ denote the vector of short-scale quantities and let δ_b denote one scalar background fluctuation. Conditional on a fixed value of δ_b, assume

$$\boldsymbol{h} \mid \delta_b \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_0 + \delta_b \mathbf{B}), \qquad \mathbb{E}\delta_b = 0, \qquad \mathrm{Var}(\delta_b) = \sigma_b^2. \tag{B1}$$

Here $\mathbf{C}_0$ is the baseline covariance when the background is zero. The symmetric matrix $\mathbf{B}$ is the covariance *response*: increasing δ_b by one unit changes the conditional covariance by $\mathbf{B}$. The number σ_b^2 measures how much the background varies across hypothetical surveys. The background need not be Gaussian for the calculation below; only its first two moments enter.

"Valid conditional covariance" means $\mathbf{C}_0 + \delta_b \mathbf{B} \succeq 0$ for every δ_b that can occur. For example, δ_b may have bounded support small enough to preserve positive semidefiniteness. This qualification matters: an arbitrary nonzero $\mathbf{B}$ combined with an unbounded background would not generally define a valid Gaussian covariance for every draw.

For a zero-mean Gaussian, Wick's theorem says that a fourth moment is the sum of its three pairings. Applying it at fixed δ_b, averaging over the background, and subtracting the three pairings constructed from the unconditional covariance $\mathbf{C}_0$ gives the connected fourth moment, or fourth cumulant,

$$\kappa_{4,ijkl} = \sigma_b^2 \left(B_{ij} B_{kl} + B_{ik} B_{jl} + B_{il} B_{jk} \right). \tag{B2}$$

Thus one coherent covariance response produces a dense connected four-point function using only one matrix. With several background variables ϵ_a, their covariance $\Lambda_{ab} = \mathrm{Cov}(\epsilon_a, \epsilon_b)$ replaces the scalar σ_b^2 and yields Equation (2).

The analogy is structural: SSC does not assert that a neural layer is literally a population of separate Gaussian universes. It borrows Equation (B2) as a compact closure shape after the exact probabilistic interpretation has been left behind.

C. RESPONSE IDENTITY, MODES, AND CAVEATS

For p background modes, let

$$\boldsymbol{h} \mid \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{m}, \mathbf{C}(\boldsymbol{\epsilon})), \quad \mathbf{C}(\boldsymbol{\epsilon}) = \mathbf{C}_0 + \sum_{a=1}^{p} \epsilon_a \mathbf{B}_a, \quad \mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbf{\Lambda} \succeq 0, \tag{C1}$$

with fixed conditional mean and valid conditional covariances. Conditional Wick expansion gives three products of entries of $\mathbf{C}(\boldsymbol{\epsilon})$. Averaging them produces the Gaussian pairings of $\mathbf{C}_0$ plus $\sum_{ab}\Lambda_{ab}B_{a,ij}B_{b,kl}$ and its two permutations. Subtracting the unconditional Gaussian pairings proves Equation (2). This is the law-of-total-cumulance origin of the closure.

Deployed SSC uses at most four modes,

$$\mathbf{B}_1 = \mathbf{C}, \qquad \mathbf{B}_2 = \mathbf{Q}/\bar{v}, \qquad \mathbf{B}_3 = \bar{v}\mathbf{I}, \qquad (B_4)_{ij} = \tfrac{1}{2}(g_i + g_j)C_{ij}, \tag{C2}$$

where $\bar{v} = n^{-1}\operatorname{tr}\mathbf{C}$, $\boldsymbol{g}$ is the standardized centered vector of ReLU gate probabilities, and $\mathbf{Q}$ is a frozen six-feature quadratic form of the live lower-order state. To state it explicitly, let $\boldsymbol{v} = \operatorname{diag}\mathbf{C}$, $t_i = \kappa_{iii}$, $K_{ij} = \kappa_{iij}$, $\mathbf{D}_m = \operatorname{diag}(\boldsymbol{m})$, and let $\mathcal{P}_{\text{off}}$ set the diagonal of a matrix to zero. At each layer the deployed form is

$$\begin{aligned}\mathbf{Q} = \mathcal{P}_{\text{off}}\big[&c_0\bar{v}\mathbf{C} + c_1(\mathbf{C}\odot\mathbf{C}) + \tfrac{c_2}{2}(\mathbf{D}_m\mathbf{K}^{\mathsf{T}} + \mathbf{K}\mathbf{D}_m)\\ &+ c_3(\boldsymbol{t}\boldsymbol{m}^{\mathsf{T}} + \boldsymbol{m}\boldsymbol{t}^{\mathsf{T}}) + c_4\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} + c_5\bar{v}^2\mathbf{1}\mathbf{1}^{\mathsf{T}}\big],\end{aligned} \tag{C3}$$

with a frozen layer-specific coefficient vector $\boldsymbol{c}$. Diagonalizing the small $\boldsymbol{\Lambda}$ converts the closure to a short sum of signed symmetric factor pairs. All required weight contractions and repeated-index slices then act directly on the response matrices.

Three caveats apply. First, SSC fits $\boldsymbol{\Lambda}$ without a positive-semidefinite constraint; an indefinite fit is algebraically useful but has no literal Gaussian-mixture interpretation. Second, a realizable covariance mixture generally generates sixth and higher cumulants, which the truncated recurrence omits. Third, the fitted tables are specific to the random-network distribution studied here: running without sampled inputs at inference does not make the method fit-free or distribution-free, and the fixed pseudorandom probes used for low-rank numerical sketching are not samples from the network's input distribution. Moment closure and response modeling are broad prior fields (Ernst et al. 2019; Uhlemann 2018); this report evaluates the particular cross-domain construction above.

D. STRUCTURED RECURRENCE AND IMPLEMENTATION DETAILS

The linear step obeys

$$\boldsymbol{\kappa}_r[\mathbf{W}\boldsymbol{h}] = \mathbf{W}^{\otimes r}\cdot\boldsymbol{\kappa}_r[\boldsymbol{h}]. \tag{D1}$$

The third cumulant is represented by canonical-polyadic (CP) columns, atoms $\operatorname{Sym}(\boldsymbol{v},\mathbf{M})$, shared-middle groups, and paired groups. Linear transport, diagonal Wick scaling, partial contraction, and repeated-index slice extraction operate directly on these carriers. This lower-order carrier algebra is inherited from Appendix S.4.3 of Wu et al. (2026); Table D1 maps the mathematical objects to their exact implementation. Compression and closure remain statistical approximations.

A schematic layer update is

$$\mathcal{T}_{\text{pre}} \leftarrow \textsc{LinearTransport}(\mathcal{T}, \mathbf{W}^{(\ell)}), \tag{D2}$$

$$(\boldsymbol{m}, \mathbf{C}, \boldsymbol{\kappa}_3) \leftarrow \textsc{ReLUWick}(\mathcal{T}_{\text{pre}}), \tag{D3}$$

$$\boldsymbol{\kappa}_3 \leftarrow \textsc{FreshCrossRepair}(\boldsymbol{\kappa}_3;\ K = 64,\ \text{retained} = 4), \tag{D4}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + \textsc{DriftCorrector}(\boldsymbol{m}, \mathbf{C}, \boldsymbol{\kappa}_3), \tag{D5}$$

$$\boldsymbol{\kappa}_4^{\text{SSC}} \leftarrow \textsc{ResponseClosure}(\{\mathbf{B}_a\}, \boldsymbol{\Lambda}) + \textsc{OneStepResponseMemory}, \tag{D6}$$

with the repair and corrector active only on their frozen layer ranges; the archived source defines the exact operational order. The cross repair ranks fresh angular CP columns in closed form, adds up to 64, and caps the transported cross-family block at four. The recurrent corrector uses 31 deterministic features and a small $31 \to 32 \to 32 \to 1$ head trained on one-step drift. The submitted stack also fixes the one-layer response-memory coefficient at -0.3 and the pair-series degrees at four for the $(1,1)$ group and two for the $(2,1)$ group. We report the stack as one deployed method rather than combining gains measured on earlier component prototypes.

Here and below, superscripts $-$ and $+$ denote the preactivation and uncorrected post-ReLU states. The $(1,1)$ and $(2,1)$ labels denote the two-index raw-moment sectors $\mathbb{E}[Y_iY_j]$ and $\mathbb{E}[Y_i^2Y_j]$, respectively, for $Y = \operatorname{ReLU}(Z)$. If

$$\omega_{m,p,i} = \mathbb{E}\left[\frac{d^m}{dz^m}\operatorname{ReLU}(Z_i)^p\right], \qquad \Delta M^{(p,q)}_{ij,m} = \frac{\omega_{m,p,i}\omega_{m,q,j}}{m!}(C^-_{ij})^m, \tag{D7}$$

Table D1. Reimplementation map for submission #327725. Each row gives the paper definition or inherited rule, followed by the exact source symbols.

Object	Definition and source
Response core and modes	Equations (C2) and (C3); `_response_k4`, `core_coefficients`
$\mathbf{\Lambda}$ coupling	Equations (2) and (3); `_response_k4`, `ssc_lambdas`
Pair series	Equation (D7), degrees 4 and 2; `Estimator.__init__`, `PAIR_DEG`, `PAIR21_DEG`
Response memory	Equation (D8), $\beta = -0.3$, age one; `_propagate`, `K4_MEM_BETAS`
Drift corrector	Equations (D9)–(D11), zero-based indices 8–30; `_cnc_correction`, `_CNC_MODEL_JSON`, `CNC_START`
Third-cumulant carriers	Wu et al., Appendix S.4.3; direct transport, Wick scaling, and $(3)/(2,1)$ slice reads; `_Carrier.contract_W`, `contract_wick_`, `_compute_dslice`, `_eval_terms`
Cross-family repair	Equation D4, at most 64 fresh columns followed by a four-column cap; `_inject_k4`, `cap_xcross_`, `XK3_TOPK`, `XK3_CAP`

then the pure pair series is the sum of these terms. Submission #327725 keeps $m \le 4$ in the $(1,1)$ sector and $m \le 2$ in the $(2,1)$ sector.

The response-memory term is also explicit. Let $(\mathbf{A}_r, s_r)$ be a signed factor of the fresh response from the preceding layer, let $\widetilde{\mathbf{A}}_r = \mathbf{W}\mathbf{A}_r\mathbf{W}^{\mathsf{T}}$, and let $\mathbf{D}_g = \operatorname{diag}(\Phi(\boldsymbol{m}^-/\boldsymbol{\sigma}^-))$ at the current ReLU. The age-one factor appended beside the new response is

$$\mathbf{A}_r^{\mathrm{mem}} = \sqrt{|\beta|}\,\mathbf{D}_g\widetilde{\mathbf{A}}_r\mathbf{D}_g, \qquad s_r^{\mathrm{mem}} = \operatorname{sign}(\beta)s_r, \qquad \beta = -0.3. \tag{D8}$$

Because the represented fourth cumulant is quadratic in $\mathbf{A}_r$, this is a coefficient-β gated memory of the previous response. Older factors are dropped.

For completeness, the drift corrector is active at zero-based layer indices $8, \dots, 30$, corresponding to post-ReLU layers 9–31 in the one-based figure labels. Its 31 inputs can be written compactly as follows. Set $\boldsymbol{\sigma}^- = (\operatorname{diag}\mathbf{C}^-)^{1/2}$, $\boldsymbol{a} = \boldsymbol{m}^-/\boldsymbol{\sigma}^-$, $\boldsymbol{g} = \Phi(\boldsymbol{a})$, $\boldsymbol{\varphi} = \phi(\boldsymbol{a})$, $\eta_i = \kappa^-_{iii}/(\sigma_i^-)^3$, and let $\mathbf{R}^+$ be the post-ReLU correlation matrix. With $K^+_{ij} = \kappa^+_{iij}$ and π_i the squared row loading of the deterministic top-eight subspace sketch of $\mathbf{C}^+$, the first 14 features are

$$\begin{aligned}\boldsymbol{f}_i^{(0)} = (&a_i, g_i, \eta_i, \log(\sigma_i^-/\overline{\sigma^-}), \langle R^+_{ij}\rangle_j, \langle (R^+_{ij})^2\rangle_j, \log(C^+_{ii}/\overline{\operatorname{diag}\mathbf{C}^+}),\\ &(\mathbf{C}^+\boldsymbol{a})_i, ((\mathbf{C}^+)^2\boldsymbol{a})_i, (\mathbf{R}^+\boldsymbol{g})_i, (\mathbf{C}^+\boldsymbol{g})_i, (\mathbf{K}^+\mathbf{1})_i, ((\mathbf{K}^+\odot\mathbf{K}^+)\mathbf{1})_i, \pi_i).\end{aligned} \tag{D9}$$

For the remaining 17, set the diagonal of the preactivation correlation $\mathbf{R}^-$ to zero. Let j_{it} be the indices of the eight largest $|R^-_{ij}|$ in row i, $r_{it} = R^-_{ij_{it}}$, and $u_{it} = |r_{it}|$. Define

$$S_i(x) = \sum_{t=1}^{8} r_{it}x_{j_{it}}, \quad U_i(x) = \sum_{t=1}^{8} u_{it}x_{j_{it}}, \quad P_{\gamma i}(x) = \frac{\sum_{j\neq i} e^{\gamma(R^-_{ij})^2}x_j}{\sum_{j\neq i} e^{\gamma(R^-_{ij})^2}}. \tag{D10}$$

Writing $\widetilde{\eta} = \operatorname{clip}(\eta, -20, 20)$ and $\ell_i = \log(C^-_{ii}/\overline{\operatorname{diag}\mathbf{C}^-})$, the second block is

$$\begin{aligned}\boldsymbol{f}_i^{(1)} = (&S_i(a), S_i(g), S_i(\varphi), S_i(\widetilde{\eta}), U_i(\varphi), U_i(\ell), \textstyle\sum_t r_{it}^2, u_{i1},\\ &P_{8i}(a), P_{8i}(g), P_{8i}(\varphi), P_{32i}(a), P_{32i}(g), P_{32i}(\varphi), r_{i1}, a_{j_{i1}}, \varphi_{j_{i1}}).\end{aligned} \tag{D11}$$

The concatenated vector $(\boldsymbol{f}_i^{(0)}, \boldsymbol{f}_i^{(1)})$ is standardized by frozen training-set means and standard deviations before entering the two tanh hidden layers.

All code pointers in Table D1 refer to the exact archived `estimator.py` for submission #327725, whose SHA-256 is `f87bf773d76244a1e480223a7da2d3b3f3af9d3f72426c47b01bbce1746f3148`. The source is the executable specification for implementation details inherited from Wu et al. or too lengthy to duplicate here.

Figure D1 instead makes one operational intervention on the final submitted stack: we rerun the exact archived source with the response constructor disabled at every layer and no other component refitted. The pair-series terms and frozen mean-drift corrector remain active, while response memory and the response-derived fourth-to-third-order repair necessarily become inactive because no response tensor exists for them to transport or consume. The comparison therefore measures the complete response branch inside the submitted estimator, interactions included.

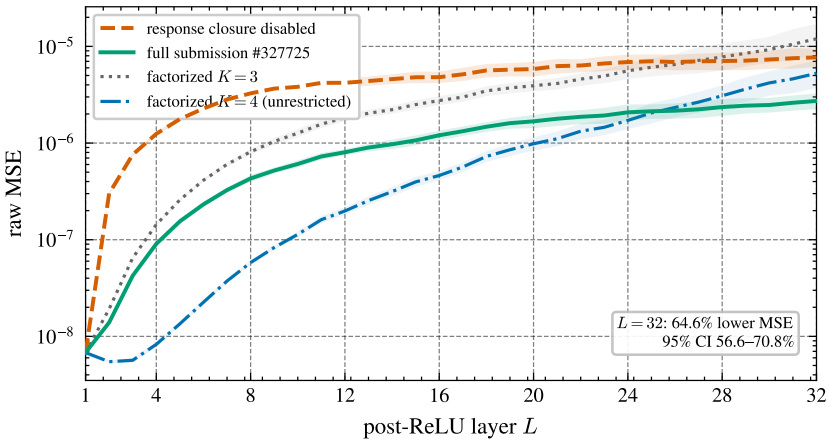

Figure D1. Matched ablation of the submitted response branch. Curves are mean raw MSE on the fixed 30-network panel from Figure 1; bands are pointwise network-bootstrap 95% intervals. The factorized references are unchanged, and full $K = 4$ remains unrestricted and out of budget. With the independent pair and drift corrections retained, disabling the response raises layer-32 MSE from 2.728×10^{-6} to 7.698×10^{-6}. The full response branch is lower by 64.6% (paired-bootstrap 95% interval 56.6–70.8%) and is lower on all 30 networks. Because the ablated arm is cheaper, this is a raw-accuracy ablation, not a compute-matched score comparison.

E. ERROR TRANSPORT AND AN OPEN MECHANISM QUESTION

Local approximation errors can accumulate even when no single update is large. If $\boldsymbol{e}^{(\ell)}$ is mean error, $\mathbf{J}_\ell$ its linearized transport, and $\boldsymbol{s}^{(\ell)}$ the local source, then schematically

$$\boldsymbol{e}^{(L)} = \sum_{\ell=1}^{L} \mathbf{J}_L \cdots \mathbf{J}_{\ell+1} \boldsymbol{s}^{(\ell)}, \qquad \frac{\partial \mathbb{E}[\mathrm{ReLU}(Z)]}{\partial \sigma^2} = \frac{\phi(\mu/\sigma)}{2\sigma}, \quad Z \sim \mathcal{N}(\mu, \sigma^2). \tag{E1}$$

They motivate the possible chain $\delta\boldsymbol{\kappa}_3 \to \delta\mathbf{C}_{\mathrm{off}} \to \delta\sigma^2 \to$ local mean source $\to$ transported deep drift. Figure D1 reports a final-stack ablation; separating its remaining error among covariance, third-cumulant, and higher-order channels is still open.

REFERENCES

Barreira, A., Krause, E., & Schmidt, F. 2018, Complete Super-Sample Lensing Covariance in the Response Approach. https://arxiv.org/abs/1711.07467

Burgess, C. P. 2007, Annual Review of Nuclear and Particle Science, 57, 329, doi: 10.1146/annurev.nucl.56.080805.140508

Calderbank, A. R., Cameron, P. J., Kantor, W. M., & Seidel, J. J. 1997, Proceedings of the London Mathematical Society, 75, 436, doi: 10.1112/S0024611597000403

Ernst, O. K., Bartol, T., Sejnowski, T., & Mjolsness, E. 2019, Deep Learning Moment Closure Approximations Using Dynamic Boltzmann Distributions. https://arxiv.org/abs/1905.12122

Frey, B. J., & Hinton, G. E. 1999, Neural Computation, 11, 193

Hilton, J., Christiano, P., & Wu, W. 2026, Announcing the ARC White-Box Estimation Challenge, Alignment Research Center blog. https://www.alignment.org/blog/announcing-the-arc-white-box-estimation-challenge/

Li, Y., Hu, W., & Takada, M. 2014, Physical Review D, 89, 083519, doi: 10.1103/PhysRevD.89.083519

Neyman, E., Lecomte, V., Wu, W., et al. 2025, Competing with Sampling, Alignment Research Center blog. https://www.alignment.org/blog/competing-with-sampling/

Takada, M., & Hu, W. 2013, Physical Review D, 87, 123504, doi: 10.1103/PhysRevD.87.123504

Uhlemann, C. 2018, Finding Closure: Approximating Vlasov–Poisson Using Finitely Generated Cumulants. https://arxiv.org/abs/1807.07274

Wright, O., Nakahira, Y., & Moura, J. M. F. 2024, in Proceedings of Machine Learning Research, Vol. 238, Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, 4087–4095. https://proceedings.mlr.press/v238/wright24a.html

Wu, W., Lecomte, V., Winer, M., et al. 2026, Estimating the Expected Output of Wide Random MLPs More Efficiently Than Sampling, doi: 10.48550/arXiv.2605.05179