Metered dead-compacted Kerdock cubature with exact compiler reductions

Konstantin Baltsat

Private Phase 1 algorithmic-contribution packet
9 August 2026
AIcrowd submission #326680

Abstract

This paper documents the estimator in successfully graded AIcrowd submission #326680. It estimates the final-layer activation mean of a width-256, depth-32, bias-free ReLU MLP under a standard Gaussian input. The angular backbone is the complete dimension-256 real Kerdock construction: 129 mutually unbiased bases, both signs of every basis vector, and hence 66,048 spherical directions. Positive homogeneity removes radius sampling exactly. A signed fast Walsh–Hadamard transform evaluates the first layer.

The principal deployed addition is a fully billed dead-compaction route. A diagonal-Gaussian recurrence proposes dead coordinates, and the first 1,652 rows of the same Kerdock carrier act as a reused pilot. A coordinate is omitted only if its analytic standardized mean is below -3 and every pilot preactivation is nonpositive. Omitted inputs are not silently set to zero: their diagonal-recurrence mean is added analytically to the next preactivation. The remaining columns are propagated through deterministic compacted routes, followed by a three-layer dead/on/kink fold. Every predict-time numerical operation is a FlopScope primitive.

A frozen target-free gate used four newly generated MLPs, two independent coordinate-sign rotations, and two independent 131,072-pair antithetic truth replicates. Relative to the un-compacted Kerdock carrier, the candidate's cross-MSE ratio was 0.98066 and its effective-compute ratio was 0.85781. The deployed source then received four exact, prediction-preserving compiler reductions: compacting pilot inputs already proved zero, reusing the existing pilot mask, and assembling each Winograd quadrant block once. On the released stack the official job completed 100/100 MLPs with zero failures, raw final-layer MSE $2.4258079705 \times 10^{-7}$, adjusted score $1.4391444175 \times 10^{-7}$, and mean effective compute $1.613677943 \times 10^{11}$. This is evidence for one legal package, not a claim of unbiasedness, fresh-private superiority, or a Phase-2 win.

1 Scope, prior art, and contribution boundary

The challenge, Phase-1 launch note, starter-kit revision, and contribution criteria are available at the following full URLs:

- `https://www.aicrowd.com/challenges/arc-white-box-estimation-challenge-2026`
- `https://discourse.aicrowd.com/t/phase-1-launch-deeper-models-and-increased-prizes/18026`
- `https://discourse.aicrowd.com/t/phase-1-update-flopscope-v0-10-0-cost-model-fixes-residual-time-safeguards-and-updated-deadlines/18125`
- `https://github.com/AIcrowd/whest-starterkit`
- `https://discourse.aicrowd.com/t/algorithmic-contribution-prize-guidelines-how-arc-judges-these-prizes-discretion-technical-writeups-llm-usage/18041`

This write-up maps to exactly one successfully graded submission: `https://www.aicrowd.com/challenges/arc-white-box-estimation-challenge-2026/submissions/326680`. The public dataset exposes targets. They were not used to construct a target lookup, to choose per-row outputs, or to bypass accounting. The promotion evidence for this estimator was frozen on newly generated MLPs before the one official calibration submission.

The closest published prior art is the structure-aware SOX estimator, submission 319341: `https://discourse.aicrowd.com/t/phase-i-submission-structure-aware-estimation-for-random-relu-mlps/18106`. SOX already discloses analytic dead/kink/on classification, pilot reclassification, structural sampling, and sparse continuation. This paper does not claim those ideas as new. Its narrower contribution is:

1. a complete Kerdock angular carrier with an algebraic construction certificate, exact radial conditioning, and a metered signed-FWHT first layer;
2. a conservative intersection of analytic and reused-carrier pilot dead tests, with analytic omitted-input fill rather than zero fill;
3. deterministic compacted continuation, a metered three-layer fold, and four exact source-level compiler reductions in one self-contained current-stack package; and
4. measured evidence for the combination, including negative results that bound what the mechanism does not solve.

Against the published SOX result, e211 is approximately 7.236% better on adjusted score and uses 17.515% less mean effective compute, while its raw MSE is approximately 11.324% worse. Thus its measured advantage is compute-driven. It should not be described as a more accurate raw estimator than SOX.

2 Estimator

2.1 Complete Kerdock angular carrier

Let $d = 256 = 2^8$. A maximal real mutually unbiased basis family in this dimension has $d/2+1 = 129$ bases. The construction uses the standard basis and 128 signed Walsh bases derived from a binary Kerdock set in a fixed $\mathrm{GF}(2^7)\oplus\mathrm{GF}(2)$ field model. The offline certificate materializes 128 alternating 8×8 binary matrices and checks all

$$\binom{128}{2} = 8{,}128$$

pairwise XOR differences. Every difference has binary rank eight, which is the required mutual-unbiasedness condition. Construction references are:

- `https://arxiv.org/abs/quant-ph/0502024`
- `https://pages.uoregon.edu/kantor/PAPERS/SFams.pdf`
- `https://doi.org/10.1112/S0024611597000403`

The finite rule contains both antipodes of every basis vector,

$$\mathcal{X} = \{\pm b : b \text{ is in one of the 129 bases}\}, \qquad |\mathcal{X}| = 2 \cdot 129 \cdot 256 = 66{,}048.$$

Antipodality cancels odd moments through degree five, and the complete MUB family satisfies the spherical fourth-moment identity

$$\frac{1}{129 \cdot 256} \sum_{x \in \mathcal{B}} \langle u, x \rangle^4 = \frac{3}{256(256+2)}$$

for unit u. This certifies a spherical 5-design. It does not make the rule exact or unbiased for a deep ReLU network, whose angular integrand is piecewise linear with network-dependent activation cells.

For a fixed bias-free ReLU MLP f_W, positive homogeneity gives

$$f_W(ru) = r f_W(u), \qquad r \geq 0.$$

Writing $G = RU$, where $G \sim N(0, I_d)$, U is uniform on $\mathbb{S}^{d-1}$, and R is independent of U, gives

$$\mathbb{E} f_W(G) = \mathbb{E} R \, \mathbb{E}_U f_W(U), \qquad \mathbb{E} R = \sqrt{2}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}.$$

The radial factor is therefore exact for this contest model. The angular average remains a finite, generally biased cubature.

2.2 Diagonal recurrence and conservative dead mask

In parallel with the carrier, the estimator runs the standard diagonal-Gaussian mean/variance recurrence. For a scalar preactivation modeled as $Z \sim N(\mu, \sigma^2)$, with $\alpha = \mu/\sigma$, it uses

$$\mathbb{E}[\max(0, Z)] = \sigma\phi(\alpha) + \mu\Phi(\alpha).$$

The corresponding closed-form second moment supplies the next diagonal variance. This approximation is cheap and target-free, but it discards cross-neuron dependence. It is used as a classifier and analytic fill, not as the final estimator.

The leading 1,652 rows of the already constructed Kerdock activation tensor are propagated as a separate billed pilot. At each prefix layer, output coordinate j is marked dead only when both tests agree:

$$\alpha_j < -3 \quad \text{and} \quad z_{rj} \leq 0 \text{ for every one of the 1,652 pilot rows.}$$

The pilot is 2.501% of the 66,048-direction carrier. It is not an additional random sample and is not chosen from target-bearing examples. The analytic test supplies breadth; the pilot veto prevents an analytic tail approximation from deleting a directionally active coordinate.

2.3 Analytic fill and compacted continuation

Let A be the retained input-coordinate set at a layer and D its omitted complement. The sampled carrier propagates only A, but the preactivation for retained output coordinates receives

$$b_{\text{dead}} = m^\top W - m_A^\top W_A,$$

where m is the diagonal-recurrence activation mean. This is the analytic mean contribution of the omitted block. It avoids treating an empirically dead sampled column as an exact zero random variable.

The route is selected only from widths available before continuation:

- use the unchanged chunked Kerdock carrier when both widths are 256;
- use an exact-zero route when the pilot-live width is at least 48 below the analytic-active width;
- use a cold compacted route when both active widths are at most 176;
- otherwise use dense compacted propagation with analytic fill.

These constants were frozen on a separate development network before the confirmation suite. They were not swept on challenge targets.

Stage	State transformation	Counted implementation
Angular rule	129 bases, both antipodes, one deterministic coordinate-sign rotation	Asset load, sign generation, stack/reshape primitives
Layer 1	Signed Walsh bases through W_1	Eight signed FWHT butterfly stages
Analytic pass	Diagonal mean, variance, and α at every layer	FlopScope normal PDF/CDF and array primitives
Pilot	First 1,652 Kerdock rows propagate to layer 29	FlopScope matrix products, comparisons, reductions
Prefix	Remove jointly certified dead coordinates; add analytic omitted means	Compacted dense/cold/exact-zero routes
Layers 30–32	Partition dead/on/kink; sample kinks and fold linear-on paths	Metered block products and subset maps

Table 1: Estimator dataflow. No predict-time numerical stage runs through raw NumPy.

2.4 Three-layer dead/on/kink fold

At layers 30–32, the diagonal α values and pilot preactivations partition each width-256 layer into dead, always-on, and kink coordinates. Only kink paths require row-level ReLU evaluation. Always-on paths are linear and are algebraically composed through weight blocks; dead paths contribute their analytic means. The final vector is the sum of the three disjoint subsets. This is the same structural family disclosed by SOX; the deployed implementation combines it with the Kerdock carrier, analytic dead-compaction, and a fully metered package.

2.5 Exact prediction-preserving compiler reductions

After the estimator mechanism and arithmetic order were frozen, four source rewrites reduced billed data movement without changing a prediction bit. First, each 1,652-row pilot product gathers only input columns whose preceding pilot ReLU is nonzero; the omitted columns were already proved exactly zero on those same billed rows. Second, the already computed vector all($z \leq 0$) is complemented to obtain the pilot-live mask, removing a redundant gather, comparison, and reduction. Third, four equal Winograd quadrants are joined by one `fnp.block` call rather than two inner concatenations followed by a third outer concatenation. NumPy's block order is identical here, but it avoids the two intermediate half-matrices.

A fourth rewrite removes advanced-index copies through identity `arange` arrays at call sites where `compact_prefix` has already materialized the activation and weight blocks in local compact coordinates. The non-identity exact-zero gather, output-alignment gather, padding, masks, routes, Winograd contractions, and reduction order are unchanged. Across 16 frozen network-orientation cases, all predictions were bitwise equal and the mean counted saving was 1.658B FLOPs.

The four official released-stack compiler submissions preserved raw MSE exactly while reducing mean effective compute from 166.897B in the direct e141 replay to 166.231B, 165.813B, 163.347B, and finally 161.368B. The last rewrite was also checked on 16 frozen network-orientation cases: every full prediction was bitwise equal, every operation ledger reconciled, and the mean saving versus its immediate predecessor was 1.658B counted FLOPs. These are ordinary FlopScope operations, not custom arithmetic or an accounting bypass.

3 Accounting, legality, and package identity

All prediction arithmetic uses `flopscope` or `flopscope.numpy`. The estimator does not import raw NumPy, inspect a budget summary to select a path, call external project modules, or exploit an accounting gap. The frozen package was validated with `whestbench 0.14.0` and `flopscope 0.10.0`. The cumulative compiler proof covers 16 frozen target-free network-orientation cases; every full output was bitwise equal to the preceding frozen candidate, every operation ledger reconciled, and all outputs were finite float32 arrays of shape 32×256. An extracted-package subprocess replay completed two additional width-256, depth-32 networks with zero failures.

The maximum result array in the fresh gate was 102,760,448 bytes, below the 100 MiB limit of 104,857,600 bytes. Across that gate, 40 independent operation ledgers reconciled. The submitted package has the following identity:

- estimator SHA-256: `dc2ee05b54f8184e938434469e54e473fddefb48284efb5fd09e592a635f69f7`;
- Kerdock asset SHA-256: `e248b73e8e13cfff73c579a6922f0ccb8b847d4fa0778a3c69f31edc008c915d`;
- three-file archive SHA-256: `ce73d7724615cfa35648fcdf05b3fba5b256124505d6bb91487444c0281e5d2d`;
- package commit: `f924b4cbd635c4a08e87dd8b40f22eb658fc2db1`.

The archive contains only `estimator.py`, `kerdock_quadratic_signs.npz`, and `manifest.json`. It was extracted into a clean directory and revalidated before the one official submission. A rejected file-mode archive that omitted the asset was never authorized for upload.

4 Frozen target-free evidence

4.1 Protocol

The confirmation suite was frozen before observing its truth:

- four newly generated width-256, depth-32 He MLPs with seeds 20260841–20260844;
- two independently seeded coordinate-sign rotations per MLP;
- two independent antithetic Gaussian truth replicates, each with 131,072 input pairs per MLP;
- the un-compacted e136 Kerdock estimator as the fixed baseline; and
- no challenge row, public label, private label, package run, or submission in the promotion gate.

For independent truth estimates T_a, T_b of the same mean and frozen prediction p, the cross loss

$$L_\times(p) = (p - T_a)(p - T_b)$$

removes each truth estimate's independent variance in expectation. Its finite-sample realization remains noisy; the two truth replicates disagreed by MSE 4.96457×10^{-7}, so small raw differences should not be overinterpreted.

4.2 Results

Frozen fresh-network metric	e136 baseline	e141 candidate	Ratio
Cross-MSE	5.1880464480e−7	5.0877052112e−7	0.980659
Mean effective compute	193.9192B	166.3466B	0.857814
Corrected expected adjusted score	1.7373849687e−7	1.4774579534e−7	0.850392

Table 2: High-precision target-free gate. The baseline adjusted value is the e136 official anchor used by the preregistered forecast; B denotes billion FLOP-equivalents.

All four network aggregates and seven of eight rotations improved in raw cross-MSE. An independent lower-precision confirmation with 32,768 antithetic pairs per replicate measured a raw ratio of 1.00520, a compute ratio of 0.85983, and a corrected adjusted ratio of 0.86183. The raw sign changed, but the compute margin preserved the promotion decision.

Mechanism-specific oracle checks measured analytic-fill drift MSE $3.320404175 \times 10^{-10}$, pilot-classification drift MSE $6.323327740 \times 10^{-10}$, and total e141-versus-e136 drift MSE $1.116707823 \times 10^{-9}$. Across all layer/rotation comparisons, the joint mask produced 1,989 false-dead decisions relative to a full-carrier oracle and zero false-alive decisions. The small output drift, rather than the literal false-dead count, is the relevant measured effect of analytic fill plus downstream mixing.

5 Official graded result

The terminal AIcrowd readback for submission #326680 is shown in Table 3. Phase-1 prizes are determined by a separate fresh private rerun, so this public readback is evidence of package behavior and current-stack accounting rather than the prize result.

Relative to the direct released-stack replay, submission 324216, e211 retained exactly the same raw MSE while reducing mean effective compute by 5.529B and improving adjusted score by 3.115%. Relative to its immediate predecessor, e204 submission 326021, e211 saved 1.979B effective compute and improved adjusted score by 0.994%. The frozen forecast was $1.4420078699 \times 10^{-7}$; the official score was 0.199% lower.

Public rank is deliberately omitted because it is mutable and does not decide Phase-1 prizes. The fresh private rerun is the authoritative generalization test, and no public target was used to select any compiler rewrite.

The all-layer MSE is not a claim about all 32 layers. The interface requires a 32×256 output, while this estimator intentionally writes zeros for layers 1–31 and its cubature estimate only in the final row.

Official metric	Value
Raw final-layer MSE	$2.425807970496408 \times 10^{-7}$
Adjusted final-layer score	$1.439144417543210 \times 10^{-7}$
Mean score multiplier	0.5932639496
Mean effective compute	161,367,794,290.68
All-layer diagnostic MSE	0.7536797071
Completed / failed MLPs	100 / 0

Table 3: AIcrowd terminal readback on WhestBench 0.14.0 and FlopScope 0.10.0, 9 August 2026.

6 Negative evidence and falsifiable limits

The campaign used explicit kills to avoid promoting attractive mechanisms on one noisy score:

Experiment	Test	Result
e142	Teacher-forced scalar-energy and conditional-product closures on a fresh MLP	Depth-32 structural bias exceeded the rank-five raw ceiling by more than 2,200 times.
e148	Pair/gate feature family with an impossible truth-informed oracle	Best cross-replicate adjusted ratio was 0.99315, far below the material gap to the lawful anchor.
e158	Two shared local nonlinear residual heads, fit without validation labels	Both regressed exact e141 by more than 3.5 times on untouched validation.
e159	Shared alpha-tail ridge, train-only fit	Untouched-validation adjusted-product ratio 1.92333; killed before reserve truth.
e160	Terminal Gaussian gate-count control	First fresh MLP had gate probability 3.5367e−5, below the frozen 10^{-4} stability floor; no truth was generated.
e161	Six-bin global alpha calibration, post-label procedural screen only	Validation raw ratio 1.01482 and 15/32 network wins; non-promotable and killed.
e205	Split/view and predicate-bitpack screen after e204	Split views saved zero FLOPs; even impossible removal of all packable predicate work was below 0.0182%.
e206	Exact Winograd recursion-stack-depth rewrite	Predictions stayed bitwise equal, but effective compute regressed 2.383% in all three paired repeats.
e207	Carrier-self-piloted dead-tail fusion on 16 frozen networks	Saved 3.762B counted FLOPs on average, but changed one layer-23 dead/live decision and failed prediction parity on 1/16 cases.

Table 4: Selected negative results after e141. Each killed family remained unpackaged and unsubmitted.

These results suggest that the remaining error is not captured by one scalar energy, a low-capacity alpha calibration, or a small local residual head. The final compiler screens also show that another micro-variant of this exact carrier is unlikely to create a safe material score change. These failures do not prove that all learned residuals or joint-distribution closures must fail. A future candidate needs a new target-free sufficient statistic, fresh synthetic held-out evidence, and a material preregistered improvement threshold.

7 Reproduction and disclosure

The historical fold and current deployed chain are pinned separately:

```
PYTHONPATH=. uvx --from whestbench==0.13.0 \
  --with flopscope==0.9.1 --with pytest pytest -q \
  tests/test_e136_kerdock_b32_three_layer_fold.py \
  tests/test_e136_kerdock_b32_three_layer_fold_package.py

PYTHONPATH=. uvx --from whestbench==0.14.0 \
  --with pytest pytest -q --import-mode=importlib \
  tests/test_e094_real_mub_kerdock.py \
  tests/test_e141_dead_compacted_sox_package.py \
  research/e199_e141_pilot_compaction/test_audit.py \
  research/e201_exact_mask_reuse/test_audit.py \
  research/e204_exact_block_join/test_audit.py
```

For e211, the additional frozen source audit was `PYTHONPATH=. uvx -from whestbench==0.14.0 -with pytest pytest -q research/e211_local_identity_gather/test_audit.py`.

The legacy pair passes 10 tests on its frozen stack; the current chain passes 16 tests on the released stack. Together they check the Kerdock rank and moment certificates, signed-FWHT parity, three-layer fold invariants, package import boundaries, frozen prediction hashes, finite float32 output, budget compliance, and exact ledger reconciliation. The self-contained source can be validated with:

```
uv run whest validate \
  --estimator submissions/e141_dead_compacted_sox \
  --class Estimator
```

The official server reported Python 3.10.20, NumPy 2.2.6, `whestbench 0.14.0`, and `flopscope 0.10.0`, matching the validated campaign stack. The durable receipts are `artifacts/e141_dead_compacted_sox_promotion_receipt.json`, `artifacts/e199_e141_pilot_compaction_result.json`, `artifacts/e201_exact_mask_reuse_result.json`, `artifacts/e204_exact_block_join_package_replay.json`, and `artifacts/e204_submission_326021_receipt.json`, `artifacts/e211_local_identity_gather_package_replay.json`, and `artifacts/e211_submission_326680_receipt.json`.

At the time of writing, this paper and the corresponding private campaign source have not been emailed or published by this packet. LLM-assisted development was used for literature search, derivations, implementation discussion, experiment design, drafting, and test review. Multiple models also proposed mechanisms that were rejected by the gates in Table 4. Numerical claims in the paper come from frozen local receipts or the terminal AIcrowd readback and were checked against the package hashes above. The author chose the experiments, owns the submission, and is responsible for every claim and limitation.

8 Limitations and reusable contribution

Complete Kerdock cubature is a deterministic finite angular rule and is generally biased for deep ReLU integration. A coordinate-sign draw supplies a finite symmetry, not Haar randomization and not an unbiasedness theorem. The diagonal recurrence ignores cross-neuron dependence. A pilot that sees no positive preactivation cannot prove global inactivity. Analytic fill reduces the effect of this approximation but does not make the compacted estimator exact.

The public result is one noisy aggregate on known Phase-1 networks. Neither the close target-free forecast nor successful execution establishes fresh-private generalization. The published SOX comparison also shows that the estimator's adjusted advantage comes from compute rather than raw accuracy.

The reusable contribution is a reproducible route from a maximal real MUB construction to a legal structural estimator: an algebraic Kerdock certificate, exact radial conditioning, signed-FWHT evaluation, conservative joint dead classification, analytic omitted-input fill, compacted continuation, late structural folding, exact prediction-preserving compiler reductions, current-stack packaging, and explicit failure boundaries. The evidence supports that mechanism and package. It does not support a broader claim that the competition has been won.