KAKEYASHARE: INCIDENCE-STRUCTURED QUERY SHARING FOR SIMULTANEOUS DIRECTIONAL RISK CERTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating local model risk along several directions can require many black-box evaluations. Overlapping neighborhoods offer reusable observations, yet exploiting them while controlling every directional error requires more than pooling queries. We introduce KakeyaShare, a simultaneous auditor for bounded loss averages over tubes in a three-dimensional input chart. The method groups locations by their exact tube memberships, so one estimated intersection mean contributes to all member directions. A geometry-only criterion selects a budget-feasible core, residual regions preserve the remaining mass, and an exact integer allocation distributes evaluations across the resulting strata. Independent volume measurements and weighted concentration yield simultaneous finite-sample intervals, while a finite-incidence bound relates the ideal sharing gain to mean tube multiplicity. At the main geometry, interpolation gives about 33.5% fewer queries than independent Monte Carlo at matched radius and 95% simultaneous confidence. Independent reference evaluation shows that median dataset RMSE decreases from 0.00619 to 0.00484; controlled ablations and cross-predictor execution expose the effects of allocation, geometry uncertainty, and rejection cost. These results show how finite tube incidence supports query reuse in simultaneous certification of local risk averages.

1 INTRODUCTION

Local model reliability has degrees: two inputs can admit an incorrect prediction nearby, yet differ substantially in how much of their neighborhood incurs loss (Han et al., 2024). Resolving this variation across directions requires many model evaluations. Under a fixed query budget, separate Monte Carlo estimates divide the evidence among directions, leaving wide uncertainty where directional comparisons are needed. This cost matters for black-box predictors whose outputs are accessible but whose internals cannot be used to integrate local risk.

Directional neighborhoods offer an additional source of information. Thin tubes through the same local input chart overlap, so a loss evaluated in their intersection informs several directional averages. Kakeya geometry studies how directional tube families occupy space through their unions and incidence patterns (Wang & Zahl, 2025). For a finite audit, the relevant quantities are the volume of each intersection and the directions sharing it. They determine which risk averages can use common observations and how much those observations contribute. The statistical challenge is to convert this measured overlap into fewer model evaluations while controlling every directional error.

Risk estimation and certification. Average-case robustness quantifies the prevalence of prediction changes around an input (Han et al., 2024). Randomized smoothing provides worst-case guarantees for a smoothed classifier (Cohen et al., 2019), and probabilistic global certification combines sampled locations with a local robustness oracle (Blohm et al., 2025). These results establish useful guarantees for their respective risk definitions. Our target is a vector of bounded loss averages over specified directional tubes. Active testing reduces the labeling cost of model evaluation through adaptive test-point selection (Kossen et al., 2021); here the expensive operation is evaluating the predictor at a chart-generated input, and the opportunity for reuse comes from overlapping integration regions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

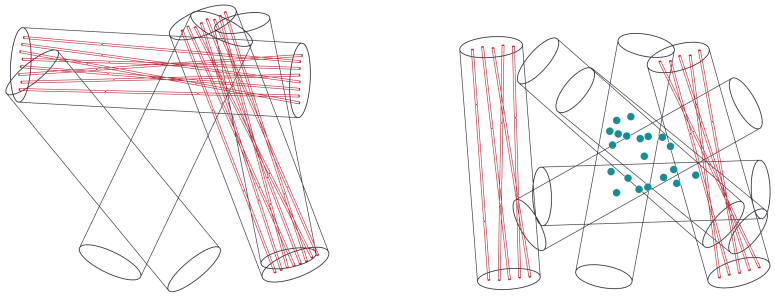


Figure 1: Tube intersections create opportunities for query reuse. The incidence schematic is adapted from Wang & Zahl (2025); teal points illustrate locations that can inform several directional risks.

Sharing samples and allocating effort. Multiple importance sampling combines proposals while preserving the desired expectation (Veach & Guibas, 1995; Owen & Zhou, 2000; Elvira et al., 2019). Optimized mixtures improve their sampling rates using pilot information (Cornuet et al., 2012; He & Owen, 2014). Multiple-frame surveys similarly account for membership in overlapping populations (Horvitz & Thompson, 1952; Lohr, 2021). Stratified sampling allocates observations to heterogeneous regions, with classical and adaptive allocation rules (Neyman, 1934; Carpentier et al., 2015; Song & Kawai, 2023). These approaches provide the statistical foundations for sample reuse. In a tube mixture, a directional observation includes both loss and a random membership indicator, together with its importance weight. Its range therefore depends on how reuse is organized. Exact membership offers a different statistical unit: within an intersection atom, the same bounded mean serves a fixed set of directions. The remaining design gap is to select shared means under a finite budget while controlling geometry and sampling errors in simultaneous intervals. Appendix A develops these connections and distinguishes proposal weighting, multiplicity strata, and exact incidence strata.

The central insight is statistical: within a region of fixed tube membership, one bounded loss mean serves every member direction. KakeyaShare uses these shared means as the units of estimation and allocation. A useful partition must retain overlap without fragmenting the budget across tiny intersections. A geometry-selected core concentrates sharing, while residual regions preserve each directional target. Integer allocation and independent volume measurements connect this representation to a finite-query guarantee. Figure 1 illustrates the geometric motivation.

We establish simultaneous risk intervals and an incidence-dependent bound on the ideal sharing gain. At the main geometry, KakeyaShare reaches the independent Monte Carlo certificate target with approximately 341 queries per direction instead of 512 at common 95% confidence, about 33.5% fewer queries. Independent reference evaluation also shows a decrease in median dataset RMSE from 0.00619 to 0.00484. The experiments connect these outcomes to core construction, allocation, and tube geometry, then examine execution across predictors and sensitivity to loss and input perturbations. Sections 2 to 4 develop the method; Sections 5 and 6 evaluate its statistical and computational behavior.

2 Problem setting and directional audit geometry

An audit fixes a predictor f, an anchor x_0, and a measurable chart $\psi_{x_0} : \Omega \to \mathcal{X}$, where $\Omega = [-1, 1]^3$. The chart maps audit coordinates to valid inputs and fixes the perturbation distributions before queries are spent. Tube geometry describes overlap in these coordinates; the chart determines which tabular or image perturbations the audit measures. The induced loss $h : \Omega \to [0, 1]$ can be $h(z) = 1 - p_y(\psi_{x_0}(z))$ for a labeled anchor. Gradients and model parameters are unnecessary. The bounded range supports finite-sample concentration (Hoeffding, 1963); independence and uniform sampling are specified in Theorem 1.

For unoriented axes $u_j \in \mathbb{S}^2/\{\pm 1\}$, radius r, and half-length L, we define

$$T_j = \{z \in \Omega : |\langle z, u_j\rangle| \le L,\ \|z - \langle z, u_j\rangle u_j\|_2 \le r\}, \qquad \mu_j = \frac{1}{|T_j|}\int_{T_j} h(z)\,dz. \tag{1}$$

The vector $(\mu_1, \ldots, \mu_D)$ records average loss along narrow directional neighborhoods, preserving chart anisotropy. We seek intervals $[L_j, U_j]$ jointly covering these D averages with probability at least $1 - \delta$, using N accepted black-box evaluations. Each queried loss counts once even when several directions use it. The chart is part of the estimand: changing it changes the input distribution being audited. Within a fixed chart, simultaneous coverage supports selecting directions after seeing the estimates, since the same event controls every reported interval.

The main geometry uses $D = 12$ projective Fibonacci axes, $r = 0.15$, and $L = 0.60$. The hemisphere construction identifies opposite axes, since $T(u) = T(-u)$. Each cylinder lies inside Ω because $\sqrt{L^2 + r^2} < 1$, giving common volume $V = 2\pi r^2 L$. Appendix C.1 gives the axis formula and exact tube sampler. The construction makes tube incidence directly measurable in the audit chart.

3 Incidence-structured sharing and budget-aware allocation

Exact incidence decomposition. A point's signature records the directions to which it belongs. For each nonempty $S \subseteq [D]$,

$$\sigma(z) = \{j : z \in T_j\}, \qquad A_S = \{z : \sigma(z) = S\}, \qquad w_{Sj} = \frac{|A_S|}{|T_j|}\mathbf{1}\{j \in S\}. \tag{2}$$

Positive-volume atoms partition the tube union. With $\theta_S = \mathbb{E}[h(Z) \mid Z \sim \mathrm{Unif}(A_S)]$, partitioning each tube integral yields

$$\mu_j = \sum_{S \neq \varnothing} w_{Sj}\theta_S. \tag{3}$$

An observation in A_S estimates the single mean θ_S, which enters every direction in S. The reconstruction coefficients depend on volume, so the loss itself remains bounded by one. The signature supplies the information needed for this reuse. Two points with the same multiplicity can belong to different directions, whereas every point in A_S contributes to exactly the same set of estimands. Estimating an atom mean therefore removes random directional membership from its observations. Its geometric coefficients determine how strongly that shared evidence enters each directional average. Figure 2 shows this sharing operation and its certification stages.

Budget-aware incidence core. Every retained atom creates a mean that needs observations. Fine partitioning can spend a finite budget initializing tiny strata before their shared estimates become precise. KakeyaShare therefore retains a core $\mathcal{C}$ of shared atoms with $|S| \ge 2$. Each direction receives an exact residual region:

$$R_j = T_j \setminus \bigcup_{S \in \mathcal{C}: j \in S} A_S, \quad t_j = 1 - \sum_{S \in \mathcal{C}} w_{Sj}, \quad \mu_j = \sum_{S \in \mathcal{C}} w_{Sj}\theta_S + t_j\theta_{R_j}. \tag{4}$$

Thus every tube retains its full probability mass. The core is selected for reusable mass, so it can omit regions with large loss. Exact residual sampling keeps those regions in the estimand and uncertainty calculation. The guarantee covers loss concentrated in the core or in the tails. Residuals for different directions may overlap; their sampling streams remain separate.

The benefit of retaining A_S is the reduction in the sum of active row norms:

$$b_S = \|w_{S:}\|_1 - \|w_{S:}\|_2, \qquad C(\mathcal{C}) = \sum_{S \in \mathcal{C}} \|w_{S:}\|_2 + \sum_j t_j = D - \sum_{S \in \mathcal{C}} b_S. \tag{5}$$

For congruent tubes, $b_S = (|A_S|/V)(|S| - \sqrt{|S|})$, favoring both mass and multiplicity. The multiplicity factor compares $|S|$ separate directional contributions with the square-root contribution of one shared mean; volume scales this benefit by the atom's probability mass. At a fixed cardinality, the largest b_S values minimize $C(\mathcal{C})$ and the continuous objective below. A design geometry stream

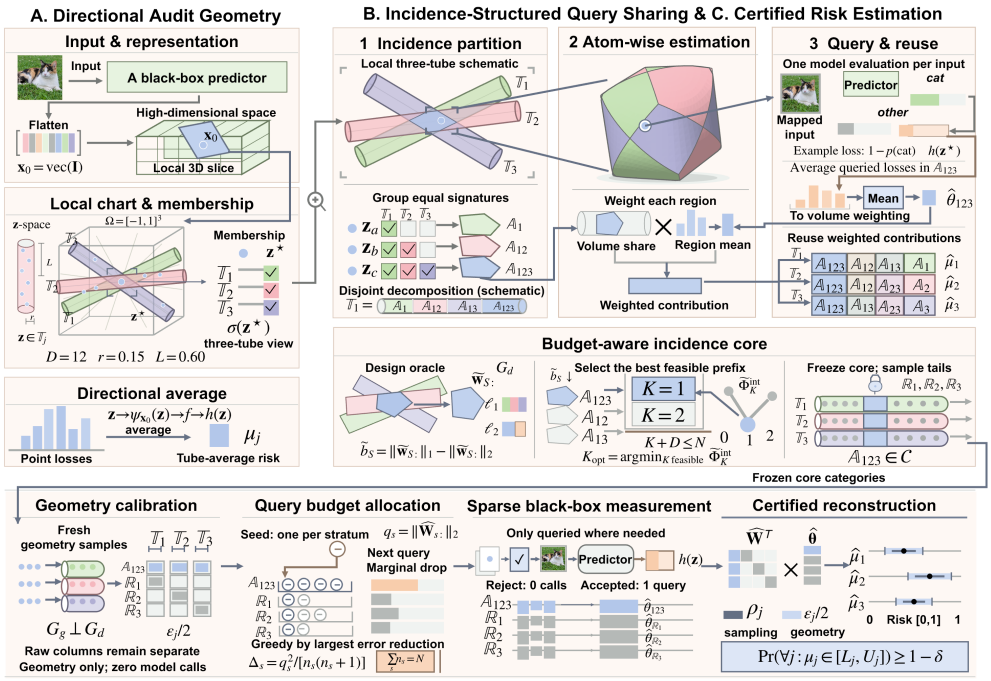


Figure 2: Core compression concentrates sharing without changing the risks being estimated; independent calibration makes the resulting reconstruction compatible with simultaneous coverage.

provides provisional weights $\widetilde{W}$ and scores $\widetilde{b}_S$. Among their sorted prefixes satisfying $K + D \le N$, we choose

$$K_{\text{opt}} = \arg\min_{K} \widetilde{\Phi}^{\text{int}}_{K}, \tag{6}$$

with ties favoring the smaller core. Design weights determine this integer sampling objective; independent calibration supplies the final allocation weights and geometry error (Appendix A).

Exact integer allocation. Let $\mathcal{S}$ contain the core atoms and residuals, and let W be their reconstruction matrix. A core row can serve several directions; a residual row serves one. Each stratum therefore affects a vector of interval widths. Summing their squared sampling terms values a query through every direction it serves. The coefficients depend only on geometry, allowing this allocation objective to be specified before any losses are observed. Allocating n_s observations to stratum s gives

$$\Phi(n) = \sum_{s \in \mathcal{S}} \frac{q_s^2}{n_s}, \qquad q_s = \|W_{s:}\|_2, \qquad \sum_s n_s = N. \tag{7}$$

The continuous solution has the same resource-allocation form as multivariate stratified sampling (Neyman, 1934; Étoré & Jourdain, 2010):

$$n_{s,\text{cont}} = N \frac{q_s}{\sum_r q_r}, \qquad \min_n \Phi(n) = \frac{(\sum_s q_s)^2}{N} = \frac{C(\mathcal{C})^2}{N}. \tag{8}$$

A small continuous allocation can round to zero, leaving a required mean unobserved. The integer problem enforces a query floor while distributing the remaining effort. Each active stratum starts with one query. Each remaining query goes to the largest marginal decrease $\Delta_s(n_s) = q_s^2/[n_s(n_s + 1)]$. These marginal sequences decrease with n_s, making the greedy allocation an exact integer minimizer. A heap costs $O((N - |\mathcal{S}|) \log |\mathcal{S}|)$. The final allocation substitutes the independently certified $\widehat{W}$ for W.

The row norm aggregates a stratum's contribution to all directional squared radii, weighting shared observations through their reconstruction coefficients. Dividing the objective by D gives the mean squared radius and leaves its minimizer unchanged. Allocation is independent of observed losses, so the entire accepted-query budget serves estimation.

4 GEOMETRY-CERTIFIED SIMULTANEOUS RISK ESTIMATION

The shared means reconstruct directional risks through volume fractions. Uncertainty in these coefficients persists even with arbitrarily many model evaluations, so the certificate must control geometry and sampling errors together.

Independent geometry measurements. The design stream draws G_d uniform points per tube to enumerate candidates, rank their benefits, select the core, and choose rejection proposals. Equal-volume atom masses are pooled across member directions for this design step. After the core is fixed, an independent stream draws G_g points per tube. Each point enters one of K_j categories: a retained atom containing j, or R_j. Their raw frequencies form the probability vector $\widehat{W}_{\cdot j}$; certification columns remain separate.

The expected multinomial L_1 error is at most $\sqrt{(K_j-1)/G_g}$. Bounded differences (McDiarmid, 1989) and a union bound yield

$$\varepsilon_j = \min\left\{2, \sqrt{\frac{K_j-1}{G_g}} + \sqrt{\frac{2\log(D/\delta_g)}{G_g}}\right\}, \qquad \Pr\big(\forall j : \|\widehat{W}_{\cdot j} - W_{\cdot j}\|_1 \le \varepsilon_j\big) \ge 1-\delta_g. \quad (9)$$

Both columns sum to one. Their error in a weighted mean of numbers in $[0,1]$ is therefore at most $\varepsilon_j/2$. Geometry depends only on canonical tube coordinates, so the same streams can support different predictors, anchors, and charts.

Separating design and certification fixes the categorical partition before its probabilities are measured. Final counts may depend on these measurements: conditional on them, fresh audit observations retain their stated uniform laws. This conditioning underlies the coverage proof. Geometry and black-box budgets address different errors; increasing model evaluations alone eventually leaves a geometry-determined width floor.

Shared observations and intervals. Uniform tube proposals are accepted on the exact signature for a core atom, or on membership in the residual for a tail. The black box is evaluated only after acceptance. Conditional on the geometry and allocation, independent uniform stratum samples give

$$\widehat{\theta}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} h(Z_{si}), \qquad \widehat{\mu}_j = \sum_s \widehat{W}_{sj}\widehat{\theta}_s, \qquad \rho_j = \sqrt{\frac{\log(2D/\delta_s)}{2}\sum_s \frac{\widehat{W}_{sj}^2}{n_s}}. \quad (10)$$

Weighted Hoeffding concentration (Hoeffding, 1963) controls sampling error in every direction. With the geometry term, the intervals are

$$L_j = \max\{0, \widehat{\mu}_j - \rho_j - \varepsilon_j/2\}, \qquad U_j = \min\{1, \widehat{\mu}_j + \rho_j + \varepsilon_j/2\}. \quad (11)$$

Theorem 1 (Simultaneous directional risk coverage)**.** *The design, certification, and audit randomization streams are independent. Conditional on the geometry and allocation, accepted audit observations are independent and uniform within their stated strata. For measurable $h \in [0,1]$,*

$$\Pr\big(\forall j \in [D] : \mu_j \in [L_j, U_j]\big) \ge 1 - \delta_g - \delta_s. \quad (12)$$

We set $\delta_g = \delta_s = \delta/2$. On the same event, the number $K_\tau = |\{j : \mu_j \ge \tau\}|$ obeys $|\{j : L_j \ge \tau\}| \le K_\tau \le |\{j : U_j \ge \tau\}|$ for every threshold $\tau \in [0,1]$. Appendix B provides the proofs.

Incidence and the ideal sharing gain. For exact W, the squared sampling RMS is $R^2_{2,\mathrm{samp}} = D^{-1}\sum_j \rho_j^2 = \log(2D/\delta_s)\Phi/(2D)$. Its continuous optimum is $\log(2D/\delta_s)C(\mathcal{C})^2/(2DN)$. Independent sampling has $C = D$. For the full partition, with $U = \bigcup_j T_j$ and $m(z) = \sum_j \mathbf{1}\{z \in T_j\}$,

$$C_{\mathrm{full}} = \frac{1}{V}\int_U \sqrt{m(z)}\, dz, \qquad \overline{m} = \frac{DV}{|U|}, \qquad \frac{\Phi_{\mathrm{ind,opt}}}{\Phi_{\mathrm{full,opt}}} \ge \overline{m}. \quad (13)$$

The inequality follows from Cauchy-Schwarz and $\int_U m = DV$. Mean incidence multiplicity therefore lower-bounds the ideal improvement in the squared-radius objective. Compression makes this

shared evidence accessible under finite query and rejection budgets. Two limiting arrangements clarify the scale of the result. Disjoint tubes have $m = 1$ and $C_{\text{full}} = D$, giving the independent-sampling objective. If all D tubes coincide, a single mean serves every direction, $C_{\text{full}} = \sqrt{D}$, and the ideal squared-radius gain is D. Intermediate incidence patterns interpolate through the row norms, while the implemented core trades this opportunity against finite sampling costs.

5 Experimental setup

Tasks and black boxes. The main audit comprises 30 TabArena classification tasks (Erickson et al., 2025), spanning 748 to 150,000 observations and 3 to 1,776 numeric features. Predictors are LightGBM, logistic regression, RealMLP (Holzmüller et al., 2024), and TabM (Gorishniy et al., 2025). We use split 0, one training seed, and up to four test anchors per model, sampled from high- and low-margin groups among correct and incorrect predictions. A metadata-selected reference subset supports independent reference integration, allocation studies, and stress tests. Further execution tests use TabICLv2 (Qu et al., 2026), TabPFN-3 (Grinsztajn et al., 2026), and CIFAR-10 checkpoints ResNet20 and ViT-B/16. Tables 3 and 17 retain unavailable models and failed cells. Appendix C gives datasets, splits, preprocessing, checkpoints, and seeds.

Shared audit protocol. The primary loss is $1 - p_y$, with the geometry of Section 2. For tabular data, ψ_{x_0} uses training-data PCA, perturbing numeric features by at most one training standard deviation per feature while holding categorical features fixed. The vision chart has maximum cube-wide pixel change $8/255$. We use $G_d = 2^{20}$ and $G_g = 2^{22}$ per tube, $\delta_g = \delta_s = 0.025$, and four audit seeds for tabular comparisons. The main budgets are $N/D \in \{128, 256, 512, 1024\}$, with independent-MC target budget $N_0 = 6144$. The reference subset extends the curve from 64 to 4096 queries per direction. Here N/D is a budget equivalent; shared samplers allocate the total N across their own strata. Pilot evaluations count toward N, while geometry proposals, preprocessing, reference integration, and elapsed time are accounted for separately.

Comparators and mechanism. Five comparators estimate the same directional means: independent Monte Carlo; union sampling with volume correction, called Union-HT (Horvitz & Thompson, 1952); equal-mixture DM-MIS (Veach & Guibas, 1995; Elvira et al., 2019); pilot-optimized OptDM-MIS (He & Owen, 2014); and multiplicity-stratified Neyman allocation (Neyman, 1934; Song & Kawai, 2023). Union-HT, OptDM-MIS, and Multiplicity-Neyman are task-specific instantiations of these principles. Appendix C.2 gives their estimators and interval assumptions, including the nominal status of the recorded adaptive MIS radius. Core and allocation ablations isolate KakeyaShare's design choices. Scrambled Sobol and Latin hypercube sampling provide point-estimation comparisons. Geometry sweeps vary D, r, and L; loss, calibration, noise, missingness, shift, chart, and split tests examine sensitivity.

Endpoints and reference uncertainty. For the unclipped half-width $w_j = \rho_j + g_j$, we report

$$R_2 = \left(\frac{1}{D} \sum_j w_j^2 \right)^{1/2}, \qquad R_\infty = \max_j w_j. \tag{14}$$

Certified query saving is $1 - N_m/N_0$ at the independent MC target $R_2(N_0)$. Tables and Figure 3 use equal $\delta_s = 0.025$; headline savings give IMC its full 0.05 error allowance for common 95% confidence (Appendix C.7). IMC has no geometry error, so its 512-query target falls from 0.08189 to 0.07765 when using the full allowance. KakeyaShare retains both error budgets; the comparison therefore accounts for the cost of calibrating its reconstruction weights. Crossings interpolate adjacent log budgets and log radii. Main-curve noncrossings are right-censored at $1024D$ and assigned capped saving -1. RMSE and MAE use independent scrambled-Sobol point references. Reference containment requires inclusion of the independent reference interval; disjoint intervals are falsified, and other cases ambiguous. Cell containment requires all directions. Table 1 reports median dataset errors and mean dataset directional containment. Other pooling rules and dataset-cluster uncertainty estimates are specified in Appendix C.6.

Table 1: Directional risk estimation and matched-radius queries at equal $\delta_s = 0.025$.

Method	$R_2 \downarrow$	RMSE↓	MAE↓	Ref. C/A/F (%)	Cert. $N/D \downarrow$
Independent MC	.0819	.00619	.00371	100/0/0	512
Union-HT	.1733	.01645	.01127	100/0/0	> 1024
DM-MIS	.2837	.01900	.01270	100/0/0	> 1024
OptDM-MIS†	.3040	.01665	.01086	100/0/0	n.a.
Multiplicity-Neyman	.1894	.01447	.00971	100/0/0	> 1024
KakeyaShare	**.0639**	**.00484**	**.00295**	99.92/.08/0	**304.8**

Errors and Ref. C/A/F (contained/ambiguous/disjoint) use the reference subset at $N/D = 512$. Certified target crossings use the main matrix. † Nominal radius, excluded from certified crossings (Appendix C.2).

6 Results and Discussion

Overlap becomes useful through structured estimation. At common 95% confidence, interpolation reduces queries per direction from IMC's 512 to KakeyaShare's 340.5 at matched radius, saving 33.5%. Under equal sampling-error budgets, R_2 falls by 22.02% at 512 queries (Table 1), and the interpolated saving is 40.5% (Figure 3). Importance and union estimators also reuse observations; wider intervals reflect larger directional observation ranges. Optimized mixtures improve RMSE over equal-mixture MIS while leaving the recorded radius wider. Estimator and uncertainty criterion must therefore be considered together.

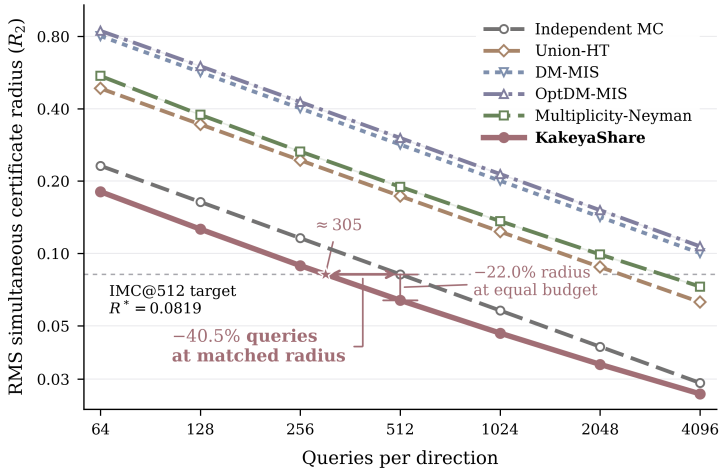


Figure 3: Query budget and interval width at equal $\delta_s = 0.025$ on the reference subset. OptDM-MIS shows its recorded nominal radius; Appendix C.2 explains the adaptive-range issue.

The main geometry fixes the KakeyaShare allocation, radius, and saving across successful audits. The 30-task matrix tests execution; the reference subset tests estimation quality. Within that subset, median dataset RMSE decreases by 21.72%, and MAE by 20.67%. Directional reference containment is 99.92%, with 0.08% ambiguous cases and no disjoint intervals. Seven synthetic losses test validity directly, yielding one familywise miss in 14,000 replicates (Table 4). Appendix D gives full curves and reference precision.

The core and allocation solve different bottlenecks. The full partition has 720 observed shared atoms and 12 residuals. Rare atoms exhaust the rejection cap at the main budget; compression makes sharing accessible. At 256 queries per direction, the budget-selected 282-atom core improves R_2 over the fixed 360-atom core (Table 2). At 1024, the fixed core has the smaller total radius: prefix selection optimizes the sampling proxy, whereas additional geometry categories increase calibration uncertainty. Within the same core, uniform allocation performs poorly; exact integer allocation improves on mass-proportional and continuous allocation with largest-remainder rounding.

Table 2: Core and allocation ablations isolate their distinct effects.

A. Core construction

Core (K)	R_2@256	R_2@512	R_2@1024
None (0)	.11581	.08189	.05791
Full (720)	Inaccessible		
Fixed (360)	.08979	.06386	**.04622**
Selected	**.08903**	**.06386**	.04662
$K = 282/360/450$ at $256/512/1024$			

B. Allocation at 512

Rule	R_2	$\min n_s$	Acc. (%)
Uniform	.24057	16	.038
Pilot† (LGBM)	.09047	8	.077
Pilot† (TabM)	.08572	8	.077
Mass-proportional	.06950	1	.285
Continuous L_2 + LR	.06433	1	.399
Exact integer L_2	**.06386**	1	.489

Radii are identical for LightGBM and TabM except pilot Neyman† (nominal; Appendix C.3). LR: largest remainder. Acceptance counts all geometry proposals.

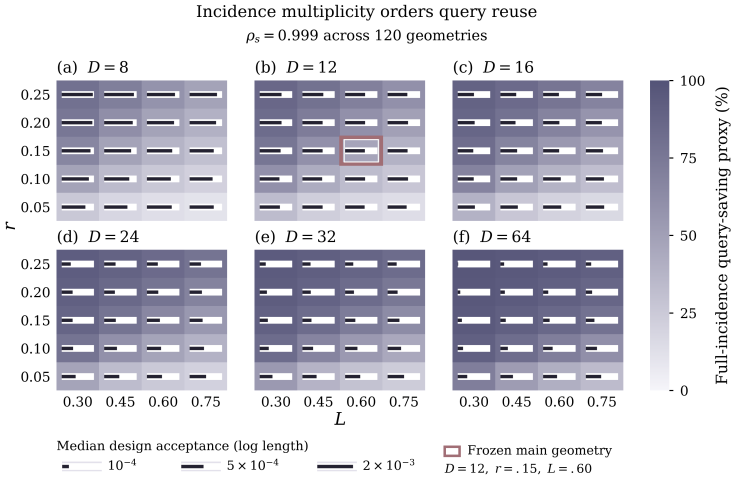


Figure 4: Finite incidence controls the ideal sharing opportunity. Color shows the full-incidence continuous saving proxy $100[1 - (C_{\text{full}}/D)^2]$; bars show design acceptance.

Geometry and computational cost. Across 120 geometries, mean multiplicity orders the full-incidence saving proxy (Spearman correlation 0.999; Figure 4). Larger radii and denser direction families increase overlap; longer tubes add direction-specific mass. Acceptance exposes sampling cost: the main core accepts 0.489% of proposals, with rarer intersections in the sweep. All configurations through $D = 64$ complete using observed signatures. Rejection and core selection add computation; Appendix D.6 reports feasibility and derives when query savings offset this cost.

Table 3: Execution across predictors and modalities.

Predictor	Successful cells	Time (%)	Predictor	Successful cells	Time (%)
LightGBM	480/480	+7.91	TabPFN-3	96/96	+0.22
Logistic	480/480	+18.48	TabICLv2	74/88	+0.67
RealMLP	480/480	+13.92	ResNet20	64/64	+3.54
TabM	416/480	+3.56	ViT-B/16	64/64	+2.31

At 512 queries/direction; time change relative to independent MC. Left: main task matrix; right: frontier subsets and CIFAR-10 (Appendix C).

Execution across predictors. Table 3 spans trees, linear and neural tabular predictors, two foundation models, and vision checkpoints. Elapsed-time overhead is 0.22–18.48% at equal query count; matched-certificate savings use different counts. TabM completes on 26 of 30 tasks; TabICLv2 completes 74 of 88 cells. Two of four frontier families execute; Table 17 details numerical failures and unavailable dependencies.

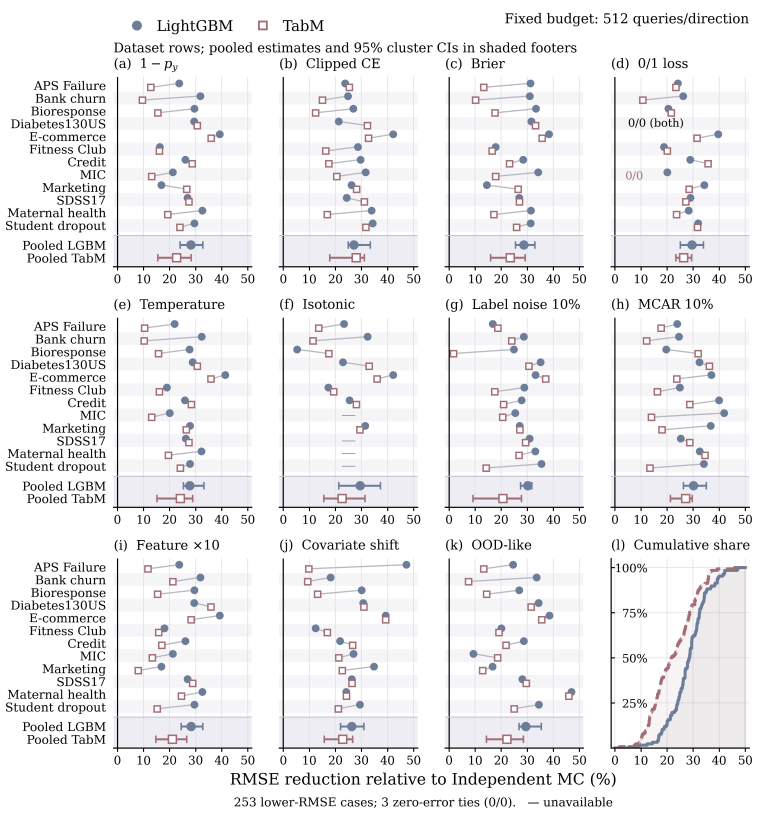


Figure 5: Error reductions under changes in loss, calibration, and inputs.

Table 4: Reference validity and sensitivity to audit choices.

A. Perturbations: MAE reduction and minimum reference containment (%)				
Family	Groups / KS cells	MAE median [range]	Dir. / cell min.	Fals. max.
Loss	8 / 3072	30.21 [26.21,31.73]	99.57 / 94.79	0
Calibration	4 / 1280	28.96 [26.42,31.55]	99.91 / 98.96	0
Data / OOD	10 / 3712	28.70 [23.98,33.04]	99.78 / 97.32	0
Overall	22 / 8064	29.41 [23.98,33.04]	99.57 / 94.79	0
B. Synthetic validity and geometry-conditioned controls				
Synthetic losses	14,000 replicates; 99.993% familywise coverage; one miss			
Random / DCT charts	2304/2304 and 128/128 cells; median $R_2/R_2^{\mathrm{PCA}} = 1.000$			
Splits 1 and 2	2560/2560 cells; certificate gain difference 0.00 percentage points			

Sensitivity and scope. Figure 5 records 253 RMSE improvements and three zero-error ties; Table 4 adds reference-containment and chart/split checks. Geometry fixes the radii; changed losses test error reduction. The guarantee covers averages in the specified three-dimensional chart. Sobol has lower point error in the tested comparisons (Appendix D.3). Geometry and rejection incur overhead; finite geometry budgets set a width floor. Empirical-Bernstein intervals are loose or inapplicable in sparse strata (Appendix D.5).

7 Conclusion

KakeyaShare turns exact tube memberships into shared bounded means and simultaneous directional certificates. At the main geometry, interpolation gives about 33.5% fewer queries at matched radius and 95% simultaneous confidence. Incidence thus guides query allocation, with geometry uncertainty and rejection cost determining the attainable benefit.

AI USE STATEMENT

In this work, we used generative AI tools to organize and revise the presentation of author-provided research materials, polish language and correct grammar, assist with the exposition and consistency checking of mathematical arguments, debug code, and arrange experimental queues. The research ideas, methods, experimental designs, and results were developed by the authors. Generative AI was not used to generate the experimental code. We reviewed all AI-assisted work, checked experimental runs and logs to ensure consistency with the original designs, and verified the mathematical arguments. We take responsibility for the final content of this work, including all text, claims, and artifacts prepared with generative AI assistance.

REPRODUCIBILITY STATEMENT

Reproducibility is supported by the theoretical derivations, experimental specification, and anonymous supplementary package. Appendix B gives the assumptions and complete proofs for the incidence decomposition, core compression, allocation, and simultaneous certificate; Algorithms 1 and 2 specify the auditor and integer allocator. Section 5 and Appendix C document dataset selection, splits, preprocessing, perturbation charts, model fitting, random seeds, query accounting, comparator definitions, reference integration, and aggregation. Appendix D reports detailed results, execution support, and computational checks. The supplementary `code/` directory supplies the installable implementation, frozen configurations, stage runners, and unit tests for the core numerical and sampling invariants. Its `REPRODUCIBILITY.md` provides installation, execution, restart, and aggregation instructions, while `environment/` records the software dependencies and hardware configuration and distinguishes the full execution environment from lightweight CPU checks. The `data/README.md` explains acquisition of public datasets and official checkpoints. The `results/tables/` directory contains 29 canonical reference tables, and `artifacts/` provides indices describing the organization of experimental outputs. The package-wide SHA-256 manifest and accompanying verification and comparison tools support integrity checks and comparison of regenerated tables with these references. The package's `KNOWN_LIMITATIONS.md` and the appendix identify unavailable model families, terminal outcomes, and the scope of the reproducibility evidence.

REFERENCES

Peter Blohm, Patrick Indri, Thomas Gärtner, and Sagar Malhotra. Probably Approximately Global Robustness Certification. In *International Conference on Machine Learning*, pp. 4570–4587, 2025.

Alexandra Carpentier, Rémi Munos, and András Antos. Adaptive Strategy for Stratified Monte Carlo Sampling. *Journal of Machine Learning Research*, 16:2231–2271, 2015.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, 2019.

Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P. Robert. Adaptive Multiple Importance Sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. Generalized Multiple Importance Sampling. *Statistical Science*, 34(1):129–155, 2019.

Cornelius Emde, Francesco Pinto, Thomas Lukasiewicz, Philip Torr, and Adel Bibi. Towards Certification of Uncertainty Calibration under Adversarial Attacks. In *International Conference on Learning Representations*, pp. 79428–79468, 2025.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. TabArena: A Living Benchmark for Machine Learning on Tabular Data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Pierre Étoré and Benjamin Jourdain. Adaptive Optimal Allocation in Stratified Sampling Methods. *Methodology and Computing in Applied Probability*, 12(3):335–360, 2010.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, pp. 77899–77935, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical Report. *arXiv preprint arXiv:2605.13986*, 2026.

Larry Guth, Hong Wang, and Joshua Zahl. A streamlined proof of the Kakeya set conjecture in $\mathbb{R}^3$. *arXiv preprint arXiv:2601.14411*, 2026.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Characterizing Data Point Vulnerability as Average-Case Robustness. In *Uncertainty in Artificial Intelligence*, pp. 1513–1540, 2024.

Hera Y. He and Art B. Owen. Optimal mixture weights in multiple importance sampling. *arXiv preprint arXiv:1411.3954*, 2014.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26577–26658, 2024.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Testing: Sample-Efficient Model Evaluation. In *International Conference on Machine Learning*, pp. 5753–5763, 2021.

Sharon L. Lohr. Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47 (2):229–263, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. In *Conference on Learning Theory*, 2009.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, pp. 148–188. Cambridge University Press, 1989.

Jerzy Neyman. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97 (4):558–606, 1934.

Art Owen and Yi Zhou. Safe and Effective Importance Sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Art B. Owen. Scrambled net variance for integrals of smooth functions. *The Annals of Statistics*, 25 (4):1541–1562, 1997.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *International Conference on Machine Learning*, pp. 50817–50847, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Chenxiao Song and Reiichiro Kawai. Dynamic Finite-Budget Allocation of Stratified Sampling with Adaptive Variance Reduction by Strata. *SIAM Journal on Scientific Computing*, 45(2):A898–A932, 2023.

Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95*, pp. 419–428, 1995.

Hong Wang and Joshua Zahl. Volume estimates for unions of convex sets, and the Kakeya set conjecture in three dimensions. *arXiv preprint arXiv:2502.17655*, 2025.

Han-Jia Ye, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later. In *International Conference on Learning Representations*, pp. 94917–94936, 2025.

A Extended related work and geometric foundations

The chart specifies the audit distribution. A local audit needs a distribution over perturbations before its average risk is defined. Average-case robustness makes this dependence explicit by measuring the prevalence of vulnerable perturbations (Han et al., 2024). In our construction, the chart ψ_{x_0} is fixed before the audit, and uniform sampling in T_j induces a distribution P_j on inputs. For every measurable input set B,

$$P_j(B) = \frac{|T_j \cap \psi_{x_0}^{-1}(B)|}{|T_j|}, \qquad \mu_j = \mathbb{E}_{X \sim P_j}[\ell(f(X), y)],$$

where $\ell \in [0, 1]$ and $h(z) = \ell(f(\psi_{x_0}(z)), y)$. This pushforward definition also covers noninjective charts, including maps that clip perturbed inputs. The induced input distribution need not be uniform on the image of the tube. All volume weights and incidence calculations remain in the canonical coordinates.

This separation lets the same tube design support different application-specific maps. The tabular PCA map, random orthogonal control, and vision maps determine which perturbations are represented; their exact definitions and scaling appear in Appendix C. Changing the chart changes the directional estimands, even when geometry, allocation, and interval widths stay fixed. Each vector resolves average loss within its chosen three-dimensional chart. A small mean still permits a small subset of large losses, and the resulting certificate concerns those averages. Its probability statement applies to each fixed audit under Theorem 1; simultaneous comparison of directions is already included in that statement.

This distinction separates the task from randomized smoothing (Cohen et al., 2019), which certifies a worst-case radius for a smoothed classifier, and from probably approximately global robustness certification (Blohm et al., 2025), which combines local robustness information over sampled inputs. It also separates the query accounting from active testing (Kossen et al., 2021). Active testing chooses examples whose labels are informative for model evaluation. In our setting, the anchor label is fixed, generated inputs are passed through the predictor, and each returned loss may inform several integration regions. The shared theme is economical risk estimation; the observation unit and the estimand determine which economy is available.

Which property is being certified? Recent certification work broadens the target beyond prediction invariance. Blohm et al. (2025) aggregate local robustness information into a probabilistic global guarantee. Emde et al. (2025) bound calibration under adversarial perturbations, deriving analytic bounds for a Brier score and optimization-based bounds for expected calibration error. These targets clarify the role of the bounded loss in our audit. The normalized multiclass Brier loss used here defines a directional average under a fixed perturbation distribution; its interval does not certify worst-case calibration or expected calibration error. Likewise, evaluating a temperature-scaled or isotonic predictor changes the audited loss function without changing the coverage argument. The connection is the need to specify the reliability property before choosing its certificate. KakeyaShare addresses simultaneous integration over overlapping regions, with sample reuse and uncertain volume fractions as the statistical design problem.

Overlapping sampling frames and proposal mixtures. Horvitz-Thompson estimation corrects unequal inclusion probabilities in finite-population sampling (Horvitz & Thompson, 1952). Multiple-frame surveys address the further issue that a unit can belong to overlapping populations (Lohr, 2021). Tube membership has an analogous accounting role: a point in an intersection represents a different fraction of each directional population than a point belonging to a single tube. Our Union-HT comparator implements a continuous-space volume correction. This name identifies the weighting principle; its estimator and concentration bound are derived directly in Appendix C.2.

Multiple importance sampling provides a general way to combine proposals (Veach & Guibas, 1995; Owen & Zhou, 2000; Elvira et al., 2019). In the tube family, the equal proposal mixture has density proportional to multiplicity. Balance weighting therefore gives an unbiased estimator with observation $Dh(z)\mathbf{1}\{j \in \sigma(z)\}/m(z)$ for direction j. This construction already shares every query. Its range can nevertheless reach D, which is costly for a range-based simultaneous certificate. Incidence stratification changes the statistical unit: within an exact membership atom, the directional membership indicators are constant, so every member direction uses the same bounded scalar mean.

Atom volumes supply the reconstruction coefficients outside that mean. Independent calibration estimates these coefficients and carries their uncertainty into the final interval. Thus the comparison concerns how shared observations enter a simultaneous concentration bound, as well as how many estimands use them.

Mixture optimization can reduce importance-sampling variance (He & Owen, 2014), and adaptive multiple importance sampling recycles information across changing proposals (Cornuet et al., 2012). A variance-improving proposal does not by itself provide a tighter finite-sample interval. When a pilot controls subsequent ranges, both the estimator and the concentration argument must account for that dependence. Appendix C.2 distinguishes the recorded adaptive MIS radius from a valid two-stage alternative. This matters specifically for certificate comparisons; the measured point errors remain empirical outcomes of the executed estimator.

Stratification and finite budgets. Neyman allocation distributes samples according to stratum mass and variability (Neyman, 1934). Adaptive stratified Monte Carlo estimates these quantities online or from pilots (Étoré & Jourdain, 2010; Carpentier et al., 2015), and finite-budget designs can optimize both allocation and within-stratum variance reduction (Song & Kawai, 2023). Our objective is the sum of squared directional Hoeffding radii. Its stratum importance is therefore the Euclidean norm of a reconstruction row. The continuous solution shares the classical square-root allocation structure, while its coefficients arise from simultaneous directional reconstruction.

Three levels of the allocation objective. The core determines which intersection means are estimated jointly and which mass remains in direction-specific residuals; allocation distributes accepted evaluations across those units. Three objectives distinguish the theoretical construction from the implemented certificate.

First, for exact geometric weights and a fixed core size K, Proposition 2 optimizes over all cores of that size. Selecting the largest b_S minimizes $C(\mathcal{C})$ and hence the continuous sampling optimum $C(\mathcal{C})^2/N$. The continuous counts have no one-query floor. This fixed-cardinality result supplies a ranking criterion before integer initialization costs are imposed.

Second, the design stream estimates $\widetilde{W}$ and ranks the observed candidate atoms by $\widetilde{b}_S$. Equation 6 searches feasible prefixes of this ranking. For each prefix, Algorithm 2 exactly minimizes the integer sampling objective with these design weights. The outer search therefore chooses the best evaluated prefix, with ties favoring smaller cores. Its search space and weights differ from the all-core, exact-weight continuous problem.

Third, after the chosen core is frozen, an independent stream estimates $\widehat{W}$ and the geometry error $g_j = \varepsilon_j/2$. The integer allocation is recomputed using $\widehat{W}$, as specified in Algorithm 1. Conditional on these quantities, write $a = \log(2D/\delta_s)/2$ and $\Phi(n; \widehat{W}) = \sum_s \|\widehat{W}_{s:}\|_2^2/n_s$. The reported unclipped radius satisfies

$$\begin{aligned} R_2^2(n) &= \frac{1}{D}\sum_j (\rho_j(n) + g_j)^2 \\ &= \frac{a}{D}\Phi(n; \widehat{W}) + \frac{2}{D}\sum_j \rho_j(n) g_j + \frac{1}{D}\sum_j g_j^2. \end{aligned} \tag{A.1}$$

The implemented allocator minimizes the first term. The cross term also depends on allocation, while changing the core changes the categories entering g_j. Full-radius minimization is therefore a distinct design problem. The fixed 360-atom core's smaller total radius at the largest ablation budget (Table 2) is consistent with this decomposition. It preserves both the exact integer guarantee for a supplied weight matrix and the observed advantage of a smaller core after calibration uncertainty is included.

Concentration and alternative integration methods. Hoeffding concentration (Hoeffding, 1963) applies without loss smoothness or variance estimation. Bounded differences (McDiarmid, 1989) separately controls the multinomial geometry measurements, and the probability-simplex constraint halves their worst-case effect on a bounded mean. Empirical Bernstein bounds (Maurer & Pontil, 2009) can exploit estimated variability, but their range correction is substantial when strata

have few samples. Confidence sequences (Howard et al., 2021) address validity under sequential monitoring through a different concentration construction. Our experiments evaluate a fixed-budget Hoeffding procedure; the available Bernstein calculations and absent vector confidence-sequence implementation do not establish an adaptive stopping guarantee.

Randomized low-discrepancy integration offers another route to accurate means. Scrambled nets can have favorable variance for smooth integrands (Owen, 1997). Their dependence structure differs from independent stratum observations, so the current Hoeffding proof cannot be transferred simply by replacing the sampler. The Sobol comparison measures point accuracy under the same accepted-query budget. Its stronger observed accuracy usefully separates our certificate objective from the broader problem of minimizing integration error.

Finite tube incidence and statistical complexity. Kakeya geometry studies the space occupied by directional families; recent three-dimensional results analyze tube unions and convex sets (Wang & Zahl, 2025; Guth et al., 2026). KakeyaShare uses the corresponding finite objects: a chosen tube family, its union U, the multiplicity $m(z)$, and the exact membership atoms A_S. Each atom defines one bounded mean. Its reconstruction row norm is $(|A_S|/V)\sqrt{|S|}$ for congruent tubes, linking the incidence pattern to the allocation objective. Summing these norms over the full partition gives

$$C_{\text{full}} = \frac{1}{V}\int_U \sqrt{m(z)}\, dz.$$

The finite-volume identity $\int_U m(z)\, dz = DV$ and Cauchy-Schwarz imply $C_{\text{full}}^2 \le D|U|/V$. Consequently, mean multiplicity $DV/|U|$ lower-bounds the ideal gain in the continuous squared sampling-radius objective. Appendix B.6 proves this elementary finite-incidence relation directly. It quantifies the opportunity before integer allocation, core compression, and geometry uncertainty enter the implemented certificate. The geometric inspiration concerns how tubes overlap; the statistical construction converts that overlap into shared bounded observations with explicit reconstruction weights.

Why heterogeneous tabular predictors are useful test cases. TabArena provides a common task collection for comparing tabular learning systems (Erickson et al., 2025). RealMLP studies strong default configurations (Holzmüller et al., 2024); TabM develops efficient neural ensembling (Gorishniy et al., 2025); and ModernNCA revisits neighborhood-based prediction with learned representations (Ye et al., 2025). Foundation models introduce a further contrast through in-context prediction: TabPFN demonstrates this approach for small tabular problems (Hollmann et al., 2025), TabICL targets larger tables (Qu et al., 2025), and their newer families extend the scale and implementation choices (Qu et al., 2026; Grinsztajn et al., 2026). These references motivate model diversity, without implying that every listed implementation completed our audit. Model-specific feasibility is reported alongside successful execution results.

B Derivations and proofs for KakeyaShare

This appendix derives the identities and guarantees used in Sections 3 and 4. Throughout, $h \in [0,1]$, $[D] = \{1, \dots, D\}$, and $|\cdot|$ denotes Lebesgue volume for measurable regions.

B.1 Exact incidence identity

Proposition 1 (Exact incidence identity)**.** *For every measurable $h : \Omega \to [0,1]$ and every direction j,*

$$\mu_j = \sum_{S \neq \varnothing} w_{Sj}\theta_S.$$

If $|S| = m$, an accepted query in A_S can be reused by all m directional estimands indexed by S.

Proof. The signature map $\sigma(z)$ assigns every point in $U = \cup_{j=1}^{D} T_j$ to exactly one nonempty subset of $[D]$. Hence the sets $\{A_S : S \neq \varnothing\}$ are pairwise disjoint and partition U. Restricting this partition

to tube T_j gives

$$T_j = \dot{\bigcup_{S:j\in S}} A_S.$$

Therefore

$$\begin{aligned}\mu_j &= \frac{1}{|T_j|}\int_{T_j} h(z)\,dz \\ &= \frac{1}{|T_j|}\sum_{S:j\in S}\int_{A_S} h(z)\,dz \\ &= \sum_{S:j\in S}\frac{|A_S|}{|T_j|}\left(\frac{1}{|A_S|}\int_{A_S} h(z)\,dz\right) \\ &= \sum_{S\neq\varnothing} w_{Sj}\theta_S.\end{aligned}$$

For $z \in A_S$, the same loss value $h(z)$ is an observation from the distribution defining θ_S, and θ_S enters every μ_j with $j \in S$. Thus one accepted evaluation is reusable by all member directions. Every member direction uses the same atom mean with its geometric coefficient. □

B.2 Exact core compression and the reuse-benefit score

For any retained core $\mathcal{C}$, the sets A_S with $S \in \mathcal{C}$ and $j \in S$ are disjoint subsets of T_j. Their complement within T_j is R_j. Hence

$$|R_j| = |T_j| - \sum_{S\in\mathcal{C}:j\in S} |A_S|,$$

and therefore

$$\frac{|R_j|}{|T_j|} = 1 - \sum_{S\in\mathcal{C}} w_{Sj} = t_j.$$

Partitioning the integral over T_j into the retained atoms and R_j immediately yields Eq. 4. Thus residual tails preserve all mass omitted from the shared core.

To derive Eq. 5, note that a retained atom contributes the row norm $\|w_{S:}\|_2$ to the active-stratum complexity. If the same atom is not shared, its directional masses remain in one-hot residual rows and contribute

$$\sum_{j=1}^{D} w_{Sj} = \|w_{S:}\|_1,$$

because $w_{Sj} \geq 0$. Moreover,

$$\sum_{j=1}^{D} t_j = D - \sum_{j=1}^{D}\sum_{S\in\mathcal{C}} w_{Sj} = D - \sum_{S\in\mathcal{C}} \|w_{S:}\|_1.$$

Consequently,

$$\begin{aligned}C(\mathcal{C}) &= \sum_{S\in\mathcal{C}} \|w_{S:}\|_2 + D - \sum_{S\in\mathcal{C}} \|w_{S:}\|_1 \\ &= D - \sum_{S\in\mathcal{C}} (\|w_{S:}\|_1 - \|w_{S:}\|_2) = D - \sum_{S\in\mathcal{C}} b_S.\end{aligned}$$

This also shows $b_S \geq 0$ because $\|v\|_2 \leq \|v\|_1$ for every nonnegative vector v.

For congruent tubes, let $a_S = |A_S|/V$. The vector $w_{S:}$ has exactly $m_S = |S|$ nonzero entries, each equal to a_S. Hence

$$\|w_{S:}\|_1 = m_S a_S, \qquad \|w_{S:}\|_2 = \sqrt{m_S}\, a_S,$$

which gives the equal-volume form of the benefit score in Section 3.

Proposition 2 (Optimal core at fixed cardinality)**.** *Assume exact geometric weights and fix the number K of retained shared atoms. The K atoms with largest b_S minimize $C(\mathcal{C})$ and therefore minimize the best continuous L2 allocation objective.*

Proof. For every core with $|\mathcal{C}| = K$, Eq. 5 gives $C(\mathcal{C}) = D - \sum_{S\in\mathcal{C}} b_S$. Since D is fixed, minimizing $C(\mathcal{C})$ is equivalent to maximizing the sum of K selected scores. This is attained by choosing the K largest b_S. Section B.3 shows that the minimum continuous objective equals $C(\mathcal{C})^2/N$, which is monotone in $C(\mathcal{C}) \geq 0$. □

B.3 L2 ALLOCATION: CONTINUOUS AND INTEGER OPTIMA

Zero-mass rows can be omitted from the exact-weight objective. For the remaining rows let $q_s = \|W_{s:}\|_2 > 0$. The continuous allocation problem is

$$\min_{n_s>0} \sum_s \frac{q_s^2}{n_s} \quad \text{subject to} \quad \sum_s n_s = N.$$

The Lagrangian

$$\mathcal{L}(n,\lambda) = \sum_s \frac{q_s^2}{n_s} + \lambda\left(\sum_s n_s - N\right)$$

has stationarity condition

$$-\frac{q_s^2}{n_s^2} + \lambda = 0,$$

so $n_s = q_s/\sqrt{\lambda}$. Enforcing $\sum_s n_s = N$ gives

$$\sqrt{\lambda} = \frac{\sum_r q_r}{N}, \qquad n_{s,\mathrm{cont}} = N\frac{q_s}{\sum_r q_r}.$$

Substitution yields

$$\min_n \Phi(n) = \frac{(\sum_s q_s)^2}{N}.$$

For the compressed KakeyaShare matrix, $\sum_s q_s = C(\mathcal{C})$, proving the relation between incidence complexity and the best continuous sampling objective.

For integer allocation, initialize $n_s = 1$. If one additional query is assigned to stratum s, its objective contribution decreases by

$$\frac{q_s^2}{n_s} - \frac{q_s^2}{n_s+1} = \frac{q_s^2}{n_s(n_s+1)} = \Delta_s(n_s).$$

The marginal sequence $\Delta_s(1), \Delta_s(2), \ldots$ is nonincreasing for every s. Therefore the integer problem is a separable resource-allocation problem with diminishing returns: among all available marginal decreases, an optimum takes the largest $N - |\mathcal{S}|$ values. Sequentially assigning each query to the currently largest $\Delta_s(n_s)$ constructs exactly this set, proving optimality of the greedy rule. A max-heap implements the procedure in $O((N - |\mathcal{S}|)\log|\mathcal{S}|)$ time.

B.4 GEOMETRY-WEIGHT CERTIFICATION

Fix a direction j. After the core is fixed, the certification oracle maps each of the G_g independent draws from T_j into one of K_j categories. Let $W_{\cdot j}$ be the true categorical probability vector and $\widehat{W}_{\cdot j}$ its empirical distribution. For $p_k = W_{kj}$, Cauchy-Schwarz gives $\sum_k \sqrt{p_k(1-p_k)} \leq \sqrt{K_j - 1}$. The variance of an empirical frequency is $p_k(1-p_k)/G_g$, so Jensen's inequality gives

$$\mathbb{E}\|\widehat{W}_{\cdot j} - W_{\cdot j}\|_1 \leq \sqrt{\frac{K_j - 1}{G_g}}. \tag{B.1}$$

Changing one categorical observation changes two empirical coordinates by $1/G_g$ and hence changes the empirical L_1 distance by at most $2/G_g$. McDiarmid's bounded-difference inequality (McDiarmid, 1989) therefore gives, for any $t > 0$,

$$\Pr\Big(\|\widehat{W}_{\cdot j} - W_{\cdot j}\|_1 - \mathbb{E}\|\widehat{W}_{\cdot j} - W_{\cdot j}\|_1 \geq t\Big) \leq \exp\left(-\frac{G_g t^2}{2}\right).$$

Setting $t = \sqrt{2\log(D/\delta_g)/G_g}$ and taking a union bound over $j \in [D]$ gives Eq. 9, with truncation at the maximal possible L_1 distance 2.

It remains to translate weight error into mean error. Let p and q be probability vectors and let $\theta \in [0,1]^K$. Since $\sum_k (p_k - q_k) = 0$, writing $A = \{k : p_k \geq q_k\}$ gives

$$\big|(p-q)^\top \theta\big| \leq \max_{0 \leq \theta \leq 1} (p-q)^\top \theta = \sum_{k \in A} (p_k - q_k) = \frac{1}{2}\|p - q\|_1.$$

Thus the geometry event implies a directional weighted-mean perturbation of at most $\varepsilon_j/2$.

B.5 Proof of the end-to-end simultaneous certificate

Proof of Theorem 1. Condition on the fixed core, the certified weights $\widehat{W}$, and the integer allocation n. For direction j, define the sampling-only centered error

$$E_j = \sum_{s \in \mathcal{S}} \widehat{W}_{sj}(\widehat{\theta}_s - \theta_s).$$

Each $\widehat{\theta}_s$ is an average of n_s independent variables in $[0,1]$. Equivalently, E_j is a weighted sum of independent bounded observations with coefficient $\widehat{W}_{sj}/n_s$ repeated n_s times. Weighted Hoeffding concentration (Hoeffding, 1963) therefore yields

$$\Pr(|E_j| \geq t \mid \widehat{W}, n) \leq 2\exp\left(-\frac{2t^2}{\sum_s \widehat{W}_{sj}^2/n_s}\right).$$

Choosing $t = \rho_j$ from Eq. 10 makes the right-hand side δ_s/D. A union bound over the D directions gives a sampling event $\mathcal{E}_s$ with probability at least $1 - \delta_s$ on which $|E_j| \leq \rho_j$ simultaneously for all j.

The design-conditional geometry bound in Eq. 9 defines a geometry event $\mathcal{E}_g$ with probability at least $1 - \delta_g$. On $\mathcal{E}_g$, Section B.4 gives

$$\left|\sum_s (\widehat{W}_{sj} - W_{sj})\theta_s\right| \leq \frac{\varepsilon_j}{2} \quad \text{for every } j.$$

Using the exact compressed identity,

$$\begin{aligned}|\widehat{\mu}_j - \mu_j| &\leq \left|\sum_s \widehat{W}_{sj}(\widehat{\theta}_s - \theta_s)\right| + \left|\sum_s (\widehat{W}_{sj} - W_{sj})\theta_s\right| \\ &\leq \rho_j + \frac{\varepsilon_j}{2}\end{aligned}$$

on $\mathcal{E}_s \cap \mathcal{E}_g$. Finally,

$$\Pr(\mathcal{E}_s \cap \mathcal{E}_g) \geq 1 - \delta_s - \delta_g,$$

which gives Eq. 12; clipping to $[0,1]$ preserves coverage because every $\mu_j \in [0,1]$. □

On the simultaneous coverage event, $L_j \leq \mu_j \leq U_j$ for every j. Hence $L_j \geq \tau$ is sufficient for $\mu_j \geq \tau$, while $\mu_j \geq \tau$ implies $U_j \geq \tau$. Counting these sets gives the threshold-count bounds in Section 4 for every $\tau \in [0,1]$ on the same event.

B.6 Incidence complexity and the Kakeya gain law

For the ideal sampling radius computed with exact W,

$$\begin{aligned} R^2_{2,\mathrm{samp}}(n) &= \frac{1}{D}\sum_{j=1}^{D}\rho_j^2 \\ &= \frac{\log(2D/\delta_s)}{2D}\sum_{j=1}^{D}\sum_{s\in\mathcal{S}}\frac{W_{sj}^2}{n_s} \\ &= \frac{\log(2D/\delta_s)}{2D}\Phi(n). \end{aligned}$$

Applying Section B.3 gives

$$\min_n R^2_{2,\mathrm{samp}}(n) = \frac{\log(2D/\delta_s)}{2DN}C(\mathcal{C})^2.$$

For independent per-direction Monte Carlo, the weight matrix has D one-hot rows, so every row norm is 1 and $C_{\mathrm{ind}} = D$.

Now consider the full atom partition with congruent tube volume V. For a point $z \in U$, let $m(z) = \sum_j \mathbf{1}\{z \in T_j\}$. On a full atom A_S, $m(z) = |S| = m_S$ and its weight row has m_S entries equal to $|A_S|/V$, hence norm $\sqrt{m_S}|A_S|/V$. Summing over atoms yields

$$C_{\mathrm{full}} = \sum_{S\neq\varnothing}\frac{|A_S|}{V}\sqrt{m_S} = \frac{1}{V}\int_U \sqrt{m(z)}\,dz.$$

By Cauchy-Schwarz,

$$\begin{aligned} C^2_{\mathrm{full}} &= \frac{1}{V^2}\left(\int_U \sqrt{m(z)}\,dz\right)^2 \\ &\le \frac{|U|}{V^2}\int_U m(z)\,dz. \end{aligned}$$

Since each of the D congruent tubes contributes volume V to the multiplicity integral,

$$\int_U m(z)\,dz = DV.$$

Therefore

$$C^2_{\mathrm{full}} \le \frac{D|U|}{V}.$$

Defining $\overline{m} = DV/|U|$ and using $\Phi_{\mathrm{full,opt}} = C^2_{\mathrm{full}}/N$ gives

$$\Phi_{\mathrm{full,opt}} \le \frac{D^2}{N\overline{m}}.$$

Independent sampling has $\Phi_{\mathrm{ind,opt}} = D^2/N$, hence

$$\frac{\Phi_{\mathrm{ind,opt}}}{\Phi_{\mathrm{full,opt}}} \ge \overline{m}.$$

This is the finite-scale Kakeya connection used by the method: the average incidence multiplicity lower-bounds the ideal improvement in the squared-radius objective under full sharing. The inequality follows from the finite-volume identity and Cauchy-Schwarz.

B.7 Exact stratum sampling and computational remarks

For a retained atom A_S, KakeyaShare chooses a proposal direction

$$j_{\mathrm{prop}}(S) = \arg\max_{j\in S}\widetilde{w}_{Sj},$$

draws $z \sim \mathrm{Unif}(T_{j_{\mathrm{prop}}(S)})$, and accepts iff $\sigma(z) = S$. For any measurable $B \subseteq A_S$,

$$\Pr(z \in B \mid \sigma(z) = S) = \frac{|B|/|T_{j_{\mathrm{prop}}(S)}|}{|A_S|/|T_{j_{\mathrm{prop}}(S)}|} = \frac{|B|}{|A_S|},$$

so accepted samples are uniform on A_S. Likewise, sampling uniformly from T_j and accepting iff the signature is not one of the retained core atoms containing j produces uniform samples on R_j. The conditional law is independent of the proposal choice.

For $D \leq 64$, a membership signature fits in one unsigned 64-bit integer; KakeyaShare enumerates only observed signatures and never materializes all 2^D subsets. With G geometry samples per direction, direct vectorized signature computation requires $O(GD^2)$ Boolean tube tests. If $R = |\mathcal{S}|$, integer allocation costs $O((N - R) \log R)$. Exactly N accepted samples are evaluated by the black box; rejected geometry proposals affect wall time but not query count. The algorithm is infeasible if $R > N$, which is explicitly prevented by the budget-aware core-size rule.

C Implementation and experimental protocol

C.1 Canonical geometry and sampling

For $j = 1, \ldots, D$, let $a_j = (j - 1/2)/D$ and $\phi_j = (j - 1)\pi(3 - \sqrt{5})$. The projective Fibonacci axis is

$$u_j = (\sqrt{1 - a_j^2} \cos \phi_j, \sqrt{1 - a_j^2} \sin \phi_j, a_j).$$

Given perpendicular unit vectors v_j, w_j and independent $U_1, U_2, U_3 \sim \mathrm{Unif}[0, 1]$, a uniform point in the cylinder is

$$Z = L(2U_1 - 1)u_j + r\sqrt{U_2}\{\cos(2\pi U_3)v_j + \sin(2\pi U_3)w_j\}.$$

Every geometry in the sweep satisfies $L^2 + r^2 < 1$, so cube truncation is unnecessary. Membership uses the axial and perpendicular inequalities in Eq. 1. A 64-bit signature records membership in at most 64 tubes.

The design pass pools an atom's volume estimates across its member directions, exploiting congruence. The certification pass counts retained-atom and residual categories separately for each direction, preserving an empirical probability column. Its category count includes every retained atom containing that direction, including categories with zero observed certification count. An empirical zero is not evidence of zero true mass. Active strata retain the allocation floor, and the L_1 bound includes the omitted probability. True zero-volume regions can be removed from the mathematical partition. If a required region cannot be sampled within the computational cap, the audit is recorded as failed and has no certified estimate.

Algorithms 1 and 2 specify the complete audit and its integer allocator. Design, certification, and audit use independent randomization streams; deterministic tie-breaking makes the geometry-only decisions reproducible.

The main design and certification budgets are 2^{20} and 2^{22} per tube. Core selection is repeated for each total query budget. Geometry is reusable across anchors and models because it is computed in canonical coordinates. Reusing a geometry event across several fixed loss functions is valid: the event bounds its error for every vector of stratum means in $[0, 1]^R$. The theorem's sampling probability, however, is per audit and simultaneous over its D directions. A joint claim over A audits would additionally allocate sampling error across those audits, for example using δ_s/A. Likewise, simultaneous selection among budget-specific cores would require accounting for their geometry events. The reported intervals concern each specified audit and budget.

C.2 Same-estimand comparators

Write $m(z) = |\sigma(z)|$, $c_U = |U|/V$, and $q_0(z) = m(z)/(DV)$. All black-box calls, including pilots, count toward the total N. Geometry-only proposals and volume calibration do not count as model evaluations. The union and multiplicity samplers allow at most 10^7 geometry proposals per cell. Core rejection uses a per-stratum cap of $\max(10^6 n_s, 4096)$ with proposal batches between 256 and 4096. Feasibility and elapsed time accompany the statistical results.

Algorithm 1 KakeyaShare: simultaneous directional risk audit

Require: Equal-volume tubes $(T_j)_{j=1}^{D}$, bounded chart loss h, budget $N \geq D$
Require: Geometry budgets G_d, G_g, errors δ_g, δ_s, rejection caps
Ensure: Estimates $\widehat{\mu}_j$ and intervals $[L_j, U_j]$, or an unavailable status

Draw G_d independent design points from each T_j; compute signatures
Estimate pooled equal-volume atom rows $\widetilde{w}_{S:}$
Sort the M observed shared atoms by $\widetilde{b}_S = \|\widetilde{w}_{S:}\|_1 - \|\widetilde{w}_{S:}\|_2$
for $K = 0, \ldots, \min(M, N - D)$ **do**
    Form the prefix core $\mathcal{C}_K$ and its residuals as in Eq. 4
    Construct $\widetilde{W}_K$; $n^{(K)} \leftarrow \textsc{IntegerAllocate}(\widetilde{W}_K, N)$
    $\widetilde{\Phi}_K^{\mathrm{int}} \leftarrow \sum_s \|\widetilde{W}_{K,s:}\|_2^2 / n_s^{(K)}$
end for
$K_* \leftarrow \arg\min_K \widetilde{\Phi}_K^{\mathrm{int}}$; resolve ties by smaller K
Fix $\mathcal{C} \leftarrow \mathcal{C}_{K_*}$ and strata $\mathcal{S} \leftarrow \mathcal{C} \cup \{R_1, \ldots, R_D\}$
for $j = 1, \ldots, D$ **do**
    Draw G_g fresh certification points uniformly from T_j
    Count its retained-atom and residual categories to form $\widehat{W}_{\cdot j}$
    Compute ε_j from Eq. 9, including zero-count categories
end for
$n \leftarrow \textsc{IntegerAllocate}(\widehat{W}, N)$
for each stratum $s \in \mathcal{S}$, using an independent audit stream **do**
    Set proposal tube to a design-selected member of s if s is an atom, or to T_j if $s = R_j$; initialize accepted losses $H_s \leftarrow [\,]$
    while $|H_s| < n_s$ and the proposal cap is unexhausted **do**
        Draw Z uniformly from the proposal tube
        if $\sigma(Z) = s$ for a core atom, or $Z \in R_j$ for a residual **then**
            Evaluate $h(Z)$ once; append it to H_s
        end if
    end while
    if $|H_s| < n_s$ **then**
        return unavailable
    end if
    $\widehat{\theta}_s \leftarrow n_s^{-1} \sum_{y \in H_s} y$
end for
$\widehat{\mu} \leftarrow \widehat{W}^{\top} \widehat{\theta}$; compute ρ_j from Eq. 10
return $\widehat{\mu}_j$, $[\max(0, \widehat{\mu}_j - \rho_j - \varepsilon_j/2)$, $\min(1, \widehat{\mu}_j + \rho_j + \varepsilon_j/2)]$ for every j

Algorithm 2 Exact integer L_2 allocation

function $\textsc{IntegerAllocate}(W, N)$
Require: R active rows of W and $N \geq R$
    $q_s \leftarrow \|W_{s:}\|_2$ and $n_s \leftarrow 1$ for each $s = 1, \ldots, R$
    Initialize a max-heap with keys $q_s^2/[n_s(n_s + 1)]$
    for $t = 1, \ldots, N - R$ **do**
        Remove a largest-key row s using fixed tie-breaking
        $n_s \leftarrow n_s + 1$; reinsert s with key $q_s^2/[n_s(n_s + 1)]$
    end for
    return n
end function

Independent Monte Carlo. Each direction receives $n = N/D$ independent uniform tube samples. Its sample mean has half-width $\sqrt{\log(2D/\delta_s)/(2n)}$. There is no estimated geometry coefficient, so its geometry term is zero. The recorded comparisons use $\delta_s = 0.025$, giving IMC at least 97.5% simultaneous coverage. For the common 95% comparison, IMC instead uses $\delta_s = 0.05$; Appendix C.7 derives the resulting target and query saving.

Union-HT. Draw from q_0 by choosing a uniform direction and then a uniform point in its tube. Accept with probability $1/m(z)$. Accepted points are uniform on U. For $Y_j = h(Z)\mathbf{1}\{j \in \sigma(Z)\}$, $\mu_j = c_U \mathbb{E}_U Y_j$. The pooled DG_g independent calibration points estimate c_U by the mean of $D/m(Z)$, whose range is $[1, D]$. With

$$g_U = (D-1)\sqrt{\frac{\log(2/\delta_g)}{2DG_g}}, \qquad \widehat{\mu}_j = \widehat{c}_U \overline{Y}_j,$$

the directional half-width is $\widehat{c}_U \sqrt{\log(2D/\delta_s)/(2N)} + g_U$. Pooling remains valid for equal numbers of independent draws from each tube, since their average expectation is the mixture expectation. The same scalar geometry event supports all directions.

DM-MIS with the balance denominator. The executed sampler randomizes the proposal index independently for each observation. Its denominator is the equal mixture of all tube densities, often called the deterministic-mixture denominator in MIS (Elvira et al., 2019). The estimator is

$$\widehat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} \frac{Dh(Z_i)\mathbf{1}\{j \in \sigma(Z_i)\}}{m(Z_i)}, \qquad Z_i \sim q_0.$$

Each summand lies in $[0, D]$ and has expectation μ_j. Its simultaneous Hoeffding half-width is $D\sqrt{\log(2D/\delta_s)/(2N)}$. No volume estimate is needed. The method name does not imply fixed per-proposal counts in this implementation.

Pilot-optimized mixture. OptDM-MIS spends $N_p = 16D$ queries on q_0, then draws $N_m = N - N_p$ queries from $q_\alpha(z) = s_\alpha(z)/V$, where $s_\alpha(z) = \sum_{j \in \sigma(z)} \alpha_j$. Mixture weights sum to one and satisfy $\alpha_j \geq 1/(4D)$. The pilot objective is the empirical average of $h(Z_i)^2/s_\alpha(Z_i)$, proportional to the sum of directional second-moment objectives under the uniform pilot mixture. SLSQP uses tolerance 10^{-10}, at most 500 iterations, and a uniform fallback.

Both stages are retained with their own proposal corrections:

$$\widehat{\mu}_j = \frac{1}{N}\left\{\sum_{i=1}^{N_p} \frac{Dh(Z_i)\mathbf{1}\{j \in \sigma(Z_i)\}}{m(Z_i)} + \sum_{i=1}^{N_m} \frac{h(Z_i')\mathbf{1}\{j \in \sigma(Z_i')\}}{s_\alpha(Z_i')}\right\}.$$

The first-stage summands are unbiased unconditionally; second-stage summands are unbiased conditional on the pilot. The combined estimator is therefore unbiased. The recorded radius is

$$\rho_j^{\mathrm{nom}} = \sqrt{\frac{\log(2D/\delta_s)}{2N^2}\left(N_p D^2 + \frac{N_m}{\alpha_j^2}\right)}. \tag{C.1}$$

Here α depends on the pilot outcomes. Conditioning on α does not preserve the pilot's original independent sampling law. Moreover, a predictable random range cannot be substituted into an optimized fixed-range Hoeffding boundary without an additional argument. We consequently label this recorded radius nominal throughout the paper.

A valid alternative follows by bounding the two centered stage sums separately. Give each stage simultaneous error probability $\delta_s/2$. Hoeffding bounds the first sum unconditionally, and bounds the second conditional on the entire pilot. A union bound then gives

$$\rho_j^{\mathrm{split}} = \frac{1}{N}\left[D\sqrt{\frac{N_p}{2}\log\frac{4D}{\delta_s}} + \frac{1}{\alpha_j}\sqrt{\frac{N_m}{2}\log\frac{4D}{\delta_s}}\right]. \tag{C.2}$$

This formula establishes a usable theoretical alternative; the experimental tables retain Eq. C.1 and contain no recomputed values for Eq. C.2.

Multiplicity-Neyman. Let $H_k = \{z : m(z) = k\}$ and $p_k = \Pr_{q_0}(m(Z) = k)$. The volume coefficient is $a_k = |H_k|/V = Dp_k/k$. Conditional on $m(Z) = k$, a mixture proposal is uniform on H_k. The directional mean is

$$\mu_j = \sum_{k=1}^{D} a_k \mathbb{E}_{H_k}[h(Z)\mathbf{1}\{j \in \sigma(Z)\}].$$

Calibration gives frequencies $\widehat{p}_k$ and $\widehat{a}_k = D\widehat{p}_k/k$. A simultaneous frequency error $\beta = \sqrt{\log(2D/\delta_g)/(2DG_g)}$ yields the conservative additive geometry term $g_{\text{mult}} = \sum_{k=1}^{D} D\beta/k$, including unseen multiplicities. Eight pilot queries per observed active stratum estimate directional variances. The remaining allocation is proportional to $\widehat{a}_k\sqrt{\overline{v}_k + 10^{-12}}$, using a weighted-deficit integer rule. Pilots count toward the budget and are excluded from the final means. Conditional on pilot and geometry, the directional sampling radius is

$$\rho_j = \sqrt{\frac{\log(2D/\delta_s)}{2} \sum_{k:n_k>0} \frac{\widehat{a}_k^2}{n_k}}.$$

The bound uses $[0, 1]$ for every indicator-weighted loss. A multiplicity stratum can contain different membership signatures, so its direction-specific indicator remains random.

C.3 Core and allocation controls

The core comparison uses no sharing ($K = 0$), fixed $K = 360$, the budget-selected prefix, and all observed shared atoms. Residuals are retained in each construction. FullAtoms is the uncompressed exact-incidence construction. It requires one sample per active region and can fail either this budget constraint or its rejection cap. At the main budget the 732 active regions fit within 6144 queries, so the observed failure is computational.

Within the same selected core, the allocation comparison uses uniform counts, mass-proportional counts, continuous L2 allocation with largest-remainder rounding, exact integer L2 allocation, and pilot Neyman allocation. The last allocates eight pilot observations to each active stratum and reuses them in the final estimate after adapting counts to their sample variances. The fixed-allocation Hoeffding argument does not cover that reuse, so its reported widths are nominal. This differs from the external multiplicity comparator, which discards its pilot from the estimator. The uniform and geometry-only allocations do satisfy the conditional independence required by Theorem 1. An unspecified row-norm heuristic has no executable definition and is reported as unavailable.

C.4 Datasets, charts, and model fitting

Table 5: Main tasks and reference-subset membership.

Dataset	Task ID	n	Numeric	Classes	Ref.
APSFailure	363616	76000	170	2	Yes
Bank_Customer_Churn	363619	10000	6	2	Yes
Bioresponse	363620	3751	1776	2	Yes
Diabetes130US	363630	71518	8	2	Yes
E-CommereShippingData	363632	10999	6	2	Yes
Fitness_Club	363671	1500	3	2	Yes
GiveMeSomeCredit	363673	150000	10	2	Yes
Is-this-a-good-customer	363682	1723	5	2	
MIC	363711	1699	17	8	Yes
Marketing_Campaign	363684	2240	16	2	Yes
SDSS17	363699	78053	9	3	Yes
anneal	363614	898	6	5	
bank-marketing	363618	45211	5	2	
blood-transfusion-service-center	363621	748	4	2	
churn	363623	5000	15	2	

Continued on next page.

Table 5 continued.

Dataset	Task ID	n	Numeric	Classes	Ref.
coil2000_insurance_policies	363624	9822	80	2	
credit-g	363626	1000	7	2	
credit_card_clients_default	363627	30000	20	2	
customer_satisfaction_in_airline	363628	129880	5	2	
diabetes	363629	768	8	2	
hazelnut-spread-contaminant-detection	363674	2400	30	2	
heloc	363676	10459	23	2	
jm1	363712	10885	21	2	
kddcup09_appetency	363683	50000	174	2	
maternal_health_risk	363685	1014	6	3	Yes
online_shoppers_intention	363691	12330	10	2	
polish_companies_bankruptcy	363694	5910	64	2	
qsar-biodeg	363696	1054	36	2	
seismic-bumps	363700	2584	11	2	
students_dropout_and_academic_success	363704	4424	19	3	Yes

The main tasks and the 12 reference tasks are identified in Table 5. We use OpenML task splits with repeat 0, sample 0, and fold 0. Within the training partition, a stratified 20% validation split uses seed 26083101. The training partition alone supplies preprocessing, PCA, calibration fitting, and model fitting; test outputs are used to select audit anchors. For tabular anchors, top-two predictive probability margins are divided into the bottom and top quartiles separately among correct and incorrect predictions. One anchor is sampled per available group using a deterministic hash of dataset, model, split, and group. Reference evaluation uses two per group, giving eight anchors per dataset and model. The split sensitivity study uses folds 1 and 2 on ten tasks.

Numeric chart inputs are median-imputed from training data. Constant features are held fixed. On the remaining standardized features, let $B \in \mathbb{R}^{p\times 3}$ contain the first three principal directions and let $\alpha = 1/\max_k \|B_{k:}\|_1$. The chart adds αBz in standardized coordinates, transforms back, and clips to training minima and maxima. The cube-wide displacement before clipping is at most one training standard deviation in each numeric feature. Categorical values remain those of the anchor. Clipping is part of the fixed measurable chart and hence part of the estimand. The random-chart control replaces B with a seeded Gaussian QR basis while keeping the same normalization.

The LightGBM adapter uses 1000 estimators, learning rate 0.05, 31 leaves, column and row subsampling parameters 0.8, and L2 penalty 1. Logistic regression uses $C = 1$, L-BFGS, and at most 3000 iterations. Their numeric pipeline uses median imputation and standardization; categorical features use most-frequent imputation and one-hot encoding with unknown categories ignored. RealMLP and TabM use their AutoGluon model adapters with the fixed model seed, no bagging, no stacking, no weighted ensemble, and one GPU. The runtime also defines a RealMLP fallback consisting of a three-layer MLP of widths 256, 256, and 128; model identity should be checked against the saved fingerprint when reproducing a particular run. The aggregate result tables do not expose a per-task fallback indicator.

The frontier adapters use classifier APIs. TabPFN-3 enables cached fitting on CUDA; TabICLv2 uses cached keys and values, batch size 32, and CUDA. Both use the main model seed. A regression checkpoint is rejected by the classification adapter. The result summaries identify model families and failures but do not identify the classifier checkpoint hashes used by every successful cell; exact checkpoint provenance therefore remains a reproducibility limitation of these summaries.

CIFAR-10 uses its 50,000 training and 10,000 test images. The two checkpoints are `cifar10_resnet20` from the `chenyaofo/pytorch-cifar-models` repository and `nateraw/vit-base-patch16-224-cifar10`. ResNet inputs use channel means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2470, 0.2435, 0.2616)$; ViT inputs use its checkpoint image processor. The three global image PCA directions are fit on training images. With $\alpha = (8/255)/\max_k \|B_{k:}\|_1$, the chart is $\mathrm{clip}_{[0,1]}(x_0 + \alpha Bz)$. Thirty-two correctly predicted anchors per model comprise 16 from each extreme margin quartile. The DCT control uses spatial frequencies $(0,0), (1,0), (0,1)$, shared across channels and orthonormalized. Two audit seeds are used for vision.

Table 6: Randomization and budget settings.

Quantity	Setting
Model training	26083101
Tabular audit	26083111, 26083112, 26083113, 26083114
Geometry design / certification	26083121 / 26083122
Geometry diagnostics	26083124 / 26083123
Dataset bootstrap	26083191
Stress manifest	26083151
Main N/D	128, 256, 512, 1024
Reference-subset N/D	64, 128, 256, 512, 1024, 2048, 4096
Frontier N/D	256, 512, 1024
Main geometry draws per tube	$G_d = 2^{20}$, $G_g = 2^{22}$
Geometry sweep design draws per tube	$G_d = 2^{18}$

C.5 Losses, calibration, and input stress

The four losses are $1 - p_y$, normalized multiclass Brier score $\frac{1}{2}\sum_c (p_c - \mathbf{1}\{c = y\})^2$, normalized clipped cross entropy $-\log(\max(p_y, 10^{-6}))/[-\log(10^{-6})]$, and zero-one prediction error. All lie in $[0, 1]$. Temperature scaling minimizes validation negative log likelihood over temperatures in $[10^{-2}, 100]$; binary isotonic regression uses validation scores and clips extrapolation. Isotonic results therefore use only the binary-eligible tasks.

Input stress includes 10% training-label corruption, 10% missing-completely-at-random numeric entries in training and test data followed by training-median imputation, feature-scale changes, a covariate-shift subset, and an OOD-like subset. Label replacements are sampled uniformly from the other classes without replacing selected indices. The scale control assigns each numeric feature a seeded factor of either 0.1 or 10, applied to training and test inputs. Covariate shift selects the top 20% of test examples by first training-principal-component score; the OOD-like subset selects the top 5% by training-based Mahalanobis distance, with ridge 10^{-10}. Each altered predictor, anchor, chart, or loss has its own reference integration. Canonical tube geometry remains unchanged.

C.6 Reference integration, metrics, and aggregation

Point references average eight independent scrambled-Sobol integrations, each with 2^{14} points per direction for tabular data and 2^{13} for vision. An independent Monte Carlo stream supplies a conservative interval for the same directional mean. Reference interval half-widths are 0.02277 and 0.03221 respectively; Table 10 also reports between-scramble standard errors and point-reference agreement. A point reference is an accurate numerical comparison target, while its independent interval supports the containment diagnostic.

Let I_j denote the audit interval and I_j^{ref} its reference. A direction is reference-contained if $I_j^{\mathrm{ref}} \subseteq I_j$, falsified if their intersection is empty, and ambiguous otherwise. A cell is reference-contained only if every direction is reference-contained, and falsified if at least one direction is falsified. These diagnostics explicitly retain uncertainty in the reference. They are not estimates of exact-truth coverage with an error-free oracle. Synthetic functions with analytic integrals give a direct check; the remaining synthetic functions use high-precision independent references.

For a paired budget curve at its stated confidence budget, let $R_0 = R_{2,\mathrm{IMC}}(N_0)$ and let adjacent budgets N_a, N_b bracket its first crossing. Log interpolation gives

$$\log N_m = \log N_a + \frac{\log R_0 - \log R_m(N_a)}{\log R_m(N_b) - \log R_m(N_a)} (\log N_b - \log N_a).$$

For eligible certified curves, noncrossings receive the main-grid cap $N_m = 2N_0$ and are marked right-censored. Curves already below the target at the smallest budget are marked as such. The OptDM-MIS radius is nominal, so its certified crossing in Table 1 is unavailable. Figure 3 and Table 11 retain that nominal curve for comparison with the executed estimator's point errors and reference containment; Eq. C.2 has not been evaluated experimentally.

Paired anchor-seed savings are aggregated within datasets, then summarized across datasets. Dataset-cluster bootstrap intervals resample datasets while retaining their internal observations. At the main geometry, the KS radius depends only on the shared geometry measurements, partition, and allocation. Its dataset bootstrap interval therefore collapses to the same value; this is a property of the endpoint and supplies no estimate of variation across predictors or datasets. The invariance statement is specific to this KS configuration; loss-dependent pilot allocations can change comparator radii.

Table 1 uses median dataset RMSE and MAE, after pooling the available directional errors within each dataset; Ref. C is the mean of dataset directional-containment percentages. Appendix budget tables instead pool directional squared and absolute errors within each model and budget, as recorded in the source summaries. These different aggregation levels explain numerical differences between the tables. Stress comparisons pair the same dataset, model, anchor, seed, and condition. Failure counts use planned cells or explicitly stated available support, and unavailable models remain in the model-level denominator.

C.7 Confidence budgets and matched-radius query savings

The simultaneous error allowance must account for all estimated components of a certificate. KakeyaShare uses $\delta_g = \delta_s = 0.025$, so Theorem 1 guarantees at least 95% simultaneous coverage. Independent MC has exact directional sampling distributions and no estimated volume coefficient. Giving it the entire error allowance, $\delta_s^{\mathrm{IMC}} = 0.05$, therefore yields the appropriate comparator for a common 95% guarantee. Equal sampling-error budgets answer a different question: they hold the sampling concentration factor fixed, but give IMC the stronger 97.5% guarantee. Both comparisons are reported explicitly.

For $n = N/D$ queries per direction and sampling error allowance η, the IMC RMS half-width is

$$r_{\mathrm{IMC}}(n;\eta) = \sqrt{\frac{\log(2D/\eta)}{2n}}. \tag{C.3}$$

At $D = 12$ and $n = 512$, this gives $r_{\mathrm{IMC}}(512; 0.025) = 0.0818901065$ and $r_{\mathrm{IMC}}(512; 0.05) = 0.0776472021$. The KakeyaShare curve retains its geometry and sampling budgets. Its recorded total RMS radii at 256 and 512 queries per direction are 0.08902805638 and 0.06386047622. Applying the log-budget/log-radius interpolation in Appendix C.6 to these adjacent points gives the crossings in Table 7. Since the geometry and allocation fix this curve, no predictor evaluations or new coverage trials are needed for this calculation.

Table 7: Confidence-budget comparisons at the main geometry. KakeyaShare always uses $(\delta_g, \delta_s) = (0.025, 0.025)$. Query counts are per direction; KS crossings are interpolated.

Comparison	IMC δ_s	Target R_2	IMC n	KS n	Saving
Equal sampling budget	0.025	0.081890	512.0	304.8	40.5%
Common 95%, IMC $n = 512$	0.050	0.077647	512.0	340.5	33.5%
Common 95%, fixed higher target	0.050	0.081890	460.3	304.8	33.8%

The headline comparison fixes the IMC budget at 512 and uses the second row: $1 - 340.5363/512 = 33.4890\%$. If instead the radius 0.0818901065 is held fixed, the analytic IMC query equivalent is

$$n_{\mathrm{IMC}} = \frac{\log(24/0.05)}{2(0.0818901065)^2} \simeq 460.3188, \tag{C.4}$$

and the saving is $1 - 304.7615/460.3188 = 33.7934\%$. This third row has the same total confidence as the second but a different target radius. The noninteger IMC count is an analytic budget equivalent; an executed equal-per-direction allocation would round it upward. Interpolated KS counts likewise summarize a curve and are not separately executed audits.

Tables 1–2 and Figure 3 preserve the recorded equal-δ_s comparisons. In particular, their 40.5% saving and 22.02% same-budget radius reduction are under that convention. The analytic IMC confidence adjustment changes interval widths and crossings, not the recorded point estimates. Reference containment percentages remain attached to the originally evaluated intervals; no recalculated

containment is implied. DM-MIS also has no geometry-estimation term and can use the full 0.05 allowance for a common-confidence comparison, whereas the geometry-dependent estimators must retain an appropriate split. The headline common-confidence claim compares KakeyaShare with IMC; the other recorded curves are not relabeled as common-confidence experiments. OptDM-MIS retains its explicitly nominal status.

D COMPLETE EMPIRICAL EVIDENCE

This appendix reports the numerical evidence behind the main comparisons, including unsuccessful configurations. Tables distinguish exact-truth checks, reference-containment diagnostics, point error, and computational feasibility. Percentages use the denominator named in each table. A nominal interval calculation is never counted as a proof of coverage.

D.1 SYNTHETIC VALIDITY AND GEOMETRY UNCERTAINTY

The synthetic audit uses 2000 repetitions per method and loss. Four losses have analytic directional integrals: the constant 0.37; $\mathrm{sigmoid}(6a^\top z)$ for $a = (1, 2, -1)/\sqrt{6}$, whose mean is $1/2$ by central symmetry; $\|z\|_2^2/(L^2 + r^2)$, whose mean is $(L^2/3 + r^2/2)/(L^2 + r^2)$; and $\mathbf{1}\{a^\top z > 0\}$, also of mean $1/2$. Three incidence-sensitive losses are a needle with value 0.95 in the radius-0.055 ball centered at $(0.18, -0.11, 0.07)$ and 0.05 elsewhere, the core loss $\mathrm{clip}((m(z) - 1)/(D - 1), 0, 1)$, and its complement. These use independent high-precision numerical references. The reference computation uses 32 scrambles with 2^{24} points per direction and a separate 2^{24}-point Monte Carlo interval with reference error allowance 10^{-6}.

Table 8 shows one KS miss and eight IMC misses, all for the discontinuous jump loss. The overall familywise containment rates are 99.993% and 99.943%. They demonstrate conservative behavior at the stated 95% target. One-sided binomial undercoverage tests with Holm correction reject none of the seven functions. A finite set of losses cannot replace the theorem's assumptions; these tests exercise discontinuity, concentrated mass, and incidence-dependent loss. FullAtoms has no successful synthetic audits under the rejection cap. At the main geometry, the geometry term contributes a median 6.553% and a 95th-percentile 6.673% of the total directional half-width.

Table 8: Synthetic familywise containment and falsification.

Loss	IMC covered	KS covered	IMC misses	KS misses
centered smooth	2000/2000	2000/2000	0	0
constant	2000/2000	2000/2000	0	0
core concentrated	2000/2000	2000/2000	0	0
jump	1992/2000	1999/2000	8	1
multiplicity antagonistic	2000/2000	2000/2000	0	0
needle	2000/2000	2000/2000	0	0
radial quadratic	2000/2000	2000/2000	0	0

Independent design seeds produce similar cores, and independent certification streams satisfy their two-stream L_1 triangle bounds at all four main budgets. Table 9 reports these diagnostics. They test numerical stability of the chosen geometry budgets; the formal geometry guarantee is the probability bound in Appendix B.4.

Table 9: Independent geometry-stream diagnostics.

N/D	Main K	Diagnostic K	Jaccard	Max. L_1 gap	Bound holds
128	213	211	0.9813	0.00339	Yes
256	282	287	0.9826	0.00375	Yes
512	360	366	0.9836	0.00402	Yes
1024	450	449	0.9845	0.00432	Yes

D.2 Reference precision and complete budget curves

Table 10: Independent reference quality.

Model	Anchors	RQMC/MC RMSE	Scramble SE	MC half-width
LGBM	96	0.00141	0.0000250	0.02277
TabM	96	0.00111	0.0000132	0.02277
ResNet20	32	0.00005	0.0000000	0.03221
ViT-B/16	32	0.00000	0.0000000	0.03221

Table 10 separates point-reference agreement from the conservative Monte Carlo interval used for containment. As the audit budget grows, its interval can become narrower than the reference interval even if its point error decreases. The resulting ambiguous outcomes should not be read as disjointness or demonstrated undercoverage. Table 11 supplies every reference-subset budget for the six primary estimators. In particular, the large ranges of importance-weighted observations remain costly at high budgets, while KS eventually encounters its fixed geometry contribution. Table 1 summarizes at the dataset level; the errors below use pooled directional observations within each model.

Table 11: Complete reference-subset budget curves.

Model	Method	N/D	R_2	RMSE	MAE	Dir. Ref-C%	Cell Ref-C%
LGBM	DM-MIS	64	0.80236	0.06183	0.03928	100.00	100.00
LGBM	DM-MIS	128	0.56735	0.04611	0.02893	100.00	100.00
LGBM	DM-MIS	256	0.40118	0.03013	0.01934	100.00	100.00
LGBM	DM-MIS	512	0.28368	0.01997	0.01266	100.00	100.00
LGBM	DM-MIS	1024	0.20059	0.01297	0.00802	100.00	100.00
LGBM	DM-MIS	2048	0.14184	0.01029	0.00615	100.00	100.00
LGBM	DM-MIS	4096	0.10029	0.00770	0.00461	100.00	100.00
LGBM	IMC	64	0.23162	0.02023	0.01081	100.00	100.00
LGBM	IMC	128	0.16378	0.01607	0.00843	100.00	100.00
LGBM	IMC	256	0.11581	0.01086	0.00576	100.00	100.00
LGBM	IMC	512	0.08189	0.00793	0.00424	100.00	100.00
LGBM	IMC	1024	0.05791	0.00527	0.00282	99.96	99.48
LGBM	IMC	2048	0.04095	0.00373	0.00201	99.50	94.79
LGBM	IMC	4096	0.02895	0.00259	0.00140	94.38	71.35
LGBM	KS	64	0.18077	0.01339	0.00720	100.00	100.00
LGBM	KS	128	0.12620	0.01049	0.00557	100.00	100.00
LGBM	KS	256	0.08903	0.00720	0.00384	100.00	100.00
LGBM	KS	512	0.06386	0.00569	0.00292	99.89	98.70
LGBM	KS	1024	0.04662	0.00410	0.00217	99.85	98.44
LGBM	KS	2048	0.03456	0.00279	0.00149	98.81	90.62
LGBM	KS	4096	0.02607	0.00211	0.00111	86.11	54.17
LGBM	Mult.-Neyman	64	0.55083	0.05535	0.03510	100.00	100.00
LGBM	Mult.-Neyman	128	0.38008	0.03350	0.02102	100.00	100.00
LGBM	Mult.-Neyman	256	0.26771	0.02251	0.01377	100.00	100.00
LGBM	Mult.-Neyman	512	0.19100	0.01522	0.00986	100.00	100.00
LGBM	Mult.-Neyman	1024	0.13746	0.01415	0.00841	100.00	100.00
LGBM	Mult.-Neyman	2048	0.09989	0.00833	0.00522	100.00	100.00
LGBM	Mult.-Neyman	4096	0.07342	0.00503	0.00317	100.00	100.00
LGBM	OptDM†	64	0.84840	0.05835	0.03714	100.00	100.00
LGBM	OptDM†	128	0.60517	0.03876	0.02456	100.00	100.00
LGBM	OptDM†	256	0.42977	0.02546	0.01613	100.00	100.00
LGBM	OptDM†	512	0.30454	0.01740	0.01087	100.00	100.00
LGBM	OptDM†	1024	0.21557	0.01368	0.00859	100.00	100.00
LGBM	OptDM†	2048	0.15251	0.00944	0.00597	99.93	99.22
LGBM	OptDM†	4096	0.10787	0.00655	0.00404	99.59	95.05
LGBM	Union-HT	64	0.48591	0.05281	0.03340	100.00	100.00
LGBM	Union-HT	128	0.34426	0.03771	0.02362	100.00	100.00
LGBM	Union-HT	256	0.24410	0.02667	0.01712	100.00	100.00

Continued on next page.

Table 11 continued.

Model	Method	N/D	R_2	RMSE	MAE	Dir. Ref-C%	Cell Ref-C%
LGBM	Union-HT	512	0.17328	0.01754	0.01142	100.00	100.00
LGBM	Union-HT	1024	0.12320	0.01317	0.00819	100.00	100.00
LGBM	Union-HT	2048	0.08779	0.01036	0.00649	100.00	100.00
LGBM	Union-HT	4096	0.06275	0.00734	0.00438	100.00	100.00
TabM	DM-MIS	64	0.80236	0.06036	0.04067	100.00	100.00
TabM	DM-MIS	128	0.56735	0.04487	0.02967	100.00	100.00
TabM	DM-MIS	256	0.40118	0.02947	0.02010	100.00	100.00
TabM	DM-MIS	512	0.28368	0.01959	0.01323	100.00	100.00
TabM	DM-MIS	1024	0.20059	0.01273	0.00836	100.00	100.00
TabM	DM-MIS	2048	0.14184	0.01011	0.00637	100.00	100.00
TabM	DM-MIS	4096	0.10029	0.00753	0.00476	100.00	100.00
TabM	IMC	64	0.23162	0.01693	0.00935	100.00	100.00
TabM	IMC	128	0.16378	0.01336	0.00719	100.00	100.00
TabM	IMC	256	0.11581	0.00880	0.00488	100.00	100.00
TabM	IMC	512	0.08189	0.00616	0.00347	100.00	100.00
TabM	IMC	1024	0.05791	0.00414	0.00227	99.98	99.74
TabM	IMC	2048	0.04095	0.00298	0.00165	99.78	97.66
TabM	IMC	4096	0.02895	0.00200	0.00111	96.46	77.08
TabM	KS	64	0.18077	0.01139	0.00631	100.00	100.00
TabM	KS	128	0.12620	0.00862	0.00469	100.00	100.00
TabM	KS	256	0.08903	0.00579	0.00319	100.00	100.00
TabM	KS	512	0.06386	0.00476	0.00256	99.96	99.48
TabM	KS	1024	0.04662	0.00336	0.00185	99.98	99.74
TabM	KS	2048	0.03456	0.00241	0.00130	99.35	93.75
TabM	KS	4096	0.02607	0.00187	0.00100	90.00	63.02
TabM	Mult.-Neyman	64	0.54449	0.05421	0.03637	100.00	100.00
TabM	Mult.-Neyman	128	0.37454	0.03224	0.02138	100.00	100.00
TabM	Mult.-Neyman	256	0.26346	0.02168	0.01405	100.00	100.00
TabM	Mult.-Neyman	512	0.18775	0.01470	0.01009	100.00	100.00
TabM	Mult.-Neyman	1024	0.13515	0.01374	0.00864	100.00	100.00
TabM	Mult.-Neyman	2048	0.09823	0.00817	0.00543	100.00	100.00
TabM	Mult.-Neyman	4096	0.07225	0.00489	0.00325	100.00	100.00
TabM	OptDM†	64	0.83838	0.05719	0.03863	100.00	100.00
TabM	OptDM†	128	0.59697	0.03780	0.02529	100.00	100.00
TabM	OptDM†	256	0.42357	0.02477	0.01677	100.00	100.00
TabM	OptDM†	512	0.30003	0.01697	0.01123	100.00	100.00
TabM	OptDM†	1024	0.21233	0.01334	0.00892	100.00	100.00
TabM	OptDM†	2048	0.15021	0.00925	0.00620	99.93	99.22
TabM	OptDM†	4096	0.10623	0.00628	0.00408	99.80	97.66
TabM	Union-HT	64	0.48591	0.05121	0.03412	100.00	100.00
TabM	Union-HT	128	0.34426	0.03666	0.02425	100.00	100.00
TabM	Union-HT	256	0.24410	0.02601	0.01773	100.00	100.00
TabM	Union-HT	512	0.17328	0.01684	0.01162	100.00	100.00
TabM	Union-HT	1024	0.12320	0.01267	0.00833	100.00	100.00
TabM	Union-HT	2048	0.08779	0.01011	0.00670	100.00	100.00
TabM	Union-HT	4096	0.06275	0.00717	0.00448	100.00	100.00

Errors pool all recorded directional observations within a model and budget. Ref-C denotes reference containment. All displayed methods have no falsified intervals; the complement of containment is ambiguous. †: recorded nominal radius, Eq. C.1.

D.3 Point-estimation controls

At 512 queries per direction, scrambled Sobol has pooled RMSE 0.001976 for LightGBM and 0.001062 for TabM, compared with KS values 0.005692 and 0.004764. Latin hypercube gives 0.004308 and 0.002611. Table 12 reports their full curves. These results place the query-sharing result in context: its advantage is measured against the explicit simultaneous interval objective and the listed range-based comparators. The data do not establish point-error superiority over randomized low-discrepancy integration. Incorporating such dependence into shared-stratum certification would require a separate uncertainty argument.

Table 12: Point-estimation controls across budgets.

Model	Sampler	N/D	RMSE	MAE	Bias
LGBM	LHS	64	0.012000	0.006432	-0.000188
LGBM	LHS	128	0.008210	0.004432	-0.000104
LGBM	LHS	256	0.005799	0.003127	-0.000057
LGBM	LHS	512	0.004308	0.002290	0.000056
LGBM	LHS	1024	0.002953	0.001591	-0.000087
LGBM	LHS	2048	0.002112	0.001146	0.000015
LGBM	LHS	4096	0.001498	0.000800	0.000006
LGBM	Sobol	64	0.008691	0.004708	0.000224
LGBM	Sobol	128	0.005485	0.002921	-0.000043
LGBM	Sobol	256	0.003253	0.001745	-0.000048
LGBM	Sobol	512	0.001976	0.001069	-0.000044
LGBM	Sobol	1024	0.001261	0.000680	-0.000031
LGBM	Sobol	2048	0.000819	0.000435	-0.000015
LGBM	Sobol	4096	0.000504	0.000271	-0.000006
TabM	LHS	64	0.007535	0.004144	0.000025
TabM	LHS	128	0.005159	0.002872	0.000036
TabM	LHS	256	0.003767	0.002041	-0.000024
TabM	LHS	512	0.002611	0.001437	-0.000004
TabM	LHS	1024	0.001861	0.001034	-0.000026
TabM	LHS	2048	0.001358	0.000745	-0.000029
TabM	LHS	4096	0.000959	0.000524	-0.000011
TabM	Sobol	64	0.004864	0.002701	0.000114
TabM	Sobol	128	0.003008	0.001655	0.000114
TabM	Sobol	256	0.001722	0.000965	0.000076
TabM	Sobol	512	0.001062	0.000583	0.000029
TabM	Sobol	1024	0.000646	0.000355	0.000013
TabM	Sobol	2048	0.000412	0.000223	0.000014
TabM	Sobol	4096	0.000247	0.000136	0.000006

D.4 What the mechanism controls explain

The selected core has $K = 282, 360, 450$ at 256, 512, and 1024 queries per direction. The fixed-core comparison changes this choice while retaining the same sampling and certification machinery. Table 13 also reports allocation-specific minimum counts and acceptance. Uniform allocation spends a large fraction of evaluations on tiny strata and raises the range-based aggregate uncertainty. Mass allocation restores much of the benefit, while row norms account for all directional coefficients jointly. Exact integer allocation consistently improves the evaluated objective over continuous allocation with largest-remainder rounding.

The mechanism evidence has three levels: the finite-volume identity predicts an ideal sharing opportunity; controlled core and allocation ablations test the implemented sampling choices; full-radius comparisons include calibration uncertainty. The separate mechanism calibration contains four distinct geometry/budget support points, repeated across 7680 recorded cells. Its sampling-gain correlation is Spearman 1.000 and Pearson 0.9997 on the log scale, whereas the corresponding full-radius correlations are -0.800 and -0.7584. The decomposition in Eq. A.1 identifies the additional geometry terms in the latter quantity. Dataset repetitions preserve the same radii and add no independent geometry support points. A mixed-effects analysis of the invariant matched-saving endpoint is consequently unidentifiable. These correlations describe the four evaluated configurations; the proofs and ablations provide the broader mechanism argument.

Core growth has a separate cost: more geometry categories increase the geometry error term. The fixed 360-atom core has total $R_2 = 0.046217$ at the largest tested ablation budget, compared with 0.046624 for the selected 450-atom core. This finding identifies a possible improvement to core selection, namely explicit optimization of the full radius including geometry uncertainty. It does not alter the optimality statements for the sampling objective. The unavailable heuristic contributes no fabricated comparison row.

Table 13: Core and allocation results at all tested budgets.

Model	Core	Allocation	N/D	R_2	Min. n_s	Accept.%
LGBM	Budget	Continuous+LR	256	0.09014	1	0.466
LGBM	Budget	Continuous+LR	512	0.06433	1	0.399
LGBM	Budget	Continuous+LR	1024	0.04682	1	0.332
LGBM	Budget	Integer L2	256	0.08903	1	0.568
LGBM	Budget	Integer L2	512	0.06386	1	0.489
LGBM	Budget	Integer L2	1024	0.04662	1	0.424
LGBM	Budget	Mass	256	0.09775	1	0.357
LGBM	Budget	Mass	512	0.06950	1	0.285
LGBM	Budget	Mass	1024	0.05033	1	0.234
LGBM	Budget	Pilot Neyman†	256	0.16090	8	0.086
LGBM	Budget	Pilot Neyman†	512	0.09047	8	0.077
LGBM	Budget	Pilot Neyman†	1024	0.06092	8	0.066
LGBM	Budget	Uniform	256	0.30602	10	0.068
LGBM	Budget	Uniform	512	0.24057	16	0.038
LGBM	Budget	Uniform	1024	0.18901	26	0.020
LGBM	None	Integer L2	256	0.11581	256	n.a.
LGBM	None	Integer L2	512	0.08189	512	n.a.
LGBM	None	Integer L2	1024	0.05791	1024	n.a.
LGBM	Fixed 360	Integer L2	256	0.08979	1	0.272
LGBM	Fixed 360	Integer L2	512	0.06386	1	0.489
LGBM	Fixed 360	Integer L2	1024	0.04622	1	0.754
TabM	Budget	Continuous+LR	256	0.09014	1	0.466
TabM	Budget	Continuous+LR	512	0.06433	1	0.399
TabM	Budget	Continuous+LR	1024	0.04682	1	0.332
TabM	Budget	Integer L2	256	0.08903	1	0.568
TabM	Budget	Integer L2	512	0.06386	1	0.489
TabM	Budget	Integer L2	1024	0.04662	1	0.424
TabM	Budget	Mass	256	0.09775	1	0.357
TabM	Budget	Mass	512	0.06950	1	0.285
TabM	Budget	Mass	1024	0.05033	1	0.234
TabM	Budget	Pilot Neyman†	256	0.15206	8	0.086
TabM	Budget	Pilot Neyman†	512	0.08572	8	0.077
TabM	Budget	Pilot Neyman†	1024	0.05766	8	0.066
TabM	Budget	Uniform	256	0.30602	10	0.068
TabM	Budget	Uniform	512	0.24057	16	0.038
TabM	Budget	Uniform	1024	0.18901	26	0.020
TabM	None	Integer L2	256	0.11581	256	n.a.
TabM	None	Integer L2	512	0.08189	512	n.a.
TabM	None	Integer L2	1024	0.05791	1024	n.a.
TabM	Fixed 360	Integer L2	256	0.08979	1	0.272
TabM	Fixed 360	Integer L2	512	0.06386	1	0.489
TabM	Fixed 360	Integer L2	1024	0.04622	1	0.754

Each row contains 1152 successful cells. †: the pilot is reused after adaptive allocation, so the recorded Hoeffding width is nominal.

D.5 Alternative confidence calculation

For a bounded stratum with sample variance s_s^2 and range one, the offline empirical-Bernstein calculation uses

$$b_s = \sqrt{\frac{2s_s^2 \log(4/\eta_s)}{n_s}} + \frac{7\log(4/\eta_s)}{3(n_s-1)}, \qquad \eta_s = \delta_s/|\mathcal{S}|.$$

Its directional radius is $\sum_s \widehat{W}_{sj} b_s + \varepsilon_j/2$. The form follows finite-sample empirical Bernstein concentration (Maurer & Pontil, 2009), with an explicit simultaneous allocation over strata. It requires $n_s \geq 2$ in every contributing stratum and an applicable independent-sample argument.

Only 20,736 of 48,384 recorded cells have a finite calculated radius, or 42.86%. The median over finite calculations is 1.08136, and no coverage evaluation accompanies them. Table 14 retains the failures to compute a radius. The no-sharing rows do improve on their matched Hoeffding radii, because their per-direction sample counts support variance estimation. The finely partitioned geometry-only allocations instead contain single-observation strata, while uniform and pilot-based allocations pay large range corrections. Pilot-reused rows have the additional dependence issue described in Appendix C.3. The vector confidence-sequence engine is unavailable; no sequential-coverage or early-stopping result is claimed.

Table 14: Finite empirical-Bernstein radii in the offline calculation.

Model	Core	Allocation	N/D	Finite/total	EB R_2	H R_2
LGBM	Budget	Continuous+LR	256	0/1152	n.a.	0.09014
LGBM	Budget	Continuous+LR	512	0/1152	n.a.	0.06433
LGBM	Budget	Continuous+LR	1024	0/1152	n.a.	0.04682
LGBM	Budget	Integer L2	256	0/1152	n.a.	0.08903
LGBM	Budget	Integer L2	512	0/1152	n.a.	0.06386
LGBM	Budget	Integer L2	1024	0/1152	n.a.	0.04662
LGBM	Budget	Mass	256	0/1152	n.a.	0.09775
LGBM	Budget	Mass	512	0/1152	n.a.	0.06950
LGBM	Budget	Mass	1024	0/1152	n.a.	0.05033
LGBM	Budget	Pilot Neyman†	256	1152/1152	1.75737	0.16090
LGBM	Budget	Pilot Neyman†	512	1152/1152	1.20528	0.09047
LGBM	Budget	Pilot Neyman†	1024	1152/1152	0.91235	0.06092
LGBM	Budget	Uniform	256	1152/1152	2.75407	0.30602
LGBM	Budget	Uniform	512	1152/1152	1.73819	0.24057
LGBM	Budget	Uniform	1024	1152/1152	1.09231	0.18901
LGBM	None	Integer L2	256	1152/1152	0.09008	0.11581
LGBM	None	Integer L2	512	1152/1152	0.04928	0.08189
LGBM	None	Integer L2	1024	1152/1152	0.02776	0.05791
LGBM	Fixed 360	Integer L2	256	0/1152	n.a.	0.08979
LGBM	Fixed 360	Integer L2	512	0/1152	n.a.	0.06386
LGBM	Fixed 360	Integer L2	1024	0/1152	n.a.	0.04622
TabM	Budget	Continuous+LR	256	0/1152	n.a.	0.09014
TabM	Budget	Continuous+LR	512	0/1152	n.a.	0.06433
TabM	Budget	Continuous+LR	1024	0/1152	n.a.	0.04682
TabM	Budget	Integer L2	256	0/1152	n.a.	0.08903
TabM	Budget	Integer L2	512	0/1152	n.a.	0.06386
TabM	Budget	Integer L2	1024	0/1152	n.a.	0.04662
TabM	Budget	Mass	256	0/1152	n.a.	0.09775
TabM	Budget	Mass	512	0/1152	n.a.	0.06950
TabM	Budget	Mass	1024	0/1152	n.a.	0.05033
TabM	Budget	Pilot Neyman†	256	1152/1152	1.66447	0.15206
TabM	Budget	Pilot Neyman†	512	1152/1152	1.15092	0.08572
TabM	Budget	Pilot Neyman†	1024	1152/1152	0.84475	0.05766
TabM	Budget	Uniform	256	1152/1152	2.74406	0.30602
TabM	Budget	Uniform	512	1152/1152	1.72902	0.24057
TabM	Budget	Uniform	1024	1152/1152	1.08557	0.18901
TabM	None	Integer L2	256	1152/1152	0.08573	0.11581
TabM	None	Integer L2	512	1152/1152	0.04621	0.08189
TabM	None	Integer L2	1024	1152/1152	0.02553	0.05791
TabM	Fixed 360	Integer L2	256	0/1152	n.a.	0.08979
TabM	Fixed 360	Integer L2	512	0/1152	n.a.	0.06386
TabM	Fixed 360	Integer L2	1024	0/1152	n.a.	0.04622

Finite means the formula is computable, with at least two observations in every required stratum. It does not assert empirical coverage or validate pilot reuse. H denotes the matched Hoeffding calculation.

D.6 GEOMETRY PHASE AND PRACTICAL SCALABILITY

The sweep evaluates $D \in \{8, 12, 16, 24, 32, 64\}$, $r \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, and $L \in \{0.30, 0.45, 0.60, 0.75\}$, giving 120 configurations. The full-incidence continuous saving proxy is computed as $100[1 - (C_{\text{full}}/D)^2]$. It has Spearman correlation 0.999 with mean union multiplicity across this grid. The inequality in Appendix B.6 predicts an opportunity for saving; it does not equate this proxy with the final certified query crossing.

Table 15 exposes the computational side of that opportunity. All configurations complete, the largest selected core has 17,717 atoms, and the minimum selected design acceptance reaches approximately 1.53×10^{-5}. The slowest recorded configuration takes 4317.88 seconds. The $D = 64$ accelerated path has a maximum of 133.56 seconds; the difference in backend prevents interpreting this as a favorable scaling comparison against the CPU path. Geometry and rejection costs should therefore be considered alongside the accepted-query objective.

Table 15: Full geometry sweep and computational costs.

D	r	L	Shared	Core	$\bar{m}$	Proxy%	Accept.%	Time (s)
8	0.05	0.30	163	118	1.264	26.60	0.101	1.9
8	0.05	0.45	165	105	1.162	18.23	0.092	1.9
8	0.05	0.60	164	88	1.117	13.84	0.114	2.5
8	0.05	0.75	160	84	1.091	11.17	0.104	2.0
8	0.10	0.30	166	132	1.719	49.14	0.158	2.2
8	0.10	0.45	164	125	1.386	34.55	0.116	2.3
8	0.10	0.60	163	115	1.264	26.60	0.103	2.0
8	0.10	0.75	164	113	1.201	21.62	0.091	2.1
8	0.15	0.30	167	143	2.619	66.93	0.189	2.0
8	0.15	0.45	167	132	1.717	49.10	0.148	2.1
8	0.15	0.60	164	129	1.457	38.38	0.122	2.1
8	0.15	0.75	163	125	1.335	31.46	0.104	2.0
8	0.20	0.30	167	149	3.796	76.45	0.239	2.0
8	0.20	0.45	167	140	2.248	61.69	0.180	2.1
8	0.20	0.60	167	133	1.717	49.08	0.154	2.0
8	0.20	0.75	165	129	1.502	40.62	0.139	2.0
8	0.25	0.30	201	179	4.545	80.30	0.258	2.3
8	0.25	0.45	168	143	3.035	71.11	0.210	2.0
8	0.25	0.60	165	138	2.091	58.78	0.173	2.0
8	0.25	0.75	166	132	1.718	49.13	0.159	2.0
12	0.05	0.30	692	308	1.360	33.71	0.064	10.1
12	0.05	0.45	684	249	1.215	23.28	0.065	9.5
12	0.05	0.60	661	230	1.153	17.75	0.060	9.6
12	0.05	0.75	663	199	1.119	14.36	0.055	9.8
12	0.10	0.30	697	423	2.114	60.35	0.060	10.3
12	0.10	0.45	691	350	1.546	43.43	0.061	10.5
12	0.10	0.60	682	306	1.361	33.80	0.064	9.9
12	0.10	0.75	673	272	1.269	27.58	0.065	9.6
12	0.15	0.30	703	484	3.747	77.15	0.070	11.9
12	0.15	0.45	694	416	2.115	60.38	0.064	10.4
12	0.15	0.60	691	365	1.659	48.02	0.062	9.8
12	0.15	0.75	685	333	1.467	39.71	0.062	9.6
12	0.20	0.30	750	559	5.322	83.60	0.079	10.0
12	0.20	0.45	699	473	3.121	72.99	0.064	9.4
12	0.20	0.60	697	418	2.114	60.36	0.063	9.6
12	0.20	0.75	695	370	1.735	50.64	0.066	9.0
12	0.25	0.30	951	701	6.499	86.48	0.080	12.7
12	0.25	0.45	712	519	4.345	80.05	0.071	9.4
12	0.25	0.60	700	468	2.828	70.38	0.063	9.6
12	0.25	0.75	695	419	2.111	60.31	0.060	9.1
16	0.05	0.30	1710	567	1.454	39.46	0.043	35.8
16	0.05	0.45	1685	451	1.263	27.38	0.045	36.9
16	0.05	0.60	1655	357	1.185	20.96	0.050	36.0
16	0.05	0.75	1609	295	1.143	16.99	0.051	31.6

Continued on next page.

Table 15 continued.

D	r	L	Shared	Core	$\bar{m}$	Proxy%	Accept.%	Time (s)
16	0.10	0.30	1757	886	2.588	68.36	0.038	41.6
16	0.10	0.45	1736	675	1.711	50.37	0.042	40.4
16	0.10	0.60	1709	550	1.454	39.44	0.046	39.1
16	0.10	0.75	1695	470	1.333	32.34	0.050	38.4
16	0.15	0.30	1811	1068	4.852	82.48	0.036	39.4
16	0.15	0.45	1765	844	2.588	68.35	0.041	42.0
16	0.15	0.60	1741	733	1.874	55.35	0.039	36.8
16	0.15	0.75	1725	621	1.598	46.13	0.044	35.4
16	0.20	0.30	2001	1296	6.887	87.47	0.043	45.8
16	0.20	0.45	1787	1033	4.080	79.40	0.034	42.4
16	0.20	0.60	1763	872	2.589	68.36	0.039	41.6
16	0.20	0.75	1756	761	1.987	58.20	0.038	42.7
16	0.25	0.30	2603	1575	8.467	89.73	0.039	75.0
16	0.25	0.45	1829	1138	5.560	84.63	0.037	45.7
16	0.25	0.60	1776	976	3.680	77.35	0.037	37.8
16	0.25	0.75	1771	863	2.589	68.37	0.039	42.6
24	0.05	0.30	5995	1227	1.635	48.18	0.029	415.1
24	0.05	0.45	5780	851	1.349	33.85	0.034	384.0
24	0.05	0.60	5641	712	1.241	26.02	0.033	328.9
24	0.05	0.75	5456	566	1.184	21.15	0.039	336.4
24	0.10	0.30	6291	2241	3.705	78.16	0.021	476.5
24	0.10	0.45	6099	1570	2.060	60.48	0.025	440.2
24	0.10	0.60	5967	1224	1.635	48.14	0.029	383.2
24	0.10	0.75	5891	1036	1.451	39.78	0.031	403.8
24	0.15	0.30	6569	3001	6.991	88.00	0.020	471.7
24	0.15	0.45	6272	2208	3.705	78.17	0.021	428.8
24	0.15	0.60	6173	1681	2.359	65.92	0.025	450.6
24	0.15	0.75	6040	1447	1.868	55.81	0.026	426.6
24	0.20	0.30	7661	3826	10.001	91.50	0.020	589.5
24	0.20	0.45	6432	2766	5.893	85.89	0.019	489.3
24	0.20	0.60	6265	2218	3.706	78.17	0.021	456.0
24	0.20	0.75	6193	1802	2.576	68.92	0.023	442.4
24	0.25	0.30	10026	4528	12.241	93.02	0.016	1042.5
24	0.25	0.45	6784	3250	8.057	89.52	0.021	533.4
24	0.25	0.60	6396	2629	5.351	84.53	0.020	442.7
24	0.25	0.75	6298	2198	3.704	78.16	0.021	483.2
32	0.05	0.30	13877	1958	1.816	54.71	0.025	2580.2
32	0.05	0.45	13286	1375	1.430	38.95	0.030	1938.6
32	0.05	0.60	12743	1056	1.291	30.10	0.034	2035.9
32	0.05	0.75	12336	833	1.220	24.48	0.040	1857.7
32	0.10	0.30	14745	3980	4.871	83.41	0.016	2616.4
32	0.10	0.45	14311	2722	2.462	67.84	0.019	2810.4
32	0.10	0.60	13869	1995	1.817	54.73	0.024	2626.6
32	0.10	0.75	13561	1613	1.564	45.57	0.027	2459.0
32	0.15	0.30	15904	5798	9.174	90.91	0.014	3751.8
32	0.15	0.45	14801	4012	4.868	83.40	0.016	2758.7
32	0.15	0.60	14409	2999	2.940	73.19	0.018	2957.8
32	0.15	0.75	14113	2345	2.159	62.90	0.022	2540.9
32	0.20	0.30	18909	7166	13.082	93.56	0.013	3812.7
32	0.20	0.45	15368	5105	7.703	89.27	0.014	3470.0
32	0.20	0.60	14778	3978	4.873	83.42	0.016	3088.3
32	0.20	0.75	14515	3244	3.281	75.92	0.018	3015.8
32	0.25	0.30	24368	7810	16.077	94.73	0.011	4317.9
32	0.25	0.45	16535	6370	10.586	92.08	0.014	3701.5
32	0.25	0.60	15188	4853	6.976	88.21	0.014	3347.1
32	0.25	0.75	14826	4001	4.871	83.41	0.015	2845.3
64	0.05	0.30	92504	6125	2.717	71.68	0.017	70.5
64	0.05	0.45	85194	3753	1.738	52.91	0.024	49.3
64	0.05	0.60	80046	2802	1.467	41.53	0.029	59.9
64	0.05	0.75	75759	2210	1.342	34.11	0.035	44.1
64	0.10	0.30	104779	13497	9.421	91.48	0.010	92.3

Continued on next page.

Table 15 continued.

D	r	L	Shared	Core	$\bar{m}$	Proxy%	Accept.%	Time (s)
64	0.10	0.45	97448	8647	4.564	82.98	0.013	80.8
64	0.10	0.60	92597	6132	2.716	71.66	0.017	52.8
64	0.10	0.75	88554	4570	2.037	61.10	0.021	76.3
64	0.15	0.30	119755	17285	17.792	95.39	0.008	133.6
64	0.15	0.45	104914	13298	9.427	91.49	0.010	98.0
64	0.15	0.60	99418	9882	5.676	86.16	0.011	62.8
64	0.15	0.75	95776	7517	3.741	79.42	0.015	85.1
64	0.20	0.30	144217	17243	25.435	96.74	0.007	49.4
64	0.20	0.45	113666	16439	15.006	94.56	0.009	53.3
64	0.20	0.60	104868	13537	9.420	91.48	0.010	55.3
64	0.20	0.75	100453	10564	6.375	87.62	0.011	57.0
64	0.25	0.30	180330	14836	31.189	97.33	0.005	51.3
64	0.25	0.45	126659	17717	20.532	95.98	0.008	54.6
64	0.25	0.60	111425	16050	13.601	94.01	0.009	55.7
64	0.25	0.75	104571	13308	9.426	91.49	0.010	56.2

Proxy is the ideal full-incidence saving $100[1-(C_{\text{full}}/D)^2]$, computed from the recorded C_{full}. Acceptance is the median selected-atom design acceptance. The $D = 64$ implementation uses an accelerator; its timings are not directly comparable with the CPU timings.

When query savings reduce total computation. The query objective counts accepted black-box evaluations. To relate it to elapsed time, consider A audits reusing a fixed canonical tube geometry, retained partition, and geometry calibration. For method m, write

$$T_m(A) = S_m + \sum_{a=1}^{A} \big[H_{m,a} + F_{m,a}(N_{m,a})\big]. \tag{D.1}$$

Here S_m is reusable setup cost; $H_{m,a}$ comprises per-audit work other than predictor evaluation, including any chart preparation, allocation, proposals, membership tests, rejection, and reconstruction; and $F_{m,a}(N_{m,a})$ is the time spent evaluating $N_{m,a}$ accepted inputs. These terms partition the executed work, so cached operations are counted once. Reuse requires the same canonical regions and calibrated categories. A change requiring new geometry or partition construction contributes additional setup work. Each fixed audit retains its own stated confidence guarantee; a joint guarantee over several audits requires allocating the sampling error probability across them.

For a matched valid certificate target, let $\Delta N_a = N_{\text{IMC},a} - N_{\text{KS},a}$, $\Delta S = S_{\text{KS}} - S_{\text{IMC}}$, and $\Delta H_a = H_{\text{KS},a} - H_{\text{IMC},a}$. If comparable inference configurations make predictor time approximately linear with a common per-query cost $c_{f,a}$ within audit a, then

$$T_{\text{KS}}(A) - T_{\text{IMC}}(A) \approx \Delta S + \sum_{a=1}^{A} \Delta H_a - \sum_{a=1}^{A} c_{f,a} \Delta N_a. \tag{D.2}$$

Under this approximation, query savings reduce total time when $\sum_a c_{f,a}\Delta N_a > \Delta S + \sum_a \Delta H_a$. For homogeneous audits this becomes $c_f \Delta N > \Delta S/A + \Delta H$: reuse amortizes setup, and more expensive predictor calls make each avoided query more valuable. Rejection cost remains relevant through ΔH and can prevent completion under a proposal cap.

The approximation requires comparable models, precision, batching, and inference backends. For nonlinear or method-dependent inference time, Eq. D.1 applies with the measured functions $F_{m,a}$ directly. The common-95% crossing of 340.5 queries per direction (and the equal-δ_s crossing of 304.8) is interpolated, not an executed integer query count or a wall-time measurement. Table 3 measures equal-query audit overhead; Table 15 measures geometry feasibility and cost; Table 19 measures warm inference throughput. These measurements expose the relevant costs but do not separately identify every term above at a matched certificate target. Equations D.1–D.2 give an accounting criterion for such a comparison, without asserting an unmeasured elapsed-time speedup.

D.7 Loss, calibration, and input robustness

The stress matrix contains 8064 successful KS cells paired with IMC, with 193,536 directional records across both methods. Table 16 gives all 22 model-condition groups. It includes 3072 loss cells, 1280 calibration cells, and 3712 input-stress cells. Binary isotonic calibration has eight eligible tasks; shifted-anchor groups have fewer available anchors. Within these named supports, all groups reduce both pooled RMSE and MAE. Median MAE reductions across groups are 30.21% for loss, 28.96% for calibration, and 28.70% for input stress. Across all groups the range is 23.98% to 33.04%.

The smallest cell-containment rate, 94.79%, occurs within the loss matrix, while the smallest directional rate is 99.57%. No group has disjoint audit and reference intervals. Since the reference itself has a nonzero interval and groups are correlated, these containment percentages are not independent binomial tests of the theorem's coverage. The 253 improved dataset-model-condition RMSE comparisons and three zero-error ties in Figure 5 summarize a different aggregation level from the pooled group reductions below.

Table 16: Paired error and reference-containment results under stress.

Condition	Model	Cells	RMSE	RMSE ↓%	MAE ↓%	Dir. Ref-C%	Cell Ref-C%
Isotonic	LGBM	256	0.00339	29.47	31.32	100.00	100.00
Isotonic	TabM	256	0.00443	22.56	26.60	100.00	100.00
Temperature	LGBM	384	0.00478	27.75	31.55	99.91	98.96
Temperature	TabM	384	0.00427	24.08	26.42	99.96	99.48
PC top 20%	LGBM	368	0.00639	26.30	29.28	99.82	97.83
PC top 20%	TabM	368	0.00470	22.79	25.78	99.93	99.18
Feature scale 10×	LGBM	384	0.00572	28.22	31.15	99.89	98.70
Feature scale 10×	TabM	384	0.00502	21.06	24.27	99.98	99.74
Label noise 10%	LGBM	384	0.00522	30.15	31.61	99.96	99.48
Label noise 10%	TabM	384	0.00401	20.67	23.98	99.98	99.74
MCAR 10%	LGBM	384	0.00526	30.17	33.04	99.98	99.74
MCAR 10%	TabM	384	0.00489	27.11	28.12	99.96	99.48
Mahalanobis top 5%	LGBM	336	0.00563	29.45	32.59	99.78	97.32
Mahalanobis top 5%	TabM	336	0.00527	22.12	25.46	99.95	99.40
Clipped CE	LGBM	384	0.00305	27.11	31.43	99.96	99.48
Clipped CE	TabM	384	0.00226	27.97	29.55	100.00	100.00
Brier	LGBM	384	0.00597	28.66	31.73	99.89	98.70
Brier	TabM	384	0.00506	23.45	26.64	99.98	99.74
$1 - p_y$	LGBM	384	0.00569	28.18	31.21	99.89	98.70
$1 - p_y$	TabM	384	0.00476	22.65	26.21	99.96	99.48
Zero-one	LGBM	384	0.00817	29.52	30.87	99.70	96.61
Zero-one	TabM	384	0.00877	26.30	27.42	99.57	94.79

RMSE and reference-containment (Ref-C) columns show KS. Reductions pair KS with IMC. Every displayed group has zero falsified intervals; all successful KS cells have $R_2 = 0.06386048$.

D.8 Model availability, charts, and data splits

Table 17: Execution completion and failed samplers at each budget.

Model	Method	N/D	Successful/total	R_2
ModernNCA	Unavailable	n.a.	0 models	n.a.
TabFM	Unavailable	n.a.	0 models	n.a.
TabICLv2	FullAtoms	256	0/88	n.a.
TabICLv2	FullAtoms	512	0/88	n.a.
TabICLv2	FullAtoms	1024	0/88	n.a.
TabICLv2	IMC	256	74/88	0.11581
TabICLv2	IMC	512	74/88	0.08189
TabICLv2	IMC	1024	74/88	0.05791

Continued on next page.

Table 17 continued.

Model	Method	N/D	Successful/total	R_2
TabICLv2	KS	256	74/88	0.08903
TabICLv2	KS	512	74/88	0.06386
TabICLv2	KS	1024	74/88	0.04662
TabPFN-3	FullAtoms	256	0/96	n.a.
TabPFN-3	FullAtoms	512	0/96	n.a.
TabPFN-3	FullAtoms	1024	0/96	n.a.
TabPFN-3	IMC	256	96/96	0.11581
TabPFN-3	IMC	512	96/96	0.08189
TabPFN-3	IMC	1024	96/96	0.05791
TabPFN-3	KS	256	96/96	0.08903
TabPFN-3	KS	512	96/96	0.06386
TabPFN-3	KS	1024	96/96	0.04662
ResNet20	FullAtoms	128	0/64	n.a.
ResNet20	FullAtoms	256	0/64	n.a.
ResNet20	FullAtoms	512	0/64	n.a.
ResNet20	FullAtoms	1024	0/64	n.a.
ResNet20	IMC	128	64/64	0.16378
ResNet20	IMC	256	64/64	0.11581
ResNet20	IMC	512	64/64	0.08189
ResNet20	IMC	1024	64/64	0.05791
ResNet20	KS	128	64/64	0.12620
ResNet20	KS	256	64/64	0.08903
ResNet20	KS	512	64/64	0.06386
ResNet20	KS	1024	64/64	0.04662
ViT-B/16	FullAtoms	128	0/64	n.a.
ViT-B/16	FullAtoms	256	0/64	n.a.
ViT-B/16	FullAtoms	512	0/64	n.a.
ViT-B/16	FullAtoms	1024	0/64	n.a.
ViT-B/16	IMC	128	64/64	0.16378
ViT-B/16	IMC	256	64/64	0.11581
ViT-B/16	IMC	512	64/64	0.08189
ViT-B/16	IMC	1024	64/64	0.05791
ViT-B/16	KS	128	64/64	0.12620
ViT-B/16	KS	256	64/64	0.08903
ViT-B/16	KS	512	64/64	0.06386
ViT-B/16	KS	1024	64/64	0.04662

TabFM and ModernNCA fail at the model/dependency level, so zero cells is not successful completion. TabICLv2 has 14 numerical failures per method and budget among 88 attempted cells on 11 datasets. All FullAtoms failures are numerically inaccessible strata. The main 512-budget comparison has 1856 successful KS cells out of 1920 planned: 480 each for LightGBM, logistic regression, and RealMLP, and 416 for TabM. This is 26 of 30 tasks for TabM. Broad comparator summaries that use a different feasible support do not fill these missing KS results. For the four frontier model families, only TabPFN-3 and TabICLv2 execute. ModernNCA and TabFM remain unavailable, so the intended three-of-four model feasibility threshold is unmet. Both successful frontier families and both vision checkpoints retain all FullAtoms failures in Table 17.

The random-orthogonal tabular chart control contains 2304 successful KS cells and the fixed-DCT vision control contains 128. Their same-budget radii have ratio one relative to the PCA-chart results, as expected from the unchanged geometry and allocation. The split control contains 2560 paired KS cells on ten tasks over folds 1 and 2; Table 18 reports completion, radii, and elapsed time. These tests support execution under the stated chart and split changes. Equal widths alone do not show that the losses or estimation errors are unchanged.

Table 18: KS results on two additional dataset folds.

Model	Fold	N/D	Successful/total	R_2	Time (s)
LGBM	1	128	160/160	0.12620	0.15

Continued on next page.

Table 18 continued.

Model	Fold	N/D	Successful/total	R_2	Time (s)
LGBM	1	256	160/160	0.08903	0.28
LGBM	1	512	160/160	0.06386	0.54
LGBM	1	1024	160/160	0.04662	1.02
LGBM	2	128	160/160	0.12620	0.16
LGBM	2	256	160/160	0.08903	0.30
LGBM	2	512	160/160	0.06386	0.57
LGBM	2	1024	160/160	0.04662	1.06
TabM	1	128	160/160	0.12620	0.10
TabM	1	256	160/160	0.08903	0.11
TabM	1	512	160/160	0.06386	0.21
TabM	1	1024	160/160	0.04662	0.34
TabM	2	128	160/160	0.12620	0.08
TabM	2	256	160/160	0.08903	0.10
TabM	2	512	160/160	0.06386	0.19
TabM	2	1024	160/160	0.04662	0.25

D.9 Inference cost and reproducible continuation

The system benchmark uses an NVIDIA RTX 4090 with 24 GB memory and requested batches of 256, 1024, 4096, and 16,384. Table 19 reports actual batches so that memory fallback is visible. A separate check compares BF16 and FP32 on 2048 identical inputs at batch size 1024. The maximum absolute difference across inputs and class probabilities is 0.053680 for ResNet20 and 0.000865 for ViT, exceeding the 10^{-4} precision check; main vision audits therefore use FP32. The largest FP32 memory measurements are 4.1965 GiB for ResNet20 and 7.832 GiB for ViT. Same-budget audit overhead and warm inference throughput answer different cost questions and are reported separately.

Table 19: Vision inference throughput and precision checks.

Model	Dtype	Requested	Actual	Queries/s	VRAM (GiB)	BF16/FP32 max. drift
ResNet20	bf16	256	256	32741.6	0.043	0.053680
ResNet20	bf16	1024	1024	94541.9	0.146	0.053680
ResNet20	bf16	4096	4096	68275.5	0.556	0.053680
ResNet20	bf16	16384	16384	53979.2	2.197	0.053680
ResNet20	fp32	256	256	59876.3	0.074	0.053680
ResNet20	fp32	1024	1024	67815.3	0.271	0.053680
ResNet20	fp32	4096	4096	61549.1	1.056	0.053680
ResNet20	fp32	16384	16384	44748.1	4.197	0.053680
ViT-B/16	bf16	256	256	674.7	1.629	0.000865
ViT-B/16	bf16	1024	1024	697.5	5.523	0.000865
ViT-B/16	bf16	4096	1024	691.2	5.523	0.000865
ViT-B/16	bf16	16384	1024	708.0	5.523	0.000865
ViT-B/16	fp32	256	256	289.9	2.207	0.000865
ViT-B/16	fp32	1024	1024	328.6	7.832	0.000865
ViT-B/16	fp32	4096	1024	408.5	7.832	0.000865
ViT-B/16	fp32	16384	1024	446.7	7.832	0.000865

Drift is one separate comparison per model on 2048 matched inputs at batch size 1024, maximized over inputs and class probabilities. The same model-level value is repeated across dtype and throughput-batch rows.

The continuation test interrupts after five committed chunks out of sixteen, with 2048 queries per chunk. Coordinates and seeds are preserved, and the resumed aggregate hash equals the uninterrupted aggregate hash. This verifies the tested interruption point and deterministic aggregation path. It does not establish numerical equivalence across hardware, precision modes, or unspecified checkpoint revisions. The result is useful operationally because expensive query evaluations can be committed without rerunning completed chunks.