MODE COLLAPSE IS CHEAP TO DETECT: A GROUND-TRUTH-FREE PRE-FLIGHT CHECK FOR NEURAL SAMPLERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural samplers are trained against an unnormalised target $\tilde{\pi} = e^{-E}$ with no samples from π, so a practitioner cannot tell whether an expensive run has silently dropped part of the target. The usual diagnostics are expectations under the model and are therefore confined to its support: we exhibit a sampler with self-normalised $\mathrm{ESS} = 0.99$ that misses 87% of the mass. Splitting the missing mass into lost support and misallocated weight shows the two established diagnostic families have *disjoint* blind spots – ESS is blind to the first, and normaliser comparison is provably blind to the second at any budget, returning 0.0000 while up to 64% of the mass is misallocated – while the sample-only kernel Stein discrepancy separates neither. We then argue that *detecting* missing mass is easier than sampling it – detection needs one point per missed basin plus local curvature – and turn this into a pre-flight check using only $E, \nabla E, \nabla^2 E$. On exactly solvable targets the check reaches 10^{-3}-level accuracy at a few percent of training cost on the harder benchmarks, while a tuned SMC reference requires 70–280% of training cost, and it applies unchanged to a controlled-SDE sampler with no tractable density. A self-diagnostic flags configurations whose estimate should not be believed, but cannot certify a run: a mode with attraction probability below $1/R$ escapes the estimator and the diagnostic together, and a stable search says nothing about the local-mass bias. We map the boundary on LJ-13, and report that on these benchmarks a mixture read off the check is itself the better mode-level sampler.

1 INTRODUCTION

A neural sampler is trained to draw from $\pi(x) \propto \tilde{\pi}(x) = e^{-E(x)}$ given only the ability to evaluate E and its derivatives (Zhang & Chen, 2022; Albergo & Vanden-Eijnden, 2025; Havens et al., 2025; Liu et al., 2025; Guo et al., 2026; Havens et al., 2026; Castro-Macías et al., 2026). The defining feature of the setting is the absence of samples from π, and that absence propagates into evaluation: the quantities practitioners monitor – the ELBO, the self-normalised effective sample size (ESS), importance-weight variance – are expectations under the model. Blessing et al. (2024) state the consequence plainly: such protocols "rely on samples from the model, restricting their evaluation capabilities to the model's support," and this "becomes especially problematic when assessing the ability to mitigate mode collapse." Wu et al. (2025) report the same gap from the other side, noting that their method "relies on oracle metrics for hyperparameter tuning" and that automating this "remains an important future direction." A practitioner who has just spent a large compute budget training a sampler has, at present, no cheap way to learn whether the result is trustworthy.

This paper makes a narrow claim and tests it hard. The claim is an asymmetry:

> *Detecting that a sampler is missing mass is cheaper than fixing it.* Detection requires only one point inside each missed basin together with a local curvature estimate; correcting the learned sampler generally requires substantially more computation, often including retraining.

If the asymmetry holds, a pre-flight check should exist – spend a few percent of the training budget, and cheaply obtain evidence about whether the run dropped part of the target. We construct such a

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.

check (§4), equip it with a diagnostic that reports when its own answer should not be believed (§6), and evaluate it against a tuned annealed SMC reference at matched cost (§7). Because a method of this kind is only as good as its stated scope, we map where it breaks on the 13-atom Lennard-Jones cluster (Appendix P).

Contributions.

- **The established diagnostics have blind spots that do not overlap.** Separating the missing mass into support loss and weight misallocation, we prove that ESS is blind to the first (Propositions 1–3) while normaliser comparison is blind to the second (Corollary 1), and exhibit both failures in one architecture on one target (§2, §8). The sample-only third option, kernel Stein discrepancy, separates neither on these targets, for the reason Bénard et al. (2023) identify (Appendix K).
- **A threshold-free estimator of the missing mass** $M = \sum_j (p_j - q_j)_+$, built from a multi-start second-order mode search, local harmonic masses and a Mahalanobis coverage assignment. It uses no samples from π and no estimate of the normalising constant, so it applies to samplers with no tractable density (§4).
- **A self-diagnostic that reports when its own answer should be disbelieved**, computable without ground truth. It flags evidence that the estimate should not be believed, and never certifies a cleared one – necessary because the estimator carries no fixed error sign under incomplete enumeration (Proposition 6) – and we measure its residual false-clear rate both in sample and on target families that had no part in setting its thresholds (§6, Appendix N).
- **A measured boundary rather than an asserted one.** The check costs 2.7% of training at $d{=}16$ and grows linearly in d; on LJ-13 the search recovers the global minimum exactly, but continuous symmetry makes the naive local mass undefined and the enumeration is sufficient only at low temperature (§7, Appendix P).

2 Background: what the usual diagnostics can and cannot see

Let q be a trained sampler with tractable density. Importance weights $w(x) = \tilde{\pi}(x)/q(x)$ give the self-normalised ESS rate $\widehat{\mathrm{ESS}} = (\sum_i w_i)^2/(n \sum_i w_i^2)$, universally reported as a health check. Its failure mode is structural rather than statistical. Suppose $\pi = \sum_k \omega_k \mathcal{N}(\mu_k, \sigma_k^2 I)$ with well-separated components and $q = \mathcal{N}(\mu_1, \sigma_1^2 I)$ exactly. On the support of q all other components are negligible, so $w(x) \approx Z\,\omega_1$ is *constant*: the weights are perfectly behaved and $\widehat{\mathrm{ESS}} \to 1$, while the missing mass is $1 - \omega_1$.

We verify this numerically. With $d = 16$, $K = 8$ and $\omega_1 = 0.1289$ the analytic control gives $\widehat{\mathrm{ESS}} \in \{0.979, 0.987, 0.996\}$ across runs with a true missing mass of 0.871. The same phenomenon occurs for genuinely trained flows and not only for the contrived control: reverse-KL-trained RealNVP flows that populate 2 of 8 components attain $\widehat{\mathrm{ESS}} \in \{0.907, 0.927, 0.965\}$ while missing 78–87% of the mass. ESS is high precisely because q fits a *subset* well; the badly trained sampler that spreads mass into low-density regions is the one ESS catches. The dangerous case is the one it does not.

The sample-only alternative is not a way out. Kernel Stein discrepancy needs only the score of π and draws from q (Gorham & Mackey, 2015; 2017), so unlike ESS it survives a sampler with no tractable density. It carries a pathology of exactly this paper's shape: Bénard et al. (2023) show the score is insensitive to the relative weights of well-separated modes, so a sample right *within* each basin and wrong *across* basins goes unpenalised. Measured on the constructed families of §8 with the kernel Gorham & Mackey (2017) recommend (Appendix K), a calibrated level-0.05 test rejects $q = \pi$ for 0–20% of draw sets on every alternative we construct, up to a missing mass of 0.79, across sample sizes 2000 and 8000 and kernel scales spanning an order of magnitude. The same test rejects every time on a control whose missing mass is *zero* and whose only defect is a 25% within-basin width error. It sees what the score carries and not what we are after.

3 RELATED WORK

The gap we are filling is stated by the evaluation literature itself. Blessing et al. (2024) build a standardised benchmark for sampling methods and observe that model-support-restricted protocols are precisely the wrong instrument for assessing mode collapse; Wu et al. (2025) record the same problem operationally, noting a dependence on oracle metrics for hyperparameter tuning. He et al. (2025) supply the empirical phenomenon: simulation-free neural samplers collapse without Langevin preconditioning, and do so while their training objectives look healthy. Our contribution is the missing instrument, not a new sampler.

The estimator is imported, not invented. Decomposing configuration space into basins around local minima and summing local contributions is the inherent-structure formalism of Stillinger & Weber (1984), and summing harmonic basin partition functions over enumerated minima is the harmonic superposition approximation used throughout the energy-landscape literature (Wales & Doye, 1997; Wales et al., 1998). We take three things from that field essentially unchanged: the enumeration by multi-start minimisation, the treatment of the zero modes generated by continuous symmetry, and – importantly – its known failure mode. The harmonic approximation to a basin free energy is understood there to be reasonable at low temperature and to break down above a system-dependent threshold, which is exactly the boundary our LJ-13 measurement recovers from the outside (Appendix P). Disconnectivity graphs (Becker & Karplus, 1997; Wales et al., 1998) visualise the same decomposition. What appears to be new here is the use: applying the construction not to predict a physical partition function but to audit a learned sampler, and pairing it with a statistic that reports when the enumeration underlying it is too incomplete to be believed.

A fuller account – why two-sample coverage metrics do not transfer, the methods that attack collapse rather than measure it, and the variational families this audit can be pointed at – is deferred to Appendix E.

4 A PRE-FLIGHT CHECK

Target quantity. Coverage is not binary: a sampler may place 1% of its mass on a mode holding 40% of π. Let the target decompose into basins with normalised weights p_j under π and q_j under q. We estimate

$$M \;=\; \sum_{j} \big(p_j - q_j\big)_+, \tag{1}$$

the mass of π that q under-weights. It is continuous, threshold-free, and controls the error of any mode-level expectation. It is useful to split it according to *how* the mass went missing:

$$M \;=\; \underbrace{\sum_{j:\, q_j = 0} p_j}_{M_{\mathrm{supp}}} \;+\; \underbrace{\sum_{j:\, q_j > 0} \big(p_j - q_j\big)_+}_{M_{\mathrm{wt}}}. \tag{2}$$

M_{supp} is collapse in the usual sense, basins the sampler never enters, and is what "mode collapse" normally denotes; M_{wt} is collapse *without* support loss, where every basin is reached but the proportions among them are wrong. §5 shows the two established diagnostic families each see exactly one of these terms.

For continuous samplers with formally full support the first term needs an operational reading: a flow with a Gaussian base assigns $q_j > 0$ to every basin, so $\{j : q_j = 0\}$ is empty as a matter of measure theory even when a basin receives no draw in 4×10^4 samples. We read M_{supp} at the resolution of the audit sample, treating a basin of negligible occupancy as unvisited; Proposition 2 makes the consequence for ESS precise. The total M, our primary estimand, is unaffected by this convention, and the estimator is in any case indifferent to the split: it compares a mass budget against an occupancy and never asks which kind of discrepancy produced the gap.

Step 1: mode search. We enumerate candidate maxima of $\log \tilde{\pi}$ by multi-start optimisation from a broad initialisation, using Adam with cosine decay followed by Levenberg–Marquardt Newton refinement. The second-order stage is not a convenience: a fixed-step ascent diverges outright on

a quartic target, and on an anisotropic one Adam alone leaves siblings of the *same* basin ~ 18 units apart, many at saddle points, which destroys both the deduplication and the coverage test downstream. With Newton refinement the search recovers exactly 32 of 32 modes on our hardest synthetic target.

Step 2: local masses. At each converged candidate $x_j^\star$ with $A_j = -\nabla^2 \log \tilde{\pi}(x_j^\star) \succ 0$ we take the Laplace mass

$$\log m_j \;=\; \log \tilde{\pi}(x_j^\star) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log \det A_j, \qquad \hat{p}_j \;=\; \frac{m_j}{\sum_k m_k}, \tag{3}$$

merging duplicates by Mahalanobis distance under A_j, since a Euclidean tolerance is not scale-invariant and fails as soon as modes are anisotropic. Note that (1) is a *ratio* of harmonic masses, so a systematic Laplace bias shared across modes cancels; §7 measures what happens when it is not shared.

Step 3: coverage. Draws from q cost no target evaluations. We draw n_q of them and assign each to the nearest discovered mode in the local Mahalanobis metric, which gives $\hat{q}_j$ with $\sum_j \hat{q}_j = 1$. A single-point test on $\log q(x_j^\star)$ was our first choice and is not robust: under anisotropy it rejects modes that q does cover. This partition is not the basin decomposition the ground truth uses, and complete enumeration does not by itself make the two agree; the disagreement is measurable rather than assumed away, and Appendix H reports it together with $\widehat{M}$ recomputed under the oracle partition.

What the estimator does not use. An earlier version normalised by $\hat{Z}_q$, the self-normalised estimate of the covered mass; for a collapsed q the weights are heavy-tailed and $\hat{Z}_q$ swung enough to move the reported answer from 0.55 to 0.38 while the reference $\log Z$ was essentially exact. Equations (1)–(3) never form $\hat{Z}_q$, and never touch samples of π. Both $\hat{p}$ and $\hat{q}$ are, however, normalised over the *discovered* modes, with $\hat{p}_j = p_j / P_S$ for $P_S = \sum_{k \in S} p_k$ and orphaned draws assigned to their nearest discovered neighbour. Under complete enumeration $P_S = 1$, and $\widehat{M}$ then estimates M up to the local-mass bias of (3) – a residual no search budget removes, so consistency additionally needs the Laplace masses to become exact or to share a common multiplicative error across modes (Proposition 8). Under incomplete enumeration both quantities are inflated and the error is **not** signed (Proposition 6). We therefore make no bound claim: $\widehat{M}$ estimates the missing mass of the landscape the search resolved, and the diagnostic of §6 exists to detect when that is not the whole one.

What the audit seeds. The audit returns mode locations, curvatures and approximate basin masses, so the Laplace mixture $q_L = \sum_j \hat{p}_j \, \mathcal{N}(x_j^\star, A_j^{-1})$ can be assembled and sampled with no further target evaluations at all. The obvious question is how good a sampler that is: if it matched the trained one, "detection is cheaper than correction" would be a claim about a correction nobody wants. Appendix G measures it, against the same mode-level M and against within-basin moments that are exact by quadrature; the answer is uncomfortable for the benchmarks rather than for the check, and §7 states it.

5 What the estimator can and cannot guarantee

Eight short results delimit the method; the discussion of what each one does and does not license is in Appendix C. Propositions 1–3 formalise the blind spot of ESS, Proposition 4 and its corollaries characterise normaliser comparison, Propositions 5–7 establish the limits of finite-query auditing and of its self-diagnostic, and Proposition 8 controls the effect of the local-mass approximation. Proofs are in Appendix A.

Proposition 1 (Subset collapse is invisible to ESS)**.** *Let S be a union of basins with $\pi(S) = \omega \in (0, 1)$ and let $q(x) = \pi(x)\mathbf{1}_S(x)/\omega$, i.e. q reproduces π exactly on S and places no mass elsewhere. Then the importance weights are constant on $\operatorname{supp} q$, so the self-normalised $\widehat{\mathrm{ESS}} = 1$ for every sample size, while the missing mass is $M = 1 - \omega$, which can be made arbitrarily close to 1.*

Proposition 2 (The empirical ESS inherits the blindness at finite sample)**.** *Let B be a basin with $q(B) = \varepsilon > 0$ and $\pi(B) = w$, and draw n i.i.d. samples from q. With probability $(1 - \varepsilon)^n$ no*

draw lands in B, and conditionally on that event the sample is i.i.d. from $q(\cdot \mid B^c)$, so the empirical ESS is exactly the one computed for a sampler that does not reach B at all. Taking $\varepsilon \le 1/n$, this happens with probability at least $(1-1/n)^n \ge 1/4$, while w may be arbitrarily close to 1.

This is what makes Proposition 1 bite for samplers whose support is formally everything: a flow with a Gaussian base assigns every basin positive mass, but a basin it reaches with probability below $1/n$ is invisible to an n-sample diagnostic, and the observed $\widehat{\mathrm{ESS}}$ is the one belonging to a strictly smaller sampler.

Proposition 3 (ESS is bounded below independently of the missing mass)**.** *Let q be supported on S and δ-accurate there, i.e. $\left|\log\{q(x)/\pi(x \mid S)\}\right| \le \delta$ for all $x \in S$. Then the population ESS rate satisfies $\mathrm{ESS}(q) \ge e^{-4\delta}$, a bound that does not involve M_{supp}.*

Weight collapse is the opposite case: q reaches every basin but with the wrong proportions.

Proposition 4 (Normalizer identity)**.** *Let q be supported on S, a union of basins. Then*

$$\mathbb{E}_q[\tilde{\pi}/q] \;=\; \int_S \tilde{\pi} \;=\; Z\,\pi(S) \;=\; Z\,(1-M_{\mathrm{supp}}), \qquad \textit{so} \qquad M_{\mathrm{supp}} \;=\; 1-\mathbb{E}_q[\tilde{\pi}/q]\,/\,Z. \quad (4)$$

Corollary 1 (Blindness to weight collapse)**.** *If $\mathrm{supp}(q) \supseteq \mathrm{supp}(\pi)$ then $\mathbb{E}_q[\tilde{\pi}/q] = Z$ exactly, however badly q misallocates mass among basins. Hence any diagnostic of the form $\widehat{M} = 1 - \widehat{Z}_q/\widehat{Z}_{\mathrm{ref}}$ has population value 0 while M_{wt} may be as large as $1-\max_j p_j$.*

Corollary 2 (The failure direction is a false negative)**.** *The same estimator returns $\widehat{M} \le 0$ whenever $\log \widehat{Z}_{\mathrm{ref}} \le \log \widehat{Z}_q$. As $\widehat{Z}_q$ is computed under a concentrated q while $\widehat{Z}_{\mathrm{ref}}$ must integrate the whole target, this is the typical situation at small reference budgets.*

Proposition 5 (No certificate from finitely many queries)**.** *Fix any algorithm that outputs an estimate of M after N evaluations of $\tilde{\pi}, \nabla\tilde{\pi}, \nabla^2\tilde{\pi}$ at adaptively chosen points, and which reads those values at finite precision $\varepsilon > 0$. For every N, every $\varepsilon > 0$ and every $\delta > 0$ there exist two targets $\tilde{\pi}_1, \tilde{\pi}_2$ whose values and first two derivatives differ by less than ε at all N queried points – so the algorithm returns the same answer on both – and for which $|M(\tilde{\pi}_1) - M(\tilde{\pi}_2)| \ge 1-\delta$. The construction is a distant Gaussian bump; explicit constants are in Appendix A.*

Proposition 6 (Incomplete enumeration is not one-sided)**.** *With S the set of discovered modes, $\widehat{M}$ can exceed or fall below M.* Over-estimation: *take $p = (0.4, 0.3, 0.3)$, $q = (0.2, 0.4, 0.4)$, so $M = 0.2$; if the search finds only the first two modes and the third mode's q-draws are assigned to the second, then $\hat{p} = (0.571, 0.429)$, $\hat{q} = (0.2, 0.8)$ and $\widehat{M} = 0.371 > M$.* Under-estimation: *take two modes of weight 0.5 with q concentrated entirely on the first, so $M = 0.5$; if the search finds only the first, $\hat{p} = \hat{q} = 1$ and $\widehat{M} = 0 < M$.*

Proposition 7 (Blind spot of the growth statistic)**.** *Model the search as R i.i.d. restarts, and let mode j have probability α_j of being reached by one restart. If mode j is missed at budget R and at the reference budget $R/4$, it contributes nothing to either the weight ratio or the count ratio, and the configuration is cleared as though it had been enumerated. This happens with probability $(1-\alpha_j)^R$, so for $\alpha_j \lesssim 1/R$ a mode carrying arbitrary weight can be missed by the estimator* and *by the diagnostic simultaneously.*

6 KNOWING WHEN NOT TO BELIEVE THE ANSWER

A diagnostic that cannot fail loudly is worse than none. We report two statistics, neither of which uses ground truth.

(a) Growth test. Let $W(R)$ be the total harmonic mass of the modes discovered by R restarts, and $N(R)$ the number of them. We report $\rho = W(R)/W(R/4)$ together with the count ratio $N(R)/N(R/4)$. If the enumeration has saturated both are ≈ 1.

(b) Singleton mass. The fraction of discovered mass held by modes that exactly one trajectory reached, in the spirit of Good–Turing.

The diagnostic clears the estimate when $\rho < 1.1$, the count ratio is below 1.1, and the singleton fraction is below 0.05. These three numbers are heuristic: they were chosen while looking at the targets of §7 and we do not claim they transfer. Appendix N reports what they do on families that had no part in choosing them.

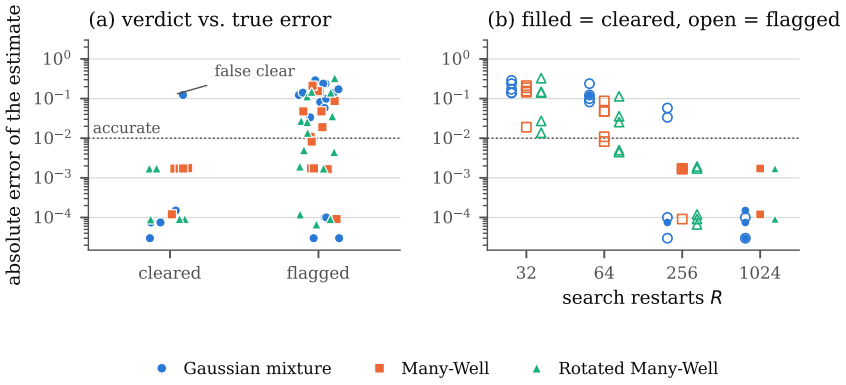

Figure 1: The self-diagnostic against ground truth it never sees, over all 60 configurations. **(a)** Error grouped by verdict: 15 of the 16 cleared configurations fall below 10^{-2} (dotted). **(b)** The same points against the search budget, filled where cleared; both the error and the verdict improve with R, which is what makes "raise the budget" the right response to a warning.

Why both ratios, and what the pair can and cannot say. Weight growth alone is insufficient, and its failure is instructive: when basins have very unequal *attraction* volumes a small search returns the same easy subset at both budgets, giving $\rho = 1.000$ and no singletons while half the modes have never been visited – we observed exactly this, with 4 of 8 modes found and the estimate off by 0.29. Discovered weight is blind to that; the discovered *count* is not, and adding the count ratio removed three of the four such failures in our sweep. What no amount of tuning removes is the one-sidedness. A search that never reaches a basin at any budget in the tested range produces no signal in either statistic, so the check is a necessary and not a sufficient condition: $\rho > 1.1$ means do not trust the estimate, while $\rho < 1.1$ means only that the search found no evidence against it. Stated as a capability, the method cheaply produces evidence of missing mass, and sometimes evidence that the search has stabilised; it does not license the stronger reading that a cleared run may be trusted.

A second blind spot, of a different kind. The one-sidedness above concerns the search. A second gap is closed by no search budget: both statistics are functions of the *set of modes discovered* (Proposition 7), so neither reads a local mass. A run that finds every mode at every sub-budget passes all three conditions and can still be wrong by whatever the anharmonicity of the basins makes the Laplace weights wrong – the channel Proposition 8 measures through η. The diagnostic covers incomplete enumeration and is silent about local-mass bias, and the two are independent; Appendix N exhibits the second with the search demonstrably complete.

What it does in practice. Figure 1 shows all 60 configurations (three targets, four search budgets, five seeds). The check is conservative: it clears 16 of 60, and 15 of those 16 have error below 10^{-2} – most of them below 10^{-3}. The single false clear sits at the smallest useful budget on the Gaussian mixture. Many accurate configurations are also flagged, which is the harmless direction of the trade: the actionable output is "increase the search budget", and panel (b) shows why that instruction is sound – error falls and verdicts flip to cleared as R grows. These rates are in sample – all sixty configurations come from three related targets – and Appendix M breaks them down by target and budget.

7 Experiments

Targets with exact ground truth. We avoid targets whose "gold" is itself an expensive approximation, since the gold would be the object under suspicion. **(i)** A Gaussian mixture, $d = 16$, $K = 8$: Z and the mode weights are analytic. **(ii) Many-Well**, $E = \sum_{i<m}(x_i^4 - 6x_i^2 - \frac{1}{2}x_i) + \sum_{i\geq m}\frac{1}{2}x_i^2$ with $d = 16$, $m = 5$, giving $2^5 = 32$ anharmonic and asymmetric modes; the energy separates over coordinates, so $\log Z$ and every sign-pattern weight reduce to one-dimensional quadrature. **(iii) Rotated Many-Well**, $E'(x) = E(Ax)$, $A = Q\,\mathrm{diag}(s)$ with Q random orthogonal

Table 1: Mean absolute error of the estimated missing mass over five seeds, against the best configuration of a tuned SMC reference. Cost is a fraction of the sampler's own training budget.

	pre-flight check		best SMC reference	
target	cost	error	cost	error
Gaussian mixture	2.7%	0.0183	70.5%	0.0154
Gaussian mixture	10.4%	0.0001	70.5%	0.0154
Many-Well	2.7%	0.0014	281.7%	0.0080
Rotated Many-Well	2.7%	0.0008	281.7%	0.1108

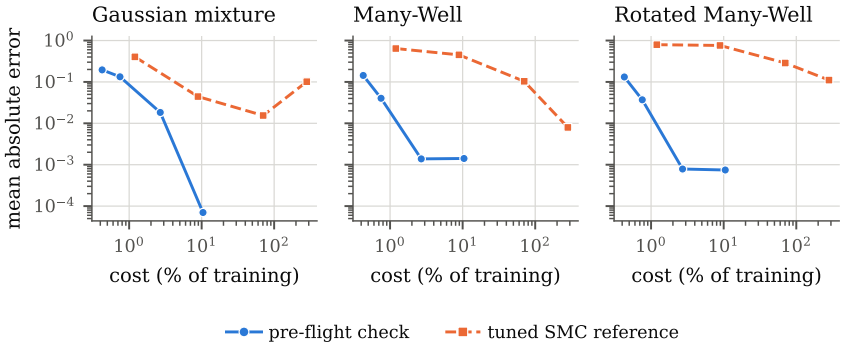

Figure 2: Error–cost frontier, five seeds. The check (blue) sits below and to the left of the tuned SMC reference (orange) on every target, and the gap widens as the target gets harder: on the rotated anisotropic target the reference is still not accurate even at $2.8\times$ the sampler's training cost.

and condition number 20, removing separability, isotropy and Hessian conditioning at once while $\log Z' = \log Z - \log|\det A|$ stays exact.

Why this reference, and what it cannot settle. We compare against an annealed SMC estimate of $\log Z$, the missing mass read off the gap to the sampler's own estimate. Proposition 4 justifies the choice: the gap *is* the missing mass due to lost support. The same analysis makes the comparison partly circular, and we say so rather than let the reader find it: Corollary 1 shows the reference is blind to M_{wt} at the population level, so beating it on weight collapse (§8) is guaranteed in advance. Only the M_{supp} comparison is a genuine contest. Appendix D adds a competitor that estimates basin weights directly and is therefore not subject to Corollary 1 at all, on all four targets and over a sweep of M. The reference is a full SMC sampler, not a straw man, and Appendix L records what upgrading it cost us.

Results. Table 1 and Figure 2 give the mean absolute error over five seeds against the best SMC configuration on each target. On the two harder targets the check is roughly two orders of magnitude cheaper while also reducing estimation error, dramatically so on the rotated target (0.0008 against 0.1108). On the Gaussian mixture the reference is marginally more accurate at $R{=}256$, at $26\times$ the cost; raising the search budget to $R{=}1024$ (10.4% of training) reduces the error to 0.0001, and – importantly – this is exactly the configuration change that the self-diagnostic of §6 calls for, without being told the answer (Table 13).

Extrapolating the observed linear trend, the check reaches parity with the training budget at $d \approx 700$ measured by wall-clock, or $d \approx 10^3$ measured by evaluation count, and is more expensive than retraining beyond that. We therefore state the claim precisely: on targets of the size used in the current sampling benchmarks the check costs a few percent of training, and this is not a statement about $d = 10^3$–10^4 problems. The standard remedy – replacing the explicit Hessian by Hessian-vector products with a Lanczos or stochastic-trace estimator of $\log\det$, which would make the per-mode cost $O(k)$ rather than $O(d)$ – applies directly here, but we have not implemented it and do not claim it.

The check costs a few percent of training at $d = 16$, and that figure is dimension-specific: Appendix O runs $d = 16$ to 128 with the sampler and search budget fixed, and finds the accuracy flat at 1–2 $\times$ 10^{-3} while the cost grows close to linearly in d, faster in wall-clock than in evaluations because a Hessian is $O(d)$ backward passes.

A target that is not kind to the Laplace step. Proposition 8 says the error is governed by the *spread* of the local-mass bias across basins, which invites the objection that our targets were chosen so that the spread is small. It is, and we can measure it: on Many-Well every basin is a copy of one quartic well, and the harmonic approximation to the exact one-dimensional basin weight differs across basins by a factor of 1.005. The rotated variant changes conditioning, not basin shape, so it inherits the same figure. We therefore build a *heterogeneous* Many-Well in which each well coordinate carries its own quartic and sextic coefficients and its own asymmetry, so that different sign patterns genuinely differ in anharmonicity while the energy still separates and $\log Z$ and every mode weight stay exact by quadrature. The bias spread rises from 1.005 to 12.57. Over five seeds the check finds all 32 modes and returns a mean absolute error of 0.0064 at R=256 (2.69% of training cost) against true missing masses of 0.33–0.50, a relative error of one to two percent. The degradation is real and in the predicted direction – a $12.6\times$ increase in spread costs roughly $5\times$ in error, from 0.0014 to 0.0064 – but it is a degradation of the third decimal, not a failure. We note that the inequality of Proposition 8 is loose here: at this spread it evaluates to a vacuous bound, so what the experiment confirms is the qualitative claim, that spread rather than magnitude is the governing quantity, and not the constant.

A sampler with no density at all. Every experiment above used a normalizing flow, which has a tractable q and therefore an ESS – so a sceptic may object that the check is only interesting because we chose to ignore a diagnostic that existed. We therefore train a Path-Integral-Sampler style controlled SDE (Zhang & Chen, 2022) on Many-Well. Two things must be kept apart. The endpoint density $q(x_T)$ exists but cannot be tractably evaluated, so no endpoint ESS, ELBO or $\widehat{Z}_q$ can be formed from the sampler's *output*. PIS does admit path-space importance weights through Girsanov and uses them to estimate Z (Zhang & Chen, 2022), but forming them needs the control and the noise realisations – the sampler's internals, not its draws. Our audit assumes draws only, the interface one has for a pretrained or third-party sampler; and where the path weights are available they give a normaliser estimate, to which Corollary 1 applies unchanged. The check is unaffected either way: its coverage step consumes draws from q and never its density. Across five seeds the trained sampler populates 1 of 32 modes with true missing mass 0.569, and the check recovers it to a mean absolute error of 0.0027 at R=256 (0.55% of training cost) and 0.0025 at R=1024 (2.21%), with a per-seed spread of 0.0017–0.0033. The self-diagnostic clears R=1024 in all five runs and flags R=256 in four of five, again erring conservatively. This is the setting the method was built for: the sampler is opaque, the target is not, and everything the check needs comes from the target.

The audit's own output is already a sampler. The check returns modes, curvatures and masses, so the Laplace mixture q_L of §4 costs nothing more, and Appendix G measures how good it is. The answer is not comfortable. At the mode level q_L wins on every target by two orders of magnitude, mean M of 0.004–0.010 against 0.51–0.82 for the flows it was read off; within basins, against exact conditional moments, it wins on the Gaussian mixture (0.002 against 0.135) and on the heterogeneous target (0.097 against 0.187), and loses only on plain Many-Well (0.047 against 0.011). Where a few dozen basins are well separated and near-ellipsoidal, enumerating the modes largely solves the sampling problem, so these benchmarks do not on their own justify a trained sampler. What q_L cannot do is exceed its own enumeration: a basin no restart reaches is absent from it exactly as from $\widehat{M}$. The claim we defend is narrower – given a sampler someone has trained, finding out how much mass it lost is cheap, and the diagnostics in use cannot tell you.

The error decomposes. Total error splits into search incompleteness and Laplace bias, and the second term is negligible because (1) is a ratio: on Many-Well, where every mode has the same quartic shape, the bias cancels almost exactly and the error is 0.001 despite the modes being manifestly non-Gaussian. What remains is search incompleteness, controlled by a single interpretable knob – the number of restarts – and reported by the self-diagnostic. The SMC reference, by contrast, never becomes accurate on the rotated target: its mean error is 0.287 at 70.5% of training cost and 0.111 at 281.7%, both far above what the check reaches for 2.7%.

Table 2: Two training runs of the same controlled-SDE architecture on the same target, with nearly equal missing mass and opposite decomposition. The M_{supp} row is the audit of §7 (five seeds); the M_{wt} row differs only in the gradient clipping applied inside the control (two seeds). ESS, the ELBO and $\widehat{Z}_q$ cannot be formed from the draws of either; path-space weights, available only with the sampler's internals, give a normaliser estimate that Corollary 1 shows is blind to the second row.

	basins reached	M_{supp}	M_{wt}	M	check: abs. error
support collapse	1/32	0.569	0.000	0.569	0.0025
weight collapse	32/32	0.000	0.559	0.559	0.0057

8 Collapse without support loss

Everything so far concerned M_{supp}. This section measures the other term, for two reasons: Corollary 1 says the SMC reference of §7 cannot see M_{wt} at any budget, so where M_{wt} occurs that competitor is not expensive but inapplicable; and M_{wt} lives at small M, a regime the earlier experiments never entered.

Table 10 gives the result and Figure 3 (Appendix J) plots it. Across twelve tilt rows M_{wt} ranges over $[0, 0.637]$ while the normalizer gap stays within 0.037 of zero and the SMC estimate is exactly 0.0000 in every one; on the mixed rows the gap reads 0.1038–0.1043 against a true M_{supp} of 0.1038 while M_{wt} grows to 0.059, so it is measuring M_{supp} and nothing else to three decimals, exactly as (4) predicts. The check tracks M throughout, with absolute error at most 0.007 over all eighteen rows.

The same architecture collapses in both ways. The controlled-SDE audit of §7 produced a sampler populating 1 of 32 basins, so its M is almost entirely M_{supp}. Retraining the same PIS-style control on the same target under a different gradient-clipping setting produces a sampler that reaches *all* 32 basins and is wrong about the proportions instead. We report this as an existence result and do not claim clipping as the general cause: it is the only setting that differs in our paired runs, but we did not sweep it. The two failures have nearly the same total missing mass and opposite composition (Table 2). Neither sampler yields an endpoint ESS, ELBO or $\widehat{Z}_q$ from its draws, since the endpoint density of a controlled SDE cannot be tractably evaluated; and for the second, Corollary 1 says that a normaliser comparison – including one built from the path-space weights, which *are* available to whoever holds the sampler's internals – would report nothing missing. The check returns 0.565 against a true 0.559 at 2.2% of training cost and clears its self-diagnostic.

The error floor at small M. Because $\widehat{M} = \sum_j (\hat{p}_j - \hat{q}_j)_+$ rectifies rather than cancels the noise in $\hat{p}_j$ and $\hat{q}_j$ where $p_j \approx q_j$, the estimator has a floor that matters precisely in this small-M regime: on the GMM tilt family at a true M of zero it reads 0.098 at $R{=}256$ and 0.006 at $R{=}1024$. The practical consequence is that an estimate small enough to be interesting should not be trusted at a budget whose convergence has not been checked, which turns the diagnostic of §6 from a guard into a requirement; Appendix B gives the measurement in full.

9 Conclusion

We tested one narrow claim, that detecting missing mass is cheaper than sampling it, and report what survived and what did not. A check costing a few percent of training reaches 10^{-3}-level accuracy on the harder exactly solvable targets, applies to samplers with no tractable density, and reports when its own answer should be disbelieved. Against that: it bounds M in neither direction under incomplete enumeration, its thresholds are uncalibrated and read no local mass, its cost advantage disappears near $d \approx 700$, and on these benchmarks the mixture the audit hands back is itself the better mode-level sampler. The finding we would keep is the unexpected one: these diagnostics' blind spots do not overlap. ESS cannot see lost support, normaliser comparison is provably blind to misallocated weight at any budget, kernel Stein discrepancy separates neither, and two runs of one architecture on one target realise the first two failures with nearly equal missing mass.

Reproducibility statement

All synthetic targets used for estimator-accuracy evaluation have closed-form or quadrature-exact ground truth: the Gaussian mixtures analytically, and Many-Well and its rotated and heterogeneous variants by one-dimensional quadrature (§7). For LJ-13 we use the published global-minimum energy as an external reference; the captured-mass statistics there are relative to our 6000-restart enumeration and are not claimed to be exact (Appendix P). Every number in the tables and figures is produced by the accompanying code at fixed seeds, with the seed count stated in each caption; the constructed-collapse family of §8 has its decomposition known in closed form by construction, so no reference sampler enters the ground truth anywhere in the paper. The same holds for the material added in the appendices: the within-basin references of Appendix G are exact conditional moments obtained by one-dimensional quadrature or in closed form, the mode counts of Appendix N are obtained by quadrature rather than by counting what a search returned, and the kernel Stein discrepancy of Appendix K is reported as a signed U-statistic, unclamped, so that estimates below zero are visible rather than hidden, and accompanied by a calibrated test rather than argued from its sign. Appendix I states which tables report a population M computed from closed-form basin weights and which report an oracle estimate from a finite draw of q, and measures the resolution of the latter.

References

Michael S. Albergo and Eric Vanden-Eijnden. NETS: A non-equilibrium transport sampler. In *International Conference on Machine Learning (ICML)*, 2025.

Oren M. Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106 (4):1495–1517, 1997.

Clément Bénard, Brian Staber, and Sébastien Da Veiga. Kernel Stein discrepancy thinning: A theoretical perspective of pathologies and a practical fix with regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Denis Blessing, Xiaogang Jia, Johannes Esslinger, Francisco Vargas, and Gerhard Neumann. Beyond ELBOs: A large-scale evaluation of variational methods for sampling. In *International Conference on Machine Learning (ICML)*, 2024.

Denis Blessing, Xiaogang Jia, and Gerhard Neumann. End-to-end learning of gaussian mixture priors for diffusion sampler. In *International Conference on Learning Representations (ICLR)*, 2025.

Diana Cai, Robert M. Gower, David M. Blei, and Lawrence K. Saul. Fisher meets Feynman: Score-based variational inference with a product of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 54228–54261, 2025.

Francisco M. Castro-Macías, Pablo Morales-Álvarez, Saifuddin Syed, Daniel Hernández-Lobato, Rafael Molina, and José Miguel Hernández-Lobato. Conditional diffusion sampling. *arXiv preprint arXiv:2605.04013*, 2026.

Junhua Chen, Lorenz Richter, Julius Berner, Denis Blessing, Gerhard Neumann, and Anima Anandkumar. Sequential controlled Langevin diffusions. In *International Conference on Learning Representations (ICLR)*, 2025.

Andrew Dennehy, Ramchandran Muthukumar, Rebecca Willett, and Nisha Chandramoorthy. Diffusion models recover accurate mixture weights despite score function insensitivity. *arXiv preprint arXiv:2607.15485*, 2026.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, 2017.

Wei Guo, Jaemoo Choi, Yuchen Zhu, Molei Tao, and Yongxin Chen. Proximal diffusion neural sampler. In *International Conference on Learning Representations (ICLR)*, 2026.

Aaron Havens, Benjamin Kurt Miller, Bing Yan, Carles Domingo-Enrich, Anuroop Sriram, Daniel Levine, Brandon Wood, Bin Hu, Brandon Amos, Brian Karrer, Xiang Fu, Guan-Horng Liu, and Ricky T. Q. Chen. Adjoint sampling: Highly scalable diffusion samplers via adjoint matching. In *International Conference on Machine Learning (ICML)*, 2025.

Aaron Havens, Brian Karrer, and Neta Shaul. Flow sampling: Learning to sample from unnormalized densities via denoising conditional processes. *arXiv preprint arXiv:2605.03984*, 2026.

Jiajun He, Yuanqi Du, Francisco Vargas, Dinghuai Zhang, Shreyas Padhy, RuiKang OuYang, Carla Gomes, and José Miguel Hernández-Lobato. No trick, no treat: Pursuits and challenges towards simulation-free training of neural samplers. In *ICLR Workshop on Frontiers in Probabilistic Inference*, 2025.

Minkyu Kim, Kiyoung Seong, Dongyeop Woo, Sungsoo Ahn, and Minsu Kim. On scalable and efficient training of diffusion samplers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Minyoung Kim. Large-scale score-based variational posterior inference for Bayesian deep neural networks. *arXiv preprint arXiv:2602.05873*, 2026.

Joohwan Ko and Justin Domke. Model-informed flows for Bayesian inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Chengkun Li, Bobby Huggins, Petrus Mikkola, and Luigi Acerbi. Normalizing flow regression for Bayesian inference with offline likelihood evaluations. In *Symposium on Advances in Approximate Bayesian Inference (AABI)*, 2025.

Guan-Horng Liu, Jaemoo Choi, Yongxin Chen, Benjamin Kurt Miller, and Ricky T. Q. Chen. Adjoint schrödinger bridge sampler. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chirag Modi, Diana Cai, and Lawrence K. Saul. Batch, match, and patch: Low-rank approximations for score-based variational inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Wasu Top Piriyakulkij, Yingheng Wang, and Volodymyr Kuleshov. Denoising diffusion variational inference: Diffusion models as expressive variational posteriors. In *AAAI Conference on Artificial Intelligence*, 2025.

Juehang Qin, Shixiao Liang, and Christopher Tunnell. FlowVAT: Normalizing flow variational inference with affine-invariant tempering. *arXiv preprint arXiv:2505.10466*, 2025.

Severi Rissanen, RuiKang OuYang, Jiajun He, Wenlin Chen, Markus Heinonen, Arno Solin, and José Miguel Hernández-Lobato. Progressive tempering sampler with diffusion. In *International Conference on Machine Learning (ICML)*, 2025.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Sebastian Sanokowski, Wilhelm Berghammer, Haoyu Wang, Martin Ennemoser, Sepp Hochreiter, and Sebastian Lehner. Scalable discrete diffusion samplers: Combinatorial optimization and statistical physics. In *International Conference on Learning Representations (ICLR)*, 2025.

Frank H. Stillinger and Thomas A. Weber. Packing structures and transitions in liquids and solids. *Science*, 225(4666):983–989, 1984.

David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

David J. Wales, Mark A. Miller, and Tiffany R. Walsh. Archetypal energy landscapes. *Nature*, 394: 758–760, 1998.

Luhuan Wu, Yi Han, Christian A. Naesseth, and John P. Cunningham. Reverse diffusion sequential monte carlo samplers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jian Xu, Delu Zeng, Qibin Zhao, and John Paisley. Diffusion bridge variational inference for deep Gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2026.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. *International Conference on Learning Representations (ICLR)*, 2022.

A PROOFS

Proposition 1. For $x \in S$ we have $w(x) = \tilde{\pi}(x)/q(x) = Z\pi(x)/\big(\pi(x)/\omega\big) = Z\omega$, a constant independent of x. Substituting $w_i \equiv Z\omega$ into $\widehat{\mathrm{ESS}} = (\sum_i w_i)^2/(n\sum_i w_i^2)$ gives $(nZ\omega)^2/(n \cdot n(Z\omega)^2) = 1$. Since q places no mass outside S and S is a union of basins, $q_j = 0$ for every basin outside S, so $M = \sum_j (p_j - q_j)_+ = \pi(S^c) = 1 - \omega$. □

Proposition 2. The draws are i.i.d., so the number landing in B is $\mathrm{Bin}(n, \varepsilon)$ and the probability that it is zero is $(1-\varepsilon)^n$. Conditioning n i.i.d. draws from q on the event that none lies in B yields n i.i.d. draws from $q(\cdot \mid B^c)$. Under $q(\cdot \mid B^c)$ every importance weight differs from the corresponding weight under q by the common factor $1-\varepsilon$, since $q(x \mid B^c) = q(x)/(1-\varepsilon)$ on B^c; the self-normalised ESS is invariant to a common rescaling of all weights, so the empirical value is identical. For $\varepsilon \le 1/n$, $(1-\varepsilon)^n \ge (1-1/n)^n$, which increases to e^{-1} and is at least $1/4$ for $n \ge 2$. Nothing constrains $w = \pi(B)$, which may be taken arbitrarily close to 1. □

Proposition 3. Write $r = \log\{q/\pi(\cdot \mid S)\}$, so $|r| \le \delta$ on S. For $x \in S$, $\pi(x) = \pi(S)\,\pi(x \mid S)$ and hence $w(x) = Z\pi(x)/q(x) = Z\,\pi(S)\,e^{-r(x)}$, which lies in $[ae^{-\delta}, ae^{\delta}]$ with $a = Z\pi(S)$. Therefore $(\mathbb{E}_q w)^2 \ge a^2 e^{-2\delta}$ and $\mathbb{E}_q[w^2] \le a^2 e^{2\delta}$, so $\mathrm{ESS} = (\mathbb{E}_q w)^2/\mathbb{E}_q[w^2] \ge e^{-4\delta}$. The bound contains no reference to $\pi(S)$, which cancels, and hence none to $M_{\mathrm{supp}} = 1 - \pi(S)$. □

Proposition 4. Since q vanishes off S, $\mathbb{E}_q[\tilde{\pi}/q] = \int_S (\tilde{\pi}/q)\,q = \int_S \tilde{\pi} = Z\int_S \pi = Z\pi(S)$. As S is a union of basins, the basins outside S are exactly those with $q_j = 0$, so $\pi(S) = 1 - \sum_{j:q_j=0} p_j = 1 - M_{\mathrm{supp}}$. □

Corollary 1. If $\mathrm{supp}(q) \supseteq \mathrm{supp}(\pi)$ then no basin has $q_j = 0$, so $M_{\mathrm{supp}} = 0$ and Proposition 4 gives $\mathbb{E}_q[\tilde{\pi}/q] = Z$ irrespective of the values q_j. Hence the population value of $1 - \widehat{Z}_q/\widehat{Z}_{\mathrm{ref}}$ is $1 - Z/Z = 0$. The bound on M_{wt} is attained by putting all of q's mass on the heaviest basin in the limit, giving $\sum_{j \ne j^\star} p_j = 1 - \max_j p_j$. □

Corollary 2. Immediate: $1 - \widehat{Z}_q/\widehat{Z}_{\mathrm{ref}} \le 0$ iff $\widehat{Z}_q \ge \widehat{Z}_{\mathrm{ref}}$. □

Proposition 5. Let $x_1, \dots, x_N$ be the points the algorithm queries when run on $\tilde{\pi}_1$, and let $Z_1 = \int \tilde{\pi}_1$. Put $\tilde{\pi}_2 = \tilde{\pi}_1 + \lambda\varphi_z$ with $\varphi_z(x) = (2\pi)^{-d/2} e^{-\|x-z\|^2/2}$ and $\lambda = Z_1(1-\delta)/\delta$, so that the bump carries a fraction $\lambda/(Z_1+\lambda) = 1-\delta$ of the total mass of $\tilde{\pi}_2$ and therefore $M(\tilde{\pi}_2) \ge 1-\delta$, while $M(\tilde{\pi}_1)$ may be 0.

Write $\rho = \|x - z\|$. Direct differentiation gives $\varphi_z(x) = (2\pi)^{-d/2} e^{-\rho^2/2}$, $\|\nabla\varphi_z(x)\| = \rho\,\varphi_z(x)$ and $\|\nabla^2\varphi_z(x)\|_{\mathrm{op}} \le (\rho^2+1)\varphi_z(x)$, so all three perturbations are bounded by $(1+\rho+\rho^2)\varphi_z(x)$. For $\rho \ge 4$ we have $3\rho^2 \le e^{\rho^2/4}$ and hence $(1+\rho+\rho^2)e^{-\rho^2/2} \le 3\rho^2 e^{-\rho^2/2} \le e^{-\rho^2/4}$. Therefore, if z is at distance at least r from every queried point and r satisfies (5), every queried value and

derivative changes by at most $\lambda(2\pi)^{-d/2}e^{-r^2/4} \le \varepsilon$. Reading at precision ε, the algorithm cannot distinguish the two targets: it follows the same execution path and returns the same estimate, which therefore errs by at least $(1-\delta)/2$ on one of them.

Such a z exists: the N excluded balls of radius r occupy volume $N\,V_d r^d$ with V_d the unit-ball volume, so any ball of radius $R_0 > r\,N^{1/d}$ has volume $V_d R_0^d > N V_d r^d$ and cannot be covered by them. □

Explicit constants for Proposition 5. It suffices to separate the two targets by a unit-variance Gaussian bump placed at distance

$$r \;\ge\; \max\Big\{4,\; 2\sqrt{\log\big(\lambda\,(2\pi)^{-d/2}/\varepsilon\big)}\Big\}, \qquad \lambda \;=\; Z_1(1-\delta)/\delta, \tag{5}$$

from every queried point, which is possible inside any ball of radius exceeding $r\,N^{1/d}$. Consequently, finite-query evidence can certify only mass associated with regions the queries actually reached; it cannot certify the absence of additional unseen mass.

Proposition 6. Both cases are direct computation. For the first, $\sum_{k\in S} p_k = 0.7$ gives $\hat{p} = (0.4, 0.3)/0.7 = (0.571, 0.429)$; assigning the third basin's q-mass 0.4 to the second discovered mode gives $\hat{q} = (0.2, 0.4{+}0.4) = (0.2, 0.8)$; hence $\widehat{M} = (0.571{-}0.2)_+ + (0.429{-}0.8)_+ = 0.371$, while $M = (0.4-0.2)_+ + (0.3-0.4)_+ + (0.3-0.4)_+ = 0.2$. For the second, S contains one mode with $p_1 = q_1$ after normalisation, so every term vanishes and $\widehat{M} = 0$, while the undiscovered mode contributes $p_2 = 0.5$ to M. □

Proposition 8. Write $M_{\mathrm{tot}} = \sum_{k\in S} m_k$, so that $p_j^S = m_j/M_{\mathrm{tot}}$. From $\hat{p}_j = c_j m_j / \sum_{k\in S} c_k m_k$ and the assumed bounds, $\hat{p}_j \le \frac{c(1+\eta)m_j}{c(1-\eta)M_{\mathrm{tot}}} = p_j^S \frac{1+\eta}{1-\eta}$ and symmetrically $\hat{p}_j \ge p_j^S \frac{1-\eta}{1+\eta}$, so $|\hat{p}_j - p_j^S| \le p_j^S \frac{2\eta}{1-\eta}$. Since $t \mapsto (t)_+$ is 1-Lipschitz and $\hat{q}$ is held fixed between the two estimates, $|\widehat{M} - \widehat{M}_\star| \le \sum_{j\in S} |\hat{p}_j - p_j^S| \le \frac{2\eta}{1-\eta}\sum_{j\in S} p_j^S = \frac{2\eta}{1-\eta}$. For the second statement, complete enumeration gives $P_S = 1$ and $p_j^S = p_j$, and $\hat{q}_j = q_j$ turns $\widehat{M}_\star$ into $\sum_j (p_j - q_j)_+ = M$. □

Proposition 7. Both statistics are functions of the set of modes discovered at the two budgets. A mode reached by no restart appears in neither set, hence changes neither the numerator nor the denominator of either ratio, and the verdict is the same as if that mode did not exist. The restarts being independent, the probability that mode j is reached by none of the R is $(1-\alpha_j)^R$, which is bounded away from zero whenever $\alpha_j R = O(1)$. □

B THE ERROR FLOOR AT SMALL M

Weight collapse occupies a range of M that the earlier experiments never visited, and there the estimator's own floor matters. $\widehat{M} = \sum_j (\hat{p}_j - \hat{q}_j)_+$ rectifies rather than cancels the noise in $\hat{p}_j$ and $\hat{q}_j$ at basins where $p_j \approx q_j$, so at the population level $\mathbb{E}[\widehat{M}] \ge M$. On the GMM tilt family at a true M of zero the estimate reads 0.098 at $R{=}256$ and 0.006 at $R{=}1024$; the mean absolute error over the family falls from 0.048 to 0.002. We are careful about the direction of the effect: the mean signed error at $R{=}256$ is positive ($+0.035$) as the rectification argument predicts, but two of six rows come out low, so at this sample size the low-budget behaviour is dominated by variance and we cannot claim a clean one-sided bias. The $M = 0$ row is the unambiguous measurement, since there non-negativity makes the whole estimate error. On Many-Well the floor is already 0.007 at $R{=}256$, so the budget required is target-dependent. The practical consequence is that an estimate small enough to be interesting – say below 0.1 – should not be trusted at a budget whose convergence has not been checked, which turns the diagnostic of §6 from a useful guard into a requirement.

C DISCUSSION OF THE GUARANTEES

The point is not that ESS is noisy but that it is *exactly* maximal in the worst case. The idealisation is not needed: if q is merely *close* to $\pi(\cdot \mid S)$ the conclusion survives with a constant.

Since δ and M_{supp} vary independently, ESS carries no information about how much mass is absent: it measures how well q fits whatever it does cover. Our measurements ($\widehat{\text{ESS}} = 0.99$ at $M = 0.87$) are the finite-sample shadow of this identity, and the spread we observe across runs (0.979–0.996) is the $\delta > 0$ correction.

Weight collapse is the opposite case. If q reaches every basin but with the wrong proportions, then $w \propto p_j/q_j$ varies across basins by construction, so the weights are not constant and ESS *does* degrade. ESS is not a broken diagnostic; it is matched to one of the two terms in (2) and blind to the other. The following results show that the natural competitor to our check has precisely the complementary blind spot.

The importance-sampling estimate of the normalising constant under a collapsed sampler is unbiased not for Z but for $Z(1 - M_{\text{supp}})$. This is the precise sense in which comparing normalising constants is the *right* competitor for our check – the missing mass is exactly the gap – and it explains the empirical behaviour of the SMC reference in §7. It also has two consequences that bound what that competitor can do.

Corollary 1 is the sharper statement. It is not about estimator variance, tuning, or budget: the estimand itself does not depend on M_{wt}, so no amount of computation makes a normalizer comparison sensitive to weight collapse. Corollary 2 explains the failure mode we measure in §7, where the reference returns exactly 0.0000 – "nothing is missing" – at budgets where its $\log Z$ estimate is still far off. Our own estimator sidesteps both: (1)–(3) never form $\widehat{Z}_q$, a choice we originally made for variance reasons (§4) and which Corollary 1 now justifies on identifiability grounds.

This says that *no* method of this family can certify its answer, and it is why §6 reports a necessary rather than a sufficient condition. It does *not* say that our particular estimator errs in a known direction – that is a separate question, settled negatively by Proposition 6. It is the same obstruction that makes simulation-free mode discovery hard (He et al., 2025): the two statements are duals, one about finding mass and one about certifying its absence.

The mechanism is the shared normalisation: dividing by $P_S < 1$ inflates $\hat{p}$, while re-assigning orphaned draws inflates $\hat{q}$, and which effect dominates depends on where the undiscovered mass sits relative to the sampler. This is the precise sense in which the estimate is conditional on the search, and it is why we report the diagnostic of §6 alongside every number rather than claiming a bound.

A third fact, proved in Appendix F, explains why an approximation as crude as Laplace suffices: the error of (1) is controlled by the *spread* of the local-mass bias across modes and not by its magnitude, so a bias shared by every mode cancels exactly in the ratio (Proposition 8). This is why Many-Well, whose modes are all copies of one quartic well, is estimated to 10^{-3} despite being manifestly non-Gaussian.

The proposition predicts precisely the failure we observe: at $R = 32$ on the Gaussian mixture, four of eight modes have small attraction volume, both budgets return the same four, the ratios are 1.000, and the estimate is wrong by 0.29. It also says what to do about it – the failure probability decays geometrically in R, which is why the configurations cleared at $R = 1024$ are accurate in every run.

D A COMPETITOR THAT DOES NOT ESTIMATE A NORMALISING CONSTANT

The reference of §7 estimates $\log Z$, and Corollary 1 bounds what that statistic can do: it is blind to M_{wt} at the population level, so the weight-collapse comparison of §8 demonstrates a property of the estimand rather than a contest. To make the comparison a contest we add a competitor that estimates the basin weights *directly* and never forms a normalising constant: parallel tempering on a 16-temperature ladder with replica exchange, reading $\hat{p}_j$ off the occupancy of the $\beta = 1$ replica, with no Hessian, no Laplace approximation and no Z. It is charged in the same currency as everything else.

Table 3 gives the comparison. Tempered MCMC is a genuine alternative route to the missing mass – unlike the normalizer gap it is not blind to either term of (2), and at 3×10^3 steps it matches the truth to four decimals. The difference is cost: it needs $15.7\times$ what the check needs to get there, and at the cheaper setting it is still 0.012 off at $5.2\times$ the cost. Mode-search auditing is not the only way

Table 3: A non-normalizer competitor on Many-Well ($K = 32$, true $M = 0.9854$). Parallel tempering is not structurally blind the way the normalizer gap is: it reaches the right answer given enough steps. It costs $15.7\times$ more to get there. Costs are relative to *this* sampler's deliberately short training budget of 1.5×10^6 evaluations and are not comparable to the 2.7% of Table 1, which is measured against the longer runs used there; what is comparable is the ratio within this table.

method	estimate	abs. error	cost
ours	0.9855	0.0001	55.7%
parallel tempering, weights read directly (10^3 steps)	0.9732	0.0122	291.7%
parallel tempering, weights read directly (3×10^3 steps)	0.9854	0.0000	875.0%

to obtain the missing mass without ground truth; on these targets it is about an order of magnitude cheaper than the tempered-MCMC route.

Broader evidence. One configuration at $M = 0.9854$ is a narrow base for a comparison, so we repeat it on all four targets with trained samplers, and on constructed samplers whose M is known in closed form and swept over its range. Costs are fractions of each sampler's own training budget.

Table 4: The non-normalizer competitor across targets, three seeds each; mean absolute error and mean cost. The R=1024 column is the same set of samplers audited at the budget the self-diagnostic asks for, and is the fair comparison: at that budget the check is more accurate than parallel tempering on every target while costing about a thirtieth as much. At R=256 the ordering reverses on the Gaussian mixture and the rotated target, which are the configurations §6 already flags. The two check columns come from separate runs of the same samplers – identical targets, seeds and training protocol – and their true M agrees to three decimals.

		check, R=256		check, R=1024		PT, 1500 steps		PT, 6000 steps	
target	mean M	err	cost	err	cost	err	cost	err	cost
GMM	0.822	0.0309	4%	0.0006	14%	0.0193	109%	0.0040	438%
Many-Well	0.688	0.0012	4%	0.0012	14%	0.0797	109%	0.0098	438%
rotated Many-Well	0.804	0.0419	4%	0.0006	14%	0.1101	109%	0.0665	438%
heterogeneous Many-Well	0.507	0.0092	4%	0.0092	14%	0.0595	109%	0.0262	438%

Table 5: The same competitor against a missing mass known in closed form, swept over its range on constructed samplers. The reference is the population M of Appendix I, not an oracle sample estimate, so the $\tau = 1$ row is a true zero. Absolute errors. The check is shown at both budgets because R=256 is the setting §6 flags on the Gaussian mixture and the comparison at that budget is not the one the method asks for; at R=1024 it is. Parallel tempering's 6000-step column costs several hundred percent of a sampler's training budget, against 14% for the check at R=1024.

sampler	true M	check R=256	check R=1024	PT 1500	PT 6000
		mean (max) over 3 restart draws			
GMM, τ=1	0.0000	0.0341 (0.0935)	0.0040 (0.0047)	0.0784	0.0204
GMM, τ=1.5	0.0737	0.0124 (0.0303)	0.0028 (0.0054)	0.0589	0.0028
GMM, τ=2.5	0.1903	0.0031 (0.0078)	0.0006 (0.0014)	0.0621	0.0023
GMM, τ=4	0.3198	0.0135 (0.0387)	0.0008 (0.0016)	0.0705	0.0028
GMM, τ=8	0.5133	0.0200 (0.0577)	0.0009 (0.0022)	0.0673	0.0022
Many-Well, τ=1	0.0000	0.0085 (0.0111)	0.0085 (0.0111)	0.1168	0.0289
Many-Well, τ=1.5	0.2541	0.0036 (0.0045)	0.0036 (0.0045)	0.1074	0.0116
Many-Well, τ=2.5	0.5010	0.0011 (0.0022)	0.0011 (0.0022)	0.1033	0.0101
Many-Well, τ=4	0.5652	0.0017 (0.0023)	0.0017 (0.0023)	0.1020	0.0087
Many-Well, τ=8	0.5709	0.0018 (0.0018)	0.0018 (0.0018)	0.1020	0.0091

E EXTENDED RELATED WORK

Why generative-model coverage metrics do not transfer. Evaluation of generative models has converged on two-dimensional fidelity/coverage measures – precision and recall (Sajjadi et al., 2018), and its k-NN refinement (Kynkäänniemi et al., 2019) – precisely because scalar scores conflate sample quality with mode coverage. These are *two-sample* statistics: they compare model draws against reference draws. In the neural-sampler setting there are no reference draws, which is the whole difficulty; the same is true of sample-based statistical distances more broadly. Our estimator is one-sample in the target: it consumes $\tilde{\pi}$ and its derivatives, never samples of π.

Stein diagnostics. The one existing family that is also one-sample in the target is the Stein discrepancy of Gorham & Mackey (2015), introduced precisely because effective sample size does not account for asymptotic bias, and made computable by the kernel form of Gorham & Mackey (2017). It needs the score and draws from q, so it is the natural comparison for the density-free setting of §7 and we treat it as one. It measures a different quantity, however: a discrepancy in the whole distribution rather than a decomposition of what is missing. Bénard et al. (2023) show why that matters here – the score is insensitive to the relative weights of well-separated modes, so a sample right within each basin escapes penalty however wrong its proportions – and it is the first of the pathologies their regularised Stein-thinning variant is built to repair. Appendix K measures the effect on our targets.

Methods that fix collapse, rather than measure it. A large recent effort attacks the failure directly: learned Gaussian-mixture priors (Blessing et al., 2025), affine-invariant tempering (Qin et al., 2025), progressive tempering across temperatures (Rissanen et al., 2025), proximal schemes on path space (Guo et al., 2026), off-policy exploration with a novelty bonus (Kim et al., 2025), and SMC-corrected transport (Chen et al., 2025; Wu et al., 2025). These are complementary to this paper in the strict sense: each produces a sampler that our check can audit, and none of them tells the practitioner whether the resulting run should be trusted. One family is outside our reach rather than complementary to it: samplers over discrete state spaces (Sanokowski et al., 2025) admit no Hessian, so the local-mass step of §4 has no analogue and the check does not apply. The closest in spirit is Dennehy et al. (2026), who analyse when score-based models do and do not recover mixture weights; they explain when the failure occurs, we measure whether it has occurred in a given run.

What the check can be pointed at. The audit needs draws from q and derivatives of $\tilde{\pi}$, and nothing else, so it is indifferent to how q was obtained. That covers the score-based variational families that have grown up alongside diffusion samplers – products of t-experts fitted by score matching (Cai et al., 2025), low-rank score-based Gaussian families (Modi et al., 2025), and their scaling to Bayesian neural networks (Kim, 2026) – as well as flow families built for posterior geometry (Ko & Domke, 2025), regression-based surrogates trained on offline log-density evaluations (Li et al., 2025), and diffusion-based variational posteriors (Piriyakulkij et al., 2025; Xu et al., 2026). Two properties of these families make the audit worth running: several are trained with mode-seeking objectives, and several – the diffusion-based ones in particular – have no tractable density, so the diagnostics of §2 are not merely misleading there but unavailable.

F LOCAL-MASS BIAS

Proposition 8 (Only the *spread* of the Laplace bias matters)**.** *Fix the discovered set S and the coverage estimates $\hat{q}$, and suppose the local mass of every mode in S is estimated with a multiplicative error $\hat{m}_j = c_j m_j$ where $c_j \in [c(1-\eta),\, c(1+\eta)]$ for some common $c > 0$ and $\eta \in [0, 1)$. Write $\widehat{M}_\star = \sum_{j \in S} \left(p_j^S - \hat{q}_j\right)_+$, with $p_j^S = p_j / P_S$, for what the same procedure returns from exact local masses. Then*

$$\left|\widehat{M} - \widehat{M}_\star\right| \;\le\; \frac{2\eta}{1-\eta}, \tag{6}$$

so the local-mass step contributes no error at all when the bias is shared ($\eta = 0$), however large that bias is. If in addition the enumeration is complete (S contains every mode, hence $P_S = 1$) and the coverage estimates are exact ($\hat{q}_j = q_j$), then $\widehat{M}_\star = M$ and the same bound holds for $\left|\widehat{M} - M\right|$.

Both extra hypotheses are needed for the second sentence and neither follows from $\eta = 0$. Take two modes of true weight $p = (0.5, 0.5)$ and a sampler $q = (1, 0)$. A search that returns only the first

mode has exact local masses on what it found, so $\eta = 0$ and $\widehat{M} = 0$, while $M = 0.5$. The error there is entirely enumeration, which is Proposition 6's subject and outside this one's. Proposition 8 is a statement about the local-mass step, not a guarantee for the estimator as a whole.

This explains an otherwise surprising measurement. On Many-Well every mode is a copy of the same quartic well, so the Laplace bias is shared and $\eta \approx 0$; the estimator reaches error 10^{-3} on a target whose modes are manifestly not Gaussian. Accuracy is governed by how *uniform* the anharmonicity is across modes, not by how severe it is.

G THE SAMPLER THE AUDIT HANDS BACK

The check ends with mode locations $x_j^\star$, curvatures A_j and local masses $\hat{p}_j$, so the Laplace mixture $q_L = \sum_j \hat{p}_j \mathcal{N}(x_j^\star, A_j^{-1})$ costs nothing further to build and nothing at all to sample: no target evaluation enters after the audit has finished. If q_L were as good as the sampler it was read off, the paper's framing would be hollow, so we measure it. Two axes are reported, on three seeds per target at $R = 1024$:

- the mode-level missing mass M under the oracle basin partition, the same quantity the estimator targets;
- a *within-basin* error, the mean relative error of $\mathbb{E}[x_i^2 \mid \mathrm{sign}(x_i)]$ over the well coordinates and both sides for the Many-Well family, and of $\mathbb{E}[\|x - \mu_k\|^2 \mid k]$ for the mixture, in both cases against exact references obtained by one-dimensional quadrature or in closed form. No reference sampler is involved.

Table 6: The Laplace mixture read off the audit, against the flow it was read off. Lower is better in every column. The mixture reproduces the mode weights by construction, so its M is what the local-mass bias and the search incompleteness leave; within basins it is a Gaussian and can only be as good as the basin is quadratic. The rotated target is omitted from the last two columns because the rotation mixes coordinates and the exact conditional moments are no longer one-dimensional.

		mode-level M		within-basin rel. error	
target	seeds	flow	Laplace mixture	flow	Laplace mixture
GMM	3	0.8218	0.0043	0.135	0.002
Many-Well	3	0.6876	0.0070	0.011	0.047
rotated Many-Well	3	0.8042	0.0070	—	—
heterogeneous Many-Well	3	0.5069	0.0099	0.187	0.097

The result deserves to be stated without softening. On these benchmarks the mixture the audit hands back is the better sampler at the mode level on all four targets, by one to two orders of magnitude. On the within-basin axis, which is exactly computable on three of them, it is also the better sampler on the Gaussian mixture and on the heterogeneous target, and loses only on plain Many-Well (0.047 against 0.011), where the flow trains well and a quartic well is far from Gaussian. The extra cost over the audit is zero.

Three things follow, and the third is the one that matters for this paper.

Its basin weights are not its component weights. It is tempting to say that q_L places weight $\hat{p}_j$ on basin B_j, and as stated that is false: every component is a Gaussian with full support, so

$$q_L(B_j) \;=\; \sum_k \hat{p}_k \Pr_{X \sim \mathcal{N}(x_k^\star, A_k^{-1})}\big(X \in B_j\big) \;\neq\; \hat{p}_j, \tag{7}$$

with equality only in the limit of basins separated far relative to the local curvature. The gap is measurable, and we measure it: over all four targets and both search budgets the total variation between q_L's empirical basin weights and $\hat{p}$ pushed onto basins is at most 0.0035 on 2×10^5 draws. What follows is therefore an approximation with an error of that order, not an identity.

It barely exceeds its own enumeration. For the same reason a basin that no restart reached is *not* absent from q_L: it receives whatever Gaussian tail mass crosses its boundary. On the Gaussian mixture at $R = 64$, where 2 of 8 basins hold no discovered mode, that leakage is 3.5×10^{-5}; on the Many-Well family it is below the resolution of 2×10^5 draws. The statement worth making is therefore quantitative rather than structural: q_L's missing mass falls short of the mass the search missed only by the tail probability of crossing a basin boundary, four orders of magnitude below the missing mass itself on these targets. The mixture is a repair for mode weights, not a substitute for the search.

Its accuracy is the estimator's accuracy. Up to (7) the mixture's basin weights are $\hat{p}$, so the η of Proposition 8 governs both. The heterogeneous target, built to make η large, is where q_L's missing mass is largest in the table, and for the same reason the estimate is least accurate there.

What this says about the benchmarks. A target with a few dozen well-separated, near-ellipsoidal basins is one where enumerating the modes essentially solves the sampling problem. That is a statement about the synthetic multimodal benchmarks this literature runs on, not about samplers: it is why the check is cheap on them, and it is the same property that fails on LJ-13 (Appendix P), where a continuous symmetry makes the local mass undefined and the enumeration is usable only at low temperature. We therefore do not claim that a neural sampler is the right tool for these particular targets. The claim is narrower and survives: given a trained sampler, on any target where the modes can be enumerated at all, finding out how much mass it lost is cheap, and the diagnostic that would otherwise be used cannot tell you.

H THE COVERAGE PARTITION AGAINST THE ORACLE PARTITION

The estimator assigns each draw to the nearest discovered mode in that mode's Mahalanobis metric; the ground truth assigns it to its basin. These are different constructions and complete enumeration does not force them to agree, so the disagreement is measured rather than assumed away. For every run of Appendix G we record the fraction of draws the two rules place together, and recompute $\widehat{M}$ with the oracle partition substituted for the Mahalanobis one, holding the Laplace weights fixed, which prices the assignment's own contribution.

Table 7: Agreement between the Mahalanobis nearest-mode partition the estimator uses and the oracle basin partition the ground truth uses, three seeds per target at $R = 1024$. The last column is the change in $\widehat{M}$ when the oracle partition is substituted with the local masses held fixed.

target	seeds	worst agreement	mean agreement	$\max \lvert\Delta\widehat{M}\rvert$
GMM	3	0.9997	0.9998	0.00025
Many-Well	3	1.0000	1.0000	0.00000
rotated Many-Well	3	1.0000	1.0000	0.00000
heterogeneous Many-Well	3	0.9969	0.9973	0.00300

Agreement is never below 0.9969 and substituting the oracle partition moves the estimate by at most 0.00300, an order of magnitude below the local-mass bias on the same targets. This is a property of basins that are close to ellipsoidal around their modes and not a general guarantee: a curved or banana-shaped basin would be cut differently by the two rules, and nothing in the method protects against that. The measurement is the claim.

I WHAT "TRUE M" MEANS IN EACH TABLE

Two different quantities are easily reported under one name, and the distinction matters for reading any accuracy claim at the 10^{-3} level. We state which one each table uses.

For a *constructed* sampler the basin weights of both π and q are available in closed form, so $M = \sum_k (p_k - q_k)_+$ is a population number and satisfies $M = M_{\text{supp}} + M_{\text{wt}}$ identically. Every constructed row in this paper now reports that population value. For the Many-Well family it is

exact: the basins are sign cells and the tilted sampler draws its sign pattern first, so $q(B_j) = v_j$ with no error. For the Gaussian mixture it is exact up to the Gaussian tails that cross the nearest-mean cell boundaries, which we measure directly and report in the last column.

For a *trained* sampler no closed form exists, and what the oracle returns is an estimate $\sum_k (p_k - \hat{q}_k)_+$ from $n = 4 \times 10^4$ draws of q. That estimate is biased upward: wherever $p_k \approx q_k$ the rectifier keeps the positive part of the Monte-Carlo noise. It is therefore bounded away from zero even when $q = \pi$ exactly, and the size of that floor is a property of n and K, not of any sampler.

Table 8: Population missing mass against the oracle sample estimate on the constructed families. The oracle column is the mean and standard deviation over 10 independent draws of 4×10^4 points. At $\tau = 1$ and at "drop to all basins" the sampler *is* π, so the population M is exactly zero and everything the oracle reports there is its own floor. The last column is the fraction of q's draws that land outside the cell of the component they came from, which is the only reason the mixture row is not exact.

q	M	M_{supp}	M_{wt}	oracle estimate	cross-cell leakage
GMM, $K = 8$					
tilt τ=1	0.000000	0.000000	0.000000	0.0050 ± 0.0022	2.1×10^{-4}
tilt τ=1.5	0.073672	0.000000	0.073672	0.0740 ± 0.0014	1.9×10^{-4}
tilt τ=2.5	0.190273	0.000000	0.190273	0.1905 ± 0.0023	1.8×10^{-4}
tilt τ=4	0.319819	0.000000	0.319819	0.3191 ± 0.0022	1.4×10^{-4}
tilt τ=8	0.513326	0.000000	0.513326	0.5131 ± 0.0013	3.5×10^{-5}
drop to 8 basins	0.000000	0.000000	0.000000	0.0053 ± 0.0013	2.4×10^{-4}
drop to 4 basins	0.322764	0.322764	0.000000	0.3227 ± 0.0000	9.8×10^{-5}
drop to 1 basin	0.791255	0.791255	0.000000	0.7913 ± 0.0000	$< 10^{-6}$
Many-Well, $K = 32$					
tilt τ=1	0.000000	0.000000	0.000000	0.0079 ± 0.0017	$< 10^{-6}$
tilt τ=1.5	0.254107	0.000000	0.254107	0.2533 ± 0.0015	$< 10^{-6}$
tilt τ=2.5	0.501037	0.000000	0.501037	0.5008 ± 0.0010	$< 10^{-6}$
tilt τ=4	0.565194	0.000000	0.565194	0.5652 ± 0.0003	$< 10^{-6}$
tilt τ=8	0.570949	0.000000	0.570949	0.5709 ± 0.0000	$< 10^{-6}$
drop to 32 basins	0.000000	0.000000	0.000000	0.0081 ± 0.0012	$< 10^{-6}$
drop to 16 basins	0.029475	0.029475	0.000000	0.0308 ± 0.0011	$< 10^{-6}$
drop to 4 basins	0.333604	0.333604	0.000000	0.3336 ± 0.0000	$< 10^{-6}$
drop to 1 basin	0.570956	0.570956	0.000000	0.5710 ± 0.0000	$< 10^{-6}$

Three consequences, in the order they affect the paper.

The zero rows carry no evidence. Reporting the oracle estimate at $\tau = 1$ as a "true M" of 0.005 to 0.008 beside $M_{\text{supp}} = M_{\text{wt}} = 0$ would be internally inconsistent, since the decomposition is an identity. The population value is zero, and what the oracle reports there is the floor described above – the same quantity Appendix B measures from the estimator's side. No accuracy claim in this paper rests on those rows.

The 10^{-3} claims are made where the oracle can resolve them. Where the population M exceeds 0.2 the oracle's standard deviation over independent draws is at most 0.0022, and on the trained samplers, whose missing mass is carried by basins that receive no draw at all rather than by noisy near ties, it is at most 0.0012 over 10 repeats. Differences of 10^{-3} are therefore resolvable in exactly the regime where we claim them, and are not resolvable near $M = 0$, which is why §7 makes no claim there and Appendix B exists.

The Gaussian mixture is not exactly decomposable. The leakage column is small but not zero, so for that target "population M" means the nearest-mean cell decomposition up to a few parts in 10^4. The Many-Well family carries no such caveat, which is a reason to read the Many-Well rows as the cleaner evidence.

Table 9: The oracle's own reproducibility on the trained samplers, over 10 independent draws of 4×10^4 points. These are the references against which every trained-sampler accuracy number in the paper is measured.

target	seeds	mean oracle M	worst sd
GMM	3	0.8217	0.0012
Many-Well	3	0.6876	0.0000
rotated Many-Well	3	0.8042	0.0000
heterogeneous Many-Well	3	0.5056	0.0011

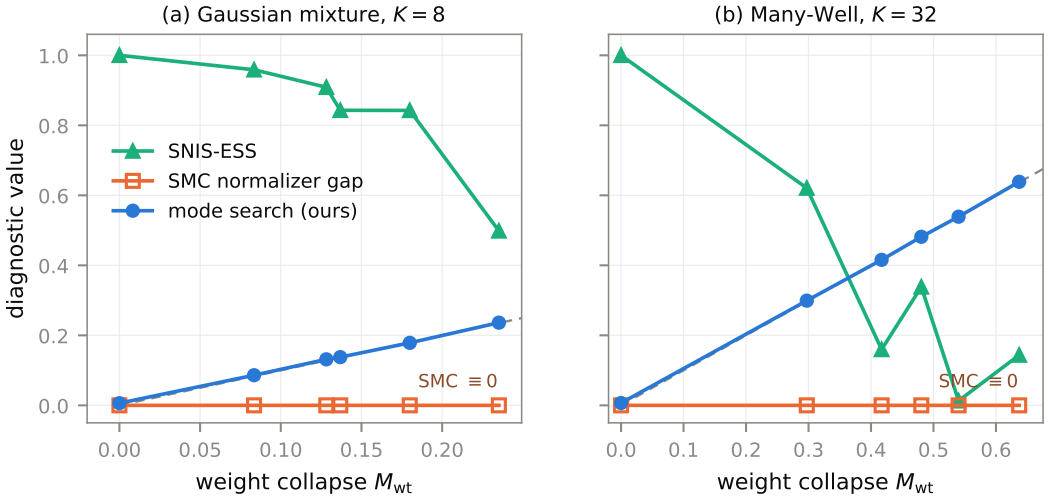

Figure 3: The blind spot of the normalizer comparison. Within each panel every sampler reaches every basin ($M_{\text{supp}} = 0$) and only the proportions are wrong. The SMC estimate built on comparing normalising constants stays pinned at zero while more than half the target mass is misallocated, as Corollary 1 requires. ESS responds; the check tracks the truth.

J Constructed collapse with a known decomposition

Constructed collapse with an exact decomposition. We build samplers whose $(M_{\text{supp}}, M_{\text{wt}})$ are known exactly. A *tilted* sampler assigns basin j probability $v_j \propto p_j^\tau$ and is exact *within* every basin, so $\text{supp}(q) = \text{supp}(\pi)$, $M_{\text{supp}} = 0$, and $M = M_{\text{wt}}$ is tuned by τ. A *mixed* sampler additionally drops the lightest quarter of the basins, making both terms nonzero. On the Gaussian mixture this is a mixture over a component subset; on Many-Well a basin is a sign pattern over the first m coordinates and we draw each coordinate from its exact within-well conditional by inverse CDF, so only the basin weights are wrong.

K The kernel Stein discrepancy on the same families

§2 claims that the one sample-only diagnostic in wide use does not separate either term of the decomposition on these targets. This appendix is the measurement behind that claim, and it is made twice: once as a raw statistic and once as a calibrated test, because the raw statistic on its own cannot support the claim.

We use the constructed families of Appendix J, whose $M_{\text{supp}}/M_{\text{wt}}$ split is known in closed form, and the inverse-multiquadric kernel $k(x, y) = (c^2 + \|x - y\|^2)^{-1/2}$ recommended by Gorham & Mackey (2017). The reported M is the population value computed from the closed-form basin weights, not an oracle sample estimate (Appendix I explains the difference and why it matters here). The last

Table 10: Constructed collapse with exactly known decomposition. In the tilt family $M_{\mathrm{supp}} = 0$, so by Corollary 1 the normalizer gap must be zero whatever M_{wt} is: it is, and the SMC estimate is exactly 0.0000 in all twelve tilt rows. The mixed rows vary M_{wt} at fixed M_{supp} and are the cleaner demonstration – on the GMM the gap reads 0.104 three times while M_{wt} grows from 0 to 0.059. ESS, blind to M_{supp}, does respond here. Search budget R=1024 on the GMM and R=256 on Many-Well, three repetitions.

target	family	M_{supp}	M_{wt}	M	ESS	norm. gap	SMC	ours
GMM	tilt τ=1	0	0.000	0.000	1.000	−0.000	0.000	0.006
	tilt τ=0.5	0	0.083	0.083	0.959	+0.001	0.000	0.086
	tilt τ=0.25	0	0.128	0.128	0.909	+0.002	0.000	0.131
	tilt τ=0	0	0.180	0.180	0.843	+0.003	0.000	0.179
	tilt τ=2	0	0.137	0.137	0.843	−0.002	0.000	0.138
	tilt τ=3	0	0.235	0.235	0.498	−0.004	0.000	0.236
	mixed τ=1	0.104	0.000	0.104	1.000	+0.104	0.058	0.104
	mixed τ=0.5	0.104	0.010	0.114	0.984	+0.104	0.060	0.116
	mixed τ=0	0.104	0.059	0.162	0.937	+0.104	0.064	0.165
Many-Well	tilt τ=1	0	0.000	0.000	1.000	−0.000	0.000	0.007
	tilt τ=0.5	0	0.297	0.297	0.621	+0.003	0.000	0.299
	tilt τ=0.25	0	0.481	0.481	0.338	+0.008	0.000	0.482
	tilt τ=0	0	0.637	0.637	0.144	+0.005	0.000	0.639
	tilt τ=2	0	0.417	0.417	0.160	−0.010	0.000	0.416
	tilt τ=3	0	0.540	0.540	0.013	−0.037	0.000	0.539
	mixed τ=1	0.008	0.000	0.008	1.000	+0.008	0.000	0.012
	mixed τ=0.5	0.008	0.265	0.273	0.651	+0.003	0.000	0.275
	mixed τ=0	0.008	0.575	0.583	0.190	−0.001	0.000	0.585

block of the mixture is a positive control: the basin weights are *exactly* right, so its population M is zero, and only the width inside each basin is wrong.

A negative statistic is not by itself a failure. An unbiased estimate of a non-negative quantity may be negative, so the sign carries no information on its own and we do not argue from it. What the table does show is that the statistic is not ordered by M: on the mixture it is *lower* at $M = 0.51$ than at $M = 0$, and the whole sweep sits inside the spread of the $M = 0$ row. To turn that into a claim about the diagnostic rather than about one estimator we calibrate it.

Read together, the two tables say something narrower and firmer than "KSD fails". At these sample sizes and over this range of kernel scales the test is powerful against an alternative the score can see, while its rejection rate against alternatives with M up to 0.79 never rises above 20% and is usually at or below the nominal 5%. That is a statement about discriminative power at a given budget, not a claim that no kernel or sample size could ever separate them: Gorham & Mackey (2017) give conditions under which the inverse-multiquadric KSD does detect non-convergence, and nothing here contradicts that. The mechanism for the gap is the one Bénard et al. (2023) identify – the Stein operator sees $\nabla \log \pi$ at the points q produces, and when basins are well separated that field is unchanged by how much weight π places on basins the sample never visits – which is why increasing n from 2000 to 8000 moves nothing.

L THE REFERENCE, AND WHAT UPGRADING IT COST

The competitor is not a straw man. A weak annealed importance sampler (fixed step, one MALA move per level, no resampling) fails so badly here – $\log Z$ off by 4.1 nats, never firing at all – that it would have suggested a $200\times$ advantage. We therefore use a full SMC sampler: geometric annealing, several MALA moves per level with dual-averaging step adaptation to 0.574 acceptance, and ESS-triggered systematic resampling. The cost accounting is generous to it, charging two target evaluations per particle per MALA move. Upgrading the reference reduced our measured advantage by an order of magnitude, and we report the post-upgrade numbers.

Table 11: Kernel Stein discrepancy against a known missing mass. The statistic is the U-statistic estimate of KSD^2 at $c = 1$, unbiased and reported signed: it can fall below zero when the true value is small relative to Monte-Carlo noise, and clamping it would hide exactly that. Five independent draw sets of 8000 points; sd is across those sets. The width control, whose M is zero, produces a statistic 39 to 449 times larger than anything the weight and support sweeps produce.

q	M	M_{supp}	M_{wt}	$10^4\,\widehat{\mathrm{KSD}^2}$	sd	$\widehat{M}$
GMM, $K = 8$						
tilt τ=1	0.0000	0.0000	0.0000	+1.11	1.69	0.0051
tilt τ=1.5	0.0737	0.0000	0.0737	+1.10	1.70	0.0763
tilt τ=2.5	0.1903	0.0000	0.1903	+1.22	1.76	0.1941
tilt τ=4	0.3198	0.0000	0.3198	+1.19	1.73	0.3220
tilt τ=8	0.5133	0.0000	0.5133	+0.84	1.03	0.5133
drop to 8 basins	0.0000	0.0000	0.0000	+0.86	1.90	0.0049
drop to 4 basins	0.3228	0.3228	0.0000	+1.31	2.19	0.3226
drop to 1 basins	0.7913	0.7913	0.0000	+1.83	2.68	0.7912
width $\times$1.25, weights exact	0.0000	0.0000	0.0000	+71.87	2.40	—
width $\times$1.5, weights exact	0.0000	0.0000	0.0000	+247.25	3.26	—
width $\times$2, weights exact	0.0000	0.0000	0.0000	+819.63	7.24	—
Many-Well, $K = 32$						
tilt τ=1	0.0000	0.0000	0.0000	−2.56	13.14	0.0056
tilt τ=1.5	0.2541	0.0000	0.2541	−2.18	11.21	0.2522
tilt τ=2.5	0.5010	0.0000	0.5010	−11.87	12.13	0.4997
tilt τ=4	0.5652	0.0000	0.5652	−10.07	10.99	0.5634
tilt τ=8	0.5709	0.0000	0.5709	−10.09	10.45	0.5692
drop to 32 basins	0.0000	0.0000	0.0000	−6.81	4.39	0.0076
drop to 16 basins	0.0295	0.0295	0.0000	−4.40	10.58	0.0295
drop to 4 basins	0.3336	0.3336	0.0000	−12.42	8.64	0.3321
drop to 1 basins	0.5710	0.5710	0.0000	−10.09	10.45	0.5692

Table 12: A calibrated level-0.05 KSD test. The null is $q = \pi$; the null distribution is obtained by reweighting the U-statistic matrix with multinomial weights (500 bootstrap replicates), and each entry is the rejection rate over 10 independent draw sets. Ranges are over kernel scales $c \in \{0.5, 1, 5\}$. Against every missing-mass alternative the rejection rate stays at or near the nominal level; against a 25% within-basin width error whose missing mass is zero the test rejects every time.

q	M	$n = 2000$	$n = 8000$
GMM, tilt τ=1	0.000	10%	0%
GMM, tilt τ=8	0.513	10%	0%
GMM, drop to 1 basin	0.791	10%	0%
GMM, width $\times$1.25	0.000	100%	100%
Many-Well, tilt τ=1	0.000	10%	20%
Many-Well, tilt τ=8	0.571	0%	0%
Many-Well, drop to 1 basin	0.571	0%	0%

M THE SELF-DIAGNOSTIC BY TARGET AND BUDGET

N THE SELF-DIAGNOSTIC OUT OF SAMPLE, AND THE FAILURE IT CANNOT SEE

§6 fixes three thresholds while looking at three closely related targets, so the clear rate and the post-clear error reported there are in sample. This appendix re-runs the audit on families that had

Table 13: Self-diagnostic by target and search budget, five seeds each. The verdict column counts how many of the five seeds were cleared.

target	restarts R	mean abs. error	cleared
Gaussian mixture	32	0.1955	0/5
Gaussian mixture	64	0.1330	1/5
Gaussian mixture	256	0.0183	1/5
Gaussian mixture	1024	0.0001	3/5
Many-Well	32	0.1434	0/5
Many-Well	64	0.0403	0/5
Many-Well	256	0.0014	1/5
Many-Well	1024	0.0014	5/5
Rotated Many-Well	32	0.1318	0/5
Rotated Many-Well	64	0.0370	0/5
Rotated Many-Well	256	0.0008	0/5
Rotated Many-Well	1024	0.0007	5/5

no part in choosing them: heterogeneous Many-Well at three heterogeneity levels, whose Laplace-bias spread can be dialled, and a mixture whose components differ in tail index. Five seeds and three budgets ($R \in \{64, 256, 1024\}$) per family. "Modes" is the number of local maxima the target actually has, computed by one-dimensional quadrature; it is not 2^m at the highest heterogeneity, where one coordinate's shallower minimum disappears.

Table 14: The self-diagnostic on target families that did not set its thresholds. "Laplace spread" is the ratio of largest to smallest per-basin harmonic bias, the η of Proposition 8 in directly measurable form. Errors are absolute errors of $\widehat{M}$. In the "cleared" count the bracketed number is the subset of cleared runs in which the search had in fact found every mode, so that a large error there cannot be blamed on enumeration.

	Laplace				worst abs. error	
family	spread	modes	runs	cleared	cleared	flagged
het. Many-Well 1	$12.6\times$	32	15	4 (4)	0.0086	0.0876
het. Many-Well 2	$32.4\times$	32	15	4 (4)	0.0192	0.0423
het. Many-Well 3	$64.7\times$	16	15	6 (6)	0.0172	0.0165
heterogeneous-tail mixture	—	10	15	4 (0)	0.1101	0.1605

Three different things happen, and the table keeps them apart.

Incomplete enumeration, silently. On the heterogeneous-tail mixture the search finds 6 of 10 modes at every budget we ran and the diagnostic clears the run anyway, with errors up to 0.110. This is not a tuning accident: the components sit at radius 14 with width of order one in $d = 16$, so the attraction probability per restart is small, and Proposition 7 says exactly that a mode missed at both budgets moves neither statistic. It is that proposition observed out of sample rather than a new failure, and it is the strongest single argument against reading a cleared verdict as a certificate.

Local-mass bias, with the search complete. On heterogeneous Many-Well every cleared run *has* found every mode, so all three conditions pass on their merits. The worst cleared error nonetheless doubles, from 0.0086 to 0.0192, as the harmonic bias goes from $12.6\times$ to $32.4\times$ – the two levels that keep the same 32 modes and so form a ladder – while the statistics do not move at all. Neither of them looks at a local mass, so neither can. The worst such run is on het. Many-Well 2: all 32 modes found, count ratio 1.000, singleton mass 0.000, every condition passed with room to spare, and the estimate off by 0.0192 – an order of magnitude above what the same budget delivers on the homogeneous targets of §7. No amount of extra search would have flagged it, because nothing was wrong with the search.

The partition and the mode decomposition stop agreeing. The highest-heterogeneity row is not a coverage failure and we are careful not to count it as one. There the asymmetry is strong enough that one coordinate keeps a single minimum, so the landscape has 16 modes and 16 basins of attraction, while the ground-truth partition we inherited from the other Many-Well variants still labels 32 cells by sign. The search finds all 16 modes; what the row measures is that the sign cells are no longer the basins, so a cell can carry mass with no local maximum inside it and $\widehat{M}$ and the reference are no longer estimating the same decomposition. A consistent reading requires fixing one basin definition on both sides, and under the attraction-basin definition this configuration is not a failure of the method at all. We report it because it marks where the harmonic description of the target stops being available, which is the boundary LJ-13 reaches from the other side (Appendix P).

The first two are not removable by tuning the three thresholds: one is one-sided by construction, the other involves a quantity that does not enter either statistic. The third is not a threshold question at all.

O How the cost scales with dimension

How the cost scales with dimension. The headline figure of a few percent of training cost is dimension-specific, and we measure how it degrades. Table 15 runs the check on Many-Well at $d = 16$ to 128 with the sampler and search budget fixed, reporting both target-evaluation counts and wall-clock. Two things happen. The accuracy is essentially flat, staying at 1–2 $\times 10^{-3}$ across the whole range, because the number of modes and the quality of the harmonic approximation are unchanged. The cost does move, close to linearly in d, because assembling a full Hessian costs $O(d)$ backward passes. Wall-clock grows faster than the evaluation count (18.7% versus 11.6% at $d = 128$), which is the concrete sense in which a Hessian evaluation and a gradient evaluation are not interchangeable units, and a reason to prefer the wall-clock column when judging the method.

Table 15: Cost of the check as the dimension grows, Many-Well with $m = 5$ (32 modes), $R = 256$, three seeds. Accuracy is flat; cost grows about linearly in d because a full Hessian needs $O(d)$ backward passes.

d	target evals (% of training)	wall-clock (% of training)	abs. error
16	2.69%	2.7%	0.0012
32	3.96%	4.4%	0.0017
64	6.49%	9.2%	0.0017
128	11.57%	18.7%	0.0017

P Where the method stops: LJ-13

Every target above has few, well-separated, non-degenerate modes. We now stress the two load-bearing assumptions on the 13-atom Lennard-Jones cluster, whose landscape has on the order of 1500 local minima and whose global minimum is the Mackay icosahedron at $E = -44.3268$ (Wales & Doye, 1997). We use the check's own search machinery, not a better optimiser, because the question is whether the check copes.

The search succeeds. From random compact configurations, 98% of 6000 restarts converge to $\|\nabla E\| < 10^{-4}$, and all five independent enumerations recover the global minimum exactly, -44.3268, matching the literature value. They find between 413 and 446 distinct minima at 6000 restarts, against a landscape holding on the order of 1500.

Continuous symmetry breaks the local mass. At the global minimum the Hessian has exactly six zero eigenvalues – three translations, three rotations – with the next eigenvalue at 42.65 and nothing in between. Equation (3) is therefore undefined on any target with continuous symmetry, which includes essentially every molecular system. The repair is standard in chemical physics: take the product over the $3N-6$ non-zero modes and treat the zero modes through the rotational partition

Table 16: LJ-13. Concentration of the harmonic weight and the fraction of it captured by a cheap search, as a function of inverse temperature. Mean ± s.d. over five independent enumerations of 6000 restarts each. Note the standard deviations in the last two columns: below $\beta = 8$ the captured fraction is not merely small but erratic.

β	heaviest mode	#modes 90%	#modes 99%	captured by 500	by 2000
1	$57.5 \pm 24.3\%$	6.8 ± 3.3	15.2 ± 7.3	$3.0 \pm 2.9\%$	$23.4 \pm 26.1\%$
2	$52.4 \pm 25.2\%$	7.2 ± 3.5	14.4 ± 6.9	$3.5 \pm 2.9\%$	$18.8 \pm 18.6\%$
3	$64.5 \pm 22.9\%$	5.0 ± 2.9	12.0 ± 5.9	$3.7 \pm 4.0\%$	$13.1 \pm 10.9\%$
5	$81.9 \pm 15.7\%$	2.0 ± 0.6	3.2 ± 1.6	$63.3 \pm 33.3\%$	$64.0 \pm 33.7\%$
8	$100.0 \pm 0.0\%$	1.0 ± 0.0	1.0 ± 0.0	$100.0 \pm 0.0\%$	$100.0 \pm 0.0\%$

function and the point-group order, as in the harmonic superposition approximation (Wales & Doye, 1997). With this correction all 426 usable minima admit a well-defined mass.

There is no usable multimodal window on LJ-13. A natural request at this point is an end-to-end demonstration on LJ-13: train a sampler, let it collapse, catch it. We cannot honestly supply one. At $\beta \geq 8$, where the enumeration is stable, the target is effectively unimodal: the global minimum carries essentially all of the harmonic weight our enumeration resolves, so there is no collapse to detect. At $\beta \leq 5$, where several minima genuinely share the weight, the captured fraction is $63 \pm 33\%$ or worse and the check must not be used. The two conditions do not overlap on this system. Our end-to-end evidence with a modern sampler class is therefore the controlled-SDE audit of §7, on Many-Well, which is one of the standard targets of the neural-sampler literature (Blessing et al., 2024; Chen et al., 2025; Rissanen et al., 2025; Guo et al., 2026); a demonstration on a rugged physical landscape remains open.

A hypothesis we had to discard. We expected the weight to concentrate so quickly that enumerating the heavy modes would suffice regardless of temperature. The middle columns of Table 16 appear to support this – eight modes hold 90% of the weight even at $\beta = 1$ – but the last two columns refute it: a 500-restart search captures only 3.2%. The heaviest modes at high temperature are the entropically favoured, low-curvature ones, and having a large harmonic weight is not the same as being easy for an optimiser to fall into. We had implicitly assumed the two coincide. They do not.

Enumeration is sufficient only at low temperature. Relative harmonic weights need only $\log w_j = -\beta E_j - \frac{1}{2}\sum \log \lambda_j^{\neq 0}$, so one enumeration serves every β. Table 16 reports how much of the weight a cheap search captures. At $\beta \geq 8$, within our 6000-restart harmonic reference, essentially all discovered weight lies on the global minimum, which the first 500 restarts already find, and the check is applicable. At $\beta \leq 3$ a 500-restart search captures only 3–4% of the weight and the check must not be used. Two further points sharpen this. First, the figures are optimistic: the denominator is the 6000-restart set, not the true $\sim$1500 minima. Second, and more damning for practice, the capture fraction is not merely small but erratic – at $\beta = 1$ a 2000-restart search captures $23 \pm 26\%$, and even at $\beta = 5$, where the mean looks usable, the spread is $63 \pm 33\%$. A quantity with that variance cannot be acted on regardless of where its mean sits, so we place the usable boundary at $\beta \geq 8$ rather than at the point where the mean first looks acceptable.

Q LIMITATIONS

The method rests on assumptions that our own experiments made load-bearing rather than decorative. **(A1)** Modes must be reachable by multi-start *second-order* search; first-order search fails at condition number 20, which is mild by physical standards. **(A2)** Discovered modes must admit a usable local mass; continuous symmetries require the zero-mode treatment of Appendix P, and anharmonicity that differs *between* modes breaks the cancellation that (1) relies on – measured in Appendix N, where dialling the per-basin harmonic bias from $12.6\times$ to $64.7\times$ degrades the estimate while the search stays complete. **(A2′)** The partition used for the estimate and the partition used for the reference must be the same object. They are basins of attraction on one side and, on the

synthetic targets, sign cells on the other; the two coincide only while every cell contains a mode. At the highest heterogeneity one coordinate keeps a single minimum and they part company (Appendix N), which is a mismatch of definitions rather than a coverage failure, and the numbers there should not be read as one. **(A3)** The estimate is conditional on the enumerated landscape; under incomplete enumeration its error has no fixed sign (Proposition 6), which is what makes the accompanying search-completeness diagnostic essential rather than decorative. **(A4)** The cost claim is dimension-limited: measured over $d = 16$ to 128 the check grows about linearly in d and would reach parity with training near $d \approx 700$ in wall-clock, so "a few percent" describes present benchmark scales and not the largest problems a neural sampler might be pointed at. All targets used for the quantitative estimator-accuracy experiments have exactly computable ground truth, which was a deliberate choice; behaviour on a rugged physical landscape is characterised only through the LJ-13 stress test, where no exact enumeration is available.

The weight-collapse evidence is narrower than the rest. Our exactly constructed families cover two targets, not three: the tilted Many-Well sampler is built coordinate-wise, which is correct only when the basin structure is axis-aligned, and applying it to the rotated target produced a sampler that does not match π at all (its ESS is 10^{-4} where the construction should be exact). We exclude those runs rather than report them. The paired controlled-SDE result rests on two seeds of one architecture, and establishes that weight collapse occurs, not how often.

The self-diagnostic covers one of the two error channels. Both of its statistics are functions of the discovered mode set (Proposition 7), so neither responds to local-mass bias. A run can find every mode at every sub-budget, clear all three conditions, and still be wrong through (A2). Its thresholds are also uncalibrated: they were chosen on the targets of §7, and Appendix N reports what they do elsewhere.

Two approximations are measured rather than justified. The coverage step partitions space by Mahalanobis nearest mode while the ground truth partitions it by basin; the two agree to four decimal places on our targets (Appendix H) and nothing guarantees they would on a basin that is not close to ellipsoidal. And the Laplace mixture of Appendix G does not place weight $\hat{p}_j$ on basin B_j: its components are Gaussians with full support, so its basin weights differ from $\hat{p}$ by whatever tail mass crosses the boundaries, measured at a total variation of 0.0035 here. Every statement we make about that mixture, including that it is on these benchmarks the better mode-level sampler – which limits what such targets can establish about neural samplers, though it leaves the check's own claim intact – is accurate only to that order.

The empirical split of M is sample-size dependent. We call a basin unvisited when no draw of q lands in it, so at 4×10^4 draws a basin holding 10^{-5} of q's mass is indistinguishable from an empty one. The decomposition (2) is exact as stated, but a basin visited with vanishing probability falls under M_{wt} by our convention and under M_{supp} under a stricter one. Nothing in the estimator depends on the split; only the attribution does.

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.