REGIONAL MASS AUDITS FOR CORRECTING NEURAL SAMPLERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural samplers can generate samples from every mode of a target distribution while producing them in the wrong proportions. We propose regional mass auditing to guide correction using target probabilities or reference data. The audit selects an under-represented region and estimates its probability deficit on data held out from selection. This deficit guides where to obtain additional samples and which examples to use for fine-tuning. Valid probability bounds determine how many samples can be added without over-filling the region at a specified confidence level. When the additional samples follow the target distribution conditioned on the region, the same bounds give a computable lower bound on the reduction in Kullback-Leibler divergence. We compare audit-guided and untargeted correction at matched sampling or training budgets on synthetic distributions, particle systems, and proteins. Audit-selected simulation starts improve particle-distance and protein free-energy estimates. Fine-tuning neural samplers with selected data improves regional probabilities in newly generated samples. For proteins, fine-tuning coarse-grained potentials with selected reference structures also improves several structural and free-energy estimates.

1 INTRODUCTION

Neural samplers learn to generate samples from a target probability distribution (Midgley et al., 2023; Akhound-Sadegh et al., 2024; Noble et al., 2025). These samples are often used to estimate quantities that would otherwise require expensive computation. In Bayesian inference, for example, the fraction of samples satisfying an event estimates its probability. In molecular modeling, the relative frequencies of different structures determine estimates of their relative free energies (Blessing et al., 2024; Majewski et al., 2023). These calculations require the model to generate different outcomes in the correct proportions. If those proportions are wrong, generating more samples estimates the model's probabilities more precisely without making them closer to the target.

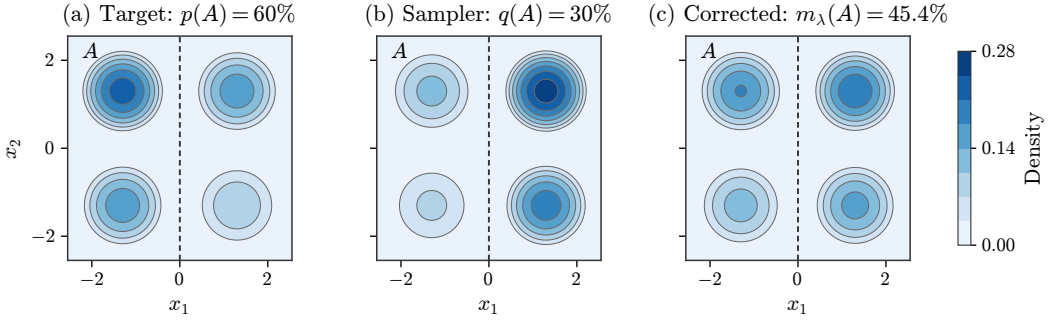

Figure 1: The sampler visits all 4 modes but under-represents $A = \{x_1 < 0\}$ (left of the dashed line). Correction adds samples from $p(\cdot \mid A)$ with weight $\lambda = \hat{d}_A = 0.22$. Illustrative densities share a color scale.

Visiting every separated high-probability region, or mode, does not ensure correct relative frequencies (Figure 1). A model for generating protein structures may produce both folded and unfolded

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.

structures, yet sample them in proportions that imply an incorrect estimate of folding stability (Majewski et al., 2023; Lewis et al., 2025). The same issue can arise when structures are obtained by simulating trajectories with a learned model or by running dynamics under a learned coarse-grained (CG) potential (Jing et al., 2024; Majewski et al., 2023).

Previous work has established the importance of recovering mode probabilities (Blessing et al., 2024; Grenioux et al., 2025) and developed tests that identify distributional discrepancies (Lopez-Paz & Oquab, 2017; Kübler et al., 2022). We study trained models used to generate samples or initialize simulations, with target probabilities or reference data available to guide correction. A reference trajectory, for example, can guide additional protein simulations or provide training structures for a learned potential. A model may reproduce some regions accurately while assigning too little probability to others. This motivates directing additional sampling and training data toward under-represented regions. We test whether this selection improves the use of a fixed sampling or training budget.

Our audit selects an under-represented region A where corrective samples can be obtained. It estimates the difference between the target and model probabilities of A, called its probability-mass deficit, using data held out from selection. Confidence bounds on these probabilities give finite-sample limits on how many samples can be added without over-filling A, under the sampling conditions in Section 4.1. Target bounds can come from numerical integration or independent target samples, for example. With molecular dynamics (MD) references, consecutive frames are often similar and some states may be rarely visited. We adjust audit estimates for frame correlations and evaluate correction against held-out reference data.

Our contributions are: (1) a regional audit that uses probability deficits to select corrective samples; (2) finite-sample limits on how many samples to add, together with a computable bound on the reduction in Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951) under conditional target sampling; and (3) matched-budget experiments testing whether regional selection improves corrected sample sets and new outputs from updated models.

2 Related work

Evaluating neural samplers. Recent benchmarks emphasize whether samplers recover correct mode probabilities (Blessing et al., 2024; Grenioux et al., 2025). Kernel Stein discrepancy (KSD) and maximum mean discrepancy (MMD) compare distributions (Liu et al., 2016; Gorham & Mackey, 2017; Gretton et al., 2012), although KSD can have low sensitivity to incorrect weights on separated modes (Liu et al., 2023). Precision/recall and coverage measures assess sample quality and diversity (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020). Other tests identify where distributions differ through learned test locations or classifiers (Jitkrittum et al., 2016; 2017; Lopez-Paz & Oquab, 2017; Kübler et al., 2022). Classifier-defined regions also permit target and model probabilities to be compared on held-out data. We guide correction by selecting regions where new samples can be obtained and estimating their probability deficits.

Statistical guarantees after region selection. Choosing a region with a large observed discrepancy can also select random sampling fluctuations. Sample splitting addresses this problem by using independent data for selection and evaluation (Cox, 1975; Wasserman & Roeder, 2009). Alternatives account explicitly for the selection process (Fithian et al., 2014) or construct bounds that hold for all candidates simultaneously (Berk et al., 2013). We use splitting to allow searches over partitions and combinations of cells: only the selected region is evaluated on held-out data, without requiring a separate analysis of each search rule.

Corrective sampling and training. Correction can reweight existing samples or generate new ones. Importance weighting changes their contributions to an estimate but cannot add samples to an empty region (Vehtari et al., 2024; Huang & Joseph, 2023). Annealed importance sampling (AIS) and transport methods obtain new samples (Neal, 2001; Hertrich & Gruhlke, 2024; Grenioux & Noble, 2026). Adaptive molecular simulation chooses starting structures for additional trajectories, using structural objectives in FAST and learned rewards in REAP (Zimmerman & Bowman, 2015; Shamsi et al., 2018; Kleiman et al., 2023). Correction also enters training: FAB uses AIS during model updates (Midgley et al., 2023), while AdaGAN establishes divergence improvement for

optimal and sufficiently accurate approximate mixture components (Tolstikhin et al., 2017). Here, finite-sample probability bounds remain valid after region selection and determine an allowable correction weight. Under conditional target sampling, they also give a computable KL improvement. The selected region can additionally guide training-data selection. Appendix A.2.4 connects classifier accuracy to regional probability differences.

3 PROBLEM SETUP

Let p be the target distribution and q the distribution of model-generated samples on a sample space $\mathcal{X}$. We use target probabilities or reference data to assess q. Test data are used only for evaluation.

To compare regional probabilities, we partition the part of the space being audited into nonoverlapping cells $C_1, \ldots, C_K$. Their union $C = \bigcup_{i=1}^{K} C_i$ is the audited domain. The partition may be specified in advance or constructed from data. When C covers only part of the sample space, we compare probabilities conditional on a sample falling in C. We write these distributions as $p_C = p(\cdot \mid C)$ and $q_C = q(\cdot \mid C)$, assuming $p(C), q(C) > 0$. When $C = \mathcal{X}$, they reduce to p and q. A candidate audit region A is a union of cells, with probability-mass deficit $d_A = p_C(A) - q_C(A)$. We seek regions with positive deficits, where the model assigns less probability than the target.

4 REGIONAL AUDIT AND CORRECTION

The audit guides where and how much to sample, and selects corrective training examples for fine-tuning.

4.1 REGION SELECTION AND PROBABILITY BOUNDS

We first estimate the target and model probabilities of each cell and combine cells with positive estimated deficits into A. Any data used to construct or choose the partition are also used only for selection. Because this search can select random fluctuations as well as genuine deficits, we assess the selected region using independent data. The domain, partition, and A are fixed before these assessment data are examined.

On the assessment data, we compute a lower confidence bound $L_p(A)$ on the target probability $p_C(A)$ and an upper confidence bound $U_q(A)$ on the model probability $q_C(A)$. These opposite directions account for uncertainty: lowering the target probability and raising the model probability both reduce the apparent shortage. When both bounds hold, their difference, $\hat{d}_A = L_p(A) - U_q(A)$, is therefore no larger than the true deficit d_A. A positive $\hat{d}_A$ establishes a shortage of at least that size; a nonpositive value is inconclusive.

Let n_p, n_q be the target and model assessment sample sizes, and k_p, k_q their counts in A. For independent samples, these counts are binomial. We use one-sided Clopper–Pearson bounds (Clopper & Pearson, 1934), with allowed error probability $\alpha \in (0, 1)$. Write $\mathrm{Beta}^{-1}(u; a, b)$ for the u quantile of a beta distribution with parameters a, b.

Theorem 1 (Probability bounds after region selection)**.** *Conditional on all selection data and decisions, suppose the assessment sets are i.i.d. from p_C, q_C, respectively, with $n_p, n_q \geq 1$ fixed before assessment. The bounds*

$$L_p(A) = \mathrm{Beta}^{-1}(\alpha/2; k_p, n_p - k_p + 1) \qquad (k_p > 0),$$
$$U_q(A) = \mathrm{Beta}^{-1}(1 - \alpha/2; k_q + 1, n_q - k_q) \quad (k_q < n_q),$$

with $L_p(A) = 0$ when $k_p = 0$ and $U_q(A) = 1$ when $k_q = n_q$, satisfy

$$\mathbb{P}\Big(d_A \geq \hat{d}_A\Big) \geq \mathbb{P}(L_p(A) \leq p_C(A),\ q_C(A) \leq U_q(A)) \geq 1 - \alpha. \tag{1}$$

Each bound has conditional failure probability at most $\alpha/2$. A union bound and averaging over selection give Equation 1 (Appendix A.1). The held-out assessment evaluates the selected region A, so candidate searches on the selection data need no separate error allowance for each candidate. The same coverage holds for any target lower bound with conditional failure probability at most

$\alpha/2$. Alternatives use known probabilities, numerical integration, or independent importance samples with a known weight bound and an empirical-Bernstein inequality (Maurer & Pontil, 2009) (Appendix A.2). For MD references, we adjust for frame correlations and check false-alarm rates between reference subsets, reporting empirical calibration (Appendix D.4.1).

4.2 Correction through additional sampling

We keep the original model fixed and use the selected region A to obtain additional samples. When conditional target sampling is available, we draw from the target distribution restricted to A. Otherwise, we sample within its cells or start simulations from structures in them, allocating samples according to the estimated cell deficits or uniformly across the selected cells.

Let r be the distribution of the additional samples on C. When simulations are used, this is the distribution of their recorded outputs, which may fall outside the starting cells. Retaining N_{out} original samples and adding B new ones assigns a fraction $\lambda = B/(N_{\mathrm{out}} + B)$ to the new samples and $1 - \lambda$ to the original ones. The corresponding mixture distribution is

$$m_\lambda = (1 - \lambda)q_C + \lambda r. \tag{2}$$

The original samples are kept, but their share of the combined set falls as B increases. The probability of the selected region is therefore $m_\lambda(A) = (1 - \lambda)q_C(A) + \lambda r(A)$. For $\lambda > 0$, this probability increases when the additional samples fall in A more often than the original model's outputs.

How many samples to add. To avoid over-filling A, the increase in its probability must not exceed the original deficit. Equation 1 bounds this deficit from below by $\hat{d}_A$, with probability at least $1 - \alpha$. The mixture changes the probability by $\lambda[r(A) - q_C(A)]$, which is at most λ because $r(A) \le 1$ and $q_C(A) \ge 0$. Choosing $0 \le \lambda \le \hat{d}_A$ therefore gives, whenever the audit bounds hold,

$$m_\lambda(A) - q_C(A) \;\le\; \lambda \;\le\; \hat{d}_A \;\le\; p_C(A) - q_C(A).$$

Hence $m_\lambda(A) \le p_C(A)$, so the correction does not over-fill A. Substituting $\lambda = B/(N_{\mathrm{out}} + B)$ gives the sample limit $B^\star = \lfloor N_{\mathrm{out}}\hat{d}_A/(1 - \hat{d}_A)\rfloor$, for $0 < \hat{d}_A < 1$. No samples are added when $\hat{d}_A \le 0$.

A larger correction uses both probability bounds. When $q_C(A) \le U_q(A)$ and $L_p(A) \le p_C(A)$, even if every new sample lies in A, its corrected probability is at most $(1 - \lambda)U_q(A) + \lambda$. Requiring this upper bound to stay below the target lower bound gives

$$(1 - \lambda)U_q(A) + \lambda \le L_p(A) \quad\Longleftrightarrow\quad \lambda \le \frac{\hat{d}_A}{1 - U_q(A)},$$

where $0 \le U_q(A) < L_p(A) < 1$. Substituting $\lambda = B/(N_{\mathrm{out}} + B)$ gives the larger sample limit $B_{\mathrm{end}} = \lfloor N_{\mathrm{out}}\hat{d}_A/[1 - L_p(A)]\rfloor$ (Appendix A.3). Both limits control the mixture probability; observed sample frequencies also vary with sampling.

After correction, we keep A fixed and assess the mixture using independent samples. A positive deficit bound on A identifies a remaining shortage; a positive bound on $C \setminus A$ identifies over-filling in A. Appendix A.5 gives the re-audit procedure and error control.

When correction also reduces overall error. The budget controls the probability assigned to A. If the additional samples follow the target distribution conditioned on A, written $r = p_C(\cdot \mid A)$, correction also reduces KL divergence over C. In this case, the probability of A moves toward its target value, the conditional KL within A cannot increase, and the relative probabilities outside A remain unchanged. The same audit bounds give a lower bound on the reduction. For probabilities a and b, we use $\mathrm{KL}(a\|b)$ as shorthand for the KL divergence between $(a, 1 - a)$ and $(b, 1 - b)$.

Proposition 1 (KL improvement)**.** *Assume Equation 1 and finite* $\mathrm{KL}(p_C\|q_C)$*. With probability at least* $1 - \alpha$*, whenever* $0 < U_q(A) < L_p(A) < 1$*, correction with* $r = p_C(\cdot \mid A)$ *satisfies, for every* $0 < \lambda \le \hat{d}_A/[1 - U_q(A)]$*,*

$$\mathrm{KL}(p_C\|q_C) - \mathrm{KL}(p_C\|m_\lambda) \ge L_p(A)\log\left(1 + \frac{\lambda}{(1 - \lambda)U_q(A)}\right) + \log(1 - \lambda) > 0. \tag{3}$$

At the largest allowed weight, the bound equals $\mathrm{KL}(L_p(A)\|U_q(A))$. Appendix A.4 gives the proof.

4.3 TRAINING WITH CORRECTIVE SAMPLES

We use the selected region to choose a corrective training dataset $\mathcal{D}_{\text{corr}}$ and fine-tune the model. These examples are obtained from target samples or existing reference data. We then generate new samples using the updated model to evaluate the effect of training. Write θ for the trainable parameters, θ_0 for the fixed original parameters, and q_0 for the original model's output distribution.

Normalizing flows. For the released FAB normalizing flow (Midgley et al., 2023), we can evaluate the density $q_\theta(x)$ and fine-tune it by maximum likelihood. We also fit samples from q_0 to retain information about regions outside A. The coefficient $\eta \in (0, 1)$ weights the corrective-data term:

$$\mathcal{L}_{\text{flow}}(\theta) = \frac{\eta}{|\mathcal{D}_{\text{corr}}|} \sum_{x \in \mathcal{D}_{\text{corr}}} [-\log q_\theta(x)] + (1-\eta)\, \mathbb{E}_{x \sim q_0}[-\log q_\theta(x)]. \tag{4a}$$

The first term fits the corrective data; the second fits the original model's distribution. Each update estimates them using a corrective mini-batch and new samples from q_0.

Diffusion models. We fine-tune the iDEM diffusion sampler (Akhound-Sadegh et al., 2024) with the denoising score-matching loss $\mathcal{L}_{\text{DSM}}$ on corrective examples. To limit changes in its predictions on original-model samples, we add a penalty $\mathcal{R}_{\text{anchor}}$: the squared difference between the updated and frozen networks' score predictions on the same noise-perturbed inputs. We minimize

$$\mathcal{L}_{\text{diff}}(\theta) = \eta\, \mathcal{L}_{\text{DSM}}(\theta; \mathcal{D}_{\text{corr}}) + (1-\eta)\, \mathcal{R}_{\text{anchor}}(\theta; \theta_0). \tag{4b}$$

The penalty uses a fixed pool of original-model samples, with added noise at each update. Appendix E.1 gives the full objective and the original model's noise schedule and loss weights.

CG potentials. For proteins, we fine-tune the general CG potential of Majewski et al. (2023), which generates samples through molecular simulation. Here $\mathcal{D}_{\text{corr}}$ consists of selected reference structures, and training updates the neural part of the dimensionless energy E_θ while keeping its physical term fixed. The gradient of the negative log likelihood compares energy gradients on reference structures with those on samples from the current potential. We approximate the latter using b_- structures x_j^- from Brownian simulation chains. The estimated gradient is

$$g_{\text{CG}}(\theta) = \frac{1}{|\mathcal{D}_{\text{corr}}|} \sum_{x \in \mathcal{D}_{\text{corr}}} \nabla_\theta E_\theta(x) - \frac{1}{b_-} \sum_{j=1}^{b_-} \nabla_\theta E_\theta(x_j^-). \tag{5}$$

We take gradient-descent steps using this estimate, evaluating the first average on a reference mini-batch and holding simulated coordinates fixed during differentiation. The finite chains make this an approximate likelihood update; Appendix E.3 gives the derivation and training details. The loss coefficient η controls the strength of corrective fitting, while λ in Section 4.2 controls the fraction of added samples. The guarantees in that section concern the explicit sample mixture.

5 EXPERIMENTS

Models and data. We use two synthetic sampling benchmarks: GMM-40, a two-dimensional mixture of 40 Gaussian components (Akhound-Sadegh et al., 2024), and Many-Well, a 32-dimensional distribution formed from 16 independent two-dimensional blocks, each containing a double-well coordinate with two energy minima (Midgley et al., 2023). These targets allow us to evaluate regional probability errors against known or numerically bounded target probabilities. We evaluate the diffusion samplers iDEM and pDEM on GMM-40, and FAB normalizing flows on GMM-40 and Many-Well. The Many-Well study uses a released checkpoint. Particle-system experiments use DW4, 4 particles in 2 spatial dimensions with a double-well interaction, and LJ13, a 13-particle Lennard–Jones system (Akhound-Sadegh et al., 2024). The DW4 supplementary-sampling study uses a newly generated Markov chain Monte Carlo (MCMC) reference under the benchmark energy; model-update studies use the references released with the original implementation. For proteins, we evaluate and fine-tune the general CG potential of Majewski et al. (2023), using protein-specific simulations as references. Additional audits examine BioEmu (Lewis et al., 2025), the 100 tetrapeptides in MDGen's test set (Jing et al., 2024), and the FAB alanine-dipeptide (ALDP) model.

Baselines. We compare the original model, untargeted correction, and audit-guided correction. Within each experiment, targeted and untargeted correction share the underlying sampling or training procedure and matched nominal budgets, but differ in how corrective samples are selected or allocated. Further controls separate region selection from within-region weighting, and corrective training from training on original-model samples alone. Reference data are treated as already available, so budgets count added samples or training examples, excluding reference-generation costs. The DW4 compute-matched comparison also accounts for basin search.

Evaluation metrics. We measure errors in regional probabilities and use L_1 and L_2 distances, or total variation (half the L_1 distance), to compare probabilities across a partition. DW4 correction is evaluated by the 1-Wasserstein distance (W_1) (Peyré & Cuturi, 2019) between particle-pair distance distributions, with errors on a separate evaluation partition and energy distributions as additional metrics. For CG proteins, the primary metric is free-energy profile error along the radius of gyration R_g, a measure of structural compactness. We report the mean absolute difference in log bin probabilities, denoted MAE_F, in thermal energy units (kT). Energy distance (ED) (Székely & Rizzo, 2013) on pairwise alpha-carbon (CA) distances and distributions of root-mean-square deviation (RMSD) from an experimental structure provide additional structural evaluations. Free-energy and structural comparisons follow the evaluation approach of Majewski et al. (2023); the R_g bins, smoothing, and averaging rule used here are specified in Appendix D.4.2.

5.1 Regional probability errors in trained samplers

On GMM-40, iDEM visits all 40 modes but assigns a group of 4 modes 6.8% probability instead of the target 10%. The audit identifies this shortage and quantifies it in probability units. Table 1a reports regional deficit bounds alongside KSD and validation 2-Wasserstein distance (W_2) (Peyré & Cuturi, 2019). Controlled Gaussian-mixture tests assess the reliability of the audit, with observed false alarms of at most 1% at nominal level $\alpha = 0.1$ (Appendices B and C.1).

For peptide CKHT, the reference and MDGen backbone-angle distributions share a prominent peak near the polyproline-II (PPII) conformation, but MDGen has much lower density near the left-handed alpha-helical (α_L) conformation (Figure 2). Across the 100 test peptides, 27 have at least one reference-observed state with no generated samples, which reweighting alone cannot recover. Appendix C gives the state definitions and molecular protocols.

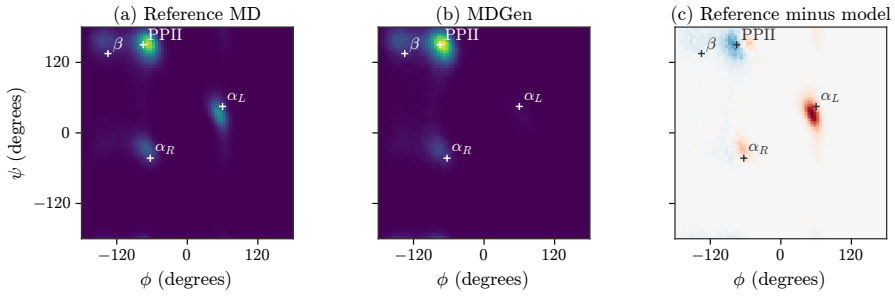

Figure 2: CKHT backbone torsion-angle distributions, pooling (ϕ, ψ) across residues. Both distributions show a prominent PPII peak, but MDGen under-represents the α_L region. In the reference-minus-model panel, red indicates higher reference density and blue higher MDGen density.

Regional comparisons also distinguish model revisions and existing corrections: BioEmu v1.2 has lower state-probability errors than v1.1, and AIS brings the ALDP flow's estimated probability in A closer to the MD reference (Table 1b,c).

5.2 Audit-guided sample correction

With the model fixed, we test whether regional selection improves supplementary sampling. GMM-40 and Many-Well check the budget rule and independent re-audit. DW4 and CG proteins test structural and free-energy improvements. Energy minimization and nearest-minimum assignment define the simulation basins. We compare equal allocation across all basins (*uniform-all*), equal

Table 1: Regional diagnostics. (a) GMM-40: bootstrap deficit lower bounds; bold marks lower KSD and W_2. (b) BioEmu reports ranges of $|\log(\hat{p}/\hat{q})|$, where $\hat{p}$ and $\hat{q}$ are reference and model state frequencies; the control compares subsets of 25 reference runs. (c) ALDP estimates include standard errors; brackets give the reference's 90% block-bootstrap interval.

(a) GMM-40

criterion	iDEM	pDEM
KSD ↓	0.057	**0.045**
deficit bound (%)	0.8	2.3
validation W_2 ↓	**3.44**	5.05

(b) BioEmu

samples	error (kT) ↓
v1.1	3.3–5.2
v1.2	1.3–1.6
ref. control	0.5–0.9

(c) FAB / ALDP

estimator	mass in A (%)
direct flow	65.437 ± 0.075
weighted AIS	65.807 ± 0.076
reference	[65.633, 65.833]

allocation across flagged basins (*uniform-flagged*), and allocation proportional to their estimated deficits (*deficit-weighted*). The three allocations share the nominal chain budget, chain length, and dynamics, separating region selection from within-region weighting.

Controlled correction and re-auditing. On GMM-40 and Many-Well, we set $B^\star$ using held-out probability bounds, add samples restricted to A, and independently re-audit the mixture. The re-audit measures the remaining deficit and checks for over-filling. GMM-40 tests iDEM, pDEM, and FAB, including larger corrections to assess over-filling detection; B_{end} would permit roughly twice as many samples. On Many-Well, no over-filling is detected at $B^\star$, and 4-cell L_2 error is 24% lower than with equally many unconditional target samples (Appendices D.1 and D.2).

DW4: allocating additional simulations. We retain 4000 iDEM outputs and select among them to initialize simulations with the Metropolis-adjusted Langevin algorithm (MALA; Roberts & Tweedie, 1996). No reference structures are added, and trajectories can leave their starting regions. We evaluate 4 empirically chosen chain counts against a separate reference subset (Figure 3). At 1172 chains of 200 steps, uniform-flagged reduces pair-distance W_1 and evaluation-partition L_1 errors by 49% each relative to uniform-all. Deficit weighting further lowers pair-distance and energy errors, with little additional change in partition error. Most of the gain therefore comes from region selection. Adding more chains eventually reverses the advantage on pair-distance and partition errors.

Table 2 tests low-count allocation, which favors basins visited less often by the model, and random selection of the same number of basins as the audit. Selecting flagged regions outperforms both, showing a benefit beyond favoring low counts or concentrating chains at random. A compute-matched control spends the basin-search allowance on extra untargeted chains.

Table 2: DW4 correction with 200 MALA steps per chain (mean ± SD, 10 simulation seeds), with original-model samples fixed. Bold marks the lowest mean error per metric. Compute matching includes basin search but excludes reference generation and pilot checks.

allocation	added chains	pair-distance W_1 ↓	partition L_1 ↓	energy W_1 ↓
original model	0	0.2034	0.3655	0.1323
uniform-all	1172	0.1567 ± 0.0023	0.2759 ± 0.0031	0.1277 ± 0.0127
low-count allocation	1172	0.1414 ± 0.0021	0.2433 ± 0.0069	0.1291 ± 0.0115
random selection	1172	0.1594 ± 0.0381	0.2833 ± 0.0747	0.1299 ± 0.0378
uniform-flagged	1172	0.0803 ± 0.0024	0.1420 ± 0.0077	0.0753 ± 0.0148
deficit-weighted	1172	$\mathbf{0.0690} \pm 0.0016$	$\mathbf{0.1395} \pm 0.0054$	$\mathbf{0.0428} \pm 0.0079$
compute-matched model	15,422	0.1653 ± 0.0017	0.3237 ± 0.0032	0.1114 ± 0.0095

CG proteins: structural accuracy after correction. We apply the same three allocations under the released general CG potential. Reference data guide region selection, while the simulations generate the additional structures. Selection and scoring use separate reference subsets. A model-seeded control starts from original-model samples. We compare final structures at matched nominal chain budgets.

On chignolin, uniform-all correction lowers free-energy error from 3.358 to 1.496 kT at $B = 4000$. Selecting flagged basins gives a further 10% reduction, and deficit weighting another 4% (Table 3a).

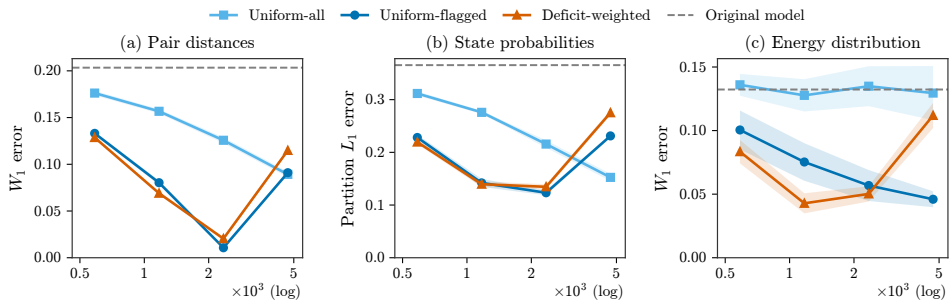

Figure 3: DW4 correction as a function of the number of added MALA chains. Each chain runs for 200 steps and contributes its final structure to the original sample set. Curves show mean errors across 10 simulation seeds, with shaded bands indicating ± 1 standard deviation. The selected region and reference data are held fixed across seeds. Dashed lines show errors before correction; lower values are better (Appendix D.3).

Table 3: Chignolin correction at nominal $B = 4000$ (mean $\pm$ SD, 5 seeds per evaluation). (a) Held-out errors; lowest means are bold. (b) Paired CA-RMSD error differences (deficit-weighted minus uniform-all); negative values indicate improvement. Allocation excludes RMSD.

(a) Regional-allocation comparison

allocation	MAE_F (kT) $\downarrow$	ED $\downarrow$	partition L_1 $\downarrow$
model-seeded	3.401 ± 0.031	11.201 ± 0.075	1.223 ± 0.001
uniform-all	1.496 ± 0.042	6.082 ± 0.272	0.907 ± 0.027
uniform-flagged	1.351 ± 0.068	4.982 ± 0.271	0.846 ± 0.028
deficit-weighted	$\mathbf{1.297} \pm 0.139$	$\mathbf{4.712} \pm 0.183$	$\mathbf{0.841} \pm 0.028$

(b) RMSD evaluation

metric	difference
histogram error	-0.041 ± 0.039
MAE_F (kT)	-0.095 ± 0.079
W_1 (Å)	-0.104 ± 0.030

Thus simulations started from the discovered basins improve the original sample set; selecting under-represented basins provides a further improvement, with a smaller additional gain from deficit weighting. Energy distance and partition error follow the same ordering. Deficit weighting also improves agreement with the reference distribution of alpha-carbon RMSD from PDB 5AWL, an observable excluded from allocation (Table 3b). On villin, it lowers partition error, while uniform-all has slightly lower free-energy error at the largest budget. Tables 12 and 13 give the full chignolin budget and RMSD comparisons.

5.3 Model updates from audit-selected data

To test whether regional selection improves fine-tuning, we compare updates with the same loss, initial parameters, optimizer, and nominal data budget. Only the selection of corrective training examples differs between the targeted and untargeted updates. New samples are evaluated against target probabilities or test data withheld from selection and training.

iDEM on GMM-40, DW4, and LJ13. We use target samples for GMM-40 and reference samples for DW4 and LJ13, with the score penalty from Section 4.3. A retain-only control trains without corrective data. On all three targets, audit-selected training gives a more accurate mean regional probability and lower partition error than the original model or untargeted training (Table 4).

FAB on Many-Well. We compare unconditional and audit-selected target data at $\eta = 0.0203$, chosen so the training mixture matches the estimated target probability in A. Selected data lower regional and 4-cell errors relative to the original flow and controls; both updates slightly increase negative log likelihood (NLL; Table 5a). At the bound-derived weight $\eta = 0.0126$, a separate 414-example comparison gives mean regional errors of 0.332 and 0.663 percentage points for audit-selected and untargeted data, respectively. Varying weight and sample count separately shows that weight has the larger effect on regional error over the tested range. Generated regional probabilities follow the training-mixture prediction (Appendix E.2.4).

Table 4: iDEM fine-tuning (mean ± SD, 3 training seeds). Held-out target/reference frequencies in A: 9.771%, 37.702%, 5.489% (left to right). Bold marks the mean probability closest to the reference and the lowest mean L_1. Paired probability changes use percentage points.

	GMM-40		DW4		LJ13	
model / data	prob. (%)	$L_1 \downarrow$	prob. (%)	$L_1 \downarrow$	prob. (%)	$L_1 \downarrow$
original	5.18	0.276	29.33	0.189	1.55	0.488
retain only	5.18 ± 0.00	0.276 ± 0.000	29.24 ± 0.17	0.189 ± 0.002	1.31 ± 0.04	0.517 ± 0.001
untargeted	5.45 ± 0.18	0.253 ± 0.013	28.07 ± 0.43	0.211 ± 0.010	1.76 ± 0.21	0.444 ± 0.018
audit-selected	$\mathbf{9.25} \pm 0.72$	$\mathbf{0.224} \pm 0.008$	$\mathbf{34.45} \pm 0.28$	$\mathbf{0.080} \pm 0.002$	$\mathbf{2.11} \pm 0.34$	$\mathbf{0.400} \pm 0.020$
audit − original	$+4.08 \pm 0.72$	-0.051 ± 0.008	$+5.13 \pm 0.28$	-0.110 ± 0.002	$+0.56 \pm 0.34$	-0.088 ± 0.020
audit − untargeted	$+3.80 \pm 0.67$	-0.028 ± 0.007	$+6.38 \pm 0.24$	-0.132 ± 0.008	$+0.35 \pm 0.18$	-0.043 ± 0.011

CG protein models. We fine-tune the general CG potential on each protein's reference data and evaluate simulations from original-model structures against a held-out reference. Deficit-weighted training lowers mean free-energy and energy-distance errors relative to the original potential on all three proteins (Table 5b). It also improves both measures over uniform-all on chignolin and villin. Both updates improve protein G, with similar mean free-energy errors. Appendix E.3 reports all four metrics and paired comparisons.

Table 5: Fine-tuning (mean ± sample SD). (a) FAB: 414 corrective samples, $\eta = 0.0203$, 8 seeds; only evaluation samples vary for the released model. NLL uses independent target samples. (b) CG: 2000 simulation steps, 5 seeds per protein. Bold marks the lowest mean per metric and comparison, not statistical significance.

(a) FAB on Many-Well

training data	region error (pp) ↓	cell L_2 (10^{-3}) ↓	target NLL ↓
released model	0.678 ± 0.039	8.16 ± 0.48	$\mathbf{27.6058} \pm 0.0270$
retain only	0.706 ± 0.086	8.44 ± 0.98	27.6442 ± 0.0263
retain + target	0.627 ± 0.137	7.66 ± 1.52	27.6613 ± 0.0275
retain + target in A	$\mathbf{0.108} \pm 0.116$	$\mathbf{2.24} \pm 0.90$	27.6609 ± 0.0285

(b) CG proteins: original and fine-tuned potentials

	MAE_F (kT) ↓			ED ↓		
protein	original	uniform-all	deficit-weighted	original	uniform-all	deficit-weighted
chignolin	2.460 ± 0.057	2.557 ± 0.063	$\mathbf{2.164} \pm 0.092$	11.660 ± 0.184	12.546 ± 0.382	$\mathbf{5.615} \pm 0.765$
villin	1.276 ± 0.074	1.295 ± 0.113	$\mathbf{1.150} \pm 0.083$	16.068 ± 0.365	11.543 ± 1.132	$\mathbf{6.956} \pm 0.791$
protein G	0.614 ± 0.032	$\mathbf{0.565} \pm 0.067$	0.568 ± 0.097	61.214 ± 1.794	51.434 ± 1.929	$\mathbf{50.057} \pm 3.945$

6 Discussion and conclusion

Regional mass auditing identifies differences between target and model probabilities and uses them to guide additional sampling and model training. Our experiments show that selecting simulation initial states and training examples can improve the resulting samples. In the main DW4 and chignolin comparisons, region selection provides most of the gain over uniform-all, with smaller additional gains from deficit weighting. When valid confidence bounds are available, the audit also determines how many samples can be added without over-filling the selected region.

As trajectories may leave their initial regions, short preliminary simulations could help identify starting states that produce useful corrective samples. Allocation could then account for both regional deficits and the likely destinations of the additional samples. For fine-tuning, a possible next step is to choose the corrective-data weight using validation samples and evaluate the updated model on separate test samples.

Ethics statement

This work involves no human subjects, personally identifiable information, or data collected from people. All experiments use public benchmarks, released model checkpoints, and molecular simulation references described in the appendix.

Reproducibility statement

Appendix A gives the assumptions and proofs. Appendices B–E describe the data sources, splits, training settings, and evaluation procedures. The supplementary material includes code for regional probability bounds, correction-budget calculations, and illustrative experiments on analytic targets. Third-party models and reference datasets are identified in the corresponding experimental sections.

AI use statement

We used generative AI tools for language editing, manuscript organization, and limited coding and debugging assistance. All AI-assisted text and code were reviewed and verified by the authors. Generative AI was not used to produce experimental results or independently determine scientific claims or conclusions. We take full responsibility for the final content of this work.

References

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Avishek Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from boltzmann densities. In *International Conference on Machine Learning*, 2024.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

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*, volume 235 of *PMLR*, pp. 4205–4229, 2024.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

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

Louis Grenioux and Maxence Noble. Diffusion-based annealed Boltzmann generators: Benefits, pitfalls and hopes. *Transactions on Machine Learning Research*, 2026.

Louis Grenioux, Maxence Noble, and Marylou Gabrié. Improving the evaluation of samplers on multi-modal targets. *arXiv preprint arXiv:2504.08916*, 2025.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Johannes Hertrich and Robert Gruhlke. Importance corrected neural JKO sampling. *arXiv preprint arXiv:2407.20444*, 2024.

Chaofan Huang and V. Roshan Joseph. Enhancing sample quality through minimum energy importance weights. *arXiv preprint arXiv:2310.07953*, 2023.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. In *Advances in Neural Information Processing Systems*, 2024.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, 2016.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, 2017.

Diego E. Kleiman, Hassan Nadeem, and Diwakar Shukla. Adaptive sampling methods for molecular dynamics in the era of machine learning. *The Journal of Physical Chemistry B*, 127(50):10669–10681, 2023. doi: 10.1021/acs.jpcb.3c04843.

Jonas M. Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, 2022.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

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*, 2019.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, pp. eadv9817, 2025.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, 2016.

Xing Liu, Andrew B. Duncan, and Axel Gandy. Using perturbation to improve goodness-of-fit tests based on kernelized Stein discrepancy. In *International Conference on Machine Learning*, 2023.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Maciej Majewski, Adrià Perez, Philipp Thölke, Stefan Doerr, Nicholas E Charron, Toni Giorgino, Brooke E Husic, Cecilia Clementi, Frank Noé, and Gianni De Fabritiis. Machine learning coarse-grained potentials of protein thermodynamics. *Nature Communications*, 14(1):5739, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009. Proc. COLT 2009.

Laurence I. Midgley, Vincent Stimper, Gregor N.C. Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. In *International Conference on Learning Representations*, 2023.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11, 2001.

Maxence Noble, Louis Grenioux, Marylou Gabrié, and Alain Oliviero Durmus. Learned reference-based diffusion sampling for multi-modal distributions. In *International Conference on Learning Representations*, 2025.

Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1), 2013. doi: 10.1063/1.4811489.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. doi: 10.2307/3318418.

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*, 2018.

Zahra Shamsi, Kevin J. Cheng, and Diwakar Shukla. Reinforcement learning based adaptive sampling: REAPing rewards by exploring protein conformational landscapes. *The Journal of Physical Chemistry B*, 122(35):8386–8395, 2018. doi: 10.1021/acs.jpcb.8b06521.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Ilya Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. AdaGAN: Boosting generative models. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25, 2024.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009. doi: 10.1214/08-AOS646.

Maxwell I. Zimmerman and Gregory R. Bowman. FAST conformational searches by balancing exploration/exploitation trade-offs. *Journal of Chemical Theory and Computation*, 11(12):5747–5757, 2015. doi: 10.1021/acs.jctc.5b00737.

A THEORY AND BOUNDS

A.1 VALIDITY AFTER REGION SELECTION

The selection stage determines the domain C, the partition, and the region A. For example, we can compare estimates $\widehat{p}_C(C_i)$ and $\widehat{q}_C(C_i)$ from selection data and combine cells with positive differences. Preliminary probability bounds can also be used to score cells, and candidate partitions can be compared on these data. The bounds below allow any such selection rule.

Let $\mathcal{F}_{\text{sel}}$ record all selection data and decisions, including any proposal distribution used to estimate target probabilities. Conditional on this information, C, the partition, A, and the proposal are fixed. The constructions below require assessment samples that are i.i.d. from q_C, p_C, or the chosen proposal, as appropriate, conditional on $\mathcal{F}_{\text{sel}}$. The number of initial draws is fixed before assessment. A proposal may therefore be fitted to the selection data and then used to draw the assessment samples.

Proof of Theorem 1. Under these sampling conditions, the Clopper–Pearson bounds in the theorem and the alternative constructions in Appendix A.2 satisfy, almost surely,

$$\mathbb{P}\big(p_C(A) < L_p(A) \mid \mathcal{F}_{\text{sel}}\big) \le \alpha/2, \qquad \mathbb{P}\big(q_C(A) > U_q(A) \mid \mathcal{F}_{\text{sel}}\big) \le \alpha/2. \tag{6}$$

The union bound adds the two failure probabilities in Equation 6:

$$\mathbb{P}\big(p_C(A) < L_p(A) \text{ or } q_C(A) > U_q(A) \mid \mathcal{F}_{\text{sel}}\big) \le \alpha/2 + \alpha/2 = \alpha.$$

Thus the joint coverage event $\mathcal{E}_A = \{L_p(A) \le p_C(A),\ q_C(A) \le U_q(A)\}$ has conditional probability at least $1 - \alpha$, without requiring independence between the two bounds. On $\mathcal{E}_A$, $\hat{d}_A = L_p(A) - U_q(A) \le p_C(A) - q_C(A) = d_A$, so $\mathcal{E}_A \subseteq \{d_A \ge \hat{d}_A\}$. Taking probabilities and averaging over the selection data proves both inequalities in Equation 1. □

The independent-sample constructions require the sampling conditions stated above. Disjoint subsets of an MD trajectory can retain correlations, so the molecular protocols adjust for correlated frames and assess calibration through reference comparisons (Appendix D.4.1). All regional guarantees concern p_C and q_C; they concern the full distributions when both assign probability one to C.

A.2 MODEL UPPER BOUNDS AND TARGET LOWER BOUNDS

Equation 6 requires a model upper bound $U_q(A)$ and a target lower bound $L_p(A)$. We obtain the model bound from independent sample counts. For the target bound, we give three alternatives: known or numerically bounded probabilities, independent target samples, and importance sampling. Only one target-side construction is needed, according to the information available.

Throughout this section, A is fixed by selection and probability statements are conditional on $\mathcal{F}_{\text{sel}}$. The same formulas apply to an individual cell by replacing A with C_i.

A.2.1 MODEL PROBABILITY FROM INDEPENDENT SAMPLES

Theorem 1 gives the one-sided Clopper–Pearson upper bound from independent model samples. If the original draws come from q, retain those in C and use their number as the sample size. Conditional on this retained count, the count in A is binomial with success probability $q_C(A)$. With no retained observations, use $U_q(A) = 1$.

A.2.2 THREE WAYS TO BOUND THE TARGET PROBABILITY

Known or numerically bounded probabilities. If $p_C(A)$ is known, use it directly. A deterministic lower bound is also sufficient when both approximation and arithmetic errors are bounded. These constructions need no statistical error allowance. Appendix D.2.3 gives the Many-Well integration bounds and their floating-point limitation: discretization and truncation errors are bounded, but roundoff is not enclosed.

Independent target samples. Use the lower bound in Theorem 1 with the target dataset's own count and sample size. If the original draws come from p, retain those in C and condition on their number as above. With no retained observations, use $L_p(A) = 0$.

Importance sampling with bounded weights. When the target energy can be evaluated, importance sampling provides another way to construct $L_p(A)$. For a dimensionless energy E with $p(y) \propto e^{-E(y)}$, write $Z_A = \int_A e^{-E(y)}\, dy$ for the total unnormalized target weight in A, and define $Z_{C\setminus A}$ in the same way for the rest of the audited domain. Their ratio gives $p_C(A) = Z_A/(Z_A + Z_{C\setminus A})$, so we can bound the two integrals separately without needing the global normalizing constant.

Let g be a proposal density that can be sampled and evaluated, positive wherever the target density is positive in C. Define $w(y) = e^{-E(y)}/g(y)$ and assume a known bound $w(y) \le W$ on C, with $0 < W < \infty$. This bound must hold throughout the audited domain. Appendix A.2.3 derives the bound used for the Gaussian mixtures.

Draw $Y_1, \dots, Y_n$ independently from g, with $n \ge 2$. Each draw contributes to one integral or the other according to whether it falls in A or $C \setminus A$:

$$H_l = w(Y_l)\mathbf{1}\{Y_l \in A\}, \qquad J_l = w(Y_l)\mathbf{1}\{Y_l \in C \setminus A\}.$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function. Both contributions are zero outside C, lie in $[0, W]$, and have means Z_A and $Z_{C\setminus A}$, respectively. Let $\bar{H}$ and $\widehat{\sigma}_H^2$ be the sample mean and sample variance of the first sequence, using denominator $n - 1$ for the variance. The confidence radius is

$$b_H(\delta) = \sqrt{\frac{2\widehat{\sigma}_H^2 \log(2/\delta)}{n}} + \frac{7W \log(2/\delta)}{3(n-1)}.$$

Define $\bar{J}$, $\widehat{\sigma}_J^2$, and $b_J(\delta)$ analogously for the second sequence. Applying the empirical-Bernstein inequality (Maurer & Pontil, 2009, Theorem 4) to $(W - H_l)/W$ and J_l/W, which lie in $[0, 1]$, and rescaling gives a lower bound $\ell_A = \max\{0, \bar{H} - b_H(\delta)\}$ on Z_A and an upper bound $u_{C\setminus A} = \bar{J} + b_J(\delta)$ on $Z_{C\setminus A}$, each with failure probability at most δ. Lowering the numerator integral and raising the complementary integral yields

$$L_p(A) := \frac{\ell_A}{\ell_A + u_{C\setminus A}} \le p_C(A)$$

when both bounds hold. The denominator is positive because $b_J(\delta) > 0$. A union bound limits failure to 2δ, so choosing $\delta = \alpha/4$ gives the target-side allowance $\alpha/2$ in Equation 6.

A.2.3 Weight bound for the Gaussian-mixture experiments

The analytic target in d_x dimensions and the importance-sampling proposal have densities

$$p(x) = \sum_m \omega_m \mathcal{N}(x; \mu_m, \sigma^2 I), \qquad g(x) = \frac{1}{K}\sum_{j=1}^{K} \mathcal{N}(x; c_j, \sigma_{\text{prop}}^2 I),$$

where ω_m are the target mixture weights, c_j are the discovered centers, and $\sigma_{\text{prop}} > \sigma$. Since the normalized target density is known, we use weights $w(x) = p(x)/g(x)$ in the ratio construction above. For each target component, let $c(m)$ be its nearest proposal center. Bounding $g(x)$ below by $K^{-1}\mathcal{N}(x; c(m), \sigma_{\text{prop}}^2 I)$, completing the square in each component ratio, and summing with weights ω_m gives $w(x) \le W$, where

$$W = K\left(\frac{\sigma_{\text{prop}}}{\sigma}\right)^{d_x} \sum_m \omega_m \exp\left\{\frac{\|\mu_m - c(m)\|^2}{2(\sigma_{\text{prop}}^2 - \sigma^2)}\right\}.$$

This value is fixed by the target parameters and the centers found before assessment.

A.2.4 Simultaneous and classifier-based bounds

The constructions above give the two probability bounds for a selected region. The following variants describe the Bonferroni baseline and the connection to classifier-based tests.

Simultaneous cell bounds. For the Bonferroni baseline, apply the same constructions to all K cells, assigning $\alpha/(2K)$ to each target and model bound. With importance sampling, each of the two integral bounds uses $\alpha/(4K)$. The resulting $2K$ probability bounds hold simultaneously with probability at least $1-\alpha$. The selected-region audit instead assesses only A on held-out data, so its error allowance does not depend on how many cells were searched.

Direct deficit bounds from classifiers. A classifier-based test (Lopez-Paz & Oquab, 2017) gives another way to bound a regional deficit. For a fixed classifier h, let $A_h = \{x \in C : h(x) = 1\}$ be the region predicted to belong to the target class. With equally likely target and model classes, the deficit satisfies $d_{A_h} = 2\,\mathrm{Acc}(h) - 1$, where $\mathrm{Acc}(h)$ is the population classification accuracy. A valid lower bound $\underline{a}$ on accuracy, evaluated independently of classifier training, therefore gives $2\underline{a}-1$ as a lower bound on d_{A_h}. It also implies $\mathrm{TV}(p_C, q_C) \geq \max(2\underline{a}-1, 0)$, since total variation is the largest absolute probability difference over measurable regions. To use the classifier's region for correction, a separate sampling procedure must generate samples within A_h; classification alone specifies membership.

A.3 Correction budgets and sample frequencies

The two budget rules. Work on the joint coverage event $\mathcal{E}_A$ defined in Appendix A.1. For any corrective distribution r on C, the probability increase is $m_\lambda(A) - q_C(A) = \lambda[r(A) - q_C(A)] \leq \lambda$. Since the deficit is at least $\hat{d}_A$, choosing $\lambda \leq \hat{d}_A$ prevents over-filling. Substituting $\lambda = B/(N_{\mathrm{out}} + B)$ gives the conservative sample limit

$$B^\star = \left\lfloor \frac{N_{\mathrm{out}} \hat{d}_A}{1 - \hat{d}_A} \right\rfloor, \qquad 0 < \hat{d}_A < 1. \tag{7}$$

Using both probability bounds gives $m_\lambda(A) \leq (1-\lambda)U_q(A) + \lambda$. Requiring this upper bound to be at most $L_p(A)$ allows $\lambda \leq \hat{d}_A/[1 - U_q(A)]$, for $0 \leq U_q(A) < L_p(A) < 1$. Substituting $\lambda = B/(N_{\mathrm{out}} + B)$ directly gives

$$B_{\mathrm{end}} = \left\lfloor \frac{N_{\mathrm{out}} \hat{d}_A}{1 - L_p(A)} \right\rfloor. \tag{8}$$

This budget is at least $B^\star$ because $\hat{d}_A/[1 - U_q(A)] \geq \hat{d}_A$. The controlled GMM-40 and Many-Well correction experiments use $B^\star$; Algorithm 1 allows any count up to B_{end}. When only the deficit lower bound is available, the limit $\lambda \leq \hat{d}_A$ is sharp over distributions with $d_A \geq \hat{d}_A$ and arbitrary corrective distributions r. Indeed, $q_C(A) = 0$, $p_C(A) = \hat{d}_A$, and $r(A) = 1$ give over-filling for every $\lambda > \hat{d}_A$.

The bounds lie in $[0, 1]$. For $\hat{d}_A \leq 0$, no samples are added. The budget formula includes $U_q(A) = 0$. If $L_p(A) = 1 > U_q(A)$, the coverage event gives $p_C(A) = 1$, so any finite correction budget is safe and we set $B_{\mathrm{end}} = +\infty$. Likewise, $B^\star = +\infty$ when $\hat{d}_A = 1$. Proposition 1 uses interior bounds for its KL statement.

Probability versus sample frequency. Let $\widehat{q}_C(A)$ and $\widehat{m}_\lambda(A)$ be the observed fractions in A before and after adding samples. If $r(A) = 1$, every corrective sample belongs to A almost surely. The discrepancy between the corrected frequency and mixture probability is then

$$\widehat{m}_\lambda(A) - m_\lambda(A) = (1-\lambda)\big[\widehat{q}_C(A) - q_C(A)\big].$$

Thus the original set's sampling error remains even when all added samples lie in A. The budgets control $m_\lambda(A)$; the independent re-audit in Appendix A.5 assesses this mixture probability.

A.4 KL improvement under conditional correction

We prove Proposition 1 by separating the regional and conditional contributions to KL divergence, examining how correction changes them, and then bounding the reduction using the audit. For

Algorithm 1: Regional audit and mixture correction

Input: partition of C; selection data; held-out data or known target probabilities; corrective sampling procedure; level α; number of retained original samples N_{out}

Output: selected region A, deficit bound $\hat{d}_A$, number of corrective samples B

Select A from the cell deficit scores on the selection data (Section 4.1);
Compute $L_p(A), U_q(A)$ so that Equation 6 holds;
$\hat{d}_A \leftarrow L_p(A) - U_q(A)$;
if $\hat{d}_A \leq 0$ **then**
  return $(A, \hat{d}_A, 0)$;
Choose a finite integer $0 \leq B \leq B_{\text{end}}$, using the boundary conventions above;
Generate B corrective samples, with output distribution r on C, and set $\lambda = B/(N_{\text{out}} + B)$;
Re-audit $m_\lambda = (1-\lambda)q_C + \lambda r$ on independent observations;
return $(A, \hat{d}_A, B)$;

probabilities $a \in [0,1]$ and $b \in (0,1)$, $\mathrm{KL}(a\|b)$ compares the binary distributions $(a, 1-a)$ and $(b, 1-b)$, corresponding to being inside or outside a region. Taking zero-weight terms to be zero,

$$\mathrm{KL}(a\|b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}.$$

Proof of Proposition 1. Work on the joint coverage event $\mathcal{E}_A$, with $0 < U_q(A) < L_p(A) < 1$ and $0 < \lambda \leq \hat{d}_A/[1 - U_q(A)] < 1$. Finite $\mathrm{KL}(p_C\|q_C)$ implies $q_C(A) > 0$. Any conditional KL term multiplied by zero target probability is taken to be zero.

First, the KL chain rule separates the discrepancy into three terms:

$$\begin{aligned}\mathrm{KL}(p_C\|q_C) = {} & \mathrm{KL}\big(p_C(A)\|q_C(A)\big) \\ & + p_C(A)\,\mathrm{KL}\big(p_C(\cdot \mid A)\|q_C(\cdot \mid A)\big) \\ & + p_C(C \setminus A)\,\mathrm{KL}\big(p_C(\cdot \mid C \setminus A)\|q_C(\cdot \mid C \setminus A)\big).\end{aligned}$$

They measure the discrepancy in the probability of A, the conditional discrepancy within A, and the conditional discrepancy outside A. The same decomposition holds with m_λ in place of q_C.

Next, consider how correction changes these terms. With $r = p_C(\cdot \mid A)$, the probability of A becomes $m_\lambda(A) = q_C(A) + \lambda[1 - q_C(A)]$. Among samples in A, the proportion from the corrective component is $\beta = \lambda/m_\lambda(A)$. Thus

$$m_\lambda(\cdot \mid A) = (1-\beta)q_C(\cdot \mid A) + \beta p_C(\cdot \mid A).$$

Convexity of KL in its second argument gives

$$\mathrm{KL}\big(p_C(\cdot \mid A)\|m_\lambda(\cdot \mid A)\big) \leq (1-\beta)\,\mathrm{KL}\big(p_C(\cdot \mid A)\|q_C(\cdot \mid A)\big).$$

The conditional KL within A therefore cannot increase. Outside A, the original probabilities are all multiplied by $1-\lambda$, so the conditional distribution on $C \setminus A$ is unchanged. The third term therefore cancels when we subtract the two decompositions. The overall decrease is at least the decrease in the first term, which we denote by G:

$$\begin{aligned}\mathrm{KL}(p_C\|q_C) - \mathrm{KL}(p_C\|m_\lambda) \geq G\big(p_C(A), q_C(A), \lambda\big), \\ G(a,b,\lambda) = \mathrm{KL}(a\|b) - \mathrm{KL}\big(a\|b + \lambda(1-b)\big).\end{aligned} \tag{9}$$

This regional decrease still depends on the unknown probabilities $p_C(A)$ and $q_C(A)$. To replace them with the audit bounds, expand G for $a \in [0,1]$, $b \in (0,1)$, and $0 < \lambda < 1$:

$$G(a,b,\lambda) = a \log\left(1 + \frac{\lambda}{(1-\lambda)b}\right) + \log(1-\lambda). \tag{10}$$

This expression increases with a and is nonincreasing in b. Replacing these probabilities by $L_p(A)$ and $U_q(A)$ therefore gives the lower bound in Proposition 1. Finally, the budget ensures $U_q(A) < U_q(A) + \lambda(1 - U_q(A)) \leq L_p(A)$. With the first probability fixed at $L_p(A)$, KL strictly decreases

as the second probability moves from $U_q(A)$ toward $L_p(A)$, so $G(L_p(A), U_q(A), \lambda) > 0$. The argument holds on the same event for every weight in the allowed interval, including one chosen from the observed bounds. Equation 1 gives this event probability at least $1 - \alpha$, completing the proof. □

At the largest allowed weight, $\lambda = \hat{d}_A/[1 - U_q(A)]$, the bound equals $\mathrm{KL}(L_p(A)\|U_q(A))$: the two probabilities in the subtracted term of Equation 9 are equal. Equality is attained when $p_C(A) = L_p(A)$, $q_C(A) = U_q(A)$, and the two conditional distributions on A agree. At this weight, the lower bound is therefore sharp given the two probability bounds.

Approximate conditional sampling. We now allow approximate conditional sampling within A. Assume finite $\mathrm{KL}(p_C\|q_C)$, $p_C(A) > 0$, $0 < q_C(A) < 1$, and $r(A) = 1$. Let ε bound the corrective sampler's conditional error: $\mathrm{KL}(p_C(\cdot \mid A)\|r) \le \varepsilon < \infty$. For $0 < \lambda < 1$, the corrective proportion within A remains $\beta = \lambda/m_\lambda(A)$. Applying the same decomposition and convexity argument gives

$$\begin{aligned}\mathrm{KL}(p_C\|m_\lambda) \le \mathrm{KL}(p_C\|q_C) - G\big(p_C(A), q_C(A), \lambda\big) \\ + p_C(A)\beta\left[\varepsilon - \mathrm{KL}\big(p_C(\cdot \mid A)\|q_C(\cdot \mid A)\big)\right].\end{aligned} \tag{11}$$

The bracket compares the corrective sampler's error bound with the original conditional KL within A. A positive bracket can offset part of the regional gain, whereas a negative bracket adds to it. The bound guarantees an overall decrease when the regional gain exceeds this final term. Setting $\varepsilon = 0$ recovers exact conditional correction.

A.5 Validity of the independent re-audit

The GMM re-audit in Appendix D.1 keeps A fixed and reuses the initial target assessment data. For $S = A$ or $C \setminus A$, let $U_m(S)$ be an upper confidence bound on $m_\lambda(S)$. The two re-audit statistics are

$$\hat{d}_{\mathrm{remain}} = L_p(A) - U_m(A), \qquad \hat{d}_{\mathrm{over}} = L_p(C \setminus A) - U_m(C \setminus A).$$

A positive value detects a remaining shortage or over-filling in A, respectively. The complement lower bound is obtained by applying the target-bound construction to $C \setminus A$ on the initial assessment data; $1 - L_p(A)$ is an upper bound and cannot serve this purpose.

The model-side bound uses an independent set of original-model samples reserved for re-auditing, together with the generated corrective pool. Here we fix this pool and take r to be the distribution that assigns equal weight to its entries. The re-audit therefore assesses the mixture formed with this corrective pool.

To justify the model-side bound, condition on the selection and initial assessment data, the chosen budget, and the corrective pool, but not on the reserved model samples. Fix an evaluation size no larger than the reserved model sample count. Each evaluation observation independently chooses the corrective component with probability λ. It then independently samples with replacement from the pool or uses a distinct reserved model sample, selected without inspecting its value. The i.i.d. model samples and independent component choices give i.i.d. observations from m_λ. For either fixed region S, the count is therefore binomial, and its Clopper–Pearson upper bound fails with conditional probability at most $\alpha/2$.

The reused target bound retains its marginal failure probability of at most $\alpha/2$; it need not retain this probability conditional on the observed budget. Averaging the conditional model-side error over the preceding data and applying a union bound gives the level-α re-audit. The budget must be fixed before the reserved model samples are examined. This reasoning applies separately to A, $C \setminus A$, and each budget; simultaneous claims require allocating error across the comparisons. The binomial argument is for random mixture draws, not a concatenation of fixed numbers from the two components.

For repeated audits, one can preassign round-wise error levels α_t with $\sum_t \alpha_t \le \alpha$ and reserve fresh assessment data for each round. Each round must satisfy the sampling conditions of Section 4.1, conditional on the preceding rounds and its current selection data. A union bound then controls the probability of any failed coverage statement across rounds by α.

B AUDIT RELIABILITY

We check whether the audit reports a deficit when the two distributions agree, and whether it detects a difference when only their mode weights change. This experiment uses a two-dimensional mixture of 20 equally weighted Gaussians with unit covariance, separate from GMM-40. We draw 3 arrangements of component centers from $\mathcal{N}(0, 64I)$. For the unequal-weight comparison, the model reduces 8 component weights by a factor of 10 and renormalizes them.

For each arrangement and $K \in \{8, 16, 32, 64\}$, 1500 samples per distribution construct the partition. Another 1500 per distribution select A, and a third independent set of the same size computes its Clopper–Pearson bounds. With equal target and model weights, we repeat the procedure 200 times per K on the first arrangement, redrawing all three datasets. A comparison that reuses the selection data for assessment isolates the effect of splitting. With unequal weights, 2×10^5 additional samples per distribution independently estimate the selected region's deficit. Table 6 shows that held-out assessment avoids the large false-alarm rates caused by data reuse and detects deficits across the tested partitions.

Table 6: Regional audits on a 20-component Gaussian mixture at $\alpha = 0.1$. Left: false-alarm rates over 200 repetitions with two independent datasets from the same target. Right: deficit bounds and independently estimated deficits for a model with reduced component weights. Ranges cover 3 arrangements of component centers. False alarms are percentages of repetitions; deficits are percentages of probability mass.

	false alarms (%)		alternative model (% mass)	
K	reused data	held-out data	lower bound $\hat{d}_A$	estimated deficit
8	28.5	≤ 1.0	4.0–9.6	10.5–13.4
16	82.0	≤ 1.0	15.3–18.1	20.9–24.4
32	100.0	≤ 1.0	17.3–22.1	21.8–26.6
64	100.0	≤ 1.0	15.6–20.5	21.8–25.6

The chignolin comparison in Table 7 asks whether assessing the selected region as a whole gives a larger deficit bound than combining separate cell bounds. It also repeats the data-reuse check on reference trajectories. Reference controls permute blocks of 25 subsampled frames, preserving dependence within each block. The energy-based alternative groups structures by the minima reached through energy minimization; the clustering-based approach assigns the original structures to their nearest centers.

Table 7: Chignolin audits comparing the released general CG model with a protein-specific reference. The audit uses 1500 subsampled frames per distribution, 5 seeds, and $\alpha = 0.1$. Deficit denotes the regional lower bound or the sum of positive cellwise lower bounds. False-alarm rates come from 200 comparisons between reference subsets. The last row assesses the region using the same data that selected it.

procedure	search	deficit (%)	false alarms (%)
Bonferroni over K cells	K=16	43–50	3.0–4.0
	K=32	31–33	2.5
	K=64	13–15	0.0–2.0
selected region	K=16	64–67	3.0–6.5
	K=32	63–67	3.0–4.5
	K=64	63–64	3.0–5.5
	energy-minimized groups	61.7	6.5
selected region, no split	K=32	67.5	99.0–100.0

C REGIONAL EVALUATION OF RELEASED SAMPLERS

C.1 GMM-40 DIAGNOSTIC COMPARISON

Table 1a reports regional deficit lower bounds alongside kernel Stein discrepancy (KSD) (Liu et al., 2016; Gorham & Mackey, 2017) and the benchmark's validation W_2. KSD uses the analytic target score and 600 generated samples, with Gaussian kernel $k(x, y) = \exp[-\|x - y\|^2/(2\ell^2)]$ and ℓ^2 equal to half the median nonzero squared pairwise distance. We estimate squared KSD from off-diagonal Stein-kernel pairs, clip negative estimates to zero, and take the square root. KSD favors pDEM and validation W_2 favors iDEM; the audit establishes a positive regional deficit for both, with a larger lower bound for pDEM.

C.2 BIOEMU VERSION COMPARISON

We compare BioEmu v1.1 and v1.2 using the same states defined from reference MD. We apply time-lagged independent component analysis (TICA) (Pérez-Hernández et al., 2013) to the sine and cosine of backbone torsion angles, using a lag of 20 frames and 6 components. We then cluster these coordinates with k-means, using $K = 4, 6, 8$. Both model versions receive only the peptide sequence, stored in A3M format. The pentapeptide reference contains 25 separate simulation runs of 500 ns each. For each state, the error is $|\log(\widehat{p}/\widehat{q})|$ in kT, where $\widehat{p}$ and $\widehat{q}$ are the observed state frequencies in the reference and generated samples. Reference controls compare state frequencies between groups of 12 and 13 trajectories. A separate permutation test uses 200 random divisions of the trajectories. Table 1b reports ranges across the 3 choices of K for both model versions and the reference controls.

C.3 MDGEN: STATES ABSENT FROM GENERATED SAMPLES

We audit all 100 tetrapeptides in the test set released with MDGen (Jing et al., 2024). For each peptide, approximately 10^5 regularly subsampled reference MD frames define 6 states. We represent each structure by the sine and cosine of its backbone torsion angles (ϕ, ψ) and fit k-means centers using reference features. Nearest-center assignment places reference structures and 10^4 generated frames in the same 6 states. For 27 peptides, at least one state observed in the reference has no generated samples; the largest combined reference frequency of these missing states is 87%. Any nonnegative reweighting of the generated frames still assigns zero frequency to those states. Consequently, its total variation distance from the empirical reference state distribution is at least their combined reference frequency. This finite-sample statement allows the model to assign positive probability to states absent from the generated set.

Figure 2 provides a qualitative comparison for CKHT. It pools backbone torsion-angle pairs (ϕ, ψ) over residues with both angles defined, whereas the 6-state audit assigns each complete peptide structure to a state. The labels locate conventional conformations in the pooled angle plot.

C.4 ANNEALED IMPORTANCE SAMPLING ON THE RELEASED ALDP FLOW

The reference and model samples are restricted to the chirality used to train the released flow. We choose A using a validation trajectory with 10^6 frames and a separate model sample. Of the 85 occupied cells in a fixed 10×10 Ramachandran grid, 19 have larger reference than model frequencies; their union defines A. Evaluation uses independent flow samples and 2×10^6 frames from the test reference trajectory. For the reference uncertainty, we bootstrap consecutive trajectory blocks of lengths 500, 2000, 10,000, and 50,000, using 400 resamples per length. Table 8 reports the interval from the smallest lower endpoint to the largest upper endpoint of these 90% bootstrap intervals.

AIS uses the released flow and its saved transition parameters, with step-size adaptation disabled and annealing target p. We run 4×10^5 independent chains through 8 intermediate distributions, with 4 leapfrog steps per Hamiltonian Monte Carlo (HMC) transition. We compare the probability of the fixed region A under direct sampling, ordinary importance weighting, and AIS. This evaluates the existing corrections, without matching their computational costs.

The AIS-weighted estimate is closer to the reference than the direct flow estimate (Table 8). Ordinary importance weighting instead concentrates almost all weight on a few samples. For weights w_l,

Table 8: ALDP probability estimates for the selected region, evaluated on held-out data. Probabilities are percentages. The uncertainty column gives the reference's 90% block-bootstrap interval and standard errors for the other estimates, on the same scale. Effective sample size (ESS) measures concentration of the importance weights.

arm	mass of A (%)	uncertainty	weight ESS
reference, held-out test trajectory	65.724	$[65.633, 65.833]$	n/a
flow samples	65.437	0.075	4.00×10^5
flow + plain importance weights	71.990	19.903	3
flow + AIS, self-normalized	65.807	0.076	3.89×10^5

the effective sample size $(\sum_l w_l)^2 / \sum_l w_l^2$ is 3 under ordinary weighting and 3.89×10^5 under AIS. The formula applies to normalized and unnormalized weights. It measures weight concentration; reference uncertainty is assessed separately by the block bootstrap.

D AUDIT-GUIDED SAMPLE CORRECTION

These experiments keep the model fixed. GMM and Many-Well use $B^\star$ to set the added sample count and re-audit the corrected probabilities. DW4 and CG compare simulation starts at several specified chain counts, with chain length fixed within each comparison. These comparisons evaluate allocation at fixed simulation budgets.

Uniform-all divides the sampling budget equally among all K cells. Uniform-flagged divides it equally among the flagged cells, provided at least one is flagged, and assigns none elsewhere. Comparing uniform-flagged with uniform-all tests region selection; comparing deficit weighting with uniform-flagged tests weighting within the selected region.

D.1 GMM SAMPLE CORRECTION AND RE-AUDITING

D.1.1 REGION SELECTION AND CORRECTIVE SAMPLING

This experiment checks correction at the recommended count $B^\star$ and then increases that count to test re-auditing. In every comparison, we keep the selected region fixed and give each allocation the same requested number of additional samples.

The energy probe searches for local energy minima before auditing. Each point is assigned to its nearest discovered minimum, so these cells cover the full space and $C = \mathcal{X}$. If a minimum is missed, nearby points are assigned to another discovered cell.

The model samples are divided into 4 independent sets of N_{out} samples each, used for region selection, the initial deficit bound and budget, re-auditing, and the model-resampling control. The target-proposal draws are split into two sets: one for selection and one for both the initial audit and re-audit. Only the selection sets determine A and the positive cell-deficit scores used to allocate samples within it.

Target probabilities are estimated by importance sampling from an equally weighted Gaussian mixture centered at the discovered minima. Its component standard deviation σ_{prop} is larger than the target component standard deviation σ. Appendix A.2.3 derives the deterministic weight bound used to assess these estimates.

Corrective draws. For a basin with center c_i and allocation b_i, we draw $20b_i$ candidates from $g_i = \mathcal{N}(c_i, \sigma^2 I)$, using the target's component scale σ. We retain candidates assigned to C_i and resample b_i with replacement, weighted by e^{-E}/g_i. This approximates the target conditional distribution within the basin. Rejecting candidates outside the cell ensures $r(A) = 1$ for deficit-weighted correction.

The two allocations share the requested count B and generate $20B$ candidates before rejection. Energy is evaluated only for candidates retained in their assigned cells, so energy-call counts can differ. If a cell has no usable candidate, its allocation is omitted and the mixture weight uses the

actual number of returned samples. The model-resampling control instead draws B entries from its own reserved model set without corrective energy evaluations. Corrective sample counts define the comparison budget; total computational costs can differ.

D.1.2 Remaining deficits and detection of over-filling

For each chosen count, we draw an evaluation sample from the mixture using the reserved model samples and the corrective pool. The initial target bound is reused because A is unchanged. A positive re-audit bound on A detects a remaining shortage; a positive bound on its complement detects over-filling in A. Appendix A.5 specifies the random mixture sampling and proves the validity of these tests.

At $B^\star$, deficit weighting leaves a smaller re-audit deficit bound than uniform-all in every paired seed across iDEM, pDEM, and FAB (Figure 4). The mean difference is 2–3 percentage points. About half of the uniform-all corrective samples enter A, compared with all deficit-weighted samples. We then repeat the comparison at 0.5, 2, 4, and 8 times $B^\star$ to examine remaining shortages and over-filling as the correction grows.

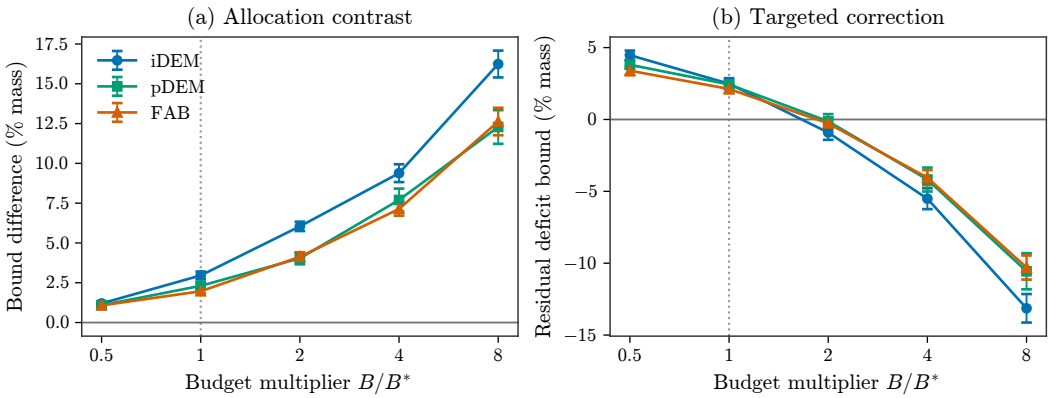

Figure 4: GMM re-audit at increasing added-sample counts, with 20 paired seeds per sampler. Left: difference in remaining-deficit bounds (uniform-all minus deficit-weighted). Right: the remaining-deficit bound under deficit weighting. Both vertical axes use percentage points; the horizontal axis is $B/B^\star$. Points show means, and error bars extend one standard error above and below them. A positive bound establishes a remaining shortage; over-filling is assessed separately on the complement of A. Counts above $B^\star$ exceed the conservative sample limit.

The conservative count leaves room for further correction: using both probability bounds gives median $B_{\text{end}}/B^\star$ ratios of 1.98, 1.84, and 2.01 for iDEM, pDEM, and FAB, respectively, including integer rounding. Complement re-audits detect no over-filling at $0.5B^\star$, $B^\star$, or $2B^\star$ (180 evaluations), and one case at $4B^\star$ (60 evaluations). At $8B^\star$, they detect over-filling in $12/20$ iDEM, $8/20$ pDEM, and $9/20$ FAB runs. These counts record detections in separate re-audits, with the same seeds reused across sample counts.

D.2 Many-Well correction on a 4-cell partition

D.2.1 Target and region selection

Many-Well provides a second controlled correction test: its target probabilities can be bounded by one-dimensional integration, without a sampled reference. The 32-dimensional target used by the released FAB model factorizes as

$$p(x) \propto \prod_{j=1}^{16} \exp\big(f(x_{2j-1}) - \tfrac{1}{2}x_{2j}^2\big), \qquad f(z) = -z^4 + 6z^2 + 0.5z.$$

Each odd coordinate has two density peaks, and each even coordinate is standard Gaussian. The signs of two odd coordinates define 4 disjoint cells. We compare all 120 pairs of odd coordinates,

selecting the pair with the largest sum of positive estimated cell deficits and resolving ties deterministically. Its cells with positive estimates form A. Their target probabilities are products of the one-dimensional sign probabilities. We estimate model cell probabilities directly from joint sample counts, allowing dependence between coordinates.

We use 2×10^6 samples from the released FAB flow, excluding one with non-finite coordinates. Each of 20 random splits uses 4 disjoint subsets for selecting A, computing its deficit bound, forming the original sample set, and re-auditing. Each subset contains approximately one fifth of the samples; the remaining fifth is unused in these comparisons. Splits reuse the same original dataset. A numerical target lower bound and a Clopper–Pearson model upper bound at level $\alpha/2$ give $\hat{d}_A$. Equation 7 then sets $B^\star$, with corrective samples accounting for 0.42–0.69% of the mixture. The numerical calculations and their error treatment are given after the results in Appendix D.2.3.

D.2.2 Correction, evaluation, and results

We sample the double-well coordinates by rejection sampling on $[-4, 4]$, using a separate piecewise-constant upper bound on the density with 20,000 intervals per half-axis. The even coordinates are independent standard Gaussians. For targeted correction, the two selected double-well coordinates are redrawn until their sign cell belongs to A. This samples the truncated target conditional on A and gives $r(A) = 1$.

The comparator samples from the same truncated target without restricting samples to A. Both allocations return the same number of corrective samples. Rejection proposal counts are recorded separately; equal sample counts do not imply equal computational cost.

The re-audit uses the independent model subset to bound the corrected mixture's probability in A, then compares it with the numerical target interval. This gives separate checks for a remaining shortage and for over-filling. The calculation below uses the mixture probability, while the 4-cell error evaluates the frequencies in the corrected sample set.

Table 9 reports the remaining deficit bound and 4-cell error. The latter is the Euclidean distance between the corrected sample frequencies and the cell-probability lower bounds, whose differences from the true probabilities are bounded by the integration interval widths. Sampling within A lowers this error by 24% relative to unconditional target correction. All 20 splits give a positive initial deficit bound, and none detects over-filling at $B^\star$. A control that audits samples from the truncated target gives one positive bound over 20 splits.

Table 9: Many-Well sample correction at $B^\star$. Entries are means over 20 random splits of the same model sample set. Both allocations add the same number of samples within each split. 4-cell L_2 errors are in units of 10^{-3}; the original sample set's error is 7.77. The deficit bound is a lower confidence bound on the remaining probability deficit, expressed in percentage points.

draw source	cell $L_2 \downarrow$	deficit bound (%)
unconditional target	7.71	0.499
target in A	5.87	0.338

D.2.3 Numerical probabilities and re-audit calculation

Target probabilities. We bound the integrals over $[-4, 0]$ and $[0, 4]$ using minima and maxima of $e^{f(x)}$ on each integration interval, including roots of f' inside it. Outside this range, $e^{f(x)}$ is bounded by $e^{-39.5x}$ for $x \geq 4$ and by $e^{-40|x|}$ for $x \leq -4$. Integrating these exponential bounds limits the omitted integral to less than 6.91×10^{-71}. Sign-probability bounds follow from the ratio construction in Appendix A.2, assigning the entire tail to the side that widens each bound. Products give cell bounds; summing over the cells in A gives the bounds for the selected region.

With 2×10^6 integration intervals per half-axis, the negative-well probability is approximately 15.5693%, with bound width below 2.01×10^{-6}. These calculations bound discretization and truncation errors but use ordinary floating-point arithmetic; machine roundoff is not enclosed. The bounds also contain the sign probabilities of the truncated target used for corrective sampling. Exact

conditional sampling in Proposition 1 therefore refers here to that truncated target; a small omitted tail probability alone does not bound forward KL error relative to the untruncated target.

Re-audit bounds. The independent model subset gives lower and upper Clopper–Pearson bounds on $q_C(A)$, each at level $\alpha/4$. Target sampling in A has $r(A) = 1$; for unconditional truncated-target sampling, the numerical target interval bounds $r(A)$. We substitute the respective lower or upper bounds into Equation 2 to bound $m_\lambda(A)$. Subtracting its upper bound from the target lower bound gives a lower bound on the remaining deficit. Over-filling is detected when the mixture lower bound exceeds the target upper bound. The model samples used here are independent of those that set the correction count.

D.3 DW4 CORRECTION BY ADDITIONAL SIMULATION

DW4 tests whether the audit can choose useful starting structures for additional dynamics. The trajectories start in the allocated cells and can move across their boundaries. We evaluate their final structures using pair-distance distributions, an independently fitted partition, and energy distributions.

Reference and region selection. The experiment uses 4000 fixed iDEM outputs and a new MCMC reference under the benchmark energy. Reference chains start from dispersed random structures, undergo random-walk Metropolis relaxation, and retain one final structure after MALA. We split $10{,}000$ reference structures into 3333 for allocation, 3333 for fitting the evaluation partition, and 3334 for scoring. The 7-cell audit selects 3 cells, with combined frequency 33.425% in the model outputs and 51.350% in the scoring reference. The latter is used only for evaluation. Pilot simulations assess the stability of the finite-run reference estimates. The dynamics and fine-tuning studies select their regions separately: this study uses the new MCMC reference, while Appendix E.1 uses the released benchmark reference.

Dynamics and allocations. Starting structures are drawn from the original model outputs in each assigned basin, with replacement when necessary. MALA uses step size 0.02 and either 50 or 200 steps under the benchmark energy; one final structure per chain is added to the original sample set. Center-of-mass translation is removed, so the target density is defined on the fixed-center-of-mass subspace. Uniform-all, uniform-flagged, and deficit weighting use the allocations in Section 5.2. Cell scores are positive differences between Clopper–Pearson reference lower and model upper bounds, using $\alpha/(2K)$ per bound. Additional controls allocate according to reference or model cell frequencies. Low-count allocation assigns weights proportional to $\max_j \widehat{q}_C(C_j) - \widehat{q}_C(C_i)$, falling back to uniform allocation if all frequencies are equal. Random selection chooses as many distinct cells as A contains, uniformly from all cells, and divides chains equally among them. 10 seeds vary starting samples, simulation noise, and random cell selection, keeping the model outputs, partition, and reference split fixed.

The sample-count sweep asks whether the allocation advantage persists as more trajectories are added. We use 1172 chains as the base count and also test half, twice, and four times that count: 586, 2344, and 4688. The base count comes from Equation 8 with $N_{\mathrm{out}} = 4000$, $L_p(A) = 0.4948$, and $U_q(A) = 0.3467$, using one-sided Clopper–Pearson limits at level $\alpha/2$. These limits reuse the data that selected A, making the count an empirical budget scale outside the held-out coverage guarantee. At each count, all allocations receive the same number of chains and steps.

Evaluation and costs. The primary error is the one-dimensional W_1 distance between the particle-pair distance distributions of the corrected sample set and scoring reference, pooling the pair distances across structures. A separate 16-cell k-means partition, fitted on the reference subset reserved for this purpose, measures the L_1 error in state frequencies. It uses the same pair-distance features as the audit, but different cells. We also report W_1 between energy distributions. Figure 5 gives the complete short-chain budget sweep, complementing Figure 3.

The model-seeded controls in Tables 2 and 10 spend the basin-search allowance on additional chains instead. Basin-based allocations cost 2.85 million force evaluations for the search, plus $B(S+1)$ for B chains of S steps. The controls use approximately the same total allowance and the same chain length, but no basin selection. Reference generation and pilot checks are excluded from both costs.

Table 10: Further DW4 allocations with 50 and 200 steps per chain. Entries give means and standard deviations across 10 simulation seeds. Each allocation adds 1172 chains, except the 50-step compute-matched control, which adds 58,172. Table 2 reports the original-model errors and the other 200-step comparisons.

steps	allocation	pair-distance $W_1 \downarrow$	partition $L_1 \downarrow$	energy $W_1 \downarrow$
50	model frequencies	0.1983 ± 0.0012	0.3614 ± 0.0014	0.1385 ± 0.0099
	uniform-all	0.1571 ± 0.0016	0.2691 ± 0.0028	0.2006 ± 0.0113
	reference frequencies	0.1597 ± 0.0014	0.2851 ± 0.0024	0.1049 ± 0.0126
	low-count allocation	0.1418 ± 0.0020	0.2414 ± 0.0057	0.2277 ± 0.0132
	random selection	0.1599 ± 0.0494	0.2881 ± 0.0834	0.1960 ± 0.0782
	uniform-flagged	0.0669 ± 0.0018	0.1436 ± 0.0075	0.0883 ± 0.0172
	deficit-weighted	0.0607 ± 0.0014	0.1397 ± 0.0078	0.0410 ± 0.0106
	compute-matched model	0.1834 ± 0.0006	0.3524 ± 0.0017	0.1457 ± 0.0104
200	model frequencies	0.1930 ± 0.0022	0.3536 ± 0.0040	0.1232 ± 0.0143
	reference frequencies	0.1593 ± 0.0016	0.2857 ± 0.0027	0.1097 ± 0.0098

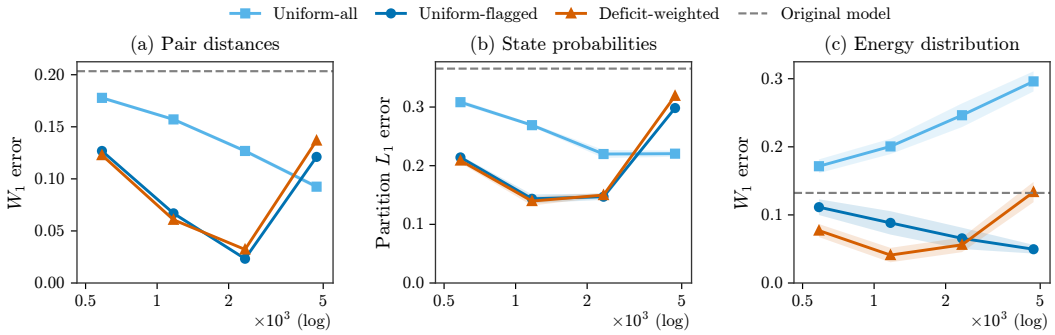

Figure 5: DW4 correction with 50 MALA steps per chain, using the same allocations, reference split, and 4 chain counts as Figure 3. The horizontal axis shows the number of added chains in thousands on a logarithmic scale. Lines show means across 10 simulation seeds; shaded bands extend one standard deviation above and below them.

Selecting flagged cells improves pair-distance and partition errors at 1172 chains under both chain lengths, including comparisons with low-count allocation and random selection. At 2344 chains, uniform-flagged has its lowest pair-distance error but exceeds the reference frequency in A by 4.02 percentage points at 50 steps and 1.67 at 200 steps. The allocation ordering changes with the added-chain count: at 4688 chains and 200 steps, low-count allocation has pair-distance error 0.0558 ± 0.0029 and partition error 0.1362 ± 0.0075, both below uniform-flagged (0.0910 ± 0.0022 and 0.2312 ± 0.0031).

Comparison using the released reference. We also repeat the dynamics comparison using the released benchmark reference. This audit selects 2 cells, $A = C_1 \cup C_2$, and gives an empirical sample limit of $B_{\text{end}} = 499$. At this count, with 200 MALA steps per chain, uniform-flagged and deficit weighting both improve on uniform-all and the compute-matched control across all three reported errors (Table 11). Deficit weighting gives a further reduction relative to uniform-flagged. The new selection controls show the same ordering here: mean pair-distance error is 0.0543 for uniform-flagged, compared with 0.0862 for low-count allocation and 0.0958 for random selection across 10 seeds.

D.4 CG PROTEIN CORRECTION BY ADDITIONAL SIMULATION

D.4.1 SAMPLING AND DATA ALLOCATION

The CG experiments test whether region selection improves the distribution of protein structures obtained by additional simulation. Each corrected set retains 15,000 original-model structures and

Table 11: DW4 correction using the released benchmark reference. Basin-based allocations add 499 chains of 200 MALA steps. Regional error is the absolute difference between corrected and reference probabilities in the selected region $A = C_1 \cup C_2$.

allocation	pair-distance $W_1 \downarrow$	partition $L_1 \downarrow$	regional error $\downarrow$
original model	0.117	0.194	0.084
model frequencies	0.111	0.188	0.081
uniform-all	0.093	0.160	0.064
reference frequencies	0.101	0.169	0.071
uniform-flagged	0.054	0.119	0.029
deficit-weighted	0.050	0.116	0.023
compute-matched model	0.079	0.153	0.064

adds the final structures from the new trajectories. The reference trajectory is divided into interleaved blocks: approximately 50% for selection and allocation, 25% for evaluation, and 25% unused here. Only the selection subset is used to define regions and allocation scores. Basin search uses the protein-specific potential for chignolin and the general potential for villin. We cluster the discovered energy minima using pairwise CA distances and choose one representative structure per cluster. New trajectories start from copies of these representatives with independent Gaussian perturbations and evolve under the released general potential. Uniform-all, uniform-flagged, and deficit weighting change only the number starting in each basin. The model-seeded control starts instead from the original model's structures.

Allocation scores are positive differences between reference lower and model upper probability bounds, with an empirical count adjustment for correlated MD frames. Appendix D.4.5 gives this adjustment. The same scores are used at each added-chain count, so the comparison changes the amount of sampling without selecting a new region.

All allocations use nominal budgets $B \in \{500, 1000, 2000, 4000\}$ chains with 500 steps each, retaining one final structure per chain. In the reported chignolin allocation batch, per-basin counts are rounded down, so realized totals can be slightly below B and differ across allocations. Basin-based allocations share the energy search; the model-seeded control starts directly from model outputs. Comparisons match nominal chain budgets and trajectory length, excluding reference-generation costs. The chignolin and villin allocation experiments and the separate chignolin RMSD evaluation each use 5 seeds. The same seeds are used across budgets.

D.4.2 Evaluation metrics

We evaluate the corrected samples against the held-out reference, using the same bins and features for every allocation. The primary metric compares free-energy profiles along R_g. The reference range, padded by 5% on each side, defines 40 equally spaced bins. Let n_j^{ref}, n_j count reference and evaluated samples in bin j, and let N_{ref}, N be the total sample counts. Adding one count per bin avoids logarithms of zero and gives smoothed frequencies $\widetilde{p}_j = (n_j^{\mathrm{ref}} + 1)/(N_{\mathrm{ref}} + 40)$ and $\widetilde{q}_j = (n_j + 1)/(N + 40)$. We score only bins whose unsmoothed reference frequency is at least 0.002, denoting their indices by $\mathcal{J}$. The free-energy relation $F/(k_{\mathrm{B}}T) = -\log p$, up to an additive constant, motivates the mean absolute difference in log bin probabilities:

$$\mathrm{MAE}_F = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} |\log \widetilde{p}_j - \log \widetilde{q}_j| \,. \tag{12}$$

The error is reported in kT. The original CG study evaluates free-energy surfaces using TICA and Markov state models (Majewski et al., 2023); we use R_g profiles with the bins and smoothing above. All samples count toward N, including those outside the bin range, so leaving the range lowers the estimated in-range probabilities. Smoothing uses the actual sample count for each corrected set.

The histogram error adds an out-of-range penalty to the L_1 distance between normalized in-range histograms. Writing $N_{\mathrm{in}} = \sum_j n_j$, the score is

$$\mathrm{Err}_{\mathrm{hist}} = \sum_{j=1}^{40} \left| \frac{n_j}{\max(N_{\mathrm{in}}, 1)} - \frac{n_j^{\mathrm{ref}}}{N_{\mathrm{ref}}} \right| + 2 \left(1 - \frac{N_{\mathrm{in}}}{N} \right).$$

This score is labeled histogram error in the tables. The RMSD evaluation uses the same formula with reference-defined RMSD bins. Evaluation-partition L_1 instead compares the ordinary frequency vectors on 16 clusters fitted to pairwise CA distances in the held-out reference.

For these descriptor vectors z and reference vectors y, the standard energy distance (ED) (Székely & Rizzo, 2013) estimates $2\mathbb{E}\|z-y\| - \mathbb{E}\|z-z'\| - \mathbb{E}\|y-y'\|$ by empirical pairwise averages, including diagonal pairs within each sample. Each side is subsampled to at most 1200 vectors with fixed evaluation randomness across arms. Here ED compares the distributions of pairwise CA distances. The one-dimensional Wasserstein distance W_1 (Peyré & Cuturi, 2019) between RMSD distributions integrates the absolute difference between the two empirical cumulative distribution functions and is reported in angstroms.

D.4.3 Region selection and within-region weighting

Table 12 separates the benefit of selecting basins from that of weighting them. Uniform-flagged reduces chignolin's mean primary error relative to uniform-all at all 4 budgets. Deficit weighting gives a smaller additional mean improvement with greater between-seed variation. Both secondary mean errors follow this ordering at $B = 4000$ (Table 3a).

For villin, deficit weighting lowers evaluation-partition error at every tested budget and seed. At $B = 4000$, uniform-all has lower mean free-energy error by 0.016 kT. Model-seeded dynamics increase this error from 1.557 to 1.588 kT.

Table 12: Chignolin R_g free-energy profile errors (kT) at 4 nominal chain budgets. Entries give means and sample standard deviations across 5 seeds; the fixed original-model error is 3.358 kT. The last three rows summarize within-seed differences, with negative values indicating improvement. The same seeds are reused across budgets. Table 3a reports secondary metrics at $B = 4000$.

	R_g free-energy error ↓			
allocation	500	1000	2000	4000
model-seeded	3.366 ± 0.009	3.375 ± 0.006	3.382 ± 0.018	3.401 ± 0.031
uniform-all	2.594 ± 0.040	2.247 ± 0.058	1.886 ± 0.038	1.496 ± 0.042
uniform-flagged	2.475 ± 0.048	2.122 ± 0.061	1.719 ± 0.066	1.351 ± 0.068
deficit-weighted	2.433 ± 0.066	2.086 ± 0.093	1.695 ± 0.102	1.297 ± 0.139
flagged − all	-0.119 ± 0.021	-0.125 ± 0.060	-0.167 ± 0.048	-0.145 ± 0.050
deficit − flagged	-0.042 ± 0.048	-0.035 ± 0.075	-0.024 ± 0.060	-0.053 ± 0.097
deficit − all	-0.161 ± 0.061	-0.161 ± 0.122	-0.191 ± 0.096	-0.199 ± 0.117

D.4.4 RMSD evaluation against an experimental structure

We evaluate whether correction improves agreement with the reference distribution of CA-RMSD to PDB 5AWL, a structural observable excluded from allocation. RMSD is computed after optimal superposition, and histogram bins are defined from the reference. Mean RMSDs are 3.518 Å in the reference and 1.974 Å in the original model samples.

At $B = 4000$, deficit weighting reduces RMSD histogram error from 1.190 to 0.848, W_1 from 1.630 to 1.112 Å, and free-energy error from 2.514 to 1.184 kT relative to the original model samples. All three errors are lower than under model-seeded dynamics at every budget in each of the 5 seeds. Table 13 compares deficit weighting with uniform-all, which also starts dynamics in the discovered basins.

D.4.5 Adjustment for correlated frames

The regional scores account empirically for repeated, similar observations along a reference trajectory. For each basin, its membership sequence has value one inside the basin and zero outside. Divide the n observations into 50 blocks of length $b = \lfloor n/50 \rfloor$, and let $\bar{x}$ and s^2_{block} be the overall frequency and sample variance of block means. The effective count is $n\bar{x}(1-\bar{x})/(bs^2_{\text{block}})$, clipped to $[2, n]$. Constant sequences, zero block variance, or $b < 2$ use n. This count and its product with $\bar{x}$ replace the sample and success counts in the beta-quantile formulas. We assess these approximate

Table 13: Errors in the chignolin CA-RMSD distribution relative to the reference, using PDB 5AWL as the comparison structure. Entries give means and standard deviations of paired error differences (deficit-weighted minus uniform-all) across 5 seeds. Negative values favor deficit weighting. The same seeds are used at each added-chain count.

added chains	histogram error	MAE_F (kT)	W_1 (Å)
500	-0.0084 ± 0.0046	-0.0712 ± 0.0323	-0.0175 ± 0.0050
1000	-0.0163 ± 0.0093	-0.0873 ± 0.0659	-0.0322 ± 0.0094
2000	-0.0285 ± 0.0186	-0.1092 ± 0.0606	-0.0596 ± 0.0177
4000	-0.0412 ± 0.0391	-0.0945 ± 0.0789	-0.1040 ± 0.0296

intervals empirically. Comparisons between reference subsets at $K \in \{6, 8, 12\}$ give no flags after adjustment, versus one or two on the larger proteins with unadjusted counts.

E MODEL UPDATES FROM AUDIT-SELECTED DATA

These experiments test whether selecting training examples improves samples generated after fine-tuning. Within each model, targeted and untargeted updates share the initial parameters, loss, optimizer, and nominal number of corrective examples (Table 14). We first describe which data enter training and which are reserved for evaluation, then report comparisons of training choices.

Table 14: Optimization settings for model updates. Targeted and untargeted training use the same settings within each model. The CG settings apply to all three proteins. Appendix E.2 gives the additional FAB weights and sample counts.

setting	iDEM diffusion	Many-Well flow	CG potential
optimizer	Adam	Adam	Adam
updates	2000	2000	400
learning rate	10^{-4}	10^{-4}	10^{-5}
batch size	512	1024	256
gradient-norm clip	none	100	10
corrective weight	$51/512$	$\eta = 0.0203$ (main comparison)	not used

E.1 iDEM FINE-TUNING ON GMM-40, DW4, AND LJ13

E.1.1 REGION SELECTION, TRAINING DATA, AND EVALUATION

Each target uses one iDEM checkpoint trained with the public implementation. We select A before fine-tuning, then compare training with data from A and with equally many untargeted examples. On GMM-40, the energy search defines nearest-center cells. Rejection of target samples outside A gives conditional corrective data; the comparator samples the full target. Both datasets are capped at $50{,}000$ examples. A separate set of $200{,}000$ target samples evaluates regional and partition probabilities.

On DW4 and LJ13, the search and assignment use sorted pairwise particle distances, which are unchanged by translations, rotations, or particle relabeling. Particle densities are defined with the center of mass fixed. We divide the $10{,}000$ released reference structures into 3333 for selecting regions, 3333 for corrective training, and 3334 for scoring. Selection uses cell bounds with nominal error allocation $\alpha/(2K)$, without accessing the training or scoring subsets. These MCMC references support empirical comparisons outside the i.i.d. target-sampling setting of Theorem 1.

The audit-selected data are the training-reference structures in A: 1292 for DW4 and 179 for LJ13. Untargeted training subsamples the same number without conditioning on A. Mini-batches sample with replacement from these fixed datasets. All training rules use the same 4000 initial model outputs for the preservation term.

Each fitted model generates 4000 new samples. We compare their frequency in A and their full-partition L_1 error with the scoring reference. The 3 optimization seeds share the initial checkpoint,

region, and reference split. For LJ13, 2 of the 1590 cells form A, whose scoring-reference frequency is 5.489%.

E.1.2 Corrective fitting and score preservation

The corrective loss fits the model to new data; the anchor limits changes in its predictions on original-model samples. For the anchor, $\widehat{q}_0$ denotes the empirical distribution giving equal weight to the 4000 fixed original-model samples. Both losses use the original noise schedule. For a clean sample x, draw $t \sim \mathrm{Unif}(0,1)$ and $\epsilon \sim \mathcal{N}(0, I)$, and set $x_t = x + \sigma_t \epsilon$, where σ_t is the noise standard deviation at t.

The network prediction is $s_\theta(x_t, t)$, and s_{θ_0} is the frozen original network. The denoising target is $u_t(x_t, x) = -\kappa(t)(x_t - x)/(\sigma_t^2 + 10^{-4})$, where $\kappa(t)$ is the original target-score scaling, or one when scaling is disabled. The small denominator offset stabilizes low-noise updates. With the original loss weight $\omega(t)$ and coordinate count d_x, the two terms in Equation 4b are

$$\mathcal{L}_{\mathrm{DSM}}(\theta; \mathcal{D}_{\mathrm{corr}}) = \frac{1}{|\mathcal{D}_{\mathrm{corr}}|} \sum_{x \in \mathcal{D}_{\mathrm{corr}}} \mathbb{E}_{t,\epsilon}\left[\frac{\omega(t)}{d_x} \|s_\theta(x_t, t) - u_t(x_t, x)\|_2^2\right], \tag{13}$$

$$\mathcal{R}_{\mathrm{anchor}}(\theta; \theta_0) = \mathbb{E}_{x \sim \widehat{q}_0, t, \epsilon}\left[\frac{\omega(t)}{d_x} \|s_\theta(x_t, t) - s_{\theta_0}(x_t, t)\|_2^2\right]. \tag{14}$$

A corrective update uses 51 corrective and 461 original-model examples, weighting their losses by these fractions. Retain-only uses 512 original-model examples and no corrective loss. The two networks use identical output scaling on the same noise-perturbed inputs, so the anchor penalizes changes in score predictions.

E.1.3 Effect of the preservation loss

Table 15 compares score anchoring with denoising score matching on original-model samples, keeping the corrective loss, region, data pools, and training budget fixed. On LJ13, audit-selected data give lower errors than untargeted data under both losses; with score anchoring, they also improve both metrics relative to the original model.

Table 15: LJ13 fine-tuning with two losses for retaining the original model's behavior: denoising score matching on original-model samples and score anchoring. Both use the same audit region and corrective data pools. Entries are means across 3 optimization seeds. Regional probability changes are relative to the original model, in percentage points; positive changes reduce the initial deficit. The original partition error is 0.4880.

	regional mass change (pp)		partition $L_1 \downarrow$	
training data	retained DSM	score anchor	retained DSM	score anchor
retain only	-0.667	-0.242	0.5429	0.5169
untargeted reference	-0.792	$+0.208$	0.5894	0.4436
audit-selected reference	-0.267	$+0.558$	0.5026	0.4004

E.2 Fine-tuning the released Many-Well flow

E.2.1 Corrective data and optimization

This study tests regional data selection when target sampling is available directly. We use the released FAB checkpoint and truncated-target sampler from Appendix D.2. Before evaluating the updated models, we fix $A = \{x_{23} > 0,\ x_{31} > 0\}$, the most frequently selected pair and cell in the preceding audit, with target probability approximately 71.286%. Each update uses a corrective dataset $\mathcal{D}_{\mathrm{corr}}$ drawn either from the full truncated target or from its restriction to A, with the same number of examples in both cases.

Both updates start from the released checkpoint and use Equation 4a with the same weight η. Each step draws 512 fresh samples from the frozen original model and 512 examples with replacement from the corrective dataset, weighting the two losses by $1 - \eta$ and η. Retain-only training instead

draws 1024 original-model samples. Thus all comparisons match optimization steps and likelihood evaluations, while the two corrective updates also match target-data counts (Table 14).

E.2.2 Evaluation of the updated model

For each of 8 optimization seeds, we generate 10^6 samples from each updated flow, using the same latent random draws for the compared models. A separate set of 20,000 target samples per seed evaluates negative log likelihood (NLL) and is never used for fitting. Regional error is zero if the generated frequency lies inside the target-probability interval, and otherwise is its distance to the nearer interval boundary. 4-cell error is the Euclidean distance between generated cell frequencies and the midpoints of their target intervals.

Table 5a reports means and sample standard deviations across seeds, including variation from training and evaluation sampling. For the fixed released model, only evaluation samples vary. At $\eta = 0.0203$, the paired differences between audit-selected and unconditional training are -0.519 ± 0.055 percentage points in regional error, $(-5.42 \pm 1.10) \times 10^{-3}$ in 4-cell error, and -0.0004 ± 0.0019 in NLL. These differences are formed within each seed before summarizing. Across all 16 double-well coordinates, the sum of absolute positive-sign probability errors remains close to the original model's value: 0.0297 with audit-selected data, compared with 0.0296 before training.

E.2.3 How strongly to fit the corrective data

We set $\eta = [\widehat{p}_C(A) - \widehat{q}_C(A)]/[1 - \widehat{q}_C(A)]$ so that the ideal training mixture matches the estimated target probability in A. The original-model probability was estimated from 600,000 samples, giving $\widehat{q}_C(A) = 0.70689$. Together with the numerical target probability of approximately 0.71285, this gives the point-estimate weight $\eta = 0.0203$. The updated model is evaluated on newly generated samples.

Table 16 also tests $\eta = 0.006$, 0.05, and 0.2, with $|\mathcal{D}_{\mathrm{corr}}| = \mathrm{round}(20{,}000\eta/(1-\eta))$. The fixed scale 20,000 sets corrective dataset sizes; original-model samples are drawn afresh throughout training. The sweep varies training weight and corrective dataset size jointly; comparisons within each row hold both fixed.

Table 16: Many-Well fine-tuning at 4 corrective-data weights and dataset sizes. Targeted and untargeted training share both settings within each row. Errors give means and sample standard deviations across 8 seeds. Regional error is measured in percentage points; 4-cell L_2 is in units of 10^{-3}. Over-filling is the percentage of audit-selected runs whose generated frequency in A exceeds the target upper bound. Table 5a reports the original-model and retain-only controls.

		regional error (%) ↓		4-cell L_2 ↓	
η (samples)	over-filling (%)	target	audit	target	audit
0.006 (121)	0	0.681 ± 0.145	0.494 ± 0.125	8.30 ± 1.57	6.12 ± 1.33
0.0203 (414)	12.5	0.627 ± 0.137	0.108 ± 0.116	7.66 ± 1.52	2.24 ± 0.90
0.05 (1053)	100	0.612 ± 0.114	0.696 ± 0.123	7.50 ± 1.33	8.36 ± 1.48
0.2 (5000)	100	0.488 ± 0.138	4.815 ± 0.139	6.12 ± 1.61	57.19 ± 1.67

The two smaller settings favor audit-selected data on regional and 4-cell errors. At the larger settings, all updated models trained on selected data over-fill A, and unconditional data give lower regional error. The retain-only control is shared across settings.

E.2.4 Separating training weight from sample count

The first sweep changes weight and dataset size together. We now vary them separately: first varying η with 414 corrective samples, then varying the dataset size from 121 to 5000 at $\eta = 0.0203$. Within each seed and data source, all weights use the same 414 samples, and smaller datasets are prefixes of the largest dataset. These runs use new corrective datasets with the same training and evaluation protocol; original-model and retain-only evaluations match Table 5a.

For audit-selected data, the training mixture assigns probability $(1-\eta)q_0(A)+\eta$ to A. We estimate $q_0(A)$ from the original model's evaluation draws for that seed and compare this prediction with the updated model's frequency. For unconditional data, the prediction replaces the final η by η times the fraction of corrective examples in A. Thus the prediction concerns the training mixture, and the comparison measures how closely the fitted flow reproduces its regional probability.

We also choose a weight from the audit bounds. In this full-domain experiment, $q_C = q_0$, and Section 4.2 permits $\eta \le \hat{d}_A/[1-U_q(A)]$ for the training mixture. Before training, a separate draw of $400{,}000$ original-model samples gives a one-sided Clopper–Pearson upper bound at failure probability 0.05. Together with the numerical target lower bound, it gives $\eta = 0.0126$. This bound controls the training mixture; the updated flow is evaluated in the same way as the other weights.

Table 17: Separating corrective-data weight and sample count in FAB fine-tuning. Entries are means across 8 seeds. Regional error is the mean absolute distance to the target probability interval, in percentage points (pp). Signed errors compare the audit-selected mixture prediction or generated frequency with the target interval; positive values indicate over-filling. The final column is the percentage of runs whose generated frequency exceeds the target upper bound. The first four rows vary weight at fixed sample count; the final three vary sample count at the shared weight 0.0203.

		regional error (pp) ↓		signed error in A (pp)		
η	samples	target	audit	predicted	generated	above target (%)
0.006	414	0.678	0.499	−0.502	−0.499	0
0.0126	414	0.663	0.332	−0.307	−0.332	0
0.0203	414	0.657	0.118	−0.082	−0.112	12.5
0.05	414	0.663	0.700	+0.791	+0.700	100
0.0203	121	0.659	0.125	−0.082	−0.116	25
0.0203	1053	0.638	0.133	−0.082	−0.126	25
0.0203	5000	0.654	0.133	−0.082	−0.114	25

At fixed data count, the mean signed regional error follows the prediction as η increases (Table 17). With the bound-derived weight, regional error falls by 51% relative to the original model, 4-cell error falls from 8.16 to 4.27 in units of 10^{-3}, and none of the 8 generated frequencies exceeds the target upper bound. At $\eta = 0.05$, both the predicted mixture and every fitted model over-fill A. At fixed $\eta = 0.0203$, mean absolute regional errors remain between 0.118 and 0.133 percentage points as the sample count changes. In this range, the weight has the larger effect on regional correction. The larger datasets modestly lower mean target NLL, from 27.6660 at 121 samples to 27.6585 at 5000; both remain above the original model's 27.6058.

E.3 CG POTENTIAL FINE-TUNING

E.3.1 DATA ALLOCATION AND FITTING PROTOCOL

Each protein is adapted separately from the released general CG potential using that protein's reference structures. Training uses half of the interleaved reference blocks, and a disjoint quarter is reserved for scoring. Only training data provide probe starting structures, reference state frequencies, and positive training examples. Uniform-all selects the same nominal number of frames per audit cell; deficit weighting selects from flagged cells in proportion to their deficit scores. Within-cell sampling uses replacement only when demand exceeds the available frames. Counts are rounded down, so realized dataset sizes can differ despite matched nominal budgets.

We initialize the neural component from the released general CG potential and keep its physical prior fixed. The dimensionless energy is $E_\theta(x) = [V_{\mathrm{prior}}(x) + V_\theta^{\mathrm{NN}}(x)]/(k_{\mathrm{B}}T)$, where V_{prior} is the fixed physical energy, V_θ^{NN} is the trainable neural energy, and $k_{\mathrm{B}}T$ is the thermal energy. Differentiating negative log likelihood gives the data-average energy gradient minus its model expectation. Equation 5 estimates these terms with reference mini-batches and Brownian-chain samples, respectively. Gradients pass through energy evaluations, with the simulated coordinates held fixed.

The Brownian chains start from the original model's structures and retain their current structures between parameter updates, continuing under the updated potential. Each mini-batch draws b_-

structures with replacement from the current chain states. Both allocation arms use the same update rule, chain procedure, and optimizer. A model-output control replaces the reference training examples with samples from its own chains.

All three proteins use 250-frame blocks, $K = 8$ audit cells, 16 evaluation cells, and a nominal corrective dataset of 4000 structures. Basin search uses the protein-specific potential for chignolin and the general potential for villin and protein G. The optimization settings are shared (Table 14). Each update advances 256 persistent chains by 20 Brownian steps at 350 K with integration step 0.001.

E.3.2 HELD-OUT EVALUATION AND RESULTS

Evaluation starts 1200 chains from the original model's structures and scores their final structures after 2000 Brownian steps. Compared potentials share starting structures and simulation noise within each seed. The scoring reference defines bins and partitions for the four metrics in Appendix D.4.2.

Table 18a reports the errors under the original potential and each training scheme. Deficit-weighted training lowers mean free-energy and energy-distance errors on chignolin and villin relative to the original potential. Chignolin's partition error increases, whereas villin improves on all four metrics. On protein G, both reference-data training schemes lower all four mean errors relative to the original potential. Deficit weighting gives lower mean energy-distance and partition errors; uniform-all gives lower mean free-energy and histogram errors. Table 18b gives the paired differences between the two allocations.

Table 18: CG evaluation from original-model starting structures: means and sample standard deviations across 5 seeds for each protein. (a) Errors after 2000 Brownian steps under each potential; the model-output control trains on its own simulation samples. (b) Paired error differences (deficit-weighted minus uniform-all); negative values favor deficit weighting.

(a) Errors under each potential

model / training data	MAE_F (kT) ↓	ED ↓	R_g hist. error ↓	partition L_1 ↓
Chignolin				
original potential	2.460 ± 0.057	11.660 ± 0.184	1.224 ± 0.011	1.222 ± 0.011
model-output control	2.427 ± 0.058	11.705 ± 0.537	1.221 ± 0.023	1.222 ± 0.012
uniform-all	2.557 ± 0.063	12.546 ± 0.382	1.295 ± 0.023	1.246 ± 0.030
deficit-weighted	2.164 ± 0.092	5.615 ± 0.765	1.207 ± 0.060	1.377 ± 0.064
Villin				
original potential	1.276 ± 0.074	16.068 ± 0.365	0.965 ± 0.008	1.184 ± 0.021
model-output control	1.260 ± 0.038	15.924 ± 1.043	0.936 ± 0.084	1.176 ± 0.047
uniform-all	1.295 ± 0.113	11.543 ± 1.132	0.947 ± 0.061	1.009 ± 0.025
deficit-weighted	1.150 ± 0.083	6.956 ± 0.791	0.716 ± 0.071	0.872 ± 0.025
Protein G				
original potential	0.614 ± 0.032	61.214 ± 1.794	0.568 ± 0.023	0.941 ± 0.018
model-output control	0.613 ± 0.029	61.200 ± 1.865	0.568 ± 0.025	0.941 ± 0.018
uniform-all	0.565 ± 0.067	51.434 ± 1.929	0.479 ± 0.024	0.922 ± 0.018
deficit-weighted	0.568 ± 0.097	50.057 ± 3.945	0.527 ± 0.121	0.919 ± 0.019

(b) Paired error differences

protein	MAE_F (kT)	ED	R_g hist. error	partition L_1
chignolin	-0.393 ± 0.084	-6.931 ± 0.795	-0.088 ± 0.063	$+0.131 \pm 0.071$
villin	-0.145 ± 0.054	-4.586 ± 1.127	-0.231 ± 0.085	-0.137 ± 0.013
protein G	$+0.004 \pm 0.123$	-1.377 ± 2.886	$+0.048 \pm 0.124$	-0.004 ± 0.005