A SYSTEMATIC STUDY OF SG-MCMC SAMPLERS FOR IN-CONTEXT ENERGY-BASED TABULAR GENERATION WITH TABPFN

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy-based models built on pre-trained tabular in-context learners sample synthetic data from a classifier-derived energy landscape, relying on Stochastic Gradient Langevin Dynamics (SGLD) without comparing it against alternative Stochastic Gradient-Markov Chain Monte Carlo schemes. Using TabPFNv2 as an energy oracle, we investigated various sampler choices, revealing that substituting SGLD with a momentum-based sampler yields substantial and consistent improvements in model fidelity. Even without injected noise, in-context gradients exhibit fluctuations driven by the underlying geometry of the energy surface. The momentum can filter it by making short-run generation a structural necessity. For coverage, we show a clean double dissociation: replica exchange alone reaches under-represented modes under biased initialization. These findings suggest a primary design heuristic for in-context energy-based tabular generation: leverage momentum to enhance fidelity and employ replica exchange to maintain coverage.

1 INTRODUCTION

Tabular data is the dominant format in medicine, finance and sciences. Augmenting a small tabular training set with synthetic points is an appealing remedy, but most generative models overfit in this regime and degrade rather than improve downstream performance (Manousakas & Aydöre, 2023; Seedat et al., 2024). A recent line of work sidesteps the training problem entirely: instead of fitting a generator, it takes a pre-trained tabular in-context learner (Müller et al., 2022; Hollmann et al., 2023; 2025), reinterprets its logit values as an energy function in the manner of Grathwohl et al. (2020), and draws synthetic points by running a Markov chain on that energy (Ma et al., 2023; Margeloiu et al., 2024). No parameters are updated, and the very few resulting generators improve downstream accuracy on small tabular datasets.

Every method in this context samples with Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011). The approaches concern SGLD's own hyperparameters rather than alternatives to SGLD; the original class-specific formulation reports that results are stable across step sizes and noise scales. This is a consequential gap because the energy being sampled is unlike the energy landscapes for which SG-MCMC samplers are typically characterized: it is produced by a forward pass through a transformer conditioned on a handful of points, and nothing guarantees that it is smooth, that its gradient is well-behaved, or that the sampler assumptions hold.

We study what the sampler choice does in this setting. Using a class-specific in-context energy as a fixed substrate, we compare SGLD against three alternatives that each add a different dynamical ingredient: momentum with constant friction (SGHMC; Chen et al. 2014), momentum with an adaptive Nosé–Hoover thermostat (SGNHT; Ding et al. 2014), and replica exchange across a temperature ladder, and we evaluate them along two axes that a single fidelity number conflates: how faithfully the synthetic points reproduce the data, and whether the sampler can reach parts of a class it was not initialized in.

Our findings are summarized in Figure 1 and are not the ones we expected. On the fidelity axis, replacing SGLD with either momentum sampler is a large and consistent improvement, and the thermostat, the ingredient designed precisely to handle unknown gradient noise, contributes nothing measurable. On the coverage axis, only replica exchange can reach an under-represented mode from

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

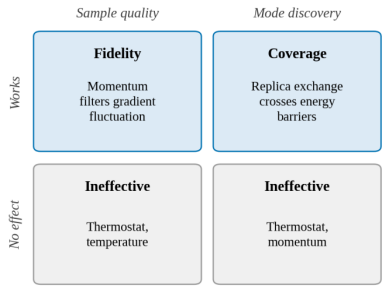

Figure 1: A 2×2 view of sampler design for in-context energy-based tabular generation. Momentum improves fidelity, replica exchange improves coverage, and the two act on independent axes.

a biased initialization, and it is the worst choice when the data already covers every mode. The two axes are hence served by different mechanisms, and neither substitutes for the other.

We make the following contributions.

- We compare four SG-MCMC samplers on a fixed in-context energy across four tabular datasets and a controlled multimodal benchmark and obtain a compact design rule: momentum for fidelity, replica exchange for coverage. Replacing SGLD with momentum raises k-NN support precision by 0.326 (95% CI [0.275, 0.378]) with significant improvements on every dataset, whereas adding a thermostat changes it by 0.001 (CI $[-0.007, +0.009]$).
- We locate the cause in the energy surface rather than in the sampler. The in-context gradient carries a super-diffusive fluctuation, with a diffusion exponent of about 0.63, that does not diminish when all injected noises are removed. Momentum filters this fluctuation; a thermostat, which regulates thermal noise, has nothing to regulate. A corollary is that short-run generation is a structural property of this landscape rather than a tuning artifact.
- We exhibit a double dissociation on the coverage axis: replica exchange is uniquely able to reach an under-represented mode under biased initialization and uniquely harmful when the data already seeds every mode. A barrier dose–response confirms the mechanism, since the advantage vanishes when the barrier does.
- We report a methodological constraint on the implementation of the family of generators: the energy is only non-degenerate if the backbone exposes unnormalized logit values.

While we do not propose a new sampler, the contribution demonstrates which of the ingredients matter in this generative setting, and why.

2 Background and Related Work

In-context models for tabular data. Prior-data fitted networks are transformers pre-trained on synthetic priors so that they approximate posterior predictive inference in a single forward pass (Müller et al., 2022). TabPFN (Hollmann et al., 2023; 2025) instantiates this for small tabular classification and is competitive with tuned gradient boosting without any parameter updates on the target data. Because "training" amounts to placing a labeled set in context, such a model can be re-conditioned on an arbitrary task at trivial cost being potentially exploited by the generators illustrated below.

Energy-based models from classifiers. Grathwohl et al. (2020) observed that the logits of a classifier already define a joint energy, $E(\boldsymbol{x}, y) = -f_\theta(\boldsymbol{x})[y]$, and that marginalizing over labels yields $E(\boldsymbol{x}) = -\operatorname{LogSumExp}_y f_\theta(\boldsymbol{x})[y]$. Sampling this energy turns a discriminative model into a generative one. Applying the idea to tabular data, TabPFGen (Ma et al., 2023) uses a single shared in-context classifier and samples the class-conditional energy, while TabEBM (Margeloiu et al., 2024) builds one energy per class from a surrogate binary task that separates that class from negative points placed far from the data, which makes the inferred densities robust to class overlap and

imbalance. Both draw samples with SGLD; TabEBM additionally reports insensitivity to SGLD's step size and noise scale, but neither compares SGLD against another sampler. We adopt the class-specific formulation as a substrate because it is the stronger of the two and because its per-class structure affects mode coverage of a single class.

Stochastic gradient MCMC. SGLD (Welling & Teh, 2011) discretizes Langevin dynamics with a stochastic gradient. SGHMC (Chen et al., 2014) adds a momentum variable with constant friction, which suppresses the random-walk behavior of first-order dynamics. SGNHT (Ding et al., 2014) replaces the constant friction with a Nosé–Hoover thermostat that adapts to the unknown magnitude of the gradient noise, and is therefore the natural candidate when the gradient comes from a black box. Replica exchange runs chains at several temperatures and swaps configurations between them (Swendsen & Wang, 1986; Earl & Deem, 2005). These samplers are well characterized on analytic targets and on Bayesian posteriors; what is not known is which of their ingredients matter when the target is an in-context energy and the goal is generation rather than posterior inference.

3 METHOD

The objective of this study is not building a new generative model but improving the sampler inside an existing one. We therefore fix the energy model and the generation protocol and vary only the SG-MCMC update rule.

3.1 CLASS-SPECIFIC ENERGY FROM AN IN-CONTEXT CLASSIFIER

We adopt the class-specific energy formulation of prior work on energy-based tabular augmentation. For each class c we build a surrogate binary task that separates the real points of that class, X_c, from a small set of negative points N_c placed at the corners of a hypercube around the data: each negative has coordinates $\pm\alpha_{\text{dist}}\sigma_d$ in dimension d, where σ_d is the per-dimension standard deviation of the training data. An in-context tabular classifier f_c is conditioned on this labeled set, with no parameter updates, and its two unnormalized logits are reinterpreted as an unnormalized joint density, giving the class-conditional energy

$$E_c(\boldsymbol{x}) = -\,\mathrm{LogSumExp}\left(f_c(\boldsymbol{x})[0],\, f_c(\boldsymbol{x})[1]\right), \tag{1}$$

so that $p(\boldsymbol{x} \mid y = c) \propto \exp(-E_c(\boldsymbol{x}))$. Generation then samples each class independently from its own energy and concatenates the results with the empirical class proportions, so that the label distribution of the synthetic set matches that of the real training split. Algorithm 1 (Appendix C) summarizes the procedure; the sampler S in its line 5 is the only component we vary.

Two implementation properties of the backbone are essential: Equation 1 requires unnormalized logits, which is why we use TabPFN v2 (Hollmann et al., 2025), and the query point must remain differentiable through the transformer (Appendix C).

3.2 SAMPLERS

We compare four samplers that differ in which dynamical ingredients they add to gradient descent on the energy. Throughout, h is the step size, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, τ is the target temperature and D is the data dimension.

SGLD the unadjusted Langevin update used by prior work, adds isotropic noise to the gradient step,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{h}{2}\nabla E(\boldsymbol{x}_t) + \sqrt{h}\,\boldsymbol{\epsilon}_t. \tag{2}$$

We set the noise scale to its theoretical value $\sqrt{h}$ rather than tuning it, so that no sampler benefits from a hand-set noise level.

SGHMC introduces a velocity with constant friction C,

$$d\boldsymbol{v} = -\nabla E(\boldsymbol{x})\,dt - C\boldsymbol{v}\,dt + \sqrt{2C\tau}\,d\boldsymbol{W}, \qquad d\boldsymbol{x} = \boldsymbol{v}\,dt, \tag{3}$$

which we integrate with a symmetric (leapfrog) splitting: a half kick on $\boldsymbol{v}$, a full drift on $\boldsymbol{x}$, a gradient re-evaluation, and a second half kick. Each half step injects noise of scale $\sqrt{Ch\tau}$, so the two half steps reproduce the variance $2Ch\tau$ of Equation 3.

SGNHT replaces the constant friction by a Nosé–Hoover thermostat: an auxiliary scalar ξ is driven by the gap between the measured kinetic energy and the target temperature and feeds back as an adaptive friction,

$$d\boldsymbol{v} = -\nabla E(\boldsymbol{x})\,dt - \xi\boldsymbol{v}\,dt + \sqrt{2A\tau}\,d\boldsymbol{W}, \qquad d\boldsymbol{x} = \boldsymbol{v}\,dt, \qquad d\xi = \Big(\frac{\|\boldsymbol{v}\|^2}{D} - \tau\Big)dt. \tag{4}$$

We use the same leapfrog splitting, maintain an independent ξ per chain, and initialize $\xi_0 = A$. The Euler–Maruyama discretization of Equation 4 was unstable in preliminary two-dimensional tests, so the leapfrog scheme is used throughout.

Replica exchange (RENHD) runs $K = 4$ chains at temperatures $\tau_k \in \{1, 2, 4, 8\}$, each evolving by Equation 4 with its own temperature, and every 20 steps proposes swaps between adjacent temperature pairs (alternating between even and odd pairings). A swap between replicas i and j is accepted with probability

$$p = \min\Big(1,\ \exp\left[(\beta_i - \beta_j)(E(\boldsymbol{x}_i) - E(\boldsymbol{x}_j))\right]\Big), \qquad \beta_k = 1/\tau_k, \tag{5}$$

applied independently to each element of the batch. Samples are read from the $\tau = 1$ replica. Note that each replica is itself a thermostatted chain, so SGNHT is the $K = 1$ special case of this sampler; any coverage advantage of RENHD over SGNHT is therefore attributable to the exchange and not to the thermostat.

Every sampler is kept at its theoretical setting ($\tau = 1$, $C = A = 1$; Appendix C). Unless stated otherwise all experiments use $h = 0.01$, $\sigma_{\text{start}} = 0.1$, $T = 50$ steps and 500 synthetic points per dataset for the fidelity experiments, and $T = 600$ steps with 300 parallel chains for the coverage experiments. For the negative distance we use $\alpha_{\text{dist}} = 1$ in the real-data fidelity experiments and $\alpha_{\text{dist}} = 5$, the value prescribed by the original formulation, in the synthetic multimodal study, whose data spread is much larger. Appendix A repeats the fidelity comparison at the prescribed value: every conclusion is unchanged and the advantage of momentum over SGLD is larger there ($+0.187$ versus $+0.152$ in precision), so reporting at $\alpha_{\text{dist}} = 1$ is the conservative choice.

4 Experimental Setup

4.1 Evaluation

Fidelity. We use the k-nearest-neighbor support estimates of precision and recall introduced by Kynkäänniemi et al. (2019). Writing $r_k(\boldsymbol{z})$ for the distance from $\boldsymbol{z}$ to its k-th nearest neighbor within its own set, the support of a set is the union of the balls $B(\boldsymbol{z}, r_k(\boldsymbol{z}))$ over its members; precision is the fraction of synthetic points that fall inside the support of the real set, and recall is the symmetric quantity with the roles of the two sets exchanged. Two details differ from the original construction, which was developed for images: we compute the estimate directly in the standardized feature space rather than in a pre-trained embedding, and we use $k = 5$ rather than $k = 3$. We additionally report the median distance to the closest record (DCR) (Zhao et al., 2021) as a memorization and privacy diagnostic, the median over synthetic points of the Euclidean distance to the nearest real record, which is small when a generator memorizes and grows as it extrapolates.

Utility. We evaluate synthetic data in the augmentation setting, the utility measure adopted in this literature (Xu et al., 2019; Kotelnikov et al., 2023): a gradient boosted tree is trained on the real training split together with the synthetic points and evaluated by balanced accuracy on the untouched real test split. This is distinct from a train-on-synthetic protocol, and the two are not comparable.

Coverage. For the multimodal study we assign each generated point to its nearest mode center and compare the realized mode proportions $\hat{\pi}$ with the target proportions π. Our primary summary is the total-variation distance $\frac{1}{2}\sum_k |\hat{\pi}_k - \pi_k|$, which penalizes over-representation as well as under-representation; a coverage measure alone cannot distinguish a sampler trapped in one mode from

one that floods a minority mode. We also report the minority coverage $\hat{\pi}_{\min}$ against its target, the escape fraction $1 - \hat{\pi}_0$ of chains that leave the initial basin, and the number of modes receiving no samples.

Statistics. Every comparison is paired: we match runs on the random seed (and on the dataset when pooling), form the per-pair difference against the baseline sampler, and report its mean with a 95% confidence interval from the t distribution. A difference is significant when the interval excludes zero and explicit in the text when an interval is narrow enough to bound an effect rather than merely failing to detect it. Unless stated otherwise all results are over five seeds. Seeds control the train/test split and the chain initialization; the sampler noise is drawn from the global generator, so the reported intervals also absorb that source of variation.

4.2 Data

The fidelity experiments use four public tabular classification datasets. Each is split 70/30 into training and test partitions with stratification, and every feature is standardized using training-split statistics only. None of the four datasets contain missing values. Following the low-sample setting in which this family of augmentation methods is normally evaluated, we subsample the training split of biodeg, steel and stock to $N_{\text{real}} = 100$, and generate $N_{\text{syn}} = 500$ points per dataset. For the imbalanced case we subsample the minority class of diabetes to a 10:1 ratio, leaving the test split untouched, which gives $N_{\text{real}} = 385$ with 35 minority instances.

The coverage experiments use a controlled two-dimensional mixture so that the correct answer is known by construction: a single class whose distribution is three unit-variance Gaussian modes at the vertices of an equilateral triangle with side $d \in \{8, 12, 16\}$ standard deviations, populated in the proportions 0.80 / 0.15 / 0.05 from 300 samples. This sets the barrier height by construction and allows us to verify it directly on the learned energy.

5 Fidelity

We first ask whether the choice of SG-MCMC sampler changes the statistical fidelity of the generated data, holding the energy model, the operating point and the fixed evaluation protocol.

We evaluate on three real tabular datasets (biodeg, steel, stock) at the operating point given above and, as a separate class-imbalanced case, on the diabetes dataset subsampled to a 10:1 imbalance. The three balanced datasets are pooled by matching on (dataset, seed), which absorbs differences in the absolute difficulty of each dataset; precision and recall are on a common $[0, 1]$ scale, so the paired differences can be averaged directly. We evaluate SGLD, SGHMC and SGNHT on this axis. Replica exchange is deferred to the coverage experiments, where the question it is designed to address arises and where its temperature ladder is numerically stable; we return to this restriction in Appendix C.4.

5.1 Replacing SGLD with Momentum

Tables 1 and 2 report the pooled result. Moving from SGLD to SGHMC raises precision from 0.486 to 0.812, a paired improvement of $+0.326$ (95% CI [0.275, 0.378]), and the effect is significant on each of the three datasets individually (biodeg $+0.348$, steel $+0.231$, stock $+0.400$). The same substitution reduces the median distance to the closest real record from 3.45 to 2.27 (-1.17, CI [-1.41, -0.93]), that is, the synthetic points move closer to the data manifold rather than farther from it. On the imbalanced diabetes case the direction is identical, with precision rising from 0.792 to 0.943 ($+0.152$, CI [0.132, 0.171]).

Adding a Nosé–Hoover thermostat on top of momentum changes nothing that we can measure. The paired difference between SGNHT and SGHMC is $+0.001$ (CI [-0.007, $+0.009$]) in precision and -0.000 (CI [-0.013, $+0.013$]) in DCR. The interval is narrow enough to bound the effect rather than merely fail to detect it: any true gain from the thermostat at this operating point is smaller than 0.009 in precision, two orders of magnitude below the gain from momentum itself.

Downstream augmentation utility is statistically indistinguishable across samplers (Appendix B), a decoupling we return to in Section 7. Whether a sampler can also find parts of a class it was not initialized in is a separate question, which we take up next.

Table 1: Fidelity by sampler, pooled over biodeg, steel and stock ($n = 15$ seed–dataset pairs, mean with 95% CI).

Sampler	Precision	Recall	DCR (median)	ESS
SGLD	0.486 [0.372, 0.601]	0.999	3.446 [2.76, 4.13]	5.85
SGHMC	0.812 [0.736, 0.889]	0.998	2.275 [1.83, 2.72]	3.34
SGNHT	0.813 [0.740, 0.886]	0.999	2.275 [1.83, 2.72]	3.22

Table 2: Paired differences between samplers on the same pooled data ($n = 15$, 95% CI on the seed-matched difference). Bold marks an interval that excludes zero.

Comparison	Precision Δ	DCR Δ	ESS Δ
SGHMC − SGLD	$\mathbf{+0.326}$ **[+0.275, +0.378]**	$\mathbf{-1.17}$ **[−1.41, −0.93]**	$\mathbf{-2.51}$ **[−3.09, −1.93]**
SGNHT − SGHMC	$+0.001$ [−0.007, +0.009]	$+0.000$ [−0.013, +0.013]	-0.12 [−0.34, +0.10]

5.2 Why Momentum Helps

The gap between SGLD and the momentum samplers is large enough to invite a mechanistic explanation, and the natural candidate, that SGLD is injecting too much thermal noise, turns out to be wrong. Three experiments locate the cause in the energy surface itself; the first, a temperature sweep, is in Appendix B.2.

The diffusion is driven by the gradient, not by the noise. Figure 2(a) repeats the long run with the injected noise switched off entirely ($\tau = 0$). Removing the noise does not reduce the spread. Starting from a median distance to the data of about 0.38, SGHMC reaches 4.55 with noise and 4.88 without it after 500 steps, and SGNHT reaches 4.89 and 4.92 respectively; if the drift away from the data were thermal in origin, the noiseless runs would barely move. Fitting a power law over $t \geq 50$ gives exponents of 0.63–0.64 across all four configurations, which lie between the 0.5 of a random walk and the 1.0 of a systematic drift. The samples therefore spread super-diffusively under a force that survives the removal of all thermal noise, which identifies the in-context gradient as the source.

The landscape is well formed; the dynamics are not. Figure 2(b) shows that the mean energy of the samples rises monotonically from about -5.0 to between -3.53 and -3.42 over the run and never approaches the mean energy of the real data (-5.36). The samples are not descending into a spurious low-energy region outside the data; they are being pushed *up* out of the correct basin. Taken with the previous result, super-diffusive component of the in-context gradient is what removes samples from the data, that momentum acts as a low-pass filter on exactly this component, and that a thermostat, a device for regulating thermal noise, has nothing to regulate. It also means that short-run generation is not a tuning artifact but a structural consequence of the landscape: the drift away from the data cannot be switched off, only slowed.

5.3 The Thermostat over Longer Runs

If the thermostat needs time to adapt, its effect should appear as the run lengthens. It does, but with the opposite sign to the one intended. Figure 4 in Appendix B reports the paired difference between SGNHT and SGHMC at $T \in \{50, 200, 500\}$: $+0.005$ (CI [-0.012, $+0.022$]) at 50 steps, $\mathbf{-0.014}$ **(CI [−0.026, −0.003])** at 200, and $\mathbf{-0.028}$ **(CI [−0.049, −0.007])** at 500, with the median DCR difference reaching $+0.329$ (CI [$+0.129$, $+0.530$]) at 500 steps. The thermostat does act on the dynamics once given enough steps, and what it does is restore the canonical temperature, which in a sampler that works because it is deliberately biased towards the data means diffusing away faster. Faithful sampling and good generation pull in opposite directions here. Nor does it buy robustness (Appendix B.4).

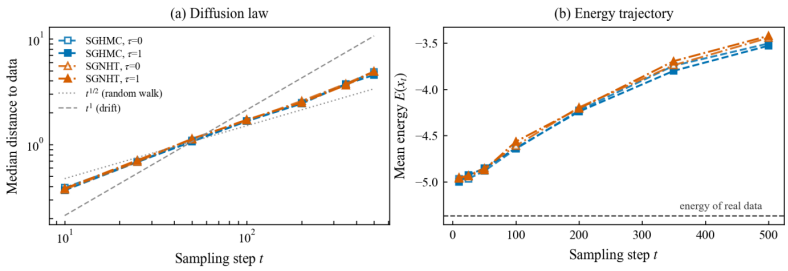

Figure 2: Why momentum helps. (a) Distance to the data grows with an exponent of 0.63–0.64, faster than a random walk, and is not reduced when the injected noise is switched off ($\tau = 0$), indicating that the diffusion originates from the gradient rather than from thermal noise. (b) The mean energy of the samples rises monotonically and never approaches the energy of the real data, so the landscape is well formed and the drift away from the data is a property of the dynamics.

6 Coverage

The fidelity axis measures whether synthetic points land where real data lives. It says nothing about whether the sampler can find parts of a class it has not been initialized in. We study that separately, on data where we control the answer.

6.1 A Controlled Multimodal Setting

We construct a single class whose distribution is a mixture of three well-separated Gaussian modes with proportions 0.80 / 0.15 / 0.05, so that the generator must place 5% of its samples in a mode that carries very little of the data. Two initializations define the two regimes we want to contrast: *data-init*, which seeds chains near randomly chosen real points and therefore places chains in every mode from the start, as in prior work; and *single-init*, which seeds every chain in the dominant mode, so that reaching the minority mode requires crossing the intervening region.

Because the whole experiment presumes that a barrier exists, we measured it before running any sampler (Appendix C.3); we use 16σ as the main configuration and 8σ, 12σ as lower-barrier controls.

6.2 A Double Dissociation

Table 3: Coverage at 16σ separation with $T = 600$ steps (five seeds, minority target 0.05). No mode is left empty under data-init for any sampler. The same sampler is the unique failure under data-init and the only one to reach the minority mode under single-init.

	Data-init		**Single-init**		
Sampler	**Minority**	**TV**	**Minority**	**TV**	**Empty modes**
SGLD	0.049	0.037	0.000	0.200	2.0
SGHMC	0.049	0.036	0.005	0.191	0.6
SGNHT	0.049	0.036	0.001	0.199	1.8
RENHD	0.073	0.050	0.024	0.138	0.0

Table 3 contains the central result of this section. Under data-init every sampler places close to the target proportion in the minority mode. SGLD, SGHMC and SGNHT all reach 0.049 with no empty modes, while replica exchange is the only method that fails, placing 0.073, i.e. 1.47 times the target (ratio error $+0.47$, CI $[+0.02, +0.91]$) and moving 3.7% more of its chains out of the dominant basin than SGLD (escape fraction $+0.037$, CI $[+0.029, +0.044]$). Its TV distance is the highest of the four, although that difference is not significant on its own.

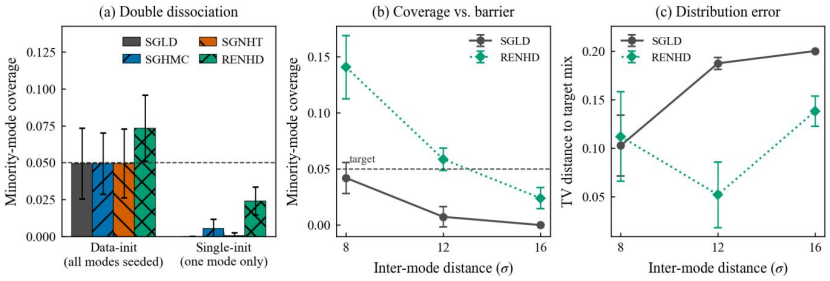

Figure 3: Coverage results at 16σ (a) and across barrier heights (b, c). (a) Double dissociation: replica exchange is the only sampler that reaches the minority mode from a biased initialization, and the only one that overshoots the target proportion when the data already seeds every mode. (b) Replica exchange overshoots the target roughly threefold at the lowest barrier, matches it at 12σ and falls to half the target at 16σ, while SGLD degrades to zero coverage. (c) In total-variation distance to the target mixture, replica exchange holds no advantage at the lowest barrier and becomes significantly better only once a barrier is present. Error bars are 95% CIs over five seeds.

Under single-init the ordering inverts completely. SGLD reaches the minority mode with exactly zero samples and leaves two of the three modes empty; SGHMC and SGNHT are barely distinguishable from it (0.005 and 0.001, neither significant against SGLD), and only replica exchange reaches it at all, placing 0.024 ($+0.024$ against SGLD, CI $[+0.015, +0.033]$). This is still only half of the target (ratio error -0.52, CI $[-0.71, -0.33]$), but it leaves no mode empty against two for SGLD (-2.0, CI $[-2.0, -2.0]$) and lowers the TV distance to 0.138 from 0.200 (-0.062, CI $[-0.078, -0.046]$). The escape fraction tells the same story directly: 0.062 of the chains leave the dominant basin under replica exchange, against 0.000 for SGLD.

The same sampler is thus the unique failure in one regime and the unique success in the other. Two consequences follow. First, the momentum, which dominates the fidelity axis, contributes almost nothing to exploration: SGHMC and SGNHT sit with SGLD in the failing group, reaching 0.005 and 0.001 of the minority mode against 0.024 for replica exchange. SGHMC does improve significantly on SGLD (TV -0.009, CI $[-0.015, -0.004]$; 1.4 fewer empty modes), but by an order of magnitude less than the exchange does, so the two axes are served by different mechanisms. Second, the reason SGLD has been adequate in prior work becomes visible: with the initialization used there, coverage is supplied by the initialization rather than by the sampler, and no sampler needs to cross anything. Replica exchange has value precisely and only where exploration is required and is actively harmful where it is not.

6.3 The Barrier Is the Cause

If replica exchange wins by crossing barriers, its advantage should disappear when there is no barrier to cross. Figure 3(b, c) varies the separation at fixed run length. In TV distance, our primary summary, the advantage over SGLD is absent at the lowest barrier ($+0.009$, CI $[-0.049, +0.067]$ at 8σ, not significant) and becomes significant only once the barrier grows (-0.135, CI $[-0.172, -0.099]$ at 12σ; -0.062, CI $[-0.078, -0.046]$ at 16σ). This is the dose–response relationship the causal claim requires, together with its converse: remove the barrier and the advantage vanishes. Appendix D.3 decomposes the same effect in minority coverage.

7 Discussion

Two axes, two mechanisms. The results support a compact account of what a sampler contributes here. Fidelity is governed by how well the sampler resists the fine-scale fluctuation of the in-context gradient, and momentum is the ingredient that does this. Coverage is governed by whether the sampler can cross the energy barrier between modes of the same class, and replica exchange is the ingredient that does so. The two are largely independent in both directions: momentum buys almost nothing on coverage, and replica exchange distorts the mixture where exploration is not needed (a 1.5-fold overshoot of the target mode proportion under data-init). Because each replica

in the replica-exchange sampler is itself a thermostatted chain, SGNHT is its single-replica special case, so the coverage advantage is attributable to the exchange rather than to any property of the underlying dynamics.

Faithful sampling is not good generation. SGNHT is, in the MCMC sense, the more principled of the two momentum samplers: it corrects the sampling temperature towards its target in the presence of unknown gradient noise. That correction is exactly what harms it here. These generators work because they are run for a short time from an initialization near the data, that is, because they are deliberately biased towards the data rather than equilibrated. Restoring the canonical temperature undoes part of that bias, which is why the thermostat's effect is null at 50 steps and significantly negative at 500. The same logic explains why short runs are not a tuning trick: the gradient's super-diffusive component cannot be switched off, only slowed, so any long run drifts away from the data. We measured the drift directly and found the mean sample energy rising monotonically away from the energy of the real data, never towards a spurious low-energy region. The landscape is sound; the dynamics are the problem.

What the sampler is for. Under the initialization used in prior work, coverage of the modes of a class is supplied by the initialization, not by the chain; every sampler we tested succeeds there. This reframes the role of the sampler in this family of generators. Its job is local refinement of points that are already in the right region, and the case for a more exploratory sampler arises only when one expects the data to under-represent a part of the class, which is precisely the situation that motivates augmentation for imbalanced data. A practitioner can act on this directly: use a momentum sampler by default, keep the run short, and switch to replica exchange only when under-represented structure is suspected, accepting a small fidelity cost and tuning the run length.

Fidelity gains did not become utility gains. A change of 0.33 in precision left downstream balanced accuracy statistically unchanged, in both the balanced and the imbalanced settings and at both negative distances. We report this as an open problem rather than a result about samplers. It may be that the support-based fidelity measures adopted in this literature are insensitive to the properties that determine downstream value, or that the augmentation benefit at this sample size is dominated by the real data; distinguishing these requires an evaluation protocol we do not have.

Limitations. Our coverage study uses a controlled two-dimensional mixture, which is what makes the barrier measurable and the target proportions known, but it leaves open whether intra-class multimodality of this kind is common in real tabular data and whether the same ordering holds there. All results use a single in-context backbone; the mechanism we identify is a property of that energy family, and while we expect it to hold for classifier-derived energies more generally, we have not verified this. We use five seeds throughout, and the sampler noise is drawn from the global generator rather than a per-run stream, so the reported intervals absorb that variation as well as the variation from the data split and the chain initialization. Finally, the imbalanced case rests on a single dataset, and our claims about the thermostat are bounded by the operating points we swept rather than universal. Replica exchange is evaluated only in the controlled multimodal setting, for reasons of numerical stability detailed in Appendix C.4.

8 Conclusion

We asked what the choice of SG-MCMC sampler does to an energy-based tabular generator built on an in-context learner, in a literature that has settled on SGLD without comparison. The answer separates into two independent axes. Replacing SGLD with momentum improves fidelity substantially and consistently, because momentum filters a super-diffusive fluctuation that the in-context gradient carries even in the absence of injected noise; the Nosé–Hoover thermostat, the ingredient designed for unknown gradient noise, adds nothing, and at longer horizons its temperature correction actively hurts. Replica exchange is the only sampler that recovers an under-represented mode from a biased initialization, and the only one that distorts the mode proportions when no exploration is needed. Sampler choice in this setting is therefore not a search for a single best method but a matching of the sampler's dynamics to the generative goal, and the short runs that these generators rely on are a structural consequence of the landscape rather than an artifact of tuning.

AI USE STATEMENT

Generative AI tools were used solely to aid or polish the writing of this manuscript. All AI-assisted language edits were reviewed by the authors, who take full responsibility for the content.

REFERENCES

Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1683–1691. PMLR, 2014.

Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems 27*, 2014.

David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005. doi: 10.1039/B509983H.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

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 32*, pp. 3929–3938, 2019.

Junwei Ma, Apoorv Dankar, George Stein, Guangwei Yu, and Anthony Caterini. TabPFGen: Tabular data generation with TabPFN. In *NeurIPS 2023 Second Table Representation Learning Workshop*, 2023.

Dionysis Manousakas and Sergül Aydöre. On the usefulness of synthetic tabular data generation. In *Data-Centric Machine Learning Research Workshop at the 40th International Conference on Machine Learning*, 2023.

Andrei Margeloiu, Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. TabEBM: A tabular data augmentation method with distinct class-specific energy-based models. In *Advances in Neural Information Processing Systems 37*, 2024.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. Curated LLM: Synergy of LLMs and data curation for tabular augmentation in low-data regimes. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986. doi: 10.1103/PhysRevLett.57.2607.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems 32*, 2019.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN: Effective table data synthesizing. In *Proceedings of the 13th Asian Conference on Machine Learning*, pp. 97–112. PMLR, 2021.

A Sensitivity to the Distance of the Surrogate Negatives

The energy of Equation 1 depends on where the surrogate negative points are placed. The original formulation puts them at $\alpha_{\text{dist}} = 5$ standard deviations from the data center, whereas the main text reports results at $\alpha_{\text{dist}} = 1$. This appendix repeats the full fidelity comparison at the prescribed value on the imbalanced diabetes case, with the same five seeds and the same operating point in all other respects.

Table 4 shows that moving the negatives outward lowers absolute fidelity for every sampler, and that it lowers it most for SGLD (-0.082 in precision, against -0.047 for SGHMC and -0.050 for SGNHT). The ranking, the significance pattern and the size relations among the samplers are unchanged.

Table 4: Fidelity at the two negative distances (diabetes, 10:1 imbalance, five seeds, mean with 95% CI).

α_{dist}	Sampler	Precision	Recall	DCR (median)	ESS
1	SGLD	0.792 [0.770, 0.814]	0.996	1.475	4.82
	SGHMC	0.943 [0.927, 0.959]	0.992	1.101	3.23
	SGNHT	0.948 [0.929, 0.967]	0.990	1.100	3.06
5	SGLD	0.709 [0.687, 0.732]	0.996	1.592	4.71
	SGHMC	0.896 [0.869, 0.924]	0.993	1.171	2.93
	SGNHT	0.898 [0.872, 0.924]	0.993	1.173	2.93

Table 5: Paired differences at the two negative distances (five seeds, 95% CI on the seed-matched difference). The two conclusions hold at both settings.

Comparison	Metric	$\alpha_{\text{dist}} = 1$	$\alpha_{\text{dist}} = 5$
SGHMC − SGLD	Precision	$+0.152$ [$+0.132$, $+0.171$]	$+0.187$ [$+0.182$, $+0.192$]
	DCR	-0.374 [-0.407, -0.341]	-0.421 [-0.450, -0.392]
SGNHT − SGHMC	Precision	$+0.005$ [-0.012, $+0.022$]	$+0.002$ [-0.001, $+0.004$]
	DCR	-0.001 [-0.034, $+0.032$]	$+0.003$ [$+0.000$, $+0.005$]

Two observations follow. First, replacing SGLD with momentum improves precision significantly at both settings, and the improvement is larger at the prescribed distance, so the choice made in the main text understates the effect. Second, the thermostat remains without effect: the precision difference between SGNHT and SGHMC is $+0.002$ with a 95% interval of width 0.005, an even tighter bound than the one reported in the main text. The only interval in this comparison that excludes zero is the DCR difference of $+0.003$, which is 0.2% of the quantity being measured and of no practical consequence. Downstream utility is null at both settings.

Finally, the reason SGLD is the sampler most sensitive to this parameter is consistent with the mechanism identified in the main text. Placing the negatives farther away flattens the energy landscape near the data, weakening the drift term; SGLD, whose only counterweight to the injected noise is that drift, loses the most, whereas the momentum samplers retain the filtering that makes them robust to a noise-dominated regime.

B Additional Fidelity Analyses

B.1 Secondary Observations

Two secondary observations are worth recording. First, recall is saturated near 0.998 for every sampler, and SGLD is very slightly higher than the momentum samplers; this is the expected precision–recall trade-off of a more diffuse sampler rather than evidence of better coverage. Second, SGLD attains the highest ESS (5.85 versus 3.2–3.4 for the momentum samplers). Because our ESS is computed on the chain-averaged trajectory, a sampler whose increments are dominated by injected noise produces nearly independent successive positions and therefore a high ESS; we caution against reading it as better mixing.

Finally, the augmentation utility of the synthetic data is statistically indistinguishable across samplers: downstream balanced accuracy lies between 0.898 and 0.904 on the three pooled datasets, with every paired interval containing zero. A change of 0.33 in precision therefore does not translate into a measurable change in downstream accuracy in this regime.

B.2 Temperature Is Not the Control Variable

We set the target temperature to $\tau \in \{1, 0.1, 0.01\}$ for both SGHMC and SGNHT. At 50 steps all twelve cells lie between 0.950 and 0.975 in precision, and at 500 steps all of them collapse to 0.09–0.15 regardless of temperature; the median DCR at 500 steps stays within 4.61–4.92 across the whole sweep. Lowering the injected noise by two orders of magnitude therefore neither prevents the long-run collapse nor changes how far the samples travel.

B.3 The Thermostat over Longer Runs

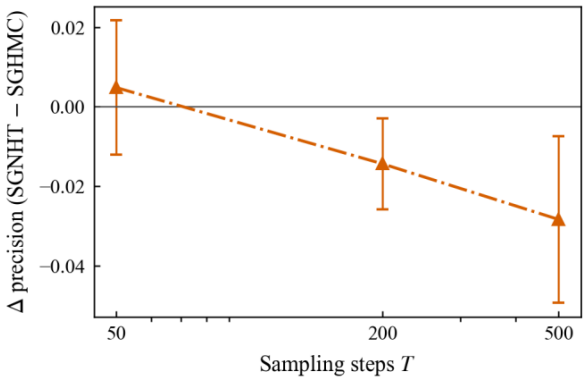

Figure 4: Effect of the thermostat as a function of run length. The paired difference between SGNHT and SGHMC is negligible for short runs and becomes increasingly negative as sampling continues. Error bars are 95% CIs over five seeds.

B.4 Robustness to Hyperparameters

A remaining possibility is that the thermostat buys robustness rather than accuracy. We evaluated both momentum samplers on a grid of step sizes $\{0.005, 0.01, 0.02, 0.05\}$ crossed with friction $\{0.3, 1, 3\}$. The two samplers behave almost identically cell by cell (for example 0.999 versus 0.997 at the mildest setting and 0.669 versus 0.657 at step 0.02 with friction 0.3), and SGNHT is in fact marginally worse in the worst cell (0.185 versus 0.266) and in grid-wide standard deviation (0.286 versus 0.268). What does stabilize both samplers is friction: at step 0.05 the precision recovers from 0.27 to 0.52 as friction increases from 0.3 to 3. Robustness in this setting is a property of friction, not of the thermostat.

Algorithm 1 Class-conditional generation with a pluggable sampler

Require: class data X_c; sampler S; steps T; step size h; negative distance α_{dist}; initial perturbation σ_{start}; number of samples n_c
Ensure: synthetic samples for class c
1: $N_c \leftarrow$ hypercube corners at $\pm\alpha_{\text{dist}}\sigma_d$ per dimension
2: condition in-context classifier f_c on $(X_c \to 1,\ N_c \to 0)$, with no parameter updates
3: define $E_c(\boldsymbol{x}) = -\operatorname{LogSumExp} f_c(\boldsymbol{x})$ from unnormalized logits
4: $\boldsymbol{x}_0 \sim \mathcal{N}(\tilde{\boldsymbol{x}}, \sigma^2_{\text{start}}\boldsymbol{I})$, $\tilde{\boldsymbol{x}}$ drawn from X_c, n_c chains in parallel
5: **for** $t = 0, \ldots, T-1$ **do**
6: $\quad \boldsymbol{x}_{t+1} \leftarrow S(\boldsymbol{x}_t, \nabla E_c)$
7: **end for**
8: **return** $\boldsymbol{x}_T$

C Additional Method Details

Sampling efficiency. We report an effective sample size (ESS) computed on the chain-averaged trajectory: we average the n_c chains at every step to obtain a single $(T+1) \times D$ path, compute a per-dimension ESS from the autocorrelation with initial-positive-sequence truncation, and average over dimensions. This measures the mixing of the ensemble mean and is not the average ESS of individual chains; the distinction matters for interpreting SGLD, whose noise-dominated increments are close to independent.

C.1 Sampler Noise Settings

Because momentum-based samplers derive their noise from τ and the friction while SGLD derives it from h, there is no single "noise" knob shared by all four. We therefore keep every sampler at its theoretical setting ($\tau = 1$, $C = A = 1$) and vary those quantities only in the dedicated ablations below. All four samplers evaluate the energy gradient by automatic differentiation through the in-context model; the leapfrog schemes use two gradient evaluations per step, and RENHD evaluates the gradient for all K replicas in a single batched call.

C.2 Backbone Implementation Requirements

Two implementation properties of the backbone are essential. First, Equation 1 requires the unnormalized logits. Recent releases of the in-context backbone normalize the logit vector so that $\operatorname{LogSumExp} f_c(\boldsymbol{x}) \approx 0$ for every input; under such a backbone E_c is identically constant and its gradient vanishes, so the sampler has nothing to follow. We verified this numerically and consequently used the last release that exposes unnormalized logits (TabPFN v2; Hollmann et al. 2025), which is also the release whose weights are redistributable. Second, the query point must remain differentiable through the transformer: we enable the differentiable-input path, fit the classifier through it, and fix inference precision to float32, since the float64 path does not carry gradients. Under these settings, the automatic-differentiation gradient of Equation 1 agrees in magnitude with a finite-difference estimate computed at a step size large enough to average over the fine-scale roughness of the surface.

C.3 Barrier Measurement for the Multimodal Benchmark

Because the whole experiment presumes that a barrier exists, we measure it before running any sampler. Profiling the energy along the path between two mode centers and separating the smooth trend from the fine-scale roughness gives a barrier height ΔE of 0.247, 0.605, 0.841 and 0.772 at separations of 8, 12, 16 and 20 standard deviations, with roughness 0.059–0.090 so that the barrier stands 4–10 times above it. Two facts from this measurement shape the design. First, the barrier saturates near 0.8: in-context energies do not produce arbitrarily high barriers no matter how far apart the modes are placed, which is consistent with the shallow gradient scale identified above. Second, at a separation of 4σ the barrier is slightly negative (-0.013), that is, the model merges the two modes into a single basin, so no barrier-crossing question can be posed there. We therefore use 16σ as the main configuration and 8σ, 12σ as lower-barrier controls. We also verified that the

surrogate negative samples must be placed far from the data, as prescribed by the underlying energy formulation: at the small negative distance used for the real-data experiments, the negatives fall on top of the modes of this synthetic mixture and distort the landscape.

C.4 Numerical Stability of the Replica-Exchange Ladder

Replica exchange requires a temperature ladder, and the flattening of the in-context energy documented earlier means that the target it induces at high temperature is not normalizable. On the lower-dimensional real datasets the hottest replica in the ladder can therefore leave the region where the gradient is informative and diverge. We consequently evaluate replica exchange only in the controlled multimodal setting of the coverage experiments above, where the surrogate negatives are placed far enough from the data to keep every replica bounded. A production implementation would need to guard against this, for example with a shallower ladder or a weak confining term in the energy.

D Additional Coverage Analyses

D.1 Coverage Metrics

We report the proportion of samples assigned to the minority mode (minority coverage, whose target is 0.05), the total-variation distance between the realized and target mode proportions (Total Variation (TV) Distance, our primary summary because it penalizes overshoot as well as shortfall), the fraction of chains that leave the starting basin (escape fraction), and the number of modes that receive no samples at all.

D.2 SGHMC versus SGNHT under Single-Init

SGHMC is marginally more exploratory than SGNHT, with an escape fraction of $+0.009$, CI $[+0.004, +0.015]$, and 1.4 fewer empty modes, which is consistent with the thermostat suppressing diffusion as observed above.

D.3 Minority Coverage across Barrier Heights

Minority coverage decomposes the same effect into its two failure modes and should not be read as a monotone advantage. SGLD degrades from 0.042 at 8σ, close to the target because it can cross a shallow barrier unaided, through 0.007 at 12σ to 0.000 at 16σ. Replica exchange declines from above rather than from below: 2.81 times the target at 8σ (0.141), 1.17 times at 12σ (0.059) and 0.48 times at 16σ (0.024). Its calibration therefore passes through the target instead of settling on it, with a ratio error falling monotonically from $+1.81$ (CI $[+1.25, +2.38]$) through $+0.17$ (CI $[-0.03, +0.37]$, the only setting at which it is on target) to -0.52 (CI $[-0.71, -0.33]$). The exchange supplies a roughly fixed amount of exploration, which is more than a shallow barrier requires and less than a high one does. The absolute gap in minority coverage consequently narrows with distance ($+0.099$, $+0.051$, $+0.024$) even as the TV-distance advantage appears, which is why we treat TV distance as the primary summary.