ReStrike: Preventing Support Collapse in On-Policy GFlowNets

Anonymous authors
Paper under double-blind review

Abstract

GFlowNets are trained to sample objects in proportion to a reward. We show that the standard on-policy balance objectives have a silent failure: the sampler can become exactly calibrated on part of the space while putting no mass on the rest. There the on-policy loss and its gradient vanish, so the collapsed sampler is a global minimum tied with the correct one, and it can even score a *lower* held-out loss than a healthy model. On Bayesian network structure learning with five variables, where the posterior over all 29,281 graphs is computed exactly, plain sub-trajectory balance without exploration ends collapsed on 17 of 20 runs across two posteriors, and trajectory balance, modified detailed balance and local search share the trap, which is metastable and, on one posterior, batch-dependent. On a fragment-based molecular benchmark the loss again prefers the sampler with the fewest modes, on two of three landscapes. The standard ε-uniform exploration floor prevents the collapse by the end of training but not the calibration error, even at four times the budget. ReStrike, which replays the distinct objects the run has found, drawn by a softmax over log-reward, recovers both posteriors at the same budget of fresh trajectories where floors, prioritized replay, a published floor-plus-replay schedule and local search fall short, and finds an order of magnitude more modes than the floor on a tiered grid. We prove that replay from inside a collapsed support cannot move it, and on one eight-node posterior replay can freeze where that schedule does better. Finally, the released code of prioritized replay training inverts its own prioritization on tiered rewards.

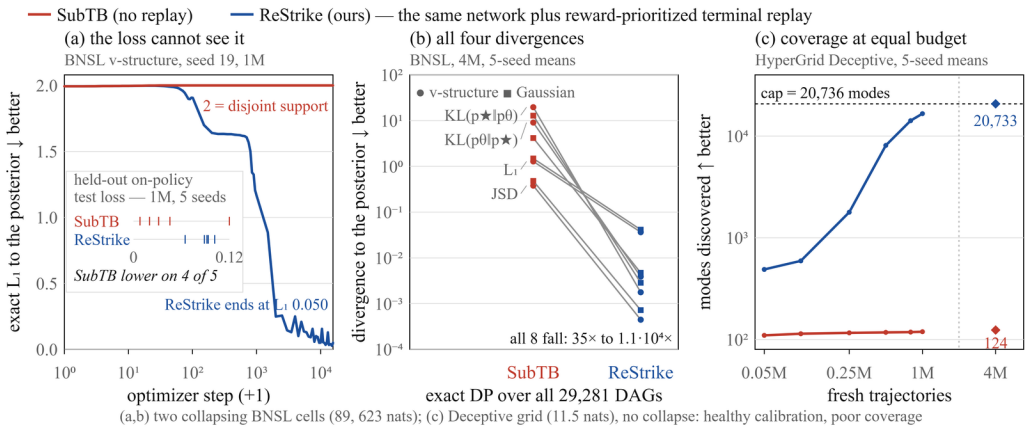

Figure 1: **The failure is invisible to the objective; one off-policy addition removes it. (a)** Two runs of one seed and network, bit-identical at initialization, differing only in reward-prioritized terminal replay. Exact L_1 stays pinned at 2 — disjoint support — while ours falls to 0.050. Inset: held-out test loss at 1M, one tick per seed; SubTB is collapsed on all five yet scores the *lower* loss on four (Obs. 1). **(b)** Exact DP over all 29,281 DAGs, 4M, 5-seed means: every divergence falls on both cells, $35\times$ to over $10^4\times$, so no metric changes the ranking. **(c)** Modes against *fresh* trajectories at equal fresh-trajectory budget; the 4M markers are separate runs. **(a,b) the two collapsing BNSL cells; (c) a grid cell that does not collapse**: healthy calibration, poor coverage (§1).

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.

1 Introduction

A Generative Flow Network (Bengio et al., 2021; 2023) learns a policy P_F that builds an object step by step and stops, so that the objects it produces are distributed in proportion to a reward R. That proportionality is the reason to train one rather than an optimizer, and it is what the standard objectives enforce: trajectory balance (TB, Malkin et al., 2022) and its sub-trajectory form SubTB (Madan et al., 2023) penalize the residuals of Eq. (1) on trajectories the policy itself draws. Training is on-policy or nearly so; where exploration is used it is an ε-uniform mixture on the same objective. Throughout, *calibration* means this proportionality — the sampler's terminal law matching $p^\star \propto R$ — and not the usual meaning of the word.

The failure. Suppose the policy puts zero probability on every action that leaves some closed region of the state space, and is calibrated inside it. Then every residual on every trajectory it can draw is exactly zero (Prop. 1): the on-policy loss is zero, so is its gradient, and the collapsed policy is a *global minimum* of the objective, tied with the true solution — nothing computed from the policy's own trajectories tells them apart (Obs. 1). In practice the collapsed model usually scores the *lower* held-out on-policy loss (Fig. 1). We call the state a restricted-support fixed point (RSFP) and the failure *calibration collapse*. It is a different thing from finding few modes while calibrated, which the same worlds also show (Fig. 1c); the two are ends of one continuum, indexed by how much of the target's mass the support carries.

Where it is shown, and how it is measured. We show it where it can be measured exactly: Bayesian network structure learning over five variables, whose $29{,}281$ DAGs are enumerated, so every divergence between sampler and posterior is computed by dynamic programming rather than estimated. A run counts as collapsed when the graphs carrying almost all of its mass carry at most half of the posterior's (Def. 1). Without exploration and at 64 fresh trajectories per update, plain SubTB is collapsed on 17 of 20 seeds after 4M trajectories on the two posteriors we study (Tables 3 and 11). The trap is metastable, not absorbing: most collapsed runs leave it if training continues, and with a larger batch the runs on one of the two posteriors no longer count as collapsed (§5.3). The failure is not SubTB's alone: TB, the modified detailed balance of the reference recipe and local-search GFlowNet share the trap; DB escapes it on one posterior, so how long it persists depends on the objective.

What repairs it, and what does not. The standard remedy, an ε-uniform floor at its default $\varepsilon{=}0.1$, prevents the collapse by the end of training but not the error: the floor ends above replay on all ten paired seeds of both posteriors (Tables 11 and 12), and on one posterior four times the budget does not lower it while the unaided sampler, escaping the trap, overtakes it (§5.3). The field's floor-plus-replay schedule — that floor annealed and combined with uniform replay (Deleu et al., 2022), here on SubTB — does not close the gap either. Replaying the run's own distinct terminals, drawn by a softmax over log-reward and reconstructed by a backward walk into the same loss (Eq. (3)), repairs both posteriors at the same budget of fresh trajectories, and at matched gradient samples repairs one and ties the floor on the other (§5.3). **Coverage is a separate axis:** on a tiered grid, where nothing collapses, the same replay finds an order of magnitude more of the reward's modes than the floor at the same budget (Table 2).

Where the repair stops. The theory states the boundary before the experiment shows it: replay drawn from inside a collapsed support cannot move it (Cor. 2). At eight nodes it is the best-calibrated sampler on the Gaussian chain and, at the reported setting, freezes on part of the v-structure, where the annealed-floor recipe is best (§5.4). Replay retains; it does not discover.

Contributions. (1) *The characterization and its testbed:* the RSFP, its silence (Prop. 1, Obs. 1; the exact finite-support form of the reverse-KL zero-forcing of Malkin et al., 2023), the exact error $2(1 - p^\star(T'))$ (Cor. 1), and an enumerable world on which every collapse verdict is decided by an exact divergence. (2) *The repair and its boundary:* own-terminal replay, drawn by a tempered softmax over log-reward, recovers the five-node posteriors on all four divergences where every floor, the published floor-plus-replay schedule, prioritized replay and local search do not, at equal fresh budget; Cor. 2 says why it can stop, and the eight-node v-structure is where it does. (3) *Where the obvious baseline breaks:* the released prioritized-replay code's closed percentile band inverts prioritization on tiered rewards, and repairing the band recovers most of that baseline's coverage gap (§4). Box 1, opening §5, defines the arms under one naming scheme; Appendix H maps each claim to the table that decides it and lists what is withdrawn.

2 SETTING

A finite terminate-anywhere DAG-MDP has states s, an exit action $\top$, and reward $R(x) > 0$ at terminals. SubTB(λ) trains (P_F, P_B, F) on a trajectory $\tau = (s_0, \ldots, s_n)$ by penalizing every sub-trajectory residual

$$A_{ij}(\tau) = \log F(s_i) + \sum_{k=i}^{j-1} \log P_F(s_{k+1}|s_k) - \log F(s_j) - \sum_{k=i}^{j-1} \log P_B(s_k|s_{k+1}), \tag{1}$$

weighted by length within the trajectory,

$$\ell_{\text{SubTB}}(\tau) = \frac{\sum_{0\le i<j\le n} \lambda^{j-i} A_{ij}(\tau)^2}{\sum_{0\le i<j\le n} \lambda^{j-i}}, \tag{2}$$

where a sub-trajectory that *exits* at $s_j = x$ substitutes $\log R(x)$ for $\log F(s_j)$ and carries one further forward factor, $\log P_F(\top|x)$, with no P_B counterpart. Every arm shares this architecture, optimizer and hyper-parameters, differing only in which trajectories enter the loss (Box 1).

Vocabulary. $p^\star \propto R$ over terminals is the target and p_θ the sampler's terminal law; $\mathrm{L}_1 = \sum_x |p_\theta(x) - p^\star(x)| \in [0, 2]$ is their distance. A *cell* is one world-and-landscape pair and an *arm* one training configuration (Box 1); *plain SubTB* is the shared on-policy sampler with no replay and no exploration, the *chassis* every arm is built on; the *reference recipe*'s ε-uniform mixture at $\varepsilon{=}0.1$ is *the ε-floor*, that floor annealed $1 \to 0.1$ with uniform replay (Deleu et al., 2022) is the *annealed-floor recipe* (*the recipe*), and a run's *verdict* is whether it meets the collapse predicate (Def. 1).

Worlds and budget. Bayesian structure learning over five-node DAGs (BNSL; all 29,281 enumerated, as in Deleu et al. (2022); exact forward-DP metrics; uniform DAG prior; BDeu on binary noisy-XOR colliders, 200 rows for the v-structure cell and a fifth of that for the sparse cell; a conjugate linear-Gaussian score on 500 rows of a planted chain for the Gaussian cell; Appendix E.1), a tiered HyperGrid (Bengio et al., 2021), the PHO4 transcription-factor oracle (Zhu et al., 2009) ($L{=}10$, exact metrics), and a non-enumerable fragment-based molecular world scored by the sEH proxy of Bengio et al. (2021). Budget is counted in *fresh* trajectories and reward evaluations: every arm's online loop draws 64 fresh trajectories per update, and replay adds up to half a batch of stored terminals to the same loss as a mixture weight (Alg. 1; PRT-full: several steps per round, a full replay batch), so those two budgets are identical across arms while gradient samples per update and wall-clock are not (two exceptions, Appendix F). The study seeds are $\{7, 19, 101, 367, 911\}$; the ten-seed replication (Table 11) adds five more, and the tuning sweep and the size ladder use the first three.

Metrics. Every calibration ranking is decided by an exact or analytic divergence, never by the trajectory-level L_1 estimate, which is biased against percentile-based baselines; differences at the sampling floor carry no ordering, and a tie we declare from a test, never from a floor (the fork ablation of Appendix F and the eight-node cells are the two sampled exceptions). Every control below was pre-registered: design and decision rules were committed before its runs, and the files ship with the code.

3 THE COLLAPSE, CHARACTERIZED

Definition 1 (Collapse predicate)**.** Let S_t be the *operating support* of the policy at step t: a smallest set of terminals with **(C1)** $p_\theta(S_t) \ge 1 - \delta$: take terminals in decreasing p_θ, ties by index, until $1 - \delta$ of the mass is accounted for. A run is *collapsed* if **(C2)** $p^\star(S_t) \le 1 - \kappa$ (that support misses $\ge \kappa$ of the true mass); it is *calibrated-collapsed* if additionally **(C3)** $\mathrm{L}_1\big(p_\theta(\cdot \mid S_t),\, p^\star(\cdot \mid S_t)\big) \le \gamma$, each law conditioned on S_t. We use $(\delta, \kappa, \gamma) = (0.05, 0.5, 0.1)$; no five-node verdict turns on the choice (Appendix A); the eight-node read-out's dependence is stated in Appendix E.1.

The verdict is (C2), invariant over the admissible cuts on all 20 study runs of the two structured cells (Appendix A); (C3) is measured, not implied (Cor. 3): every collapsed endpoint meets it, its restricted divergence at most 0.0912 while the total error runs to 2.0.

3.1 The endpoint is a fixed point, not a bad local optimum

Proposition 1 (Restricted-support fixed points)**.** *Let M be a finite terminate-anywhere DAG-MDP with $R > 0$ and let T' be any terminal set realizable as the termination support of a policy with reachable states V'. If (P_F, P_B, F) satisfy the interior balance identity on V' and the exit balance identity on T', P_F assigns probability zero to every action leaving V' or exiting outside T', and P_B to every parent outside V', then for* every *trajectory in the support of the policy, every SubTB residual is exactly zero. The sampled loss and gradient are zero almost surely, so (P_F, P_B, F) is a stationary point of the on-policy stochastic optimization — for any batch size, step size and sub-trajectory weighting.*

Corollary 1 (Exact error)**.** *The terminal distribution is $p = R\mathbf{1}_{T'}/Z_{T'}$ and $\mathrm{L}_1(p, p^\star) = 2(1 - p^\star(T'))$.*

T' is the *termination* support, a different object from S_t; the largest residual over the eight collapsed endpoints is 0.0080 (a second exact signature is in Appendix A).

Observation 1 (Silence)**.** *The expected on-policy loss J is ≥ 0 everywhere, and every restricted solution has zero loss for almost every sampled trajectory, so $J = 0$: every restricted-support solution, including fully collapsed ones, is a* global minimum *of the on-policy loss, exactly tied with the true solution.*

Empirically the ordering is worse than a tie (v-structure at 4M: held-out on-policy test loss 0.0134 for a collapsed seed, 0.0272 for the one healthy seed, 0.0339 for the repaired twin; ten seeds in §5.3): *the best-looking loss belongs to the broken model.*

3.2 How the support dies, and why it is not a local minimum

Lemma 1 (Appendix A) isolates one component of the exit force at initialization: for a sampled trajectory ending at a terminal whose reward deficit dominates, the terminal residuals are all positive, and the step on them *lowers* that terminal's exit logit, to first order in that logit (exactly for a tabular policy), whenever its reward sits below the current flow scale; softmax renormalization pushes the freed mass deeper. Every background terminal thus *votes* to suppress its own exit — a vote, not a sum: the logit also sits in the softmax denominator of its continuations, an opposing term of the same order that the lemma does not bound; that exits die is measured (Fig. 7; Appendix G). Lock-in is a *retention race*: collapse takes hold when re-samples of good terminals are too rare during the transient (Appendix G). **Metastable, not absorbing.** The exact RSFP needs infinite logits, so a run occupies a leaky neighborhood of it: a long-lived trap, not a terminal state, whose escapes jump rather than drift (Fig. 2a, §5.3); every claim below concerns it (Appendix G).

Falsification. The collapsed point is a global minimum of the *on-policy* loss (Obs. 1) but not even a local minimum of the same objective on *full support*: there the gradient is $1030\times$ the on-policy one. **But that descent is not itself the repair**: under a fixed full-support behavior the exact L_1 falls only $2.000 \to 1.549$, then drifts back to 1.710 while the loss keeps falling (Fig. 8); what moves the support with the loss is off-policy data anchored at reward-bearing exits (§4). We claim the measurement, not a barrier (§7(x)), and the gap is not a detector (Appendix G). A one-way pre-flight flag, the initial reward deficit, is in Appendix B.

4 Method, and why the obvious baseline is not enough

ReStrike. A single off-policy addition to plain SubTB (Alg. 1): a buffer of the run's own *distinct* terminals, capped at $10{,}000$ with the lowest-reward entry evicted; a draw over it, $\Pr(x) \propto \exp(\beta_{\mathrm{r}} \log R(x))$ with $\beta_{\mathrm{r}} = 0.5$; and replay of the drawn terminals, reconstructed by a backward walk under the learned P_B (a uniform walk does as well, §5), into the same SubTB loss at a mixture weight annealed $0.1 \to 0.5$ over training (a schedule §5 withdraws, kept as run): with B fresh trajectories and $n_t = \mathrm{round}(f_t B)$ replayed ones at step t,

$$\mathcal{L}_t = \frac{B - n_t}{B}\,\bar{\ell}_{\text{SubTB}}(\text{fresh}) + \frac{n_t}{B}\,\bar{\ell}_{\text{SubTB}}(\text{replayed}), \tag{3}$$

each term a batch mean of Eq. (2). The fresh-trajectory and reward-evaluation budgets are every baseline's (§2); batch 64, width 256 and $\lambda = 0.9$ are plain SubTB's, fixed across every world.

Algorithm 1 ReStrike, one optimizer step (B=64, cap 10,000, $\beta_r = 0.5$, f_t linear $0.1 \to 0.5$).

1: Sample B fresh trajectories on-policy from P_F; evaluate R at their terminals.
2: $n \leftarrow \text{round}(f_t B)$; draw n buffer terminals, $\Pr(x) \propto \exp(\beta_r \log R(x))$, without replacement once $|D| \geq n$.
3: Reconstruct each drawn terminal by a backward walk under P_B, reversed; its stored reward is reused (no oracle call).
4: $\mathcal{L} \leftarrow$ Eq. (3) on the fresh and the replayed batch; one Adam step.
5: Add the fresh batch's new distinct terminals to D; above the cap, evict the lowest reward.

Box 1: The arms. All share plain SubTB's network, optimizer and hyper-parameters (exceptions: AT-teacher's second network, PRT-full's rounds).

Arm	Definition
plain SubTB	SubTB(λ=0.9), on-policy, no replay, no exploration (ε=0). *TB*, *DB*, *modified DB* (Deleu et al., 2022): the same configuration with the objective swapped (Tables 14, 17).
ε-floor 0.1, 0.01	plain SubTB with an ε-uniform behavior policy (0.1 is the reference recipe's default); *annealed floor*: ε annealed $1 \to 0.1$.
annealed-floor recipe	the annealed floor with uniform replay (Deleu et al., 2022), on SubTB or ("(DB)") on DB; *floor + ReStrike draw*: with ReStrike's draw instead.
ReStrike	plain SubTB plus replay of the run's own distinct terminals, drawn by $\Pr(x) \propto \exp(\beta_r \log R(x))$, β_r=0.5, at a fraction annealed $0.1 \to 0.5$ (Alg. 1). *Constant fraction*: held at 0.5 (the recommended schedule, §5.3); *constant fraction, r=1*: with the draw proportional to reward (β_r=1); *uniform draw*: β_r=0; *top-k draw*: the k highest rewards.
PRT-rule, PRT-full	the prioritized replay of Shen et al. (2023): its draw rule alone on our loop, and its full round structure; *PRT-rank*: with the rank split its authors describe; *PRT-ϵ0*: without its online exploration.
AT-teacher	the explorer-teacher of Kim et al. (2025), a co-trained second GFlowNet whose proposals are replayed.
local search	local-search GFlowNet (Kim et al., 2024b): seven rounds of destroy-and-reconstruct per update inside the same 64 scored trajectories; + *ε-floor* 0.1: with the floor as the behavior policy of every draw.

Replay supplies reward-anchored gradient to *already-discovered* exits off-policy, a channel the persistence bound of Appendix A does not count — for exits the run has drawn at least once, and only for those:

Corollary 2 (Closure over own-terminal replay)**.** *Let (P_F, P_B, F) be as in Prop. 1 and let the replay buffer hold terminals $D \subseteq T'$. Every replayed trajectory — a backward walk under P_B from some $x \in D$, reversed — has every SubTB residual zero, so the point is stationary for the replay-augmented objective as well, at any replay fraction and any draw rule over D: replay from inside the support revives nothing, and an exit the run has never drawn stays invisible to it.*

Why not a percentile rule? The leading replay baseline (PRT, Shen et al., 2023) splits the buffer at $q_{0.9}$ and draws *uniformly within each side*, half the batch from each. In the released code both sides are closed intervals at a *value* threshold, not the rank split its authors describe, so a reward *tier* that straddles $q_{0.9}$ is drawn from both halves and prioritization is inverted exactly where it is most needed (Appendix E, Fig. 5). It is a tiered-reward failure and no more: on every continuous cell the band is healthy, and PRT-rank calibrates like the value split on the DAG cells (Table 1), so our margins there are not the band's. **On the tiered grid that same split is the straddle's own test, and it passes** (pre-registered; Table 2): PRT-rank finds 9,407 modes at 1M where PRT-full's released rule finds 427, so most of the baseline's coverage gap is the closed band rather than prioritized replay, and 9,407 against our 16,635 is what our buffer and draw add once the band is repaired. A softmax-over-log R draw cannot degenerate this way: two terminals differing by u in log-reward are drawn in ratio $\exp(\beta_r u)$ whatever the tier population.

5 Experiments

5.1 Calibration on the two collapsing DAG cells, and coverage at equal budget

Every number in this section comes from one configuration fixed across all four worlds (Tables 9 and 16), nothing tuned per cell. On the two structured BNSL cells, exact dynamic programming over all 29,281 DAGs (mass conservation gated per run) gives Table 1. Plain SubTB satisfies Def. 1 on $5/5$ seeds at 1M and $4/5$ at 4M — one seed escaping unaided, not one seed never collapsing (§3). ReStrike beats every other arm of Table 1 on L_1, JSD and both KL directions *simultaneously, on both cells* — by $35\times$ to over $10^4\times$ over plain SubTB, and, on the five-seed means, by a margin

Cell	Arm	L_1	JSD	$KL(p_\theta \| p^\star)$	$KL(p^\star \| p_\theta)$
v-structure	plain SubTB	1.2820±0.7979	0.3788	9.0625	19.6004
	ε-floor 0.1	0.0698±0.0112	0.0012	0.0050	0.0050
	annealed-floor recipe	0.0813±0.0150	0.0016	0.0062	0.0063
	ReStrike	**0.0365±0.0202**	**0.0004**	**0.0039**	**0.0018**
	PRT-rule	0.0956±0.0339	0.0019	0.0075	0.0076
	PRT-full	0.0766±0.0210	0.0013	0.0052	0.0053
	PRT-rank	0.0727±0.0131	0.0012	0.0047	0.0047
	AT-teacher	0.0832±0.0266	0.0018	0.0071	0.0074
Gaussian	plain SubTB	1.4904±0.8205	0.4798	4.1619	12.8628
	ε-floor 0.1	0.3229±0.1023	0.0234	0.0979	0.0994
	annealed-floor recipe	0.3260±0.1449	0.0246	0.1134	0.0974
	ReStrike	**0.0412±0.0385**	**0.0007**	**0.0047**	**0.0029**
	PRT-rule	0.5006±0.1661	0.0524	0.2549	0.2115
	PRT-full	0.3901±0.0492	0.0357	0.1807	0.1390
	PRT-rank	0.4416±0.0904	0.0430	0.2029	0.1742
	AT-teacher	0.5227±0.0710	0.0583	0.2983	0.2403

Table 1: Exact forward-DP divergences to the posterior over all 29,281 five-node DAGs, 4M, 5 seeds; lower is better; arms as in Box 1.

Arm	800K (PRT's published budget)	1M	4M
plain SubTB	118	119	124±27
ε-floor 0.1	1,285	1,369	—
ε-floor 0.01	220	240	—
ReStrike	14,200	16,635	20,733±2
ReStrike, constant fraction	16,273	17,979	—
ReStrike, constant fraction, r=1	**16,713**	**18,271**	**20,733±1**
annealed-floor recipe	1,285	1,298	—
PRT-rule	253	273	10,358±7,724
PRT-full	328	427	2,537±2,787
PRT-ϵ0	38	39	49±9
AT-teacher	215	223	930±708
PRT-rank	7,138	9,407	—

Table 2: Modes discovered on the tiered Deceptive cell (cap 20,736); five-seed means (4M with seed spread; 800K and 1M are checkpoints of the 1M runs).

against every baseline at every metric. On ten seeds per cell (Table 11) plain SubTB is collapsed on 17 of 20, and ReStrike beats the reference recipe's ε=0.1 floor on $10/10$ seeds of both cells (paired t: $p = 0.0002$, sign test $p = 0.002$).

Coverage at equal budget. On the tiered HyperGrid (Table 2) ReStrike finds 16,635 modes at 1M (14,200 at the 800K budget Shen et al. (2023) publishes) against 119 for plain SubTB, and at 4M reaches $20{,}733 \pm 2$ of the 20,736 cap, the strongest baseline $10{,}358 \pm 7{,}724$. **Neither the ε-floor nor the annealed-floor recipe closes this gap**: 1,369 and 1,298 modes at 1M. The same buffer at the constant fraction finds 17,979, and at r=1 18,271: the schedule, not the draw's temperature, is what the reported arm understates (§5.3).

5.2 The mechanism, twin-controlled (one seed)

Against the collapsing run with the same seed, network and fresh sampler (seed 19, Fig. 2), replay changes exactly the quantities the theory names: fresh trajectories keep terminating — at step 25 the twin's largest depth-wise exit probability stands at nine times plain SubTB's — the discovered set D_t *keeps growing* instead of stalling (2202 versus 744), and the exit residual at the MAP DAG repairs from -60 to 0.12 nats: rehearsal of *ordinary* terminals keeps exits alive, and the reward-weighted draw converts one-shot discoveries into a permanent repair signal (Appendix G).

5.3 Controls: what else repairs the collapse, and what does not

Table 3 holds the five-node controls, all pre-registered (per seed: Table 12; in full: Appendices E.2, E.3).

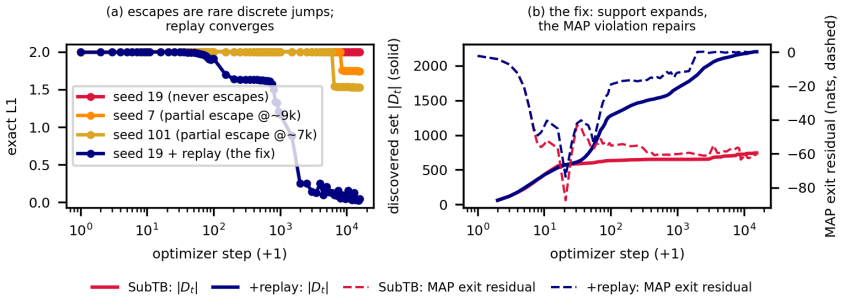

Figure 2: **The fix, twin-controlled** (BNSL v-structure, seed 19). Panel (a) also shows plain SubTB unaided: seeds 7 and 101 leave the $L_1 = 2$ plateau in single jumps (only 101 clears the predicate by 4M), seed 19 never.

The standard ε-floor prevents the verdict, not the gap. Plain SubTB under the ε-floor collapses on no seed of either cell, nor under an annealed floor ($1 \to 0.1$ over training): the collapse is a property of the on-policy sampler, not of the objective under every exploration. What the floor does not buy is calibration (Table 1): ReStrike wins every paired seed at both ε, and $10/10$ on ten seeds (§5.1). **Nor is the trap SubTB's alone** (Table 14): TB collapses on $4/5$ and $5/5$ seeds at 4M and DB on $0/5$ and $3/5$, escaping on the v-structure by 4M; the trap's persistence depends on the objective. **Nor does the floor combined with replay**, the recipe's shape (Deleu et al., 2022): the annealed floor with a uniform replay draw is worse than ReStrike on $5/5$ seeds of both cells, and on top of ReStrike's own draw on $4/5$ and $5/5$. **The draw's temperature is a per-cell choice.** A plain reward-proportional draw (r=1) from the same buffer, at the constant fraction, is not worse than the reported arm on every headline cell but the six-node Gaussian one (on ten five-node Gaussian seeds, a tie), and freezes on no eight-node seed (Tables 11, 8, 4); no single r is recommended, and the reported arm keeps r=0.5 as run.

Neither budget nor gradient samples close the floor's gap (Appendix E.3). Continued to 16M, the ε-floor ends at 0.048 and 0.344 and ReStrike (constant fraction) at 0.020 on both cells, already at 0.039 and 0.037 by 2M; seven of the eight chassis seeds collapsed at 4M escape between 5M and 14M (one never does), so the verdict is a 4M statement. At 96 fresh per update plain SubTB meets the predicate on the Gaussian cell only ($5/5$), the floor at 96 stays above ReStrike on 9 of 10 paired seeds, and ReStrike at 32 fresh + 16 replayed repairs the Gaussian cell and ties the floor on the v-structure. **Modified detailed balance shares the trap** on $10/10$ seeds at ε=0, and **the draw's spread matters**: a top-k draw is worse on $5/5$ Gaussian seeds (Tables 17, 19).

Nor does local search leave the trap (Appendix E.3): local-search GFlowNet (Kim et al., 2024b), at the study's 64 scored trajectories per update, collapses on $4/5$ Gaussian seeds and freezes on a partial support on three v-structure seeds, and on the ε-floor is worse than the floor alone on the Gaussian cell; ReStrike (constant fraction) is lower on 9 of 10 paired seeds against either. **Nor is the floor's gap residual scaling**: clipping its residual gradient at one nat removes 0.14 of its error on either cell.

And the objective cannot tell them apart (Table 3, loss column): under the ε-floor the held-out on-policy loss is *lower* than ReStrike's on 9 of 10 v-structure seeds while the exact L_1 is worse on every seed, and collapsed plain SubTB carries the lowest Gaussian loss of all (Obs. 1).

The draw carries the repair; the anneal does not. The prioritized draw carries a measurable part of the five-node calibration (Tables 3, 13, 17), while a constant replay fraction of 0.5 beats the annealed one on both cells (0.0337 and 0.0239), so the schedule is withdrawn and kept only as run. **No learning rate un-collapses the baseline** in a 108-run sweep by its pre-registered criterion (Table 10), and **a uniform backward walk repairs both cells as well** (Table 7): the claim is reward-prioritized terminal replay, not a learned backward policy.

5.4 Six and eight nodes: the verdict splits by cell

Table 4 summarizes both rungs (Appendix E.1). At six nodes the Gaussian trap deepens (plain SubTB and local search at 1.9996 on every seed) and ReStrike, at either fraction, beats every floor, recipe, PRT and local-search arm on $5/5$ seeds, while the added node flattens the v-structure posterior and only local search collapses (one seed). At eight nodes (7.8×10^{11} DAGs) the cells split. *Gaussian*

Cell	Arm	L_1 (mean±sd)	JSD	$KL(p_\theta \| p^\star)$	$KL(p^\star \| p_\theta)$	loss	collapsed
v-structure	plain SubTB	1.2820±0.7979	0.3788	9.0625	19.6004	0.022	4/5
	ε-floor 0.01	0.0819±0.0312	0.0020	0.0079	0.0086	0.024	0/5
	ε-floor 0.1	0.0698±0.0112	0.0012	0.0050	0.0050	0.019	0/5
	annealed floor	0.0785±0.0258	0.0018	0.0072	0.0073	0.018	0/5
	ReStrike, uniform draw	0.1145±0.0408	0.0084	0.0264	0.2827	0.019	0/5
	ReStrike, constant fraction	**0.0337**±0.0171	**0.0003**	**0.0038**	**0.0013**	0.047	0/5
	annealed-floor recipe	0.0813±0.0150	0.0016	0.0062	0.0063	0.021	0/5
	floor + ReStrike draw	0.0684±0.0158	0.0010	0.0041	0.0041	0.008	0/5
	ReStrike	0.0365±0.0202	0.0004	0.0039	0.0018	0.048	0/5
Gaussian	plain SubTB	1.4904±0.8205	0.4798	4.1619	12.8628	0.094	4/5
	ε-floor 0.01	0.8784±0.2771	0.1737	0.6530	8.4335	0.657	2/5
	ε-floor 0.1	0.3229±0.1023	0.0234	0.0979	0.0994	0.275	0/5
	annealed floor	0.4966±0.2995	0.0615	0.2492	0.3370	0.543	0/5
	ReStrike, uniform draw	0.4450±0.1343	0.0439	0.2021	0.1859	0.549	0/5
	ReStrike, constant fraction	**0.0239**±0.0025	**0.0002**	**0.0018**	**0.0007**	1.027	0/5
	annealed-floor recipe	0.3260±0.1449	0.0246	0.1134	0.0974	0.258	0/5
	floor + ReStrike draw	0.3089±0.2298	0.0282	0.1664	0.1075	0.602	0/5
	ReStrike	0.0412±0.0385	0.0007	0.0047	0.0029	0.332	0/5

Table 3: **The ε-exploration and composition controls**: two collapsing cells, five seeds, 4M, exact forward-DP metrics; *loss* = held-out on-policy test loss on the arm's own draws; *collapsed* = seeds meeting (C2); bold = best, except the loss column, whose lowest Gaussian value is collapsed plain SubTB's; per seed: Table 12.

Arm	v-structure, d=6	v-structure, d=8	Gaussian, d=6	Gaussian, d=8
plain SubTB	0.0595 (0/5)	2.28 (0.56)	1.9996 (5/5)	11.59 (0.00)
ε-floor 0.1	0.0797 (0/5)	1.39 (0.76)	0.8710 (0/5)	8.52 (0.05)
annealed-floor recipe	0.0558 (0/5)	**0.28** (0.94)	0.4574 (0/5)	4.07 (0.35)
annealed-floor recipe (DB)	0.0601 (0/5)	**0.27** (0.94)	0.5913 (0/5)	2.38 (0.42)
PRT-full	0.1092 (0/5)	0.57 (0.94)	0.8686 (0/5)	2.45 (0.47)
floor + ReStrike draw	0.0608 (0/5)	**0.28** (0.94)	0.4085 (0/5)	3.40 (0.38)
ReStrike	0.0488 (0/5)	2.21 (0.58)	0.0737 (0/5)	**0.77** (0.55)
ReStrike, constant fraction	**0.0292** (0/5)	1.05 (0.80)	**0.0616** (0/5)	0.84 (0.55)
local search	0.3904 (1/5)	2.39 (0.52)	1.9996 (5/5)	11.55 (0.00)
ReStrike, constant fraction, r=1	0.0480 (0/5)	0.75 (0.84)	0.0934 (0/5)	**0.71** (0.55)

Table 4: **Six and eight nodes, by cell.** Six nodes: exact L_1 over all 3,781,503 DAGs (seeds meeting (C2) in parentheses); eight nodes: the edge-marginal L_1 over 56 pairs from 200k draws (the sampled support's exact mass at a million draws in parentheses); five seeds, 4M; bold = the lowest per column and any within 10% of it; per-seed values: Table 15.

chain: plain SubTB is a calibrated collapse (Def. 1) on four of five seeds, local search on five; at a million draws the ε-floor's draws cover 0.05 of the posterior, the recipe's 0.35, PRT-full's 0.47 and ReStrike's 0.55, and ReStrike calibrates best on the marginals, on $5/5$ seeds against the floor, the recipe on either objective and PRT-full. *v-structure:* on a posterior whose MAP carries 0.03, plain SubTB freezes on 0.35–0.44 of the mass on $3/5$ seeds and ReStrike on $2/5$, worse than the recipe on $5/5$ — the case Cor. 2 allows. At the constant fraction and r=1 the same buffer freezes on no seed and is lower on both cells (0.75 and 0.71), still above the recipe on the v-structure, where every arm with a floor finds the support.

5.5 Losses stated in advance, and the silence on a molecular benchmark

Where we lose, stated in advance. On the PHO4 oracle (Table 8) we win total variation at $\beta \in \{10, 16\}$ and *lose* JSD and both KL directions at every β; at β=4 we lose all four. At r=1 the same buffer beats the reported arm on all but one metric–β pair, and we beat plain SubTB on every metric at every β; BNSL, not PHO4, is our divergence-robust result. On the sparse DAG cell, which does not collapse, the exact L_1 ties at 0.0259.

The silence reproduces on a real molecular benchmark: a non-enumerable fragment-based space scored by the sEH proxy (Bengio et al., 2021), four arms, five seeds, three landscapes at 4M (Table 16); the cell tests the objective, not the repair (§7(vi)). On Deceptive SubTB ends on a median

of 20 Tanimoto-distinct modes against our 2810 and scores the **lowest** held-out loss of any arm (7.68 vs. our 13.87), below ours on 4 of 5 seeds: a loss monitor would rank the 20-mode sampler first. Smooth repeats it; on Peaked the objective ranks PRT-rule lowest: the silence holds on two of three landscapes; coverage is reported, never claimed (Appendix E).

What to run. Where replay of what the run has found suffices — the two collapsing five-node cells, both six-node cells and the eight-node Gaussian chain — ReStrike is the best-calibrated arm and every floor arm calibrates worse, at every budget we ran; its constant 0.5 fraction is the better setting at five and six nodes, and the draw's temperature is a per-cell choice (§5.3). Where discovery is the bottleneck — the eight-node v-structure — the recipe calibrates best. Watch the support, not the loss: the loss cannot tell a frozen support from a calibrated one (Obs. 1); a support read-out can (Appendix E.1).

6 Related work

Terminal replay is standard; the draw and the characterization are ours. Buffering terminals and replaying them through a backward walk is in the reference library (Viviano et al., 2023), in Shen et al. (2023) and He et al. (2025), and replay buffers aid mode discovery (Vemgal et al., 2023; Jain et al., 2022). **Exploration** by intrinsic rewards (Pan et al., 2023), Thompson sampling (Rector-Brooks et al., 2023) or temperature conditioning (Kim et al., 2024a) changes what the sampler visits; da Silva et al. (2024) train on divergences. **Nearest neighbors.** Shen et al. (2023), our primary baseline, introduces the prioritized replay whose band §4 analyses; Sendera et al. (2024) and Kim et al. (2024b) get their off-policy signal from local search (the latter run in §5.3); Jang et al. (2024) trains the backward policy, which our reconstruction control separates from the draw; He et al. (2025) is the closest kernel match (goal-conditioned); Wang et al. (2026) and Malek et al. (2026) are the separate-proposer family. None of these characterizes the failure or is evaluated on an exactly enumerable posterior. da Silva et al. (2025) study when a GFlowNet samples the intended distribution from *residual imbalance*; ours is the complementary case, exact balance on what it draws; Lei et al. (2026) bound total variation by a balance loss controlled *uniformly over trajectories*, the complement of Obs. 1.

7 Limitations and honest scope

This is a *regime* claim, not dominance. (i) On the PHO4 oracle we lose JSD and both KL directions at every β. (ii) On the *grid* cells, where nothing collapses, we claim no calibration result. (iii) No claim concerns early training. (iv) The 108-run sweep varied the learning rate only; no draw temperature is best on both cells at 4M (Table 17). (v) The precondition is on the *failure*: Lemma 1 needs an exit action at non-terminal states (no claim about autoregressive generators), and Prop. 1 covers TB, DB with a learned state flow and SubTB, not modified detailed balance, whose collapse is measured (§5.3). (vi) The molecular world supports a *loss* claim only: three landscapes at $n{=}5$ on a proxy reward, oracle spend unmatched (Appendix E). (vii) Two arms spend outside the accounting of §2 (Appendix F). (viii) A batch-diversity term would separate the collapsed and the true solution but bias the optimum (untested). (ix) Size-robustness on the grid is claimed at scaled budget only (Appendix D); the six- and eight-node DAG cells cover the two collapsing posteriors; the minimal model is scored by a *proxy* for Def. 1 (Appendix C). (x) That a diffuse full-support behavior does not repair the collapse holds at *this budget*; that no statistic which *can* see the collapse leaves the target where it was is an argument, not a theorem.

8 Conclusion

On-policy SubTB can settle near an exactly calibrated sampler of the wrong support, and the training loss not only fails to reveal this but can *prefer* it. The endpoint is a restricted-support solution, not a bad optimum: replay of discovered terminals repairs it where the run has found the support, and from inside a collapsed support it cannot (Cor. 2); the ε-floor prevents the collapse by the end of training, not the calibration gap. Open: why DB escapes where TB, SubTB, modified DB and local search stay, and why the batch matters on the v-structure. The practical stake is the monitor: the loss a deployed sampler reports can prefer the broken model.

AI USE STATEMENT

We used a generative AI assistant substantially in this work, and disclose it here in the categories the policy names. *Required disclosure:* the assistant helped develop conceptual framing, formulate mathematical claims, and formulate and refine hypotheses, contributed ingredients to proofs and to the writing of proofs, provided feedback on experimental design and pre-registration, implemented methods, the synthetic environments and analysis code, and assisted in interpreting results. *Recommended disclosure:* it drafted parts of this paper, produced figures, surveyed literature, and proposed structure. **The verification apparatus is the substance of this disclosure.** Every numeric table is emitted by a generator script directly from on-disk artifacts; the running text and the two boxes are written by hand and then gated, with a build-blocking audit that binds each of their numeric literals to a generated table or to the derived-facts file, and refuses any literal that binds to nothing. That audit sees decimal literals, so a small whitelist of protocol constants and the exponents of numbers written in scientific notation are checked by reading rather than by the gate. Beyond it: every exact-DP evaluation passes a mass-conservation gate on every run; bit-identity checkers verify that instrumentation does not perturb training; and decision rules for every pre-registered control were committed to version control before those experiments ran. In the course of preparing this work the assistant produced multiple confident but incorrect statements, each caught by returning to the primary artifact; that experience is why the controls above exist and are reported. The authors have reviewed all AI-assisted work and take full responsibility for the content, claims and artifacts of this paper.

REPRODUCIBILITY STATEMENT

The proofs of Prop. 1, Cors. 1–3, Obs. 1 and Lemma 1 are in Appendix A with all assumptions stated. Every numeric table and figure is regenerated from the released artifacts by the five scripts the supplement's README names: the table generator, which also emits Fig. 5; a second generator for Table 16; the figure script, for the six raster panels, and a second one for Fig. 6; and a Node step for Fig. 1, drawn as SVG, rasterized for inclusion, and shipping its own numeric and geometry audits. The exact-DP evaluator, its mass-conservation gate, the collapse predicate checker, and the instrumented probe behind Fig. 1a, Fig. 2 and Fig. 7 are all in the supplement, as are the seed set, budgets and the pre-registration files. Anonymized source is provided as supplementary material.

ETHICS STATEMENT

This work studies the training dynamics of generative samplers on synthetic and public scientific benchmarks; it involves no human subjects and no personally identifying data. Improved samplers for molecular design carry the dual-use considerations common to that field; the molecular result we retain is a diagnostic read-out of a training objective against a public proxy oracle, and enables no capability beyond it.

REFERENCES

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. GFlowNet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Tiago da Silva, Eliezer de Souza da Silva, and Diego Mesquita. On divergence measures for training GFlowNets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tiago da Silva, Rodrigo Barreto Alves, Eliezer de Souza da Silva, Amauri H Souza, Vikas Garg, Samuel Kaski, and Diego Mesquita. When do GFlowNets learn the right distribution? In *International Conference on Learning Representations (ICLR)*, 2025.

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence (UAI)*, 2022.

Haoran He, Can Chang, Huazhe Xu, and Ling Pan. Looking backward: Retrospective backward synthesis for goal-conditioned GFlowNets. In *International Conference on Learning Representations (ICLR)*, 2025.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya A Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with GFlowNets. In *International Conference on Machine Learning (ICML)*, 2022.

Hyosoon Jang, Yunhui Jang, Minsu Kim, Jinkyoo Park, and Sungsoo Ahn. Pessimistic backward policy for GFlowNets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Minsu Kim, Joohwan Ko, Taeyoung Yun, Dinghuai Zhang, Ling Pan, Woo Chang Kim, Jinkyoo Park, Emmanuel Bengio, and Yoshua Bengio. Learning to scale logits for temperature-conditional GFlowNets. In *International Conference on Machine Learning (ICML)*, 2024a.

Minsu Kim, Taeyoung Yun, Emmanuel Bengio, Dinghuai Zhang, Yoshua Bengio, Sungsoo Ahn, and Jinkyoo Park. Local search GFlowNets. In *International Conference on Learning Representations (ICLR)*, 2024b.

Minsu Kim, Sanghyeok Choi, Taeyoung Yun, Emmanuel Bengio, Leo Feng, Jarrid Rector-Brooks, Sungsoo Ahn, Jinkyoo Park, Nikolay Malkin, and Yoshua Bengio. Adaptive teachers for amortized samplers. In *International Conference on Learning Representations (ICLR)*, 2025.

Zengxiang Lei, Ananth Shreekumar, Jonathan Rosenthal, Ruoyu Song, Alvaro A Cardenas, Daniel J Fremont, Dongyan Xu, Satish Ukkusuri, and Z Berkay Celik. Stable GFlowNets with TV monitoring and probabilistic guarantees. *arXiv preprint arXiv:2605.01729*, 2026.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning GFlowNets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning (ICML)*, 2023.

Idriss Malek, Aya Laajil, Abhijith Sharma, Eric Moulines, and Salem Lahlou. Loss-guided auxiliary agents for overcoming mode collapse in GFlowNets. In *AAAI Conference on Artificial Intelligence*, 2026.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in GFlowNets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Nikolay Malkin, Salem Lahlou, Tristan Deleu, Xu Ji, Edward J Hu, Katie Everett, Dinghuai Zhang, and Yoshua Bengio. GFlowNets and variational inference. In *International Conference on Learning Representations (ICLR)*, 2023.

Daniel E Newburger and Martha L Bulyk. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 37(Database issue):D77–D82, 2009.

Ling Pan, Dinghuai Zhang, Aaron Courville, Longbo Huang, and Yoshua Bengio. Generative augmented flow networks. In *International Conference on Learning Representations (ICLR)*, 2023.

Jarrid Rector-Brooks, Kanika Madan, Moksh Jain, Maksym Korablyov, Cheng-Hao Liu, Sarath Chandar, Nikolay Malkin, and Yoshua Bengio. Thompson sampling for improved exploration in GFlowNets. In *ICML Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

Marcin Sendera, Minsu Kim, Sarthak Mittal, Pablo Lemos, Luca Scimeca, Jarrid Rector-Brooks, Alexandre Adam, Yoshua Bengio, and Nikolay Malkin. Improved off-policy training of diffusion samplers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Max W Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho, and Tommaso Biancalani. Towards understanding and improving GFlowNet training. In *International Conference on Machine Learning (ICML)*, 2023.

Tomi Silander and Petri Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 445–452, 2006.

Jin Tian and Ru He. Computing posterior probabilities of structural features in Bayesian networks. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 538–547, 2009.

Nikhil Vemgal, Elaine Lau, and Doina Precup. An empirical study of the effectiveness of using a replay buffer on mode discovery in GFlowNets. In *ICML Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

Joseph D Viviano, Omar G Younis, Sanghyeok Choi, Victor Schmidt, Yoshua Bengio, and Salem Lahlou. torchgfn: A PyTorch GFlowNet library. *arXiv preprint arXiv:2305.14594*, 2023.

Xi Wang, Wenbo Lu, and Shengjie Wang. Rooted absorbed prefix trajectory balance with submodular replay for GFlowNet training. In *International Conference on Machine Learning (ICML)*, 2026.

Cong Zhu, Kelsey JRP Byers, Rachel Patton McCord, Zhenwei Shi, Michael F Berger, Daniel E Newburger, Katrina Saulrieta, Zachary Smith, Mita V Shah, Mathangi Radhakrishnan, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research*, 19 (4):556–566, 2009.

A Proofs

Throughout, M is a finite terminate-anywhere DAG-MDP with $R > 0$ on terminals; V' is the set of states reachable under P_F and $T' \subseteq V'$ its termination support. The interior balance identity is $F(s)P_F(s'|s) = F(s')P_B(s|s')$ and the exit balance identity is $F(s)P_F(\top|s) = R(s)$.

On the predicate. (C1) constructs S_t; it is not a test on the run — the verdict is (C2), invariant over the admissible cuts on all 20 study runs of the two structured cells (five seeds, 1M and 4M); S_t is not the discovered set D_t, which only grows (Fig. 7d). The predicate separates support deficit (C2) from calibration *on* that support (C3); (C2) reads $p^\star$ and so costs the partition function.

Proof of Prop. 1. Any sub-trajectory of a realized τ lies inside V' and, if it ends with $\top$, ends at some $s \in T'$. For an interior sub-trajectory,

$$A_{ij} = \log F(s_i) - \log F(s_j) + \sum_{k=i}^{j-1} \big[\log P_F(s_{k+1} \,|\, s_k) - \log P_B(s_k \,|\, s_{k+1})\big],$$

and each bracket equals $\log F(s_{k+1}) - \log F(s_k)$ by the interior balance identity, so the sum telescopes to $\log F(s_j) - \log F(s_i)$ and $A_{ij} = 0$. For a terminal sub-trajectory the same telescoping leaves $\log F(s_n) + \log P_F(\top \,|\, s_n) - \log R(s_n) = 0$ by the exit balance identity — the middle term being the extra forward factor of §2. Hence every residual of every realizable trajectory vanishes: the loss is identically 0 on the support, and so is its sampled gradient — for any batch size, step size and sub-trajectory weighting, since all three act only as non-negative weights on terms that are individually zero. Nothing in the argument uses λ, and TB and DB are the weightings that retain only the full-length and only the length-one sub-trajectories respectively (for TB, the learned $\log Z$ plays the role of $\log F(s_0)$ and equals $\log Z_{T'}$ at such a solution), so the statement covers those objectives verbatim. □

Proof of Cor. 1. Telescoping the interior identities along a realizable trajectory τ into x and closing with the exit identity gives $F(s_0)P_F(\tau) = R(x)P_B(\tau|x)$. Sum over the realizable trajectories into x: the interior identity forces P_B to vanish on edges of V' that P_F never takes, and the hypothesis on P_B forbids parents outside V', so those trajectories carry all of the backward mass from x and $p(x) = R(x)/F(s_0)$ on T'; normalizing, $F(s_0) = Z_{T'} := \sum_{x \in T'} R(x)$ and $p = R\mathbf{1}_{T'}/Z_{T'}$.

(Without the hypothesis on P_B one gets only $p \propto R\rho$ with $\rho(x)$ the backward mass that stays inside V', and $F(s_0) \le Z_{T'}$.) With $p^\star = R/Z$, $Z = \sum_x R(x)$,

$$\mathrm{L}_1(p, p^\star) = \sum_{x \in T'} R(x)\Big(\tfrac{1}{Z_{T'}} - \tfrac{1}{Z}\Big) + \sum_{x \notin T'} \tfrac{R(x)}{Z} = \big(1 - \tfrac{Z_{T'}}{Z}\big) + \big(1 - \tfrac{Z_{T'}}{Z}\big) = 2\big(1 - p^\star(T')\big),$$

using $p^\star(T') = Z_{T'}/Z$. □

Proof of Cor. 2. P_B vanishes on parents outside V', so the walk stays in V' and its reverse ends at $x \in T'$; by the interior balance identity $F(s)P_F(s'|s) = F(s')P_B(s|s')$ with $F > 0$ on V', $P_B(s|s') > 0$ forces $P_F(s'|s) > 0$, so the reversed walk is a trajectory in the policy's support and Prop. 1 applies to it. The replay term is the same squared residual evaluated on that trajectory, hence zero with zero gradient, whatever its weight. A uniform backward walk (the control of Table 7, panel (b)) can step to a parent outside V' and is not covered. □

Corollary 3 (Support-deficit lower bound)**.** *For any terminal set S and any p supported on S, $\mathrm{L}_1(p, p^\star) \ge 2(1 - p^\star(S))$, with equality iff $p \ge p^\star$ pointwise on S. If instead p places mass η outside S, then $\mathrm{L}_1(p, p^\star) \ge 2(1 - p^\star(S)) - 2\eta$; under **(C1)** of Def. 1, $\eta \le \delta$.*

Proof. Both $\sum_{x \in S} |p - p^\star|$ and $\sum_{x \notin S} |p - p^\star|$ are bounded below by $|1 - p^\star(S) - \eta|$, so their sum is at least $2|1 - p^\star(S) - \eta| \ge 2(1 - p^\star(S)) - 2\eta$; $\eta = 0$ gives the first claim. For the equality case, $\eta = 0$ makes the second sum exactly $1 - p^\star(S)$, so equality needs it in the first as well: $\sum_{x \in S} |p - p^\star|$ meets $|\sum_{x \in S}(p - p^\star)|$ only when the summands share a sign, and that sign is non-negative because they sum to $1 - p^\star(S) \ge 0$. □

Tightness, and what a match does not prove. Reward-proportionality is sufficient, not necessary. Once $p^\star(T')$ is small the condition is nearly free, so a measured match to Cor. 1 certifies the support deficit and *not* the calibration inside it — why §3 reports the restricted divergence as its own measurement. The three endpoints anchoring the mechanism are $2(1 - 0.00142) = 1.9972$ against a measured 1.9971; $2(1 - 0.0000) = 2.0000$ against 2.0000; and $2(1 - 0.39323) = 1.2135$ against 1.2057 — the only one of the three whose residual is visible at all, and a cell whose support is still slowly growing at the measurement point. For the second signature, on the *tournament attractor* — the frozen sub-support of complete five-node DAGs that Fig. 7's seed ends on — $\log F(s_0)$ settles at -37.28 against that support's own $\log Z_{T'} = -37.39$.

What the divergence bound does and does not license. Cor. 3 bounds the divergence of *any* policy whose operating support is S_t: $\mathrm{L}_1 \ge 2(1 - p^\star(S_t)) - 2\eta$, where η is the mass the policy places outside S_t and **(C1)** caps it at δ. A collapsed run therefore satisfies $\mathrm{L}_1 \ge 2\kappa - 2\delta = 0.9$ — *not* $\mathrm{L}_1 \ge 1$, the $\eta = 0$ special case, which silently drops the very term **(C1)** exists to bound. Contrapositively, $\mathrm{L}_1 < 0.9$ certifies **safe** outright — *seed by seed*, since Cor. 3 bounds one policy, not an average of five. So read, it decides the six non-BNSL safe rows of Table 5, whose worst *single seed* is 0.843, clear of the bound by 0.057; the seventh, BNSL sparse, has an enumerated support and is decided by Def. 1 like the other two BNSL rows. Read instead on a cell's five-seed mean the rule fails in the direction that matters — a mean below 0.9 is compatible with a *majority* of seeds above it, the aggregation the BNSL rows are decided at — so the generator certifies each seed and refuses any row carrying one at or above the bound. The printed L_1 column is the five-seed mean, and its own margin (0.805, clear by 0.095) is not the certificate's. Three further readings of Table 5 belong here rather than in its caption. The row where this bound is tightest, PHO4 β=16, is a 4M row, so the column's budget mix carries no verdict. The two verdict columns are independent, and so non-circular, only where neither was fitted: on the three BNSL rows the threshold was fitted against the observed verdicts, so those rows test nothing, while on the six others it was not. That buys less than it sounds. Below the threshold *predicted* withholds a verdict rather than certifying safety, since reach collapses sub-threshold landscapes unaided (Appendix C), and *every* out-of-sample row of Table 5 is sub-threshold — so on none of them is there a prediction for *observed* to agree with; the four six- and eight-node cells added later are all flag-side and one of them does not collapse (Appendix B). What the seven sub-threshold rows record is the rule's silence: not a tautology, but not a test either, and not a second certificate. Two cautions on the estimate itself. It is a mean of $-\log R$ over a possibly heavy-tailed distribution — on BNSL Gaussian the seed-to-seed spread alone exceeds the whole range the safe cells occupy — so we compute it over $20{,}000$ trajectories per seed and would

not trust a few hundred near the threshold. And being an expectation under π_0, it is a property of the landscape *and* the sampler, so it must be measured through the code that will train: for Table 5, the runs' own environment, preprocessor, width, estimator and seeds. Because Def. 1 cuts S_t at exactly $1-\delta$, this η is *attained* — measured across the structured runs it reaches 0.0498, so the realized bound, 0.900, is the worst case. That is the reason to cut there: read the operating support at any looser level and η shrinks with the cut rather than with the run, $\mathrm{L}_1 \geq 0.9$ becomes unreachable slack, and **(C1)** stops being a statement about the policy at all.

The bound runs *one way only*. Its converse — a large L_1 implying (C2) — needs the equality case of Cor. 3 ($p \geq p^\star$ pointwise on S_t with $\eta = 0$), which the verdict cannot presuppose: a two-way test would assume the structure the verdict is meant to establish. We therefore never do. The verdicts this paper rests on — the two collapse rows of Table 5, and the $5/5$ and $4/5$ seed counts of §5 — are decided by Def. 1 itself: **(C1)** fixes the operating support off each run's enumerated terminal distribution and **(C2)** is read on it; those runs are independently in the RSFP form, their restricted divergence measured at 0.0078–0.0471 at 4M (at most 0.0912 at 1M, §3). The bound and the predicate agree on all 20 structured study runs (five seeds, two cells, two budgets), but only the predicate licenses a collapse verdict, and only it is used for one.

S_t and T' are different objects and take different cuts. S_t is the tightest set **(C1)** admits, which is what makes $\eta \leq \delta$ attained rather than an artifact of a loose cut. Cor. 1, by contrast, is an identity on the support the RSFP actually carries, so T' takes the widest cut we compute (99% of p_θ); read at the operating cut it discards real support mass and its residual over the eight collapsed endpoints inflates fivefold, $0.0080 \to 0.0396$ — a measurement of the cut, not of the endpoint.

The cut does not carry the verdict, and we checked rather than asserted it. $p^\star(S^q)$ is non-decreasing in q, so if two levels return the same (C2) verdict every level between them does. Scored at $q = 0.90$ and at $q = 0.99$ the verdicts agree on all 20 structured study runs — 18 collapsed either way — which pins the whole interval, the operating cut $q = 1-\delta$ included. Only at $q = \frac{1}{2}$, far outside the range **(C1)** admits, does one run move (the gauss seed that recovers by 4M): that cut is degenerate for any calibrated policy, which sits on the (C2) boundary there. The generator re-derives this on every build.

Nor does κ carry it, and γ is met at both budgets. Over all 60 endpoints the post-hoc evaluation scores — three cells, two budgets, plain SubTB and ReStrike, five seeds — the true mass of the operating support is at most 0.377 on every run the predicate calls collapsed and at least 0.921 on every run it does not, so any κ from 0.079 to 0.623 returns the same 60 verdicts. γ enters only the calibrated-collapsed label, never the collapse verdict, and every collapsed run meets (C3): the restricted divergence is at most 0.0471 at 4M (§3) and 0.0912 at 1M, so the label survives any γ down to 0.0912. The generator derives these bounds too.

Two other places say "collapse" under a different rule, and we flag them rather than let the word travel. The minimal model of Appendix C is scored by an explicit proxy for Def. 1, which that appendix states and bounds. The tuning sweep behind Table 10 is scored by its own pre-registered criterion, which is looser than the bound above and which we have not re-scored against Def. 1; its verdict is therefore a statement about that criterion, not about the predicate.

Proof of Obs. 1. $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\ell(\tau;\theta)]$ with $\ell \geq 0$ pointwise, so $J \geq 0$ for every θ. By Prop. 1 a restricted solution has $\ell(\tau;\theta) = 0$ for π_θ-almost every τ, hence $J = 0$. The true full-support solution also attains $J = 0$. Therefore every restricted-support solution — including fully collapsed ones — is a global minimizer of J, exactly tied with the true solution: the value of J does not separate them (its local geometry is not claimed to coincide; Appendix G measures which statistics do and do not). □

Relation to the reverse-KL view. Malkin et al. (2023) show that the expected on-policy TB gradient is, up to a constant, that of a mode-seeking KL, so such objectives can reach a low loss without finding every mode; Prop. 1 and Obs. 1 are the exact limit case: at an RSFP every residual is zero for TB, DB and SubTB at every λ, the sampled loss itself vanishes, and the divergence equals $\log(Z/Z_{T'}) > 0$.

Lemma 1 (One component of the exit force, per trajectory, at initialization)**.** *For a sampled τ terminating at x with $\log R(x) = -\Delta_x$, the terminal residuals are $\Delta_x + O(|\tau| \log B)$ for branching factor*

B, all positive once Δ_x dominates, and $\partial \ell_\tau / \partial z_\top(x) = \sum_j 2 w_j A_j (1 - P_F(\top|x)) > 0$, summed over the terminal sub-trajectories of τ — the only ones of that *trajectory carrying the logit, since a DAG trajectory visits x once. The step on those terms therefore* lowers *the exit logit of the sampled terminal, to first order in that logit (exactly for a tabular policy), whenever its reward sits below the current flow scale; softmax renormalization pushes the freed mass deeper.*

Proof of Lemma 1, per-trajectory at initialization. At standard initialization $\log F \equiv O(1)$ and P_F is near-uniform, so for a sampled τ terminating at x with $\log R(x) = -\Delta_x$ the terminal residual is $\Delta_x + O(|\tau| \log B)$, positive once Δ_x dominates. The exit logit $z_\top(x)$ enters ℓ_τ only through terminal sub-trajectories, each contributing $w_j A_j^2$ with $\partial A_j / \partial z_\top(x) = 1 - P_F(\top \mid x)$, so $\partial \ell_\tau / \partial z_\top(x) = \sum_j 2 w_j A_j (1 - P_F(\top|x)) > 0$ whenever the A_j are positive. A gradient step therefore lowers that exit logit, and softmax renormalization redistributes the freed mass to continuation actions. The sign flips only for terminals whose reward exceeds the current flow scale, which is why background terminals suppress their own exits while modes do not. The force is a *vote* rather than a sum: $z_\top(x)$ also sits in the softmax denominator of x's continuations, so a batch trajectory passing *through* x contributes $-2 w_j A_j P_F(\top \mid x)$: smaller per trajectory by the factor $P_F(\top|x)/(1-P_F(\top|x)) \approx 1/B$ but drawn about B times as often, so of the same order in expectation and of the opposite sign at initialization. The net direction is the measurement of Fig. 7, not the lemma. □

Corollary 4 (Balanced exit rate)**.** *At any solution satisfying the interior balance identity on V' and the exit balance identity on T', with P_B supported on V', the exit probability at $s \in T'$ is $P_F(\top \mid s) = R(s)/F(s)$. Flow conservation gives $F(s) = \sum_{x \succeq s} F_s(x)$, where $F_s(x) \le R(x)$ is the flow routed from s to terminal x and $F_s(s) = R(s)$; hence*

$$K(s) \;:=\; \frac{F(s)}{R(s)} \;=\; 1 + \sum_{x \succ s} \frac{F_s(x)}{R(s)} \;\le\; 1 + \sum_{x \succ s} \frac{R(x)}{R(s)},$$

with $K(s) > 1$ exactly when some child of s carries flow — at the full-support solution, whenever s has a child at all. The routed flow is $F_s(x) = R(x) \Pr(x \rightsquigarrow s)$ for $x \in T'$, the chance the P_B walk back from x visits s: $K(s)$ is therefore fixed by R, P_B and T' together and not by the landscape below s, of which only the upper bound is a property. At s_0, which no trajectory avoids, $F(s_0) = Z_{T'}$ and $K(s_0) = Z_{T'}/R(s_0)$ for every P_B supported on V' — the landscape's own $Z/R(s_0)$, and so the bound, exactly at full support. Where $K(s) > 1$, under a softmax over $\{\top\} \cup \mathrm{ch}(s)$, the correct exit logit sits $\log(K(s)-1)$ below the log-sum-exp of its continuation logits; equivalently, and without the caveat, $\log R(s)$ sits $\log K(s)$ below $\log F(s)$. (A childless s has $K(s) = 1$ and exits with probability one.)

Proof. The first identity is the exit balance condition $F(s) P_F(\top|s) = R(s)$. Summing the interior identity over the children of s gives $F(s) = R(s) + \sum_{s'} F(s \to s')$, and unrolling it along the finite DAG expresses $F(s)$ as the total flow it routes to terminals at or below s, of which x receives $F_s(x) \le R(x)$ (equality iff every trajectory reaching x passes through s). Substituting $F(s \to s') = F(s') P_B(s|s')$ and iterating back from the terminals of T' identifies that flow as $F_s(x) = R(x) \Pr(x \rightsquigarrow s)$; every trajectory passes s_0, so $\Pr(x \rightsquigarrow s_0) = 1$ for $x \in T'$ and $F(s_0) = Z_{T'}$. For the logit form, $P_F(\top \mid s) = 1/K(s)$ under a softmax means $\sum_c e^{z_c} = (K(s) - 1)\, e^{z_\top}$, i.e. $z_\top = \log \sum_c e^{z_c} - \log(K(s) - 1)$. □

What the corollary says about larger problems. The landscape below s sets a *ceiling* on the funnel rather than its depth: it fixes $\bar{K}(s) = 1 + \sum_{x \succ s} R(x)/R(s)$, and $K(s)$ lies in $[1, \bar{K}(s)]$, with P_B — trained (§2), and the object the reconstruction control of §5 varies — deciding where. There is no landscape-only bound the other way: where s's descendants are reachable without passing s, routing can return $K(s)$ toward 1 with arbitrarily much reward below it. (On $s_0 \to \{a, b\} \to c$ with $R \equiv 1$ and ceiling $\bar{K}(a) = 2$, $P_B(a \mid c) = \frac{1}{10}$ gives $K(a) = \frac{11}{10}$ and $\frac{9}{10}$ gives $\frac{19}{10}$ — two exact GFlowNets on one landscape, spanning the band.) Where the ceiling is an identity, P_B drops out: at s_0, which no trajectory avoids, $K(s_0) = Z_{T'}/R(s_0)$, which at full support is the landscape's own $Z/R(s_0)$ — so the funnel the *root* must dig deepens exactly with the rewarded mass of the whole space. It is the same identity as the $\log F(s_0) \to \log Z_{T'}$ signature of §3, read at full support. Suppressing background exits is therefore the right response to the support the sampler holds — Lemma 1 is that response starting from the floor of K's range at initialization, where a near-uniform policy exits as though K were the branching factor rather than the far larger value the reward below

s licenses. What the corollary does *not* say is that a large K causes collapse; a correctly-supported sampler has large K too. The size claim is the empirical one of Appendix B, and it is about reach.

Escape channel (conditional). With finite logits the off-support exit mass per trajectory is $\mu_t > 0$. Let b be the batch size. If the dynamics suppress fast enough that along every history with no repair event $\sum_t \mu_t \le M_0$ almost surely, and $\mu_t \le \frac{1}{2}$ from a deterministic step t_0 on with $\mu_t \le \bar{\mu} < 1$ before it, then $P(\text{no repair ever}) = \mathbb{E}[\prod_t (1-\mu_t)^b] \ge c\, e^{-2bM_0} > 0$, where $\log(1-\mu) \ge -2\mu$ supplies the exponential over $t \ge t_0$ and $c = (1-\bar{\mu})^{bt_0} > 0$ collects the finitely many earlier steps — which the constant must carry, since μ_t is $O(1)$ while the funnel is still open. So collapse persists on that event. The premise — an almost-sure bound on cumulative off-support mass — is what the measured runs exhibit (exit probability is below $2{\cdot}10^{-4}$ at every depth by step 200, Appendix G) but is *not* derived from the coupled dynamics, and we label it empirical. **The bound counts on-policy draws only**: replay reaches an off-support exit by a route the product above does not model, so it does not lower μ_t — it voids the premise. An ε-floor contributes about $(\varepsilon/B)^d$ per step to μ_t for a repair d exits away, so the finite-mass premise fails at every depth. Read as a heuristic — the product with μ_t of that order — persistence over a budget of N steps is of order $e^{-bN(\varepsilon/B)^d}$: negligible when the repair is one exit away, close to one for a deep needle ($d = d^\star$), so ε is a partial, regime-dependent palliative rather than a fix. On the two production cells the repairs *are* one exit away, so the heuristic predicts no persistent collapse under any floor, and the pre-registered control of Table 3 mostly agrees: all but 2 of its thirty ε-floor runs clear Def. 1. The 2 that do not, at ε=0.01 on the Gaussian, are what the heuristic cannot see: a repair event the floor draws is not always *retained*, the race of §3, which the product above does not model. And every ε run keeps a calibration error that replay removes on every paired seed but one against the floor-only arms, but two once the floor-plus-replay arms are counted (§5).

B The deficit predictor, and why it is one-way

Two quantities govern the race: the initial reward deficit $\mathbb{E}_{\pi_0}[-\log R]$, setting how fast the funnel closes, and the reach of the reward's modes, setting whether repair events arrive first. Reach is a property of the landscape alone; the deficit is an expectation under the untrained policy, so it depends on the sampler too; neither depends on the training run or its outcome. Only the deficit is computable *in practice*: reach asks where the modes are, the question the sampler is being trained to answer.

We fit the rule on the three BNSL cells only, then applied it unchanged to the six remaining production cells (Table 5). The separation is wide — the two collapsing cells sit at 623 and 89 nats, every non-collapsing cell at 25.9 or below, a $3.4\times$ gap with nothing between — and the rule flags exactly the two cells that collapse. **All six out-of-sample rows sit at** 11.5 **nats or below, so they exercise only the side the rule withholds on**, for the reason the next paragraph measures. The size rungs (Appendix E.1) add four flag-side rows out of sample, 121 and 823 nats at six nodes and 253 and 1169 at eight (v-structure, Gaussian; five seeds): three collapse and the six-node v-structure does not (only local search, on one seed) — the rule's first false flag, in the direction its one-way statement allows.

The second axis is real, and separable. The reach axis needs a setting holding deficit fixed: a minimal needle lattice — branching B, background depth Δ_{bg}, one reward needle at depth $d^\star$ — trained with the production objective on a tabular policy (Appendix C). Sweeping the two axes independently (Fig. 3) gives a joint threshold, not a product: at fixed deficit a deep enough needle still collapses, at fixed reach a large enough deficit collapses regardless. **Every** Δ_{bg}=20 **cell measures below the** 30**-nat threshold fixed in the next paragraph**, so read as a clearance the rule is wrong on 9 of its 12 cells there, against $12/12$ right at Δ_{bg}=80: **that is why the check runs one way only.** Widening the *golden corridor* — the set of repair paths, Appendix C — at unchanged Δ_{bg} monotonically rescues the run.

The size axis. Size enters this law through reach (Appendices C, D): registered beforehand and measured on the lattice and on an enumerable grid ladder, collapse is indexed by budget and the repair by budget per mode, neither by size there. The DAG cells at six and eight nodes (Appendix E.1) add the limit the ladder does not reach: a support the run never finds, which replay cannot supply.

Cell	scale (nats)	baseline L_1	predicted	observed
BNSL Gaussian	623	1.490	**collapse**	**collapse**
BNSL v-structure	89	1.282	**collapse**	**collapse**
BNSL sparse	25.9	0.026	no flag	safe
HyperGrid Deceptive	11.5	0.124	no flag	safe
HyperGrid Hard	0.0	0.060	no flag	safe
HyperGrid GMM	9.2	0.431	no flag	safe
PHO4 β=16	7.1	0.805	no flag	safe
PHO4 β=10	4.5	0.663	no flag	safe
PHO4 β=4	1.8	0.316	no flag	safe

Table 5: **The predictor, applied before training.** *Scale* is the initial deficit $\mathbb{E}_{\pi_0}[-\log R]$ in *every* row, measured over $20{,}000$ trajectories per seed under the untrained policy of the runs' own network; the Hard landscape's untrained policy already sits at $R \approx 1$, so its deficit reads 0.0. The L_1 column mixes budgets: BNSL and PHO4 rows are 4M, the three HyperGrid rows 1M, the only grid runs evaluated analytically. *Predicted* reads the deficit alone, with the threshold fixed on the three BNSL rows; *observed* reads the measurement alone — Def. 1 per seed, majority of five, on the BNSL rows, and the one-way bound of Cor. 3 (safe below $L_1 = 0.9$, never collapse) on the six others. **The columns are not read alike:** *observed* certifies, *predicted* only flags, and below the threshold it returns no verdict to agree with. Appendix A states how each verdict is taken and why the two flagged rows are the only place the columns can be compared.

Using it, one way only. The practical form is a pre-flight check: compute $\mathbb{E}_{\pi_0}[-\log R]$ under the untrained policy against the 30-nat threshold the three BNSL cells fix. **Above it, treat the landscape as at risk; below it, conclude nothing.** The cut is on the *absolute* deficit, not on the deficit normalized by the reward's dynamic range, which does not separate our nine cells. None of our seven sub-threshold rows contradicts the rule, but what certifies them is Cor. 3, which does not read the deficit at all; the estimate itself is a mean over a possibly heavy-tailed distribution and is sampler-dependent (Appendix A). Within those limits the flag costs nothing and applies to landscapes we have never seen.

The cut is stated on a reward scale, and which one. $\mathbb{E}_{\pi_0}[-\log R]$ is not invariant to $R \to cR$: it shifts by $-\log c$. What fixes the scale is the sampler, whose initial flow sits at $\log F \equiv O(1)$ (Lemma 1), so the deficit is the initial log-reward read against the initial flow — and every one of the nine cells presents its reward with a maximum of order one ($\log R_{\max}$ is zero on the BNSL and PHO4 cells and within one and a half nats of zero on the three grids), the convention the 30-nat threshold assumes. A rescaled reward is not a different landscape but a different initialization of the same race: $R \to cR$ acts on every residual exactly as lowering $\log F$ by $\log c$ at initialization does, which is the counterfactual of Appendix G, where it leaves the onset where it was and moves the escape. On a reward not normalized this way the check is therefore read on $\mathbb{E}_{\pi_0}[\log(R_{\max}/R)]$, or the cut re-fit: the threshold is a property of the reward's shape under this sampler's initialization, not of its units.

C Minimal model: construction, gates, and sweeps

Construction. A terminate-anywhere lattice of depth D on the two-coordinate state (k, g) — depth $k \in \{0, \dots, D\}$ and a flag for whether the trajectory is still on the golden path — so the space is $2D{+}1$ states, every one able to exit. On the path, B continuation actions plus the exit are legal ($B{+}1$): m stay on the path, the remaining $B - m$ leave it; off the path a single continuation plus the exit remain. Aggregating the off-path states by depth is what makes the DP exact, and it makes the lattice a DAG rather than the expanding tree it stands in for: an off-path state at depth $k{+}1$ has two parents, the on-path and the off-path state at depth k, reached by $B - m$ and by one edge respectively, so P_B there is a learned, fractional tabular head trained alongside P_F and entering the SubTB residuals as it does on the DAG world, where on a tree it would be identically one and drop out. The lattice's P_B dynamics are therefore a DAG's, not a tree's; the loss-match gate below is taken on the chain, where P_B is trivial, and on the lattice the gates are the DP's. Terminals carry background reward $e^{-\Delta_{\text{bg}}}$ with two exceptions: the *single* on-path terminal at depth $d^\star$ carries reward 1 (the needle), and the two depth-D terminals, on- and off-path, carry $e^{-\Delta_{\text{bg}}/10}$ — a forced dead end well above background, mirroring the plateau of the DAG world. That plateau is not a detail: it is why $p^\star(\text{needle})$ falls away from 1 as the deficit shrinks rather than staying near 1 — what the scoring paragraph below turns on, and

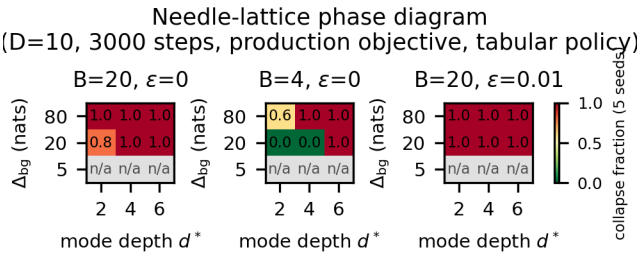

Figure 3: **Phase diagram of the minimal model** (collapse fraction over 5 seeds). The boundary is joint in deficit and reach; neither axis alone predicts the outcome — the deficit alone included, which is why the pre-flight check of Appendix B is claimed one way only. B is the size axis, entering through the uniform reach $(m/(B{+}1))^{d^\star}$ — $(B{+}1)^{-d^\star}$ at the $m{=}1$ swept here; the full sweep over it is Fig. 4. **The $\Delta_{\text{bg}}{=}5$ row carries no collapse verdict** and is greyed out: the score is a proxy whose fixed threshold sits *above* the calibrated needle mass at that deficit, so it is no collapse measure there (Appendix C). Both halves of the joint boundary are visible without it.

where its three values are given. The sweeps use $D{=}10$. Because the DP state is (depth, on-path), B costs nothing to evaluate and enters only through the uniform reach $(m/(B{+}1))^{d^\star}$ of the needle: a controlled proxy for what size does to reach, not a state count — the state-count axis is Appendix D. The space is small enough for exact dynamic programming to compute the terminal distribution, the state-visitation measure and every SubTB residual — no sampling enters any reported quantity except where a sampled gradient is the object of study. Training reuses the production objective, optimizer, learning rates and λ; the policy, however, is a *tabular* lookup over the $2D{+}1$ states rather than the production MLP, so this lattice differs from the real cells in parameterization as well as landscape.

Scoring, and what it is a proxy for. Each configuration runs for 3,000 steps over 5 seeds, scored by an exact-DP *proxy* for Def. 1 rather than Def. 1 itself: the model's terminal mass on the needle, thresholded at $\frac{1}{2}$. These are different tests — **(C2)** reads the *true* mass of the model's support, the proxy the *model's* mass on one state — and the proxy's threshold is landscape-independent while the quantity it stands in for is not. With the plateau at $e^{-\Delta_{\text{bg}}/10}$, the needle carries $p^\star = 0.9993, 0.7870$ and 0.4284 at $\Delta_{\text{bg}} = 80, 20, 5$. At $\Delta_{\text{bg}} \in \{20, 80\}$ the threshold sits well below the calibrated value and the proxy is a usable, if not identical, stand-in. At $\Delta_{\text{bg}}{=}5$ it sits *above* it: a correctly calibrated needle mass would itself score "collapsed", so that column is not a collapse measure at all. We log the needle mass, the dead-end masses and L_1 but not the full terminal pmf, so those verdicts cannot be re-derived under **(C2)** without re-running — we do not claim they are wrong, only that they are uninterpretable. **We therefore report no collapse verdict at $\Delta_{\text{bg}}{=}5$**, reading that row by the two predicate-free measures instead, mean L_1 and median mode visits. This is a post-hoc correction: the sweep as run, including the registration below, used the proxy throughout, and the released artifact carries the original numbers unchanged.

Gates. The module's self-test (`--validate`) checks three things, on two fixed lattices that are *not* swept cells: that its closed-form per-state loss matches the library's SubTB loss, that the DP terminal mass equals 1 to within 10^{-6}, and that a Monte-Carlo re-estimate of the terminal distribution matches the DP result. **The swept probes themselves are not gated**: each normalizes the DP pmf before scoring it, so a mass defect there would be divided out rather than caught. That is weaker than the DAG evaluator of the main text, which gates every run, and we state the discipline this lattice has rather than the one that world has.

Sweeps. Fig. 3 sweeps Δ_{bg} against $d^\star$ at $B \in \{4, 20\}$ with $\varepsilon = 0$, and repeats the $B{=}20$ panel at $\varepsilon = 0.01$. The boundary is joint: neither axis alone predicts the outcome. Cell by cell, at fixed deficit a deep enough needle still collapses ($\Delta_{\text{bg}}{=}20$, $B{=}4$: $0/5$ seeds at $d^\star{=}2$ and at $d^\star{=}4$, $5/5$ at $d^\star{=}6$), and at fixed reach a large enough deficit collapses regardless ($B{=}4$, $d^\star{=}2$: $0/5$ at $\Delta_{\text{bg}}{=}20$, $3/5$ at 80) — the two halves Appendix B reads off this figure. The ε variant shifts exactly two boundary cells (one per branching factor; the figure shows the $B{=}20$ one), one seed each, in opposite directions — consistent with ε acting on one-exit-away repairs only. A separate corridor-width sweep holds $(\Delta_{\text{bg}}, d^\star) = (20, 4)$ at the collapsing corner, at *one* rewarded terminal throughout, and varies the golden-branch multiplicity $m \in \{1, 2, 4, 8\}$ (`m_gold` in the released code) — that is, the uniform

reach $(m/(B{+}1))^{d^\star}$, which rises from $5.1\cdot10^{-6}$ to $2.1\cdot10^{-2}$ across the four. The *trained policy's* mass on the needle rises $0.001 \to 0.036 \to 0.336 \to 0.617$, against a true target of $p^\star = 0.787$, so even at $m{=}8$ it is still short; the collapse fraction falls $1/1/1/0$. Reach, not deficit, is what that sweep moves, and it moves the outcome. Its deficit is one where the proxy above is usable, so these four verdicts stand.

Scoring the deficit law on this lattice. The law of Appendix B cuts $\mathbb{E}_{\pi_0}[-\log R]$ at 30 nats. That is *not* Δ_{bg}, which is the lattice's background log-reward depth: the log-reward table takes only $-\Delta_{\text{bg}}$, $-\Delta_{\text{bg}}/10$ and 0, so $\mathbb{E}_{\pi_0}[-\log R] \le \Delta_{\text{bg}}$ for every initial policy — a bound that settles the *safe* side and cannot license the *flag* side, since ≤ 80 is compatible with < 30. We therefore measure it, per cell, by exact DP for the uniform initial policy through the released environment's own masks and transitions (terminal mass 1 to machine precision). It never exceeds 19.97 at $\Delta_{\text{bg}}{=}20$ and never falls below 79.21 at $\Delta_{\text{bg}}{=}80$, so the cut lands where the construction suggests — but now by measurement. Scoring the two deficits whose collapse verdicts this appendix admits — over the $\varepsilon{=}0$, replay-free cells at all four branching factors, under the majority-of-five aggregation Table 5 uses — at $\Delta_{\text{bg}}{=}80$ the rule flags risk and is right on $12/12$; at $\Delta_{\text{bg}}{=}20$ it reads *no flag*, and 9 of those 12 cells collapse anyway. The failures are exactly where reach is short, at the deeper needles and the larger branching factors: the joint boundary of Fig. 3 restated as a scoring error. This is the measurement behind the one-way reading of Appendix B: the flag is claimed and the clearance is not.

The size axis (pre-registered). The B sweep was widened to $B \in \{4, 8, 20, 40\}$ after the fact but before looking, under a registration that fixed the prediction (collapse fraction non-decreasing in B) and what would falsify it (a drop of two seeds or more — coarser than the one-seed granularity of a five-seed estimate — at ≥ 2 of the nine $(\Delta_{\text{bg}}, d^\star)$ cells). Result: **zero cells drop** — at all nine in fact, but we count only the six at $\Delta_{\text{bg}} \in \{20, 80\}$, where the collapse score is usable, so the registered rule is decided on six cells rather than the nine it was written against. The registration's second measure needs no predicate and is untouched: the mean L_1 increases in B at **nine of nine**, three of them already at the L_1 ceiling, where the rise is below the printed precision. The direction is visible at the mildest deficit swept, read off the predicate-free measures at $\Delta_{\text{bg}}{=}5$, $d^\star{=}4$ over $B = 4, 8, 20, 40$ — median mode visits falling and mean L_1 rising across the four, both given in the caption of Fig. 4. Re-running the sweep leaves the 36 previously computed cells byte-identical (the resume key carries $(\Delta_{\text{bg}}, d^\star, B, \varepsilon, \text{replay})$), which is checked rather than assumed. An $\varepsilon{=}0.01$ floor reproduces the two-cell, one-seed shift reported above, now across four branching factors rather than two.

The corners at production budget, and what it costs the lattice (pre-registered). The sweep's collapse fractions are "not escaped within 3,000 steps". Registered with the B sweep, we re-ran the four collapsing corners ($B{=}20$, $\Delta_{\text{bg}} \in \{20, 80\} \times d^\star \in \{4, 6\}$, five seeds) for 15,625 steps — the 1M budget, a quarter of the 62,500 the enumerable cells train at — predicting that replay would clear the mode-mass threshold in at least two of them. **The prediction is void, because the control clears all four**: plain SubTB goes from $5/5$ seeds collapsed at every corner to $0/5$, with L_1 $1.785 \to 0.376$, $1.787 \to 0.302$, $1.999 \to 0.047$ and $1.999 \to 0.101$, and median mode visits rising from 0–21 to 224,502–672,886. The arms then split: replay reaches $\mathrm{L}_1 = 0.000712$ and 0.001094 at the two $\Delta_{\text{bg}}{=}20$ corners — two orders below plain SubTB — and *loses* at both $\Delta_{\text{bg}}{=}80$ corners (0.334 against 0.047, 0.365 against 0.101), where it also draws *fewer* mode visits, consistent with a reward-prioritized buffer whose best stored terminal is still background rehearsing the wrong support while the needle is undiscovered. Two conclusions we hold ourselves to. The lattice cannot testify about the repair: at the budget its cells are comparable at, no collapse remains to repair and the replay result is a $2:2$ split. And every collapse fraction in this appendix, including the B sweep above, is a statement about a fixed budget — unlike the enumerable cells, where SubTB is still collapsed on $4/5$ seeds after 62,500 steps. The lattice is where the two axes separate, not where the phenomenon is durable.

D The enumerable size ladder

Where the lattice varies reach by hand, this varies the space. The tiered HyperGrid's top tier is a fractional band, so as the height grows the mass it carries stays in a narrow range instead of tracking the space, while the state space grows $12.4\times$ and the mode set $16\times$ from 32^4 to 60^4. It is *not* constant — we measured it at every height rather than assuming it, and it runs 0.042, 0.051, 0.058, 0.051 — so

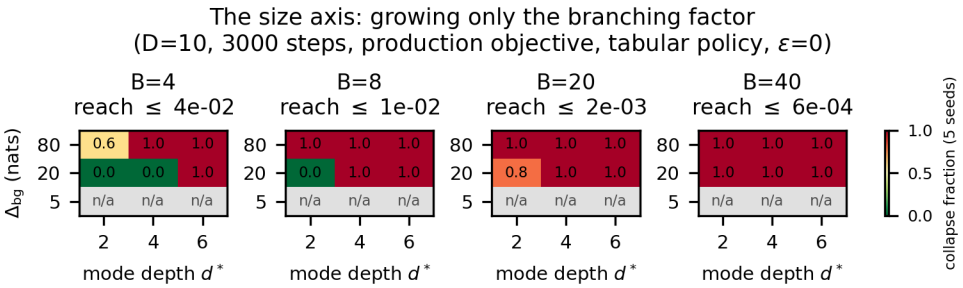

Figure 4: **The size axis of the minimal model.** Growing only the branching factor, at a fixed reward landscape, moves every cell toward collapse: no cell's collapse fraction falls with B, and the mean L_1 rises in B at nine of nine cells (at three, already at the ceiling, by less than the printed precision). At Δ_{bg}=5 — far below the threshold the deficit law is fit at — the collapse score carries no verdict and the row is greyed (Appendix C); the direction there is read instead off median mode visits ($51{,}863 \to 10{,}483 \to 4 \to 0$) and mean L_1 ($1.022 \to 1.049 \to 1.934 \to 1.948$).

this is size at a nearly held reward geometry, with one saving grace we did not arrange: the residual does not order with height, and runs *against* the result below. The landscape carrying the most top-tier mass (h=48) is covered worse by plain SubTB than the one carrying the least (h=32). Two arms — plain SubTB and the reported arm — three seeds at every height, registered before the runs.

No single ladder isolates size, so we report both and say what each confounds. *Budget scaled with the problem* fixes trajectories per mode at 192.9, the rate the h=60 4M run spends, so the optimizer-step count varies from 3,906 to 62,500. *Budget fixed at 1M* fixes the step count at 15,625, so trajectories per mode vary from 772 to 48. Every point is read from the per-trajectory mode traces at indices fixed in the registration, so no run is trained twice and the h=48 and h=60 heights reuse the runs already reported.

			budget scaled with the problem			budget fixed at 1M		
Grid	states	modes	traj	SubTB	**ReStrike**	traj/mode	SubTB	**ReStrike**
32^4	1,048,576	1,296	0.25M	30 (2.3%)	357 (27.6%)	772	31 (2.4%)	1,296 (100.0%)
40^4	2,560,000	4,096	0.79M	100 (2.4%)	4,077 (99.5%)	244	100 (2.4%)	4,094 (99.9%)
48^4	5,308,416	10,000	1.93M	123 (1.2%)	10,000 (100.0%)	100	123 (1.2%)	9,833 (98.3%)
60^4	12,960,000	20,736	4.00M	140 (0.7%)	20,734 (100.0%)	48	134 (0.6%)	16,442 (79.3%)

Table 6: **The size axis on an enumerable space.** Modes discovered (fraction of the cap), 3 seeds. The space and the mode set grow; the tier mass stays in a narrow range and is not held exactly (above). Three seeds is a descriptive ladder and not a test — the exact two-sided sign floor at n=3 is 0.25 — so no entry here is called significant.

Two readings of Table 6, and one correction to our own registration. **Plain SubTB fails harder as the space grows**: its absolute mode count barely moves ($31 \to 134$ at the fixed budget) while the mode set grows 16×, so coverage decays with size, $2.4\% \to 2.4\% \to 1.2\% \to 0.6\%$. **The repair is limited by budget per mode, not by size**: with the budget scaled it holds 99.5%, 100.0% and 100.0% at $h = 40, 48, 60$ — flat across a 5× larger space — and at the fixed budget it degrades only as trajectories per mode do, 100.0%, 99.9%, 98.3%, 79.3% against 772, 244, 100 and 48 per mode. The correction: our registered prediction was $\geq$ 95% at *every* height, and the 32^4 height of the scaled ladder returns 27.6%. The registration's own secondary rule says why. Scored on the traces, reaching 95% costs 332.5 trajectories per mode at 32^4 against 79.9 at 60^4, so the scaled ladder's 192.9 sits below what the *smallest* grid needs. **The per-mode cost falls as the space grows** (h=60 needs 0.24× what h=32 needs), the reverse of the direction that rule was written to catch, and why the shortfall appears at the small end. The registered falsifier — coverage declining across $\geq$ 3 consecutive heights with a total drop $\geq$ 10 points — did not fire in any window of the scaled ladder it was registered on; read against the fixed-budget column it would, 100.0% to 79.3% across the four heights, the trajectories-per-mode decline described above. The $\geq$ 95% clause failed, and we report it as failed rather than restating the prediction.

Held fixed across every height, and therefore untested: the network width, and the buffer capacity (10,000), which at h=60 is already below the mode count. The cap is not what bounds coverage at

that height — with the budget scaled, the same capped buffer reaches 100.0% of the $20{,}736$ modes — but whether a larger buffer would lift the fixed-budget height is a run we have not made, and the 79.3% there confounds the two.

E The percentile band, full results, tuning sweep, and the molecular world

The figure and the tables below are quoted from §4 and §5 and live here because they are boundary rather than headline evidence: the band fraction is the measured half of an argument §4 also makes analytically, the ablation's verdict is that the fork changes nothing surviving correction, and PHO4 is our disclosed Pareto, not the divergence-robust cell. Their numbers are stated in the text either way.

Three analytic evaluations exist at 1M. The one on Deceptive (0.089 against plain SubTB's 0.124) backs Fig. 1c's reading of that cell as healthy calibration and claims nothing; the two behind the withdrawal of §5 on the grid cells are these: on GMM plain SubTB sits well *above* ours and above PRT-full, and on Hard above ours too — 0.060 against 0.049, four of five seeds, $p = 0.0445$, short of our Bonferroni level. Neither licenses an ordering in our favour, which is why the claim is withdrawn rather than read as a tie. The deficit law does not flag PHO4 either, but it is measured exactly rather than at a floor, and it does not tie.

The distinct-state count is not a ranking metric. A perfectly calibrated i.i.d. sampler of the tiered cell would touch $911{,}151$ distinct terminals in 1M draws *in expectation*; seven of eight arms exceed it (plain SubTB $925{,}304$, ours $918{,}082$), because the metric accumulates over a run starting near-uniform rather than sampling the converged policy — and the uniform policy is the one that maximizes it. Overshoot is therefore the expected behavior, not a guaranteed one — PRT-full falls below it. The reference bounds nothing and a higher count is evidence of neither better sampling nor worse calibration.

The straddle is the released code's rule. In its authors' words the rule focuses on "the top 10% ranked by reward and randomly sample[s] among them to be 50% of the batch, with the bottom 90% ranked comprising the remainder of the batch". The closed-interval reading of §4 is not an interpretation of the baseline's paper but of its implementation: `__biased_sample_xs` takes `np.percentile(rewards, 90)` as a *value* threshold and forms the two sides as `r >= threshold` and `r <= threshold`, both closed, then draws with replacement from each. A rank-based split at the 90th-percentile index would leave the sides disjoint and could not straddle; it is not what the code does. Our PRT arms replicate the value rule (`torch.quantile`, the same interpolation) rather than a rank rule, and the band fraction of Fig. 5 is measured on that code path. A terminal *at* the threshold falls in both halves and is drawn from each, $1 + n_{\text{top}}/n_{\text{bot}}$ times as often per element as a strictly higher-reward terminal above it.

Configuration	replay	Lightning	BNSL v-structure	BNSL Gaussian
(a) component ablation — 4M, test-L_1				
SubTB (no replay)	—	—	1.2824	1.4918
ReStrike (ours)	✓	—	0.0401	0.0594
ReStrike + Lightning warm-up	✓	✓	**0.0295**	**0.0589**
(b) reconstruction control — 1M, exact forward-DP, budget-matched				
learned P_B reconstruction	✓	—	**0.0499**	0.1233
uniform P_B reconstruction	✓	—	0.0515	**0.0840**

Table 7: Component ablation and the pre-registered reconstruction control. The two panels are at different budgets and metrics and must not be compared across. Panel (a)'s bolding is *nominal*: no cell separates the Lightning row from the replay-only arm under multiplicity correction.

E.1 Six and eight nodes

Three registrations (`_preregister_bnsl_size.md`, `_preregister_bnsl_size8.md`, `_preregister_bnsl_size_composite.md`), each written before its runs and shipped with the code, take the two collapsing posteriors of Table 1 to six and eight nodes with the study's

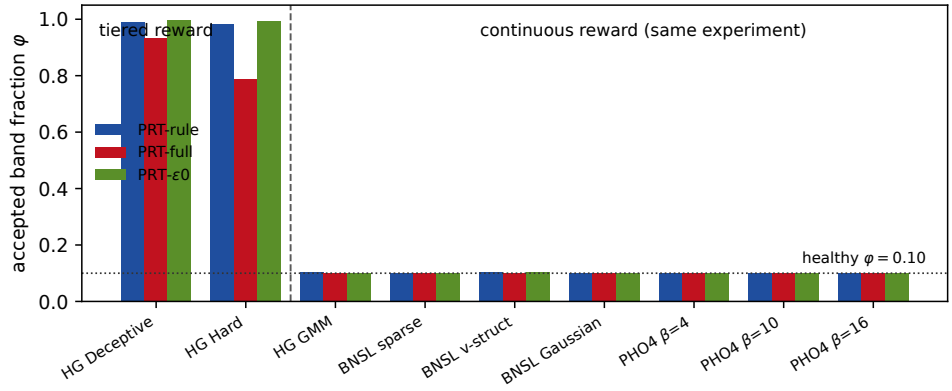

Figure 5: **The degeneracy, measured, with a within-experiment control.** Realized band fraction φ for the three percentile arms. On tiered rewards φ runs 0.789–0.996; on *continuous* rewards — same runs, same code path — it pins at the healthy 0.100–0.104. The split is binary, with no overlap. The analytic half of the argument — a straddling tier drawn $1 + n_{\text{top}}/n_{\text{bot}}$ times as often per element against a softmax ratio $\exp(\beta_r u)$ for a log-reward gap u, independent of tier population — is in the text above and §4 and does not rest on this figure.

β	Arm	L_1	JSD	$\mathrm{KL}(p_\theta \| p^\star)$	$\mathrm{KL}(p^\star \| p_\theta)$
4	plain SubTB	0.3158	0.0202	0.0794	0.0859
	ReStrike	0.2581	0.0132	0.0537	0.0535
	PRT-rule	**0.2455**	**0.0117**	**0.0473**	**0.0472**
	annealed-floor recipe	0.2610	0.0135	0.0548	0.0547
	ReStrike, constant fraction, r=1	0.2481	0.0124	0.0502	0.0502
10	plain SubTB	0.6629	0.0875	0.3833	0.3846
	ReStrike	0.5219	0.0607	0.2673	0.2606
	PRT-rule	0.5440	0.0571	0.2514	0.2331
	annealed-floor recipe	0.5402	0.0628	0.2770	0.2691
	ReStrike, constant fraction, r=1	**0.4756**	**0.0529**	**0.2283**	**0.2295**
16	plain SubTB	0.8045	0.1414	0.6258	0.8532
	ReStrike	0.5725	0.0897	0.4472	0.4609
	PRT-rule	0.6222	0.0813	0.3866	**0.3477**
	annealed-floor recipe	0.6305	0.0961	0.4728	0.4663
	ReStrike, constant fraction, r=1	**0.5192**	**0.0804**	**0.3707**	0.4719

Table 8: PHO4, exact metrics, 4M, 5 seeds — a disclosed Pareto, not a win. The oracle is the empirical PHO4 binding surface, read off UniPROBE 8-mer E-scores (Newburger & Bulyk, 2009; Zhu et al., 2009) and made reverse-complement symmetric; it is not a learned surrogate.

Cell	SubTB	**ReStrike**	RS+warm-up	RS-DISCO	PRT-r	PRT-f	PRT-ϵ0	AT
HyperGrid Deceptive	1.9511	1.9134	1.9136	1.9137	1.9324	1.9484	1.9511	1.9500
HyperGrid Hard	1.5169	1.5161	1.5157	1.5161	1.5176	1.5189	1.5182	1.5180
HyperGrid GMM	0.8916	0.8883	0.8840	0.8900	0.8888	0.9297	0.9103	0.9187
BNSL v-structure	1.2824	0.0401	**0.0295**	0.0305	0.0966	0.0813	0.1149	0.0852
BNSL Gaussian	1.4918	0.0594	0.0589	**0.0443**	0.5047	0.3939	0.4434	0.5249
BNSL sparse	**0.0448**	0.0453	0.0510	0.0475	0.0503	0.0549	0.0540	0.0599
PHO4 β=4	0.3158	0.2581	0.2619	0.2593	**0.2455**	0.3076	0.3015	0.2892
PHO4 β=10	0.6629	0.5219	**0.5172**	0.5198	0.5440	0.7247	0.7232	0.5840
PHO4 β=16	0.8045	0.5725	0.5684	**0.5643**	0.6222	0.7977	0.7941	0.7565

Table 9: All nine enumerable cells, 4M, 5 seeds — one metric across all for comparability, *not* a ranking table. PHO4 rows are exact; the rest are the 200k fresh on-policy estimate. The Hard row's nominal ordering is withdrawn (§7(ii)). **The three HyperGrid rows, greyed, sit at that estimate's sampling floor and no ordering may be read off them**: the analytic policy-L_1 for those cells is several times smaller (Table 5), and where we ran it against the same runs — GMM at 1M — it *reversed* this estimate's ordering, putting PRT-full first at 0.305 against 0.318 for ours. We have no analytic evaluation of the 4M runs, so the ordering printed here is unverified rather than merely noisy — the point, and exactly the bias §2 warns of. §7(ii) *withdraws* any claim on the three HyperGrid rows rather than calling them ties: at this estimate's floor they carry no ordering, and on GMM the analytic evaluation contradicts the one printed here. RS+warm-up and RS-DISCO are the two Lightning arms — ReStrike with the fork warm-up, and the discovery variant whose extra reward evaluations are disclosed in Appendix F, which defines both.

Question	Answer
Does any swept learning rate un-collapse the baseline?	**No**
Best tuned baseline, structured DAG (exact L_1)	1.0487
Ours, same cells (exact L_1)	0.0263 / 0.0229
Diffuse cell at lr 10^{-2}: collapse *spreads*	0.0254 $\rightarrow$ 0.8248
Tiered grid, modes: best tuned baseline vs. ours	294 vs. 20,734

Table 10: Baseline tuning robustness: 3 seeds $\{7, 19, 101\}$, 108 runs. The sweep varied the learning rate only; network capacity was not swept. ReStrike is re-read on the same three seeds, so its entries differ from the five-seed Tables 1 and 2.

Cell	Arm	seed 23	seed 59	seed 131	seed 503	seed 977	mean	collapsed
v-structure	plain SubTB	1.3594	1.1893	0.0308	1.2057	2.0000	1.1570	4/5
	plain SubTB, held-out loss	0.0107	0.0212	0.0255	0.0035	0.0058	0.0133	—
	ε-floor 0.1	0.1104	0.0594	0.0826	0.0579	0.0584	0.0737	0/5
	ε-floor 0.1, held-out loss	0.0268	0.0177	0.0170	0.0104	0.0118	0.0167	—
	ReStrike	0.0233	0.0229	0.0439	0.0342	0.0164	0.0281	0/5
	ReStrike, held-out loss	0.0310	0.0242	0.0613	0.0206	0.0190	0.0312	—
	annealed-floor recipe	0.0992	0.0769	0.0930	0.0935	0.1120	0.0949	0/5
	annealed-floor recipe, held-out loss	0.0309	0.0200	0.0189	0.0176	0.0240	0.0223	—
	floor + ReStrike draw	0.0634	0.0790	0.0534	0.0417	0.0609	0.0597	0/5
	floor + ReStrike draw, held-out loss	0.0080	0.0177	0.0073	0.0047	0.0102	0.0096	—
	ReStrike, constant fraction, r=1	0.0136	0.0295	0.0188	0.0266	0.0240	0.0225	0/5
	ReStrike, constant fraction, r=1, held-out loss	0.0227	0.0218	0.0427	0.0370	0.0137	0.0276	—
Gaussian	plain SubTB	1.9971	1.1373	1.9971	0.9526	1.9971	1.6162	5/5
	plain SubTB, held-out loss	0.0047	0.1429	0.0042	0.2310	0.0040	0.0774	—
	ε-floor 0.1	0.6497	0.7268	0.2068	0.4327	0.2696	0.4571	0/5
	ε-floor 0.1, held-out loss	0.4891	0.7946	0.1343	0.2167	0.1871	0.3644	—
	ReStrike	0.0256	0.0253	0.0287	0.0252	0.0243	0.0258	0/5
	ReStrike, held-out loss	1.8860	0.0889	0.1078	0.0042	0.0501	0.4274	—
	annealed-floor recipe	0.1881	0.3531	0.4259	0.1891	0.2327	0.2778	0/5
	annealed-floor recipe, held-out loss	0.0626	0.3240	0.6456	0.0893	0.1370	0.2517	—
	floor + ReStrike draw	0.2861	0.3668	0.2621	0.1199	0.3270	0.2724	0/5
	floor + ReStrike draw, held-out loss	0.3720	0.3300	0.1032	0.0418	0.0965	0.1887	—
	ReStrike, constant fraction, r=1	0.0249	0.0536	0.0245	0.0395	0.0296	0.0344	0/5
	ReStrike, constant fraction, r=1, held-out loss	0.2976	0.0644	0.0539	0.7792	1.4644	0.5319	—

Table 11: **Five new seeds for the headline floor comparison**, pre-registered (`_preregister_seeds_beta.md`, `_preregister_floorreplay_tenseeds.md`): plain SubTB, the reference recipe's ε=0.1 floor, ReStrike and the two floor-plus-replay arms of Table 3 on seeds $\{23, 59, 131, 503, 977\}$, the two collapsing cells, 4M, exact forward-DP L_1 and held-out on-policy loss per seed; *collapsed* counts the seeds meeting (C2) of Def. 1. The study's five seeds (Tables 1–3) are not recomputed; the pooled ten-seed pairing and its tests are quoted in §5.

Cell	Arm	seed 7	seed 19	seed 101	seed 367	seed 911	mean
v-structure	SubTB, no replay, ε=0	1.2057	2.0000	0.0350	1.2056	1.9635	1.2820
	ε-floor 0.01	0.0635	0.0637	0.1371	0.0699	0.0756	0.0819
	ε-floor 0.1	0.0696	0.0529	0.0787	0.0666	0.0810	0.0698
	annealed floor	0.0483	0.0907	0.0973	0.1030	0.0530†	0.0785
	ReStrike, uniform draw	0.1064	0.1128	0.1837	0.0888	0.0809	0.1145
	ReStrike, constant fraction	0.0304†	0.0240	0.0218	0.0636	0.0288†	0.0337
	annealed-floor recipe	0.0725	0.0821	0.0610	0.0946	0.0964	0.0813
	floor + ReStrike draw	0.0470	0.0631	0.0750	0.0901	0.0669†	0.0684
	plain SubTB, 96 fresh per update	0.0317†	0.1944	0.0324	0.0952	0.6052	0.1918
	ε-floor 0.1, 96 fresh per update	0.1175	0.0393	0.0467	0.0492	0.0596†	0.0624
	ReStrike, constant fraction, 32 fresh + 16 replayed	0.1472	0.0558	0.0329	0.0471	0.0696†	0.0705
	modified DB, ε=0	1.8162	1.6475	1.8247	1.9454	1.8162	1.8100
	modified DB + ε-floor 0.1	0.0629	0.0358	0.0631	0.0821	0.1121	0.0712
	modified DB + annealed floor	0.1913	0.0849	0.0906	0.0676	0.0912	0.1051
	ReStrike, top-k draw	0.0181†	0.0308	0.6036	0.0255†	0.0241†	0.1404
	local search	0.6180	0.6180	1.8045	0.0278†	0.6180	0.7373
	local search + ε-floor 0.1	0.0805	0.0571	0.0530	0.0570	0.0434†	0.0582
	ε-floor 0.1, Huber residual	0.0744	0.0575	0.0444	0.0662	0.0565†	0.0598
	ReStrike, constant fraction, r=1	0.0301†	0.0240	0.0206†	0.0196†	0.0110†	0.0211
	ReStrike, constant fraction, r=2	0.0235†	0.0146†	0.0373	0.0277†	0.0103†	0.0227
	ReStrike, constant fraction, r=1, 32 fresh + 16 replayed	0.0280†	0.0298	0.0605	0.0234†	0.0491†	0.0382
	annealed-floor recipe, 32 fresh	0.1173	0.0691	0.1262	0.0819	0.0660†	0.0921
	ReStrike	0.0344	0.0234	0.0211	0.0324	0.0712	0.0365
Gaussian	SubTB, no replay, ε=0	1.3496	0.1111	1.9971	1.9971	1.9971	1.4904
	ε-floor 0.01	0.4542	1.1688	0.8619	0.8285	1.0785	0.8784
	ε-floor 0.1	0.2298	0.3114	0.3386	0.4876	0.2474	0.3229
	annealed floor	0.5465	0.9385	0.5635	0.2056	0.2287	0.4966
	ReStrike, uniform draw	0.4407	0.3018	0.3246	0.6029	0.5552	0.4450
	ReStrike, constant fraction	0.0232†	0.0228†	0.0278	0.0245†	0.0210†	0.0239
	annealed-floor recipe	0.5723	0.2183	0.3177	0.3002	0.2213	0.3260
	floor + ReStrike draw	0.7084	0.1954	0.2962	0.1989	0.1459	0.3089
	plain SubTB, 96 fresh per update	1.9971	1.9971	1.1259	1.9971	1.9971	1.8229
	ε-floor 0.1, 96 fresh per update	0.4865	0.4999	0.3732	0.3244	0.1598	0.3688
	ReStrike, constant fraction, 32 fresh + 16 replayed	0.0301	0.0483	0.0636	0.0358	0.0380†	0.0432
	modified DB, ε=0	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
	modified DB + ε-floor 0.1	0.1669	0.2365	1.4683	0.3722	0.1750	0.4838
	modified DB + annealed floor	0.1424	0.2387	1.8622	0.3703	0.1664	0.5560
	ReStrike, top-k draw	0.0356	0.0325	0.0590	0.0292	0.1156	0.0544
	local search	1.9971	1.9971	1.9971	0.4951	1.9971	1.6967
	local search + ε-floor 0.1	0.8430	0.4294	0.3129	0.2093	0.8178	0.5225
	ε-floor 0.1, Huber residual	0.2440	0.2922	0.2323	0.3521	0.2684	0.2778
	ReStrike, constant fraction, r=1	0.0228†	0.0320	0.0226	0.0208†	0.0269†	0.0250
	ReStrike, constant fraction, r=2	0.0290	0.0236†	0.0403	0.0328	0.0280†	0.0307
	ReStrike, constant fraction, r=1, 32 fresh + 16 replayed	0.0506	0.0402	0.0378	0.1449	0.2188	0.0985
	annealed-floor recipe, 32 fresh	0.3849	0.3679	0.3093	0.3856	0.6680	0.4231
	ReStrike	0.0235	0.0258	0.0194	0.0277	0.1099	0.0412

Table 12: **Per-seed exact L_1 of the arms of Tables 3 and 17 on the study's five seeds**, 4M (the arms at 96 fresh per update at 6M fresh, ReStrike at 32 fresh + 16 replayed and the recipe at 32 fresh at 2M); a dagger marks a seed on which the arm's exact L_1 is below ReStrike's, so the paired counts of §5 — and Box 2's "all paired seeds but seven" — are read here (the seven floor, recipe or local-search pairs are the annealed ε floor, the floor with ReStrike's draw, the ε-floor at 96 fresh per update, the floor with clipped residuals, local search, the floor with local search and the recipe at 32 fresh, one v-structure seed each; the daggers of plain SubTB at 96 fresh per update and of the ReStrike variants are outside that clause).

Cell	β_r=0	β_r=0.25	β_r=0.5	β_r=1	β_r=2
v-structure	0.3141	0.0780	0.0488	0.0384	**0.0332**
Gaussian	0.7049	0.0574	0.1396	**0.0511**	0.0647

Table 13: **The draw's temperature.** ReStrike with the replay fraction held at 0.5 (the constant-fraction control of Table 3) and the softmax temperature β_r of the draw over log-reward varied; β_r=0 is the uniform draw. Seeds $\{7, 19, 101\}$, 1M, mean exact L_1; bold marks the lowest per cell. Pre-registered with Table 11.

Cell	Objective	Budget	seed 7	seed 19	seed 101	seed 367	seed 911	mean L_1	loss	collapsed
v-structure	SubTB	4M	1.2057	2.0000	0.0350	1.2056	1.9635	1.2820	0.0220	4/5
	TB	4M	1.5346	0.7634	1.9458	1.9230	1.5781	1.5490	0.1184	4/5
	DB	4M	0.0750	0.0286	0.0373	0.0307	0.0413	0.0426	0.0223	0/5
	SubTB	1M	1.7386	2.0000	1.5259	1.6710	2.0000	1.7871	0.0441	5/5
	TB	1M	1.5409	1.6273	1.9470	1.9590	1.9974	1.8143	0.0357	5/5
	DB	1M	0.8970	2.0000	1.1762	1.5980	0.6383	1.2619	0.0790	3/5
Gaussian	SubTB	4M	1.3496	0.1111	1.9971	1.9971	1.9971	1.4904	0.0945	4/5
	TB	4M	1.9971	1.9971	1.9971	1.9971	1.9971	1.9971	0.0838	5/5
	DB	4M	1.9971	1.9971	0.0640	0.0560	1.1459	1.0520	0.2916	3/5
	SubTB	1M	1.9971	1.9971	1.9971	1.9971	1.9971	1.9971	0.4234	5/5
	TB	1M	1.9971	1.9971	1.9971	1.9971	1.9971	1.9971	1.2055	5/5
	DB	1M	1.9971	1.9971	1.9971	1.9971	1.9971	1.9971	0.0316	5/5

Table 14: **TB and DB at ε=0 on the two collapsing cells**, pre-registered (`_preregister_tb_db.md`): plain SubTB's configuration with only the objective swapped, trajectory balance (a learned $\log Z$) and detailed balance (the length-one residuals of SubTB, the same $\log F$ head), the study's five seeds, optimizer and learning rates as plain SubTB's (listed in the registration), exact forward-DP L_1 per seed at 4M and at 1M; *loss* is the mean held-out on-policy loss of the arm's *own* objective and is not comparable across objectives; *collapsed* counts the seeds meeting (C2) of Def. 1 at the endpoint. The SubTB rows are the study's runs (Table 1).

generator, data seed, sample counts, seeds, budget and network. *Six nodes* (3,781,503 labeled DAGs, $129\times$ the study's space) keeps every exact read-out of the five-node cells: the enumeration is held as int8 and the forward DP, the predicate and the divergences are the ones of Table 1; the five-node path is bit-identical before and after the change. The generator plants the colliders $(0, 1, 2)$ and $(2, 3, 4)$ as at five nodes; the sixth node is marginally independent noise, which flattens the posterior (top-five mass $0.53 \to 0.37$) and removes the trap for every arm but local search (one seed, Table 15), so "the trap vanishes at six nodes" is a statement about that flatter posterior, not about size alone; the Gaussian chain runs $0 \to \cdots \to 5$. *Eight nodes* (7.8×10^{11} DAGs) cannot be enumerated, but with modular scores the sum over DAGs factorizes by inclusion–exclusion over sink sets (Tian & He, 2009), $O(3^d)$ from the per-node local scores under the same uniform prior: $\log Z$, the posterior of any DAG, every edge marginal (the same recursion with the parent sets of one node restricted) and the MAP (a max-plus dynamic programme over sink sets, Silander & Myllymäki, 2006, which needs no inclusion–exclusion) are exact; $\log Z$ and every marginal agree with the six-node enumeration to 10^{-8}, the MAP's score exactly and on the same DAG. The generator plants three colliders, $(0, 1, 2)$, $(2, 3, 4)$ and $(4, 5, 6)$, the eighth node independent; at the study's sample count the posterior identifies only the last: the planted edges carry marginals 0.26–0.82, the middle collider's reversed orientations up to 0.47, the MAP has six edges with that collider oriented $2 \to 3 \leftarrow 4$ and carries 0.03 of the posterior, and the expected edge count is 7.37; the Gaussian chain $0 \to \cdots \to 7$ gives a 13-edge MAP and 17.20 expected edges. The reward is $\exp(\text{score} - \text{score}_{\text{MAP}})$ as on every cell; a state's id is its adjacency key; no sampled L_1 curve exists during training and no mode count. The read-out draws 200k fresh on-policy trajectories per model at ε=0 (the study's held-out recipe), scores every distinct DAG exactly, and reports the edge-marginal L_1 over the 56 ordered pairs against the exact marginals (each term in $[0, 1]$: a marginal metric, weaker than the joint divergences of the enumerable rungs) and Def. 1 at its cut with S_{95} taken on the empirical pmf and $p^\star(S_{95})$ exact. That estimate is biased one way, as the registration states: S_{95} from N draws under-covers a diffuse policy, so $p^\star(S_{95})$ is biased downward (over-inclusion at the cut is possible, but under-coverage dominates once N is far below the effective support) and a "collapse" verdict for a diffuse arm certifies nothing, while a collapsed policy (a handful of DAGs) is resolved exactly. At six nodes, where the exact DP exists, the sampled read-out matched it on $50/50$ (C2) verdicts with $p^\star(S_{95})$ within 0.03, including the diffuse floor arms; at eight nodes the bias is visible on the Gaussian cell, so Table 15 prints the sampled mass at 200k and at a million draws instead of a count: ReStrike rises from 0.31 to 0.55, the annealed floor with uniform replay from 0.20 to 0.35, the ε=0.1 floor from 0.02 to 0.05 (at most 0.09 on any seed), and plain SubTB stays at 0 (its S_{95} is 32–43 DAGs, restricted L_1 0.02–0.13: the calibrated collapse of Def. 1 on four of the five seeds, the fifth failing (C3) by 0.03); on the v-structure no value moves by more than 0.024. The verdict of a diffuse eight-node arm therefore depends on κ and on the draw count (ReStrike's Gaussian verdict flips between 0.31 and 0.55), which is why Def. 1's invariance is claimed at five nodes only; what is certified at eight nodes is plain SubTB's collapse on both cells, the floor's and the recipe's draws covering only 0.02–0.09 and 0.24–0.45 of the Gaussian posterior at a million draws (lower bounds,

still rising with the draw count), and the calibration ordering on the marginals. On the v-structure the collapsed seeds hold 0.35–0.44 of the posterior (plain SubTB on three seeds, ReStrike on two); the MAP appears in none of 200k draws on one collapsed ReStrike seed and in 0.06 of them on the other, so replay does not miss a needle so much as keep a partial support, the case Cor. 2 allows and the lattice's Δ_{bg}=80 corners showed (Appendix C). The pre-flight deficit of Appendix B flags all four cells, the six-node v-structure included (its numbers are there). Both registrations' pre-worded outcome is the partial one, stated by cell; the eight-node registration's clause that a floor arm itself meets the predicate holds for the ε=0.1 floor on the Gaussian at both draw counts, with the caveat above. The composite arm (§5) is the floor-plus-ReStrike-draw arm of Table 3 run at both sizes under its own registration. At eight nodes the composite is the best arm on neither cell: 0.2812 and 3.3979, better than the recipe on 2 and 3 of 5 seeds and than ReStrike on 5 and 0, with 0 and 5 seeds meeting the sampled predicate. At six nodes it ends at 0.0608 and 0.4085 (exact L_1), better than the recipe on 3 and 3 of 5 seeds and than ReStrike on 1 and 0.

The two baselines a practitioner would run, at size. A fourth registration (`_preregister_bnsl_size_baselines.md`) adds PRT-full (Table 1's arm, its ε=0.01 online batch included) and the annealed-floor recipe on detailed balance (the recipe arm with the objective swapped) to both rungs. At eight nodes: on the v-structure neither freezes ($0/5$; sampled support mass 0.90–0.92 at 200k draws, 0.94 at a million), the DB recipe is the best arm (0.27, better than the SubTB recipe on $3/5$ seeds) and PRT-full ends at 0.57, between the recipe and ReStrike — the registration's rule 4(a) antecedent holds (a mean above the recipe's) but its pre-worded reading, that own-terminal replay of either draw rule freezes here, does not describe the run and is withdrawn: a 0.01 online floor discovers the support on this cell, which the heuristic of Appendix A under-predicts. On the Gaussian chain ReStrike remains best on the marginals against both (0.77 against 2.45 and 2.38, $5/5$ seeds each: rule 1) and on the sampled support's mass at a million draws (0.55 against 0.47 and 0.42), and the DB recipe calibrates better than the SubTB recipe there (2.38 against 4.07). At six nodes (exact L_1, Table 15) PRT-full ends at 0.1092 and 0.8686 and the DB recipe at 0.0601 and 0.5913 (v-structure, Gaussian), against ReStrike's 0.0488 and 0.0737 and the SubTB recipe's 0.0558 and 0.4574; ReStrike is lower than PRT-full on 4 and 5 of 5 seeds and than the DB recipe on 4 and 5; no seed of either arm collapses on either cell.

The molecular world: what we withdraw, and what replaces it. An earlier version of this appendix reported 120 runs (8 arms $\times$ 3 landscapes $\times$ 5 seeds) as a coverage table. **We withdraw those numbers.** An implementation defect made the replay loss unsound in this world and in this world only: reconstructing a forward trajectory from a backward walk copies the backward actions into the forward slots, which is valid where a forward and a backward action index the same object — we verified it empirically on HyperGrid, BitSequence, PHO4 and BNSL — but not for graph actions, where a forward ADD-NODE carries the fragment *class* and its backward counterpart has no such field. Every replayed trajectory was therefore scored as though fragment 0 had been added, whatever molecule it held. This touched the seven arms that replay and left the SubTB baseline, which has no buffer, untouched — so the affected comparison was precisely the one the table was for. Repaired (the added fragment is recoverable from the states, where it was never lost), the same configuration finds 2810 modes where the defective one found 126: the old table was wrong by more than an order of magnitude, not by a margin, and no part of it is salvageable.

What replaces it is narrower and sound. Table 16 reports 4 arms $\times$ 5 seeds at 4M on each of the three landscapes of the real proxy, trained with the repair. It carries the claim the old table could not: the *loss* read-out, which needs no enumeration and is computed by the recipe the enumerable worlds use. The full 8-arm $\times$ 3-landscape grid is not claimed here.

E.2 The five-node controls, in full

The numbers behind §5.3, arm by arm (Table 3; per seed, Table 12; ten seeds, Table 11; TB and DB, Table 14). A published explorer-teacher (Kim et al., 2025) also repairs the collapse but spends outside our accounting (§7(vii)) and itself *replays* the teacher's proposals, so it isolates neither route: it ends at an exact L_1 of 0.0832 and 0.5227 against plain SubTB's 1.2820 and 1.4904 (Table 1). Under the ε-uniform floor no seed of either cell collapses at ε=0.1, and at ε=0.01 none on the v-structure and 2 of 5 on the Gaussian; the floor's exact L_1 is 0.0698 and 0.0819 on the v-structure and 0.3229 and 0.8784 on the Gaussian (ε=0.1 and 0.01), against ReStrike's 0.0365 and 0.0412, ReStrike winning

Cell	Arm	d	seed 7	seed 19	seed 101	seed 367	seed 911	mean	loss	collapsed / $p^\star(S_{95})$
v-structure	plain SubTB	6	0.0393	0.1100	0.0492	0.0722	0.0267	0.0595	0.0617	0/5
	ε-floor 0.1	6	0.0890	0.0646	0.0662	0.0872	0.0917	0.0797	0.0227	0/5
	annealed-floor recipe	6	0.0572	0.0600	0.0515	0.0567	0.0536	0.0558	0.0117	0/5
	annealed-floor recipe (DB)	6	0.0439	0.0740	0.0431	0.0661	0.0733	0.0601	0.0042	0/5
	PRT-full	6	0.1123	0.1172	0.1086	0.1020	0.1060	0.1092	0.0401	0/5
	floor + ReStrike draw	6	0.0499	0.0510	0.0425	0.0916	0.0692	0.0608	0.0095	0/5
	ReStrike	6	0.0296	0.0408	0.0324	0.0312	0.1100	0.0488	0.0500	0/5
	ReStrike, constant fraction	6	0.0307	0.0265	0.0249	0.0349	0.0290	**0.0292**	0.0740	0/5
	local search	6	0.0328	0.0401	1.4992	0.0507	0.3292	0.3904	0.0470	1/5
	ReStrike, constant fraction, $r{=}1$	6	0.0319	0.0262	0.0840	0.0486	0.0493	0.0480	0.0458	0/5
	plain SubTB	8	3.0640	2.9986	1.1183	1.1140	3.1048	2.2799	0.1451	0.55→0.56
	ε-floor 0.1	8	1.6812	0.7267	2.3659	1.6388	0.5489	1.3923	0.0903	0.75→0.76
	annealed-floor recipe	8	0.1914	0.2440	0.4434	0.1927	0.3403	0.2824	0.0457	0.92→0.94
	annealed-floor recipe (DB)	8	0.2396	0.1526	0.3865	0.2524	0.2982	**0.2659**	0.0135	0.92→0.94
	PRT-full	8	0.6676	0.4568	0.8194	0.3184	0.5982	0.5721	1.6976	0.91→0.94
	floor + ReStrike draw	8	0.2810	0.3032	0.1762	0.3085	0.3370	0.2812	0.0447	0.92→0.94
	ReStrike	8	1.6804	1.5054	1.6327	3.1248	3.1098	2.2106	0.2300	0.57→0.58
	ReStrike, constant fraction	8	0.1936	0.3520	1.4883	0.1343	3.0821	1.0501	0.2861	0.78→0.80
	local search	8	3.1665	0.1360	3.1694	0.8365	4.6488	2.3914	0.2391	0.51→0.52
	ReStrike, constant fraction, $r{=}1$	8	0.2484	0.1055	0.2911	1.5954	1.5287	0.7538	0.4868	0.82→0.84
Gaussian	plain SubTB	6	1.9996	1.9996	1.9996	1.9996	1.9996	1.9996	0.1633	5/5
	ε-floor 0.1	6	0.6211	0.7161	1.0664	0.9704	0.9809	0.8710	1.4171	0/5
	annealed-floor recipe	6	0.5629	0.4615	0.4113	0.4210	0.4302	0.4574	0.2268	0/5
	annealed-floor recipe (DB)	6	0.4799	0.5599	0.4211	0.8068	0.6885	0.5913	0.1614	0/5
	PRT-full	6	0.9406	0.7900	0.6865	1.2915	0.6344	0.8686	2.8709	0/5
	floor + ReStrike draw	6	0.3580	0.3673	0.3434	0.5131	0.4609	0.4085	0.3884	0/5
	ReStrike	6	0.0642	0.0520	0.1401	0.0502	0.0619	0.0737	0.5337	0/5
	ReStrike, constant fraction	6	0.0756	0.0589	0.0623	0.0650	0.0460	**0.0616**	1.9592	0/5
	local search	6	1.9996	1.9996	1.9996	1.9996	1.9996	1.9996	0.1787	5/5
	ReStrike, constant fraction, $r{=}1$	6	0.0555	0.0937	0.2023	0.0642	0.0512	0.0934	0.6516	0/5
	plain SubTB	8	11.5381	11.5652	11.6687	11.6065	11.5954	11.5948	0.4624	0.00→0.00
	ε-floor 0.1	8	11.1165	6.3639	8.3161	7.3154	9.5125	8.5249	5.2233	0.02→0.05
	annealed-floor recipe	8	5.3174	2.3176	5.4380	2.2670	5.0086	4.0697	6.0880	0.20→0.35
	annealed-floor recipe (DB)	8	1.7242	2.2995	2.9968	2.5891	2.2835	2.3786	0.2587	0.26→0.42
	PRT-full	8	3.0151	2.1056	2.5979	2.2436	2.2870	2.4499	1.4892	0.24→0.47
	floor + ReStrike draw	8	3.8992	2.6784	3.1123	3.6851	3.6146	3.3979	5.8729	0.23→0.38
	ReStrike	8	0.7643	0.8105	0.6624	0.8853	0.7059	0.7657	0.8644	0.31→0.55
	ReStrike, constant fraction	8	1.0865	0.6933	0.7431	0.9711	0.6892	0.8366	0.7104	0.31→0.55
	local search	8	11.5647	11.5484	11.5572	11.5469	11.5548	11.5544	2.2254	0.00→0.00
	ReStrike, constant fraction, $r{=}1$	8	0.9591	0.6310	0.6456	0.6310	0.6665	**0.7067**	50.3347	0.31→0.55

Table 15: **Six and eight nodes**, pre-registered: the two collapsing posteriors on six nodes (exact forward-DP L_1 over all $3{,}781{,}503$ DAGs) and on eight (7.8×10^{11} DAGs; the edge-marginal L_1 over the 56 ordered pairs against the exact marginals, each term in $[0, 1]$, from 200k fresh on-policy draws per model), the study's five seeds, 4M; the eight arms of Table 4. *Loss* is the mean held-out on-policy SubTB loss. The last column counts the seeds meeting (C2) of Def. 1 at six nodes, where it is exact; at eight nodes it gives the exact posterior mass of the sampled S_{95} at 200k→a million draws, a lower bound for a diffuse policy (plain SubTB's 32–43-DAG support is exact). Bold marks the lowest mean within a cell and size.

Landscape	Arm	Modes (median)	range	Held-out loss	$<$ ours
Deceptive	SubTB (no replay)	20	16–21	**7.68**	4/5
	ReStrike (ours)	2810	1745–6820	13.87	—
	PRT-rule	1959	1018–2263	16.82	2/5
	ReStrike, uniform draw	1261	1011–2598	15.07	3/5
Peaked	SubTB (no replay)	22	11–31	18.81	3/5
	ReStrike (ours)	1736	847–1894	18.63	—
	PRT-rule	1676	1292–2853	**14.77**	3/5
	ReStrike, uniform draw	1255	526–1573	17.48	3/5
Smooth	SubTB (no replay)	11	7–18	**4.01**	4/5
	ReStrike (ours)	658	317–951	21.33	—
	PRT-rule	1009	595–1331	11.79	4/5
	ReStrike, uniform draw	742	294–847	22.34	2/5

Table 16: Molecular calibration cells, post-repair: the three landscapes. Tanimoto-distinct modes above the decoy-aware gate, and the held-out on-policy SubTB loss (mean over 1562 batches, the enumerable worlds' recipe). "$<$ ours" counts seeds on which the arm scores a *lower* loss than ReStrike. On Deceptive plain SubTB wins the loss on $4/5$ seeds while finding $140\times$ fewer modes — the silence of Obs. 1, on a non-enumerable reward; Smooth repeats it, Peaked ranks PRT-rule lowest. Held-out evaluation costs a median of 105 proxy calls on Deceptive, is excluded from the budget accounting of §2, and is reported separately. **No coverage claim is made from these cells**: n=5, and the arms touch different numbers of distinct molecules — ours by more than an order of magnitude over plain SubTB — so the oracle spend is not matched. The uniform-draw arm is a prioritization deconfound, not a baseline from the literature. Estimator-based columns (L_1 restricted to a run's own support, and its mass on high-reward terminals other arms reached) are omitted: their estimator is gated against forward-sampled ground truth and the gate rejected 14 of the 20 Deceptive runs at both fidelities tried.

every paired seed at both ε; on ten seeds per cell the ε=0.1 comparison is $10/10$ on both cells (paired t: $p = 0.0002$, sign test $p = 0.002$) and plain SubTB is collapsed on 17 of 20. TB collapses on $4/5$ and $5/5$ seeds at 4M and DB on $0/5$ and $3/5$ ($3/5$ and $5/5$ at 1M), DB escaping on the v-structure by 4M (0.0426). The annealed floor with a uniform replay draw ends at 0.0813 and 0.3260, worse than ReStrike on $5/5$ seeds of both cells, and placed on top of ReStrike's own draw it ends at 0.0684 and 0.3089, worse on $4/5$ and $5/5$ ($10/10$, $10/10$ and $9/10$, $10/10$ on ten seeds, Table 11). Under the ε=0.1 floor the held-out on-policy loss is lower than ReStrike's on 9 of 10 v-structure and 4 of 10 Gaussian seeds while the exact L_1 is worse on every seed; on the Gaussian collapsed plain SubTB carries the lowest loss, 0.094 against ReStrike's 0.332. The constant replay fraction of 0.5 ends at 0.0337 and 0.0239, better than the annealed arm on both cells (Tables 3 and 13); at size the two tie on the Gaussian chain, and on the v-structure the constant fraction improves the annealed arm without repairing it (1.05, one seed frozen; Table 4).

E.3 Budget, batch, objective, draw and local search: nine further controls

Five controls registered on 19 September 2026 and four on the 20th, each before its runs and read out under its committed rules (`_preregister_batch_control.md`, `_preregister_floor_budget.md`, `_preregister_modified_db.md`, `_preregister_draw_controls.md` and `_preregister_local_search.md` in the supplement; the sentences of §5.3 follow the ones those files pre-worded). Every BNSL arm is evaluated by the study's exact forward-DP path and Def. 1: the batch, objective and draw arms paired by seed against ReStrike's study run (Table 17), the budget arms read along the study's own runs continued (Table 18, Fig. 6); the grid arms are read at the study's checkpoints against the reported arm's means (Table 19).

The floor's gap against budget. The study's 4M runs of plain SubTB, the two fixed floors and the constant-fraction ReStrike arm were continued to 16M with the weights archived every 1M; the continued runs reproduce the study's 4M values to the last digit (30 checked, 0 failed). The ε=0.1 floor goes from 0.070 to 0.048 on the v-structure and from 0.323 to 0.344 on the Gaussian cell, the ε=0.01 floor from 0.082 to 0.042 and from 0.878 to 0.445; from 10M on the Gaussian ε=0.1 floor stays between 0.27 and 0.43, so four times the budget removes at most half of a floor's error (rule 1 of the registration: structural). ReStrike ends at 0.020 and 0.020 and does not degrade (rule 5);

the same arm at r=1, continued the same way, collapses on no seed at any budget and ends at 0.014 and 0.027 (Table 18). Of the eight chassis seeds meeting (C2) at 4M, four per cell, seven leave it by 14M — on the Gaussian cell all four (seeds 101 and 367 by 5M, seed 7 by 6M, seed 911 by 10M; the cell mean falls from 1.490 to 0.043, first below 0.05 at 14M), on the v-structure three (seed 367 by 5M, seed 7 by 10M, seed 911 by 14M, still at 0.749 at 16M) — while seed 19 stays at the total-collapse value 2.000 through 16M; no seed re-enters the predicate once out (rule 4: metastable, and measured). ReStrike is at 0.039 and 0.037 at 2M.

Gradient samples. ReStrike's replayed trajectories add gradient samples to each update (§2). The control keeps the study's 62,500 updates and draws 96 fresh trajectories per update (6M fresh) for plain SubTB and the ε=0.1 floor, and runs ReStrike at 32 fresh + 16 replayed (48 gradient samples, 2M fresh). Plain SubTB at 96 meets (C2) on $0/5$ v-structure seeds — three of the five end on a partial support, $p^\star(S)$ of 0.86, 0.91 and 0.66 (exact L_1 0.19, 0.10 and 0.61), against about 0.95 for a calibrated run — and on $5/5$ Gaussian seeds, four of them at 1.997: rule 2, the confound is real on the v-structure and closed on the Gaussian cell, consistent with the escape budgets above (6M fresh reaches the v-structure escape window; the Gaussian escapes come at 5M–10M at batch 64 and not within 6M fresh at batch 96). The floor at 96 ends at 0.062 and 0.369 against 0.070 and 0.323 at 64 (rule 3: the floor's gap is not a gradient-sample effect). ReStrike at 48 samples collapses on no seed, repairs the Gaussian cell (0.0432 against the floor's 0.323) and on the v-structure ends 0.0008 above the floor (0.0705 against 0.0698), a tie the registered sense of "repairs" does not admit (rule 4). The tie is the temperature's, not the protocol's: at r=1 the same 48 samples end at 0.0382 against the floor's 0.0698 on the v-structure and at 0.0985 on the Gaussian cell, below the floor on both (Table 17).

The modified detailed balance. The objective of Deleu et al. (2022) — torchgfn's `ModifiedDBGFlowNet`, forward and backward policies only, no state flow — replaces SubTB on the unchanged chassis (arm suffix `_objMDB`), at ε=0, under the ε=0.1 floor and under the recipe's annealed floor, on the two collapsing cells. At ε=0 it collapses on $10/10$ seeds, deeper than SubTB: v-structure mean 1.810 ($p^\star(S)$ from 0.03 to 0.17), Gaussian 2.000 on every seed with $p^\star(S) = 0$, the whole mass on DAGs of zero posterior mass (rule 1). Under the fixed floor it ends at 0.071 and 0.484, under the annealed floor at 0.105 and 0.556, each with one Gaussian seed (101) still meeting (C2), at 1.468 and 1.862; ReStrike is lower on $10/10$ paired seeds against each arm (rule 4: neither floor twin is within 10% of ReStrike). Prop. 1 does not reach this objective (§7(v)); the collapse is a measurement.

The draw. A top-k draw replays the n highest-reward buffer entries instead of the softmax draw. On the v-structure it is lower than the reported arm on seeds 7, 367 and 911, higher on seed 19, and frozen on seed 101 on a partial support ($p^\star(S) = 0.66$, exact L_1 0.60, within the support 0.0085), the cell mean 0.140 against 0.0365 (rule 2); on the Gaussian cell it is worse on $5/5$ seeds, 0.054 against 0.041 (rule 3): the spread of the draw matters, and top-k is the failure mode it avoids. On the tiered grid (the Deceptive cell of Table 2, 1M) a uniform draw from the same buffer finds 14,837 modes, 0.89 of the reported arm's 16,635 (rule 5, "partial"), and a top-k draw 17,447, 1.05: the buffer with prioritized eviction carries most of the coverage headline and the softmax draw the last tenth; on the BNSL cells the same top-k draw is the one that freezes.

Local search, and the floor with clipped residuals. Local-search GFlowNet (Kim et al., 2024b) runs on the unchanged chassis (arm suffix `_lsgfn`: torchgfn's `LocalSearchSampler`; eight fresh trajectories and seven rounds of destroy-and-reconstruct per update, each round destroying half of a trajectory's edge actions with the learned P_B and rebuilding with P_F, a candidate kept as the next round's parent when its reward is at least the parent's, every candidate trained on whether kept or not), so an update scores the study's 64 trajectories and a run spends its 4M evaluations. At ε=0 it collapses on $4/5$ Gaussian seeds at 1.997 ($p^\star(S) = 0.0014$, the chassis's total-collapse value) and ends on a partial support on the fifth (0.495, $p^\star(S) = 0.72$); on the v-structure three seeds end on a partial support of $p^\star(S) = 0.66$, calibrated within it (restricted L_1 0.014–0.028; exact L_1 0.618 each), one collapses (1.804) and one calibrates (0.028, below the reported arm on that seed); means 1.697 and 0.737 (rule 1 on the Gaussian cell, rule 2 on the v-structure). Under the ε=0.1 floor (the mixture on every draw of the search) no seed collapses and it ends at 0.058 and 0.522, above the floor alone on $3/5$ Gaussian seeds (rule 2 on both cells); ReStrike-constant is lower on $4/5$ and $5/5$

paired seeds against either arm. The floor with its residual gradient clipped at one nat (`_huber_1`: the squared residual becomes Huber for the training loss, the held-out loss keeps the square) ends at 0.060 and 0.278, removing 0.14 of the floor's error on each cell, lower than the floor on $3/5$ seeds of each, and neither within 10% of ReStrike-constant nor lower on four of five paired seeds (rule 4 does not fire: the gap is not residual scaling). The constant-fraction arm at 4M ends at 0.021 and 0.025 for $r{=}1$, 0.023 and 0.031 for $r{=}2$, against 0.034 and 0.024 at $r{=}0.5$: no temperature is lowest on both cells (rule 6), so $r{=}0.5$ stays the constant-fraction arm the text compares against and the temperature is a per-cell choice; the 1M sweep's ordering (Table 13) does not hold at 4M. At size (the registration's stage 2; Tables 4, 15) local search collapses with the chassis: on the six-node Gaussian cell on $5/5$ seeds at 1.9996 with $p^\star(S) = 0$, on the eight-node chain on $5/5$ at a marginal L_1 of 11.55 (the chassis 11.59), and on the eight-node v-structure on $3/5$ (mean 2.39, within 10% of the reported arm's 2.21 and above the constant-fraction arm's 1.05 on $4/5$ seeds); on the six-node v-structure, where the flatter posterior removes the trap for every other arm, it collapses on one seed (1.4992) and ends on a partial support on another (0.3292), mean 0.3904 against ReStrike's 0.0488 (rule 7: rule 1 on three cells, rule 2 on the fourth).

Cell	Arm	L_1	JSD	$\mathrm{KL}(p_\theta \Vert p^\star)$	$\mathrm{KL}(p^\star \Vert p_\theta)$	loss	collapsed	ReStrike lower
v-structure	plain SubTB (study)	1.2820	0.3788	9.0625	19.6004	0.022	4/5	—
	plain SubTB, 96 fresh per update	0.1918	0.0338	0.1040	1.7099	0.018	0/5	4/5
	ε-floor 0.1, 96 fresh per update	0.0624	0.0011	0.0043	0.0044	0.014	0/5	4/5
	ReStrike, constant fraction, 32 fresh + 16 replayed	0.0705	0.0018	0.0124	0.0073	0.100	0/5	4/5
	modified DB, ε=0	1.8100	0.5454	2.6566	55.2863	0.002	5/5	5/5
	modified DB + ε-floor 0.1	0.0712	0.0013	0.0051	0.0051	0.023	0/5	5/5
	modified DB + annealed floor	0.1051	0.0031	0.0125	0.0126	0.030	0/5	5/5
	ReStrike, top-k draw	0.1404	0.0241	0.0766	1.0676	0.034	0/5	2/5
	local search	0.7373	0.1787	0.6904	8.9143	0.036	1/5	4/5
	local search + ε-floor 0.1	0.0582	0.0008	0.0033	0.0034	0.012	0/5	4/5
	ε-floor 0.1, Huber residual	0.0598	0.0008	**0.0032**	0.0032	0.013	0/5	4/5
	ReStrike, constant fraction, $r{=}1$	**0.0211**	**0.0003**	0.0038	**0.0011**	0.025	0/5	1/5
	ReStrike, constant fraction, $r{=}2$	0.0227	0.0004	0.0041	0.0013	0.034	0/5	1/5
	ReStrike, constant fraction, $r{=}1$, 32 fresh + 16 replayed	0.0382	0.0009	0.0098	0.0040	0.087	0/5	2/5
	annealed-floor recipe, 32 fresh	0.0921	0.0020	0.0079	0.0080	0.027	0/5	4/5
	ReStrike (study)	0.0365	0.0004	0.0039	0.0018	0.048	0/5	—
Gaussian	plain SubTB (study)	1.4904	0.4798	4.1619	12.8628	0.094	4/5	—
	plain SubTB, 96 fresh per update	1.8229	0.6006	5.4055	16.7939	0.455	5/5	5/5
	ε-floor 0.1, 96 fresh per update	0.3688	0.0323	0.1425	0.1416	0.389	0/5	5/5
	ReStrike, constant fraction, 32 fresh + 16 replayed	0.0432	0.0006	0.0055	0.0024	2.341	0/5	4/5
	modified DB, ε=0	2.0000	0.6931	68.9862	63.6770	0.002	5/5	5/5
	modified DB + ε-floor 0.1	0.4838	0.0886	0.4120	1.8547	6.073	1/5	5/5
	modified DB + annealed floor	0.5560	0.1319	0.9015	3.0889	0.198	1/5	5/5
	ReStrike, top-k draw	0.0544	0.0041	0.0151	0.0716	1.126	0/5	5/5
	local search	1.6967	0.5691	5.2970	14.3366	0.137	4/5	5/5
	local search + ε-floor 0.1	0.5225	0.0738	0.3634	0.3636	2.962	0/5	5/5
	ε-floor 0.1, Huber residual	0.2778	0.0159	0.0660	0.0657	0.163	0/5	5/5
	ReStrike, constant fraction, $r{=}1$	**0.0250**	**0.0004**	**0.0026**	**0.0015**	0.460	0/5	2/5
	ReStrike, constant fraction, $r{=}2$	0.0307	0.0007	0.0040	0.0035	0.939	0/5	3/5
	ReStrike, constant fraction, $r{=}1$, 32 fresh + 16 replayed	0.0985	0.0040	0.0173	0.0180	2.273	0/5	5/5
	annealed-floor recipe, 32 fresh	0.4231	0.0406	0.2050	0.1601	0.445	0/5	5/5
	ReStrike (study)	0.0412	0.0007	0.0047	0.0029	0.332	0/5	—

Table 17: **The batch, objective, draw and local-search controls** beside the study's chassis and ReStrike rows: two collapsing cells, five seeds, the study's $62{,}500$ updates (4M fresh trajectories; 6M for the arms at 96 fresh per update, 2M for ReStrike at 32 fresh + 16 replayed; the local-search arms score 64 trajectories per update, eight of them fresh), exact forward-DP metrics; columns as in Table 3, plus the seeds of the pair on which ReStrike's exact L_1 is the lower; per seed: Table 12.

F The optional Lightning fork

The fork is the second element in the method's name and the one we do not claim. It is included because it ships in the released code and two of the eight arms exercise it; this appendix says exactly what it does, so neither the arm labels nor the name *ReStrike* is left undefined, and it is not the mechanism: no cell's calibration difference against the replay-only arm survives multiplicity correction (the verdict below). The sampler is `src/gfn/samplers_lightning.py` in the supplement.

The two budget exceptions of §2. Two arms spend outside the fresh-trajectory and reward-evaluation accounting (the two of §7(vii)): the discovery variant (RS-DISCO) draws trees outside that count, disclosed below, and the explorer-teacher baseline draws one uncounted rollout every second round for its own buffer.

Cell	Arm / seed	1M	2M	4M	6M	8M	10M	12M	14M	16M
v-structure	plain SubTB	1.787 (5)	1.478 (5)	1.282 (4)	1.104 (3)	0.972 (3)	0.811 (2)	0.774 (2)	0.626 (1)	0.567 (1)
	ε-floor 0.1	0.130 (0)	0.090 (0)	0.070 (0)	0.062 (0)	0.053 (0)	0.043 (0)	0.044 (0)	0.044 (0)	0.048 (0)
	ε-floor 0.01	0.192 (0)	0.089 (0)	0.082 (0)	0.060 (0)	0.045 (0)	0.056 (0)	0.063 (0)	0.041 (0)	0.042 (0)
	ReStrike, constant fraction	0.054 (0)	0.039 (0)	0.034 (0)	0.021 (0)	0.016 (0)	0.017 (0)	0.026 (0)	0.017 (0)	0.020 (0)
	ReStrike, constant fraction, r=1	0.038 (0)	0.028 (0)	0.021 (0)	0.026 (0)	0.021 (0)	0.012 (0)	0.016 (0)	0.019 (0)	0.014 (0)
	plain SubTB, seed 7	1.739*	1.211*	1.206*	1.206*	1.187*	0.399	0.229	0.186	0.037
	plain SubTB, seed 19	2.000*	2.000*	2.000*	2.000*	2.000*	2.000*	2.000*	2.000*	2.000*
	plain SubTB, seed 101	1.526*	1.008*	0.035	0.024	0.020	0.025	0.018	0.032	0.017
	plain SubTB, seed 367	1.671*	1.206*	1.206*	0.330	0.050	0.024	0.018	0.043	0.033
	plain SubTB, seed 911	2.000*	1.963*	1.964*	1.959*	1.604*	1.604*	1.604*	0.867	0.749
Gaussian	plain SubTB	1.997 (5)	1.997 (5)	1.490 (4)	0.656 (1)	0.506 (1)	0.162 (0)	0.101 (0)	0.041 (0)	0.043 (0)
	ε-floor 0.1	0.968 (0)	0.706 (1)	0.323 (0)	0.574 (1)	0.515 (0)	0.289 (0)	0.312 (0)	0.323 (0)	0.344 (0)
	ε-floor 0.01	1.624 (5)	1.240 (3)	0.878 (2)	0.614 (0)	0.518 (0)	0.475 (0)	0.573 (0)	0.439 (0)	0.445 (0)
	ReStrike, constant fraction	0.114 (0)	0.037 (0)	0.024 (0)	0.097 (0)	0.031 (0)	0.021 (0)	0.017 (0)	0.017 (0)	0.020 (0)
	ReStrike, constant fraction, r=1	0.070 (0)	0.082 (0)	0.025 (0)	0.109 (0)	0.021 (0)	0.020 (0)	0.018 (0)	0.016 (0)	0.027 (0)
	plain SubTB, seed 7	1.997*	1.997*	1.350*	0.086	0.055	0.029	0.055	0.037	0.020
	plain SubTB, seed 19	1.997*	1.997*	0.111	0.063	0.042	0.035	0.082	0.032	0.059
	plain SubTB, seed 101	1.997*	1.997*	1.997*	0.566	0.290	0.279	0.056	0.031	0.034
	plain SubTB, seed 367	1.997*	1.997*	1.997*	0.569	0.149	0.149	0.033	0.044	0.058
	plain SubTB, seed 911	1.997*	1.997*	1.997*	1.997*	1.997*	0.317	0.279	0.060	0.042

Table 18: **Exact L_1 against the fresh-trajectory budget, 1M to 16M**: five-seed means with the seeds meeting (C2) in parentheses, and plain SubTB per seed (a star marks (C2)); the 16M runs are the study's 4M runs continued, the weights archived every 1M.

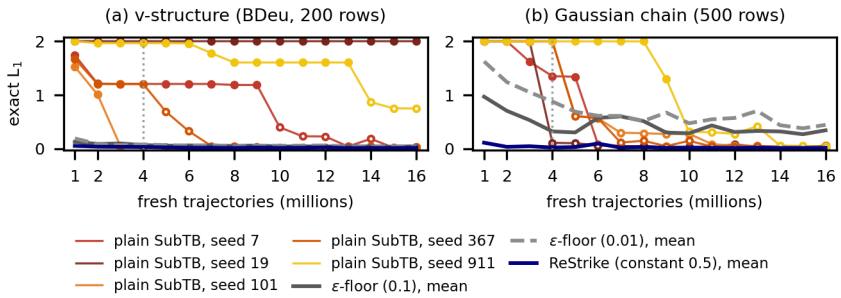

Figure 6: **The same curves** (Table 18): plain SubTB per seed (a filled marker meets (C2), an open one does not), the two fixed floors and the constant-fraction ReStrike arm as five-seed means; the dotted line is the study's 4M budget.

Arm	modes at 800K	modes at 1M	ratio at 800K	ratio at 1M
plain SubTB (study)	118.4	119.2	—	—
ReStrike (study)	14,199.8	16,634.8	1.000	1.000
ReStrike, uniform draw	11,315.4 $\pm$ 368	14,837.0 $\pm$ 263	0.797	0.892
ReStrike, top-k draw	15,450.8 $\pm$ 90	17,446.8 $\pm$ 68	1.088	1.049

Table 19: **The draw on the tiered grid** (the Deceptive cell of Table 2): modes discovered by ReStrike's buffer under a uniform and a top-k draw, five-seed mean $\pm$ std at the 800K and 1M checkpoints, and the ratio to the reported arm's study mean.

The tree. Given a seed trajectory, the fork state is the step whose one-step detailed-balance residual is largest, chosen among the seed's non-terminating steps. From it, one probe is rolled per non-exit action — the seed itself serves as the probe for the action it already took, and the rest are forced-action rollouts sharing the seed's prefix by reference — and the remaining sampling bundle is divided among the branches in proportion to $\exp(\log R)$ of their probes, with a floor of one trajectory per branch and largest-remainder rounding. Each branch then recurses with its own sub-bundle, choosing a new fork state strictly deeper than its parent's. When a bundle is smaller than the branching factor the fork stops there and ordinary policy rollouts complete the remainder. The allocation is detached, so it steers sampling without contributing gradient, and the leaves are assembled into one padded batch. Where that batch goes is what separates the two arms.

The two arms. The *warm-up* arm runs the tree with no external seed: each iteration builds one tree per first action, forbids forking at the root so the trees stay disjoint at depth one, and *on the enumerable worlds* scores every tree with the same SubTB loss as any other batch, taking a single optimizer step over all of them. It stops once fresh trajectories reach 1% of the budget on the grid and 5% on the DAG and PHO4 cells, whose warm-up step packs more trees, after which training is exactly the reported arm. There the probes are themselves leaves, so a tree costs exactly its bundle and this arm spends what every other arm spends. **The molecular implementation is a different arm.** It is a separate sampler (`src/molecular_body.py`), and there the tree's leaves go to the replay buffer and nowhere else — never the loss, never `traj_count` — exactly like the discovery arm's, so they are extra reward evaluations we do not charge (§2, §7(vii)). That, and not a change of currency, is why this arm was the larger oracle spender in the withdrawn molecular table (Appendix E); it is not among the arms retained there. The *discovery* arm instead fires a tree occasionally throughout training, under a controller keeping the firing period short while a firing yields a terminal the run has not seen and backing it off geometrically when it does not. Its leaves enter the replay buffer and nothing else — never the loss, the metrics or the trajectory count — so replay is their only path to the policy. **Those trees are extra reward evaluations outside the budget accounting**, which is why the arm is reported but never quoted (§7).

The auxiliary KL, and why it is off. Each split can also emit a loss pulling the forward policy at the fork state toward the reward-weighted distribution over its branches. In the shipped arms it is disabled outright, and where it is enabled its weight decays linearly to zero across the warm-up window. Three independent checks agree it should go: paired against its own no-KL twin it is inert ($|\Delta \mathrm{L}_1| \sim 10^{-4}$, $p > 0.4$); its Monte-Carlo target is self-referential and sits well off the true flow, so persisting it *raises* the asymptotic error floor; and on the GMM cell, where the error is diffuse rather than concentrated at a few states, correcting fork states changes nothing. The latter two are why we did not try to repair it.

Verdict. On calibration the fork does nothing that survives correction. Paired against the replay-only arm on the metric of Table 9 (fresh on six cells, exact on PHO4) for each of the nine enumerable cells, 7 are indistinguishable uncorrected; of the other two, one is a gain and one a *loss*, and at the Bonferroni level this paper applies ($\alpha = 0.0056$) neither survives — the smallest p over all nine cells is 0.044. On the enumerable worlds its mode counts sit at the cap or below the reported arm's; the one axis it leads there is the raw distinct-state count, which Appendix E shows is not a ranking metric at all and which reverses sign on the GMM cell. Its one real coverage result was the molecular world, where all three landscape winners were fork arms — but that table is withdrawn (Appendix E), the fork arms are not among the arms of the cells we retain, and they were in any case the two largest oracle spenders there. We therefore report the fork as an exploration ablation rather than a component: it may buy real coverage in a non-enumerable space, and this evidence cannot separate that from the extra spend.

G Anatomy, falsification and initialization counterfactuals

Fig. 1a gives the headline of the collapse — calibration pinned while the loss looks converged — on the two runs it compares. This appendix carries the mechanism behind it, which needs four panels rather than one. That exits die is measured: exit probability is below $5 \cdot 10^{-3}$ at every depth below the forced dead end at step 25 and below $2 \cdot 10^{-4}$ at step 200, where the support stops growing appreciably (Fig. 7). Plain off-policy descent from the collapsed parameters is barrier-free: the held-

out full-support loss (its own held-out batch; the detector comparison below scores another) falls monotonically $828.8 \to 78.8 \to 22.0 \to 4.2$, and the calibration does not follow it down (§3, Fig. 8). The escapes of §3 bracket the trap from both ends. The deepest *recovery* we see was calibrated-collapsed at 1M — meeting both verdict clauses of Def. 1, its operating support carrying 0.00137 of the posterior at a restricted divergence of 0.0242 — and still ends at $L_1 = 0.111$ at 4M; the deepest collapse does not leave at all, seed 19 sitting at $p^\star(S_t) = 7 \cdot 10^{-18}$. On the twin-controlled repair of §5 we do not certify that exit-rate recovery precedes the buffer's first mode: the instrumented probe logs the model's mode *mass*, not the buffer's contents.

The gap is not a detector, and we checked. If it were, the repaired twin would show a small gap; it shows a *larger* one — gradient ratio $1208\times$ against the collapsed run's $1030\times$, full-support loss 1828 against 890 — while on a cell the law does not flag the gradient ratio is $97\times$. The loss *ratio* does order the two correctly, and we state it rather than omit it, but far too weakly to use: $5.4{\cdot}10^4$ collapsed against $4.7{\cdot}10^4$ repaired is a factor of 1.15, against $4.2{\cdot}10^2$ on the safe cell — two orders of magnitude between *cells* against 1.15 between a broken model and a fixed one. It tracks the reward scale, not the failure, and the uniform detailed-balance residual runs the same way (748 against 393). One statistic separates them usefully — the residual over transitions the model routes no flow through (2798 against 886) — but each model is scored on its *own* complement, so there is no threshold, and locating those states means enumerating the space, which the setting needing a detector denies.

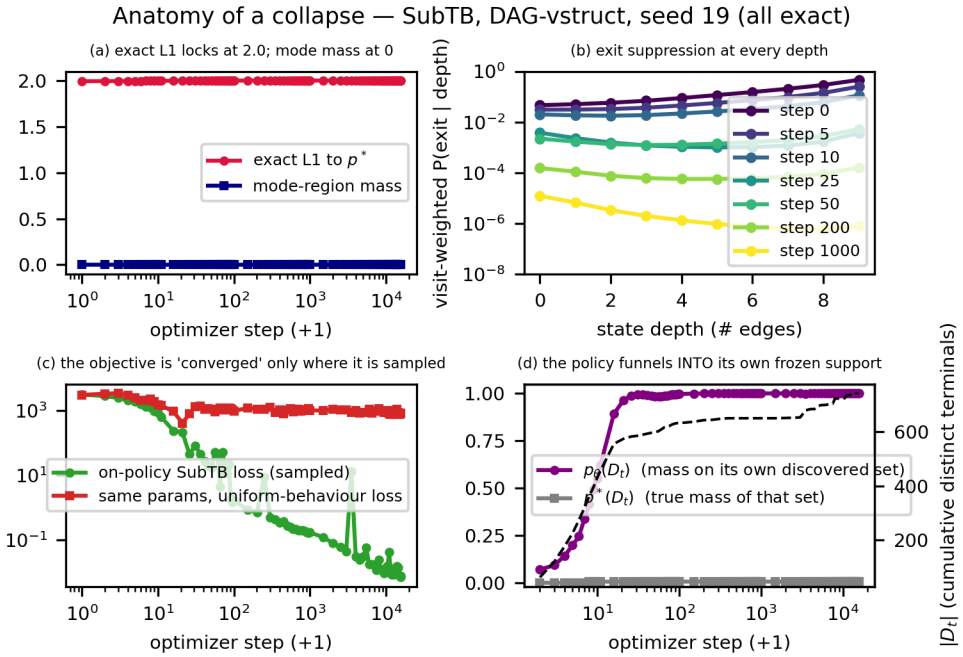

Figure 7: **Anatomy of a collapse** (SubTB, BNSL v-structure, seed 19) — the mechanism behind Fig. 1a, whose red curve is panel (a) here. Exit probability dies at every depth within ~ 50 steps; the discovered set D_t (panel d, cumulative and therefore non-decreasing) stalls at ~ 650 of $29{,}281$ states by step 200, growing only marginally after; and the exact L_1 pins at 2.0 while the on-policy loss goes to zero. Panel (c) adds what Fig. 1a's inset cannot show at that size: the *same parameters* scored under a full-support behavior keep a loss $5.4{\cdot}10^4\times$ larger (Appendix G), so the convergence is local to what the policy samples. (log $F(s_0)$ settling at the restricted partition function is Fig. 9b, not a panel here.)

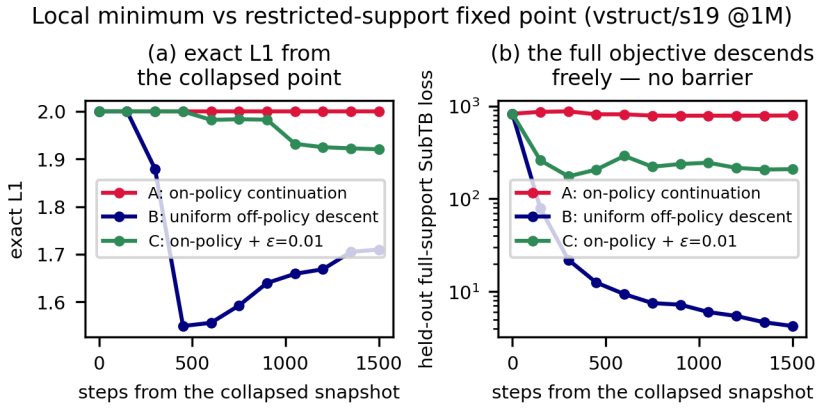

Figure 8: Collapse is a fixed point of the sampling-coupled dynamics, not a local minimum of the objective. Panel (b) is the full-support loss, descending freely; panel (a) is the calibration, which does not follow it down — off-policy descent alone leaves the sampler well inside the collapsed regime. Arm C continues the same snapshot on-policy under an ε=0.01 floor and moves only part of the way on both axes.

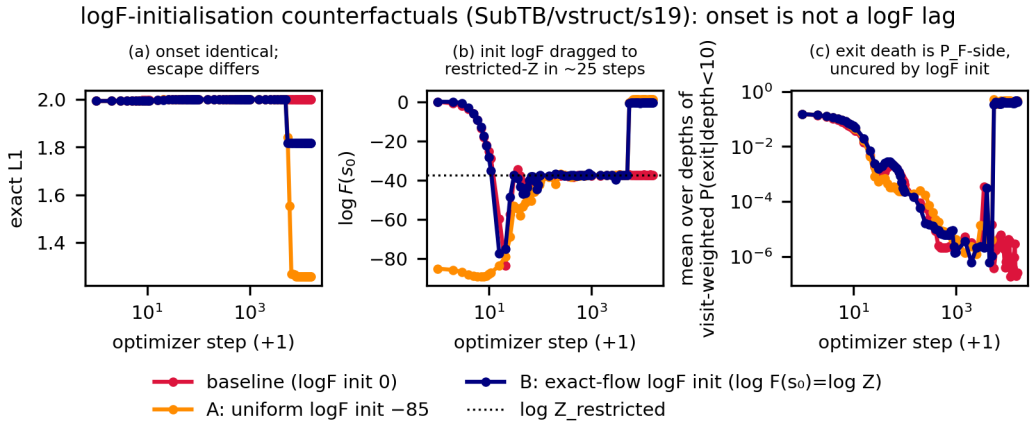

Figure 9: Two initialization counterfactuals: onset is not a log F lag — but escape is sensitive to it: panel (a) shows both counterfactual runs leaving the plateau where the baseline does not, panel (b) is where $\log F(s_0)$ settles at $\log Z_{T'}$, and panel (c) shows the depth-averaged exit probability (visit-weighted within each depth) dying on the same timescale under all three initializations: exit death is on the P_F side, and no log F initialization cures it.

H Claims map

Box 2 maps the main text's claims to the tables that decide them, with their status: proven, empirical, pre-registered, regime-limited, withdrawn, or a loss stated in advance.

Box 2: What is claimed, where it is decided, and what is withdrawn.

Claim, and where it is decided	Status
The RSFP: zero sampled loss and gradient, a global minimum tied with the truth, error $2(1-p^\star(T'))$, the support-deficit bound $\mathrm{L}_1 \geq 2(1-p^\star(S))$ for any policy supported on S, and stationarity under replay from a buffer inside the support (Prop. 1, Obs. 1, Cors. 1–3).	proven
Plain SubTB collapses on both structured five-node cells (17 of 20 seeds collapsed at 4M, seven of the eight collapsed study seeds escaping by 14M); TB, modified DB and local search share the trap, DB escapes on the v-structure by 4M; at 96 fresh per update only the Gaussian verdict stands (Tables 1, 11, 14, 17, 18).	empirical
The loss prefers the broken model: under the ε-floor the worse-calibrated arm scores the lower held-out loss on 9 of 10 v-structure seeds, collapsed plain SubTB scores the lowest Gaussian loss of all, and the fewest-mode sampler does so on two of three molecular landscapes (Tables 3, 11, 16).	empirical
ReStrike recovers both five-node posteriors on all four divergences at once at equal fresh-trajectory budget (on exact L_1 by 2M at the constant fraction), beats every floor, floor-plus-replay and local-search arm on all paired seeds but seven (and both fixed floors at 16M), and needs no learned backward walk (Tables 1, 3, 12, 17, 18, 7).	empirical
On the tiered grid ReStrike finds 16,635 of 20,736 modes at 1M against 119 for plain SubTB, 1,369 for the floor and 1,298 for the annealed-floor recipe; the constant fraction finds 17,979; a uniform draw from its buffer recovers 0.89 of them (Tables 2, 19).	empirical
The closed 0.9-quantile band of prioritized replay inverts prioritization on tiered rewards; the rank split its authors describe finds 9,407 modes where PRT-full's released rule finds 427 (Table 2, Fig. 5).	pre-registered
At six nodes ReStrike beats every floor, recipe and PRT arm on $5/5$ Gaussian seeds and only local search collapses on the v-structure; at eight nodes ReStrike calibrates best on the Gaussian chain and freezes on $2/5$ v-structure seeds, where the recipe calibrates best; the same buffer at $r{=}1$ freezes on none and is lower on both cells (Tables 4, 15).	regime-limited
The annealed replay fraction (a constant 0.5 is better on both cells; kept as run), the grid cells' calibration orderings, the 120-run molecular coverage table (a replay-loss defect) and the Lightning fork as a mechanism (Tables 3, 9; Appendices E, F).	withdrawn
PHO4 on JSD and both KL directions at every β, and to the same buffer drawn at $r{=}1$ on all but one metric–β pair (Table 8).	stated loss