ARCHITECTURE-DEPENDENT INFORMATION FLOORS FOR LEARNED BOLTZMANN SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned Boltzmann samplers often fit a tractable proposal q and correct the remaining mismatch to a target $\mu \propto e^{-\beta U}$ by importance reweighting. This raises a basic question: before training, what sampling efficiency is attainable by a given model class? We study the architecture-dependent information floor $\Phi_{\mathcal{Q}}(\mu) = \inf_{q \in \mathcal{Q}} \mathrm{D}(\mu \| q)$ for classes specified by conditional-independence structure. For factorised, finite-context autoregressive, and graph-Markov families, $\Phi_{\mathcal{Q}}$ admits exact target-only expressions in terms of total correlation, conditional mutual information, and the entropy gap to an I-projection. Combining these identities with asymptotic importance-sampling thresholds yields class-level lower bounds on the number of proposal draws required for reweighting. Under a uniform interface-information condition, the floor is extensive for bounded-size partitions and therefore induces exponential scaling with system size. We verify the identities against numerical optimisation and transfer-matrix calculations, compare the two-dimensional Ising measurements with Onsager's exact solution, and evaluate molecular examples in regimes where the floors are small. Finally, a pair-marginal certificate provides a data-driven lower bound without fitting a proposal. The resulting quantities separate optimisation error from structural mismatch and can be evaluated before training.

1 INTRODUCTION

Learned samplers for Boltzmann distributions are increasingly used to accelerate sampling in molecular systems, lattice models, and related energy-based problems. Representative approaches include normalising flows, equivariant and stochastic flows (Köhler et al., 2020; Wu et al., 2020), coarse-grained and autoregressive Boltzmann generators (Klein & Noé, 2024), flow-augmented Monte Carlo and replica exchange (Gabrié et al., 2022; Invernizzi et al., 2022), and lattice field-theory samplers (Albergo et al., 2019). These methods differ substantially in parameterisation and training, but many share a common evaluation pattern: a learned proposal is corrected against the target by importance reweighting. Performance is therefore determined not only by optimisation quality, but also by the mismatch that remains even for the best member of the model class.

We isolate the part of this mismatch that is imposed by conditional-independence structure. The central object is the best achievable forward divergence within a class, $\inf_{q \in \mathcal{Q}} \mathrm{D}(\mu \| q)$. For several structured families this infimum has an exact expression that depends only on the target distribution and the structural constraints of the model class. This makes it possible to distinguish structural approximation error from optimisation and finite-data effects before a model is trained.

Forward KL links fitting and reweighting. Let $\mathcal{X}$ be the configuration space, U the potential, β the inverse temperature, and $\mu(dx) = Z^{-1} e^{-\beta U(x)} dx$. Suppose a learned sampler provides a tractable proposal q, draws $x_1, \ldots, x_n \sim q$, and uses self-normalised importance weights $w_i \propto e^{-\beta U(x_i)}/q(x_i)$. In this setting the forward divergence $\mathrm{D}(\mu \| q) = \mathbb{E}_\mu[\log(\mu/q)]$ has two complementary roles:

- **Maximum-likelihood fitting.** Training on reference samples minimises $-\mathbb{E}_\mu \log q = \mathrm{H}(\mu) + \mathrm{D}(\mu \| q)$, so the excess population NLL above the target entropy is exactly the forward KL.

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

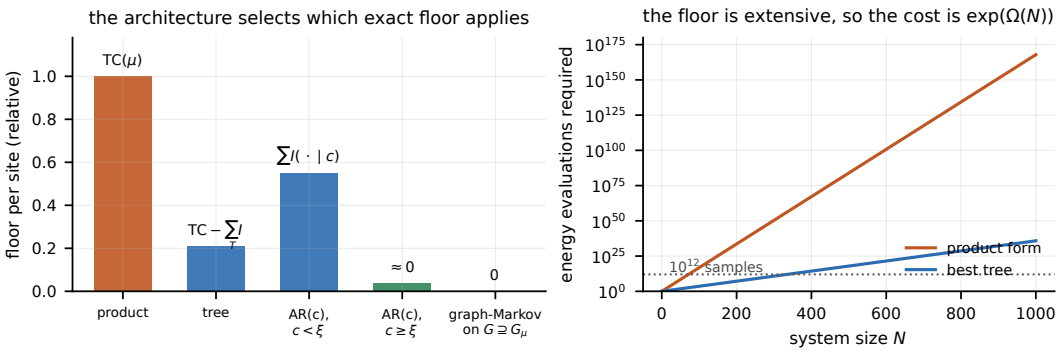


Figure 1: **Architecture-dependent information floors. Left:** for each structured model class, $\inf_{q\in\mathcal{Q}} \mathrm{D}(\mu\|q)$ reduces to an information functional of the target distribution (Theorems 2–4). **Right:** under the interface-information assumption of Theorem 6, bounded-size partitions produce an extensive floor, and the corresponding importance-sampling threshold scales exponentially with system size. Richer dependency structure can reduce the exponential rate.

- **Importance-sampling complexity.** Chatterjee & Diaconis (2018) show, under general conditions, a cutoff in the logarithmic sample size around $\mathrm{D}(\mu\|q)$: accurate importance sampling requires a number of proposal draws of order $\exp(\mathrm{D}(\mu\|q))$ in the asymptotic regime they analyse.

The first identity turns a structural approximation floor into an irreducible maximum-likelihood loss. The second translates the same floor into a reweighting-cost barrier for proposals restricted to that class.

Contributions.

1. **Exact information floors.** For factorised, finite-context autoregressive, and graph-Markov classes, we derive closed-form expressions for $\inf_{q\in\mathcal{Q}} \mathrm{D}(\mu\|q)$ in terms of total correlation, conditional mutual information, and an I-projection entropy gap (Theorems 2–4).
2. **Sampling implications.** Combining the floors with the importance-sampling threshold of Chatterjee & Diaconis (2018) gives a class-level lower bound on the proposal draws required by reweighting (Corollary 5).
3. **Extensive structural mismatch.** Under a uniform lower bound on interface information, bounded-size partitions incur a floor that grows linearly with system size (Theorem 6). We validate the scaling on solvable chains and compare two-dimensional Ising measurements with Onsager's exact result.
4. **Annealed decompositions and data-driven certificates.** We characterise stagewise reweighting in terms of consecutive divergences and derive a pair-marginal certificate that lower-bounds the factorised floor from reference samples (Section 5).
5. **Empirical validation.** Numerical optimisation, transfer-matrix calculations, Ising simulations, and molecular examples test the identities and show how the floors distinguish structurally easy and difficult regimes.

2 Exact floors

Let $x = (x_1, \ldots, x_d)$ with x_i in a Polish space, μ a probability measure with finite differential entropy (all statements hold verbatim for densities and for discrete measures with H the Shannon entropy).

Definition 1 (model classes)**.** For a partition $\mathcal{S} = \{S_1, \ldots, S_k\}$ of $\{1, \ldots, d\}$ let $\mathcal{Q}_{\mathrm{prod}}(\mathcal{S}) = \{\bigotimes_i q_i : q_i \text{ on } \mathcal{X}_{S_i}\}$. For an ordering π and context length c let $\mathcal{Q}_{\mathrm{AR}}(c) = \{q(x) = \prod_t q_t(x_{\pi(t)} \mid x_{\pi(t-c)}, \ldots, x_{\pi(t-1)})\}$. For an undirected graph G let $\mathcal{Q}_G$ be the set of distributions that are Markov with respect to G.

Theorem 2 (factorised floor = total correlation)**.** $\inf_{q\in\mathcal{Q}_{\mathrm{prod}}(\mathcal{S})} \mathrm{D}(\mu\|q) = \mathrm{TC}_{\mathcal{S}}(\mu) := \sum_{i=1}^{k} \mathrm{H}(\mu_{S_i}) - \mathrm{H}(\mu)$*, attained uniquely at* $q^\star = \bigotimes_i \mu_{S_i}$.

Proof. For $q = \bigotimes_i q_i$, $\mathrm{D}(\mu\|q) = -\mathrm{H}(\mu) - \sum_i \mathbb{E}_\mu \log q_i(x_{S_i}) = -\mathrm{H}(\mu) + \sum_i \big[\mathrm{H}(\mu_{S_i}) + \mathrm{D}(\mu_{S_i}\|q_i)\big]$, because $\mathbb{E}_\mu \log q_i$ depends on μ only through the marginal μ_{S_i}. Each $\mathrm{D}(\mu_{S_i}\|q_i) \geq 0$ with equality iff $q_i = \mu_{S_i}$. □

The decomposition above makes the structural source of the floor explicit: the cross-block dependence in μ remains after each marginal factor has been optimised. By contrast, the product-class optimum under the reverse divergence $\mathrm{D}(q\|\mu)$ is the mean-field variational problem and generally has no analogous closed-form expression (Section 5.1).

Theorem 3 (autoregressive floor = predictive-information deficit)**.** *With the ordering suppressed,*

$$\inf_{q\in\mathcal{Q}_{\mathrm{AR}}(c)} \mathrm{D}(\mu\|q) \;=\; \sum_{t=1}^{d} \mathrm{I}_\mu\left(x_t\,;\, x_{1:t-c-1} \,\middle|\, x_{t-c:t-1}\right), \tag{1}$$

attained by the truncated conditionals $q_t^\star = \mu(x_t \mid x_{t-c:t-1})$*. The right-hand side is non-increasing in* c *and vanishes iff* μ *is a* c*-th order Markov chain in the ordering.*

Proof. By the chain rule $\mathrm{D}(\mu\|q) = \sum_t \mathbb{E}_\mu\big[\mathrm{D}\left(\mu(\cdot \mid x_{1:t-1}) \,\|\, q_t(\cdot \mid x_{t-c:t-1})\right)\big]$; each term is minimised over q_t at the corresponding conditional of μ given the *available* variables, and the resulting value is $\mathbb{E}_\mu[\log \mu(x_t \mid x_{1:t-1}) - \log \mu(x_t \mid x_{t-c:t-1})]$, which is the conditional mutual information in (1). □

Theorem 4 (graph-Markov floor = I-projection deficit)**.** $\inf_{q\in\mathcal{Q}_G} \mathrm{D}(\mu\|q) = \mathrm{H}(\mu_G) - \mathrm{H}(\mu)$*, where* μ_G *is the* I*-projection of* μ *onto* $\mathcal{Q}_G$*. If* G *is decomposable with cliques* $\{C\}$ *and separators* $\{S\}$*, then* $\mathrm{H}(\mu_G) = \sum_C \mathrm{H}(\mu_C) - \sum_S \mathrm{H}(\mu_S)$*, so the floor is a finite combination of low-dimensional marginal entropies.*

Theorem 4 contains the previous two: $\mathcal{Q}_{\mathrm{prod}}$ is the empty graph and $\mathcal{Q}_{\mathrm{AR}}(c)$ the c-th power of a path. Its practical content is that for the decomposable case the floor is computable from marginals of dimension at most the clique size, i.e. from quantities that a reference simulation already resolves.

Corollary 5 (importance-sampling cost within a class)**.** *Let* $\mathcal{Q}$ *be any class in Definition 1 and let* $\Phi_{\mathcal{Q}}(\mu)$ *denote its information floor. For every* $q \in \mathcal{Q}$*,* $\mathrm{D}(\mu\|q) \geq \Phi_{\mathcal{Q}}(\mu)$*. Consequently, in the asymptotic regime of Chatterjee & Diaconis (2018), the logarithmic sample-size threshold for importance sampling with proposal* q *is no smaller than* $\Phi_{\mathcal{Q}}(\mu)$ *up to lower-order terms. Optimisation, parameter count, and training data can change* q *within* $\mathcal{Q}$ *but cannot reduce this structural floor.*

3 Extensivity of the structural floor

Theorem 6 (extensive floor under interface information)**.** *Let* μ_N *be a translation-invariant Gibbs measure on* N *sites. Assume that the mutual information contributed across each broken bond of a block interface is bounded below by a constant* $\iota > 0$*. For a partition into bounded-size blocks with* $m(N) = \Theta(N)$ *broken bonds,*

$$\mathrm{TC}_{\mathcal{S}}(\mu_N) \;\geq\; \iota\, m(N) \;=\; \Omega(N), \qquad \textit{hence} \qquad \exp\big(\mathrm{TC}_{\mathcal{S}}(\mu_N)\big) = \exp(\Omega(N)). \tag{2}$$

For the one-dimensional q*-state Potts chain with single-site blocks, the corresponding slope equals the per-bond mutual information and is available in closed form from the transfer matrix.*

For the zero-field Ising chain at $\beta J = 1$, the site-factorised floor is linear in L over $L = 4, \ldots, 14$ with maximum residual 9.5×10^{-11}. Its slope, 0.110944, agrees with the independently computed per-bond mutual information to 1.4×10^{-9}. Over $L = 2, \ldots, 12$, the measured importance-sampling cost has slope $d \log n^\star / dL = 0.153$, compared with $d\ \mathrm{TC}\,/dL = 0.111$ ($R^2 = 0.93$). The larger empirical slope is consistent with the stronger sensitivity of variance to χ^2 divergence.

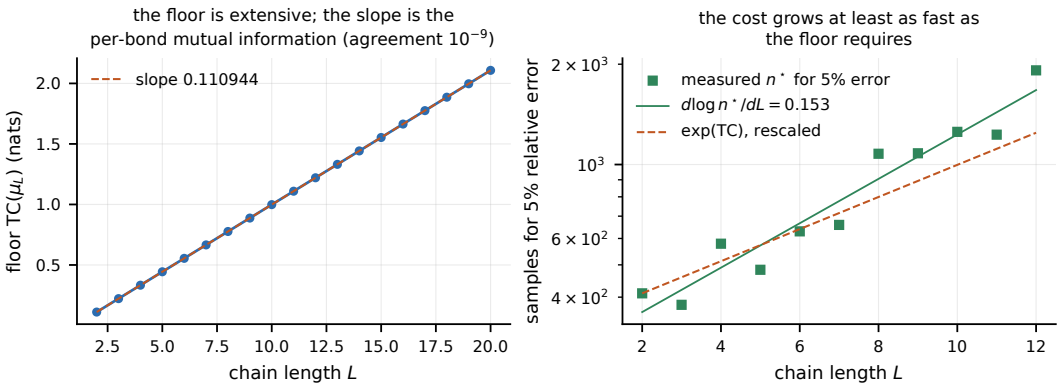


Figure 2: **Left:** site-factorised floor of the zero-field Ising chain at $\beta J = 1$, computed from transfer-matrix entropies. The fitted slope is 0.110944, agreeing with the per-bond mutual information to 1.4×10^{-9}; the maximum linear-fit residual is 9.5×10^{-11}. **Right:** measured sample count required for 5% relative error on an observable when using the product projection as the proposal.

Remark 7 (scope of the extensivity result)**.** Theorem 6 concerns a sequence of bounded-size partitions satisfying a uniform interface-information lower bound. If block size or graph structure changes with N, the floor can scale differently. The result therefore quantifies the cost of a fixed structural restriction rather than asserting that every local architecture has the same asymptotic behaviour.

4 Annealed decompositions and stagewise cost

Theorem 8 (stagewise importance sampling)**.** *Let $\mu_0 = q$ and $\mu_1 = \mu$, and let μ_λ define a path with $0 = \lambda_0 < \cdots < \lambda_L = 1$. If each stage is corrected by importance sampling, the stagewise sample-size threshold is controlled by $\mathrm{D}(\mu_{\lambda_\ell} \| \mu_{\lambda_{\ell-1}})$. Thus a ladder whose consecutive divergences are $O(\log N)$ yields polynomial per-stage thresholds; the total cost is then polynomial whenever the number of stages is polynomial. For sufficiently small steps, $\mathrm{D}(\mu_{\lambda_\ell} \| \mu_{\lambda_{\ell-1}}) = \frac{1}{2}(\Delta\lambda_\ell)^2 \partial_\lambda^2 \log Z_{\lambda_\ell} + O(\Delta\lambda_\ell^3)$, so approximately equal divergence increments use steps proportional to the inverse square root of the local Fisher information.*

This analysis provides an information-theoretic interpretation of sequential and annealed Boltzmann samplers (Matthews et al., 2022): intermediate targets can replace one difficult reweighting problem by a sequence of better-overlapped stages.

5 Estimating the floors from simulation data

Because the floors are entropy differences of low-dimensional marginals (Theorem 4), they are estimable from a reference trajectory, with the usual caveats of entropy estimation from finite samples (Paninski, 2003). Two properties of the estimator decide what such an estimate means.

Monotonicity makes a discretised floor a rigorous statement. Total correlation is monotone under deterministic per-coordinate maps: for any binning functions f_i, $\mathrm{TC}(f_1(x_1), \ldots, f_k(x_k)) \leq \mathrm{TC}(x_1, \ldots, x_k)$. So every binning yields a *certified lower bound* on the continuum floor, and the bound increases under refinement. We therefore report the floor as a function of resolution rather than as a single number.

Corollary 9 (a certificate from pair marginals)**.** *Let μ_G be the maximum-entropy law with the same G-clique marginals as μ. Then $\mathrm{TC}(\mu_G) \leq \mathrm{TC}(\mu)$ for every graph G, with equality exactly when $\mu \in \mathcal{Q}_G$.*

Proof. The I-projection onto $\mathcal{Q}_G$ preserves the clique marginals, hence the single-site ones, so $\sum_i \mathrm{H}(\mu_i)$ is the same for μ and μ_G; and μ_G has the largest entropy of any law with those clique marginals, so $\mathrm{H}(\mu_G) \geq \mathrm{H}(\mu)$. Subtract. □

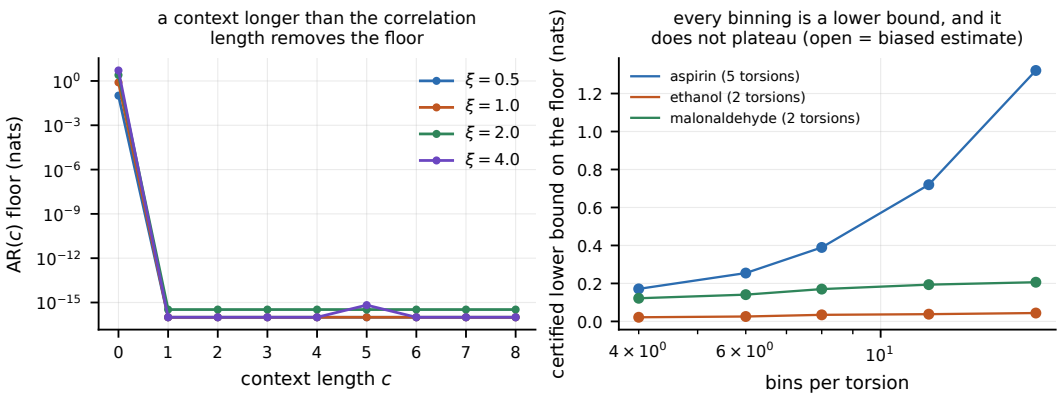


Figure 3: **Left:** AR(c) floor for Gaussian targets with exponentially decaying correlations of range ξ. The floor decreases sharply once the available context reaches the correlation scale. **Right:** product-class floor for torsion distributions from three MD17 molecules as a function of angular resolution. Each per-coordinate discretisation gives a lower bound by monotonicity of total correlation; open markers denote estimates rejected by the finite-sample diagnostic.

For a path or tree G, the left-hand side depends only on pair marginals. A reference trajectory can therefore provide a lower bound on the factorised floor without estimating the full joint law or fitting a proposal. Discretisation and graph coarsening both decrease the resulting quantity, so a binned pairwise estimate remains a lower bound on the continuum floor.

Finite-sample entropy estimation. When a discretised joint distribution is undersampled, plug-in entropy can be biased downward and the resulting floor biased upward. We use the effective support e^{H} together with a split-sample diagnostic: an estimate is rejected when its value on a random half of the frames differs from the full-sample estimate by more than three bootstrap standard errors.

5.1 The reverse divergence, and why we do not use it

Energy-based training of Boltzmann generators often minimises the reverse divergence $\mathrm{D}(q\|\mu)$. Its product-class optimum is the mean-field variational problem and generally lacks the closed-form structural identities derived above. Reverse KL can also remain small when q undercovers regions of target mass, whereas importance sampling depends directly on the target-to-proposal mismatch. We therefore focus on $\mathrm{D}(\mu\|q)$, which is both the population maximum-likelihood excess loss and the divergence entering the importance-sampling threshold of Chatterjee & Diaconis (2018).

6 Experiments

Exact numerical checks. We compare the closed-form floors with direct numerical optimisation over the corresponding model classes, obtaining agreement to 10^{-11} for the Gaussian cases and 10^{-16} for discrete tables. Randomly perturbed product and AR(c) models remain above the predicted floors, and independent transfer-matrix calculations reproduce the entropy identities. A direct cost experiment gives $d \log n^\star / d\ \mathrm{D} = 1.29$ with $R^2 = 0.966$ over $\mathrm{D} \in [0, 2.04]$.

Ising-chain scaling. Figure 2 compares the exact site-factorised floor with measured importance-sampling cost on the Ising chain. The floor slope equals the per-bond mutual information, while the empirical cost slope is $d \log n^\star / dL = 0.153$ versus $d\ \mathrm{TC}\ / dL = 0.111$ ($R^2 = 0.93$).

Trained models approach the structural floors. We fit maximum-likelihood models from each structural class to discretised aspirin torsions using an 80/20 split. To reduce estimator mismatch, we compare each held-out NLL with that of an unconstrained model evaluated on the same split. With 6^5 cells, the unconstrained model itself has a small finite-sample generalisation gap, so this matched

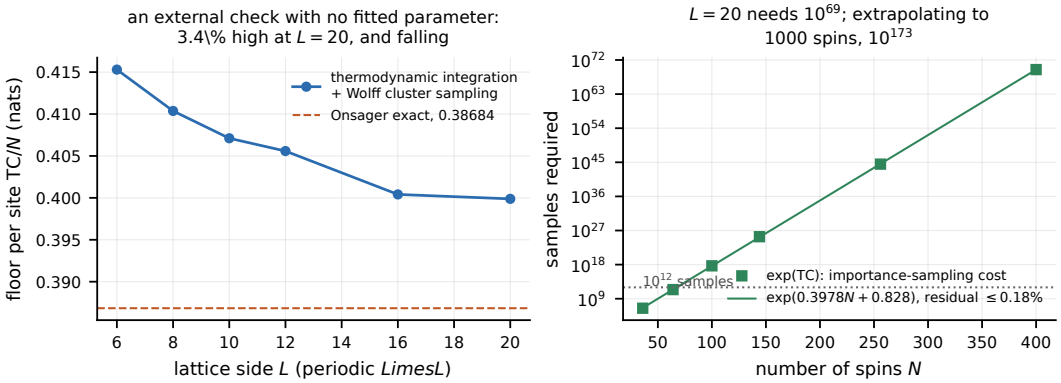


Figure 4: **Left:** measured product-class floor per site on the critical square lattice, compared with Onsager's exact infinite-lattice value of 0.38684 nats. **Right:** $\exp(\mathrm{TC})$ as a function of system size, together with a linear fit to TC in N (maximum relative residual $\leq 0.18\%$). The dotted line marks 10^{12} proposal draws as a reference scale.

comparison is more stable than subtracting a plug-in entropy estimate obtained from a different sample.

class	held-out NLL	excess	floor	achieved / floor
product	6.1542	0.2388	0.2546	**0.938**
AR(1)	6.0459	0.1305	0.1464	**0.892**
AR(2)	5.9400	0.0247	0.0386	0.640
AR(3)	5.9155	0.0002	0.0073	0.022
unconstrained	5.9153	0	0	—

For the product and AR(1) classes, the observed excess NLL reaches 94% and 89% of the corresponding floor, respectively. For AR(2) and AR(3), the floors (0.039 and 0.007 nats) are small relative to the finite-sample uncertainty of this comparison, so the achieved-to-floor ratios are not informative. The fitted product model also shows the expected $n^{-1/2}$ decay in the measured importance-sampling error (0.146 at $n = 30$ and 0.0089 at $n = 10^4$), consistent with a proposal whose remaining mismatch is small.

Critical two-dimensional Ising model. Direct plug-in estimation of the joint entropy is infeasible at the lattice sizes considered here. Instead, for the Gibbs measure we obtain the entropy thermodynamically from $\log Z(\beta) = N \log q - \int_0^\beta \langle E \rangle_{\beta'} d\beta'$ and $S = \beta \langle E \rangle + \log Z$. At zero field, symmetry gives the uniform single-site marginal, so $\mathrm{H}_1 = \log 2$ and $\mathrm{TC} = N \log 2 - S$.

We use the Wolff cluster algorithm (Swendsen & Wang, 1987; Wolff, 1989) for the critical Ising simulations. In a one-dimensional validation at $\beta = 0.8$, $L = 12$, single-spin Metropolis exhibits a 12.3% discrepancy in the target statistic, while the Wolff estimates are within 0.03–0.79%. The two-dimensional calculation also has an independent reference: at $\beta_c = 0.4407$, Onsager's free energy (Onsager, 1944) gives $S/N = 0.30631$ and hence $\mathrm{TC}/N = \mathbf{0.38684}$ nats per site (Appendix C).

The measured floor per site decreases with lattice size and reaches 0.3999 nats at $L = 20$, within 3.4% of the infinite-lattice value. Over the simulated sizes, TC is approximately linear in N with slope 0.3978 nats per site and a maximum fit residual of 0.2% of the largest value, consistent with the extensive scaling predicted by Theorem 6.

At $L = 20$ the site-factorised floor is 160.0 nats. The corresponding importance-sampling threshold is therefore on the order of $e^{160} \approx 10^{69}$ proposal draws in the asymptotic scaling considered above. Extrapolating the measured linear slope illustrates how rapidly this structural mismatch grows with system size; the molecular examples below provide the complementary regime in which the measured floors are small.

Tree-structured models. Chow–Liu trees provide a tractable example of how additional dependence reduces the structural floor (Chow & Liu, 1968): $\mathrm{D}(\mu\|q_T) = \sum_i \mathrm{H}(x_i) - \sum_{(i,j)\in T} I(x_i; x_j) - \mathrm{H}(\mu)$. The optimal tree is the maximum-weight spanning tree with edge weights $I(x_i; x_j)$. We estimate the pairwise mutual informations from the same Wolff samples and use the Onsager entropy for $\mathrm{H}(\mu)$.

Across five lattice sizes ($N = 36$–256), the best tree removes a median of **81%** of the product-class floor (range 79.4–86.4%; Appendix B). The residual tree floor remains approximately linear in N, with fitted slope $\mathbf{0.0825}$ nats per site and maximum relative residual 2.4%, compared with 0.38657 for the product class. Thus the tree substantially reduces the exponential rate while retaining extensive residual mismatch over the sizes studied.

We also compute unrestricted Chow–Liu trees using all $\binom{N}{2}$ pairwise mutual informations. For $L = 6, 8, 10$, the resulting maximum-weight spanning trees have the same floors as the lattice-edge-restricted solutions (1.894, 4.667, and 7.595 nats), indicating that the long-range pairwise terms do not alter the optimum at these sizes.

For comparison, subtracting all $2N$ nearest-neighbour mutual informations as though the lattice were loop-free produces a negative value (-49.1 nats at $L = 16$ and negative values at every tested size). The pairwise terms therefore cannot simply be summed on a loopy graph: doing so double-counts dependence and does not define a valid bound.

Molecular examples. On MD17 we extract the rotating torsions of aspirin (5), ethanol (2), and malonaldehyde (2), and compute floors for discretised torsion distributions using 2×10^5 to 10^6 frames. At the coarsest resolution the site-product floor is 0.171 nats for aspirin and 0.022 nats for ethanol, much smaller than in the critical Ising example. The AR(1) floor of both two-torsion molecules is zero to machine precision, as expected from the dimensionality of the problem.

An exactly solvable molecular-chain benchmark. The rotational-isomeric-state model of an n-alkane (Abe et al., 1966) provides an exactly solvable molecular-chain benchmark with nearest-neighbour interactions. At 300 K, using $\Delta E_{\mathrm{gauche}} = 0.8$ and a 3.0 kcal/mol pentane-interference penalty, the per-torsion floor is 0.04076 nats and agrees with the per-bond mutual information to 0.02%. The corresponding exponential scale for a thousand torsions is approximately 10^{18}. At 200 and 500 K, the per-torsion floors are 0.01742 and 0.06094 nats, respectively.

We next evaluate the pair-marginal certificate of Corollary 9 against an exact reference. We run Metropolis-adjusted Langevin dynamics on the torsions of the OPLS-AA alkane potential (Jorgensen et al., 1996) with an explicit g^+g^- interference term, using a null and a coupled setting.

In the zero-coupling control, the true floor is zero. The measured trans population is 0.6941, compared with 0.6938 from quadrature over the same continuous basins, and the estimated floor is 3.6×10^{-7} nats.

With the coupling enabled, the pair-marginal lower bound is 0.905 nats over 32 torsions. For a nine-torsion chain, summing the grid transfer matrix over all 3^9 basin-label sequences gives an exact floor of 0.23878 nats. The certificate computed from the exact pair law recovers 97.3% of this value, while the estimate from 1.47×10^7 sampled pairs recovers 99.7%. Across chain lengths, the exact total correlation is approximately linear with slope 0.02896 nats per torsion and largest residual 7.8×10^{-4} (Figure 5).

7 Related work

Importance-sampling complexity. Chatterjee & Diaconis (2018) establish a logarithmic sample-size threshold governed by $\mathrm{D}(\mu\|q)$ under general conditions. Related effective-sample-size analyses are often expressed through χ^2 divergence (Kong et al., 1994). Our use of these results is to connect the proposal-target divergence to exact structural approximation floors for specific model classes.

Boltzmann generators and reweighting. Since Noé et al. (2019), learned Boltzmann sampling has expanded to flow-based, sequential, transferable, and physics-informed constructions. Reweighting provides a standard correction mechanism for imperfect proposals (Nicoli et al., 2020) and is closely

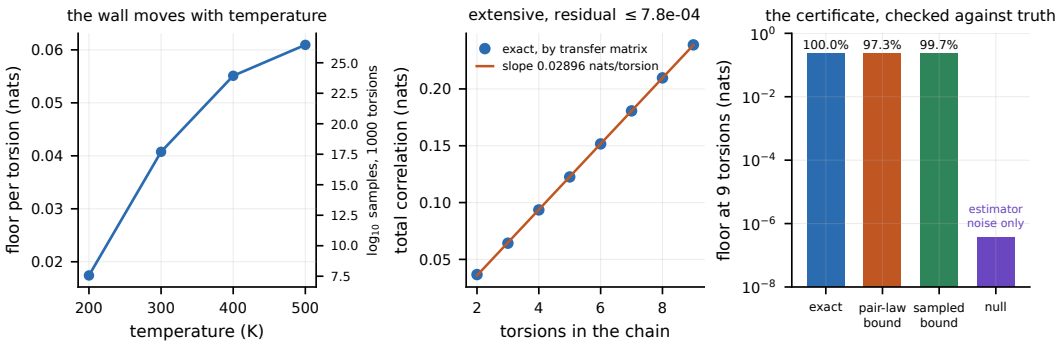


Figure 5: **Left:** exact per-torsion floor of the RIS n-alkane as a function of temperature; the right axis shows the corresponding e^{Φ} scale for a thousand-torsion chain. **Middle:** exact total correlation of the basin-label chain as a function of length. **Right:** pair-marginal certificate from Corollary 9, compared with the exact value and with a zero-coupling null control.

related to annealed importance sampling (Neal, 2001) and nonequilibrium work methods (Jarzynski, 1997; Crooks, 1999). Recent work targets larger molecular and materials systems (Tan et al., 2025a;b; Schebek et al., 2025; Grenioux et al., 2025; Rehman & Friesecke, 2026). Our analysis complements these algorithmic advances by isolating the portion of proposal-target mismatch imposed by structural constraints on the model class.

Expressivity lower bounds. Existing limitations for normalising flows include results based on smoothness and Lipschitz constraints, often stated in total variation or related metrics (Verine et al., 2022; Kong & Chaudhuri, 2020). The results here concern a different restriction: conditional-independence structure. For the classes we analyse, the forward-KL optimum is an exact identity rather than a generic expressivity bound, so the two perspectives are complementary.

Post-hoc proposal diagnostics. Pareto-smoothed importance sampling (Vehtari et al., 2024) uses the tail of the importance weights to diagnose a particular fitted proposal. Our quantities answer a complementary pre-training question: they bound the best forward-KL mismatch attainable within a structural class. Post-hoc weight diagnostics and class-level information floors can therefore be used together.

Information-theoretic structured approximation. Total correlation and I-projections are classical (Watanabe, 1960; Csiszár, 1975), including the product-family identity $\inf_{\text{product}} \mathrm{D}(\mu\|q) = \mathrm{TC}(\mu)$. We use these objects to connect conditional-independence restrictions in learned Boltzmann samplers with importance-sampling complexity, extensivity, and data-driven structural diagnostics.

8 Conclusion

We introduced architecture-dependent information floors for learned Boltzmann sampling. For factorised, finite-context autoregressive, and graph-Markov model classes, the best achievable forward KL can be written as a target-only information quantity. These identities separate structural approximation error from optimisation and finite-data effects, and they connect directly to the logarithmic sample-size scale of importance reweighting. Under a uniform interface-information condition, bounded partitions produce extensive floors; richer dependency structures can reduce the rate, as illustrated by the tree-structured Ising example. The same quantities can be estimated or lower-bounded from reference data, providing a pre-training diagnostic for model-class selection. Extending these calculations to broader structured generative families and to adaptive intermediate-target constructions is a natural direction for future work.

AI use statement

GPT was used only to polish the writing of this manuscript.

REPRODUCIBILITY STATEMENT

The theoretical statements are accompanied by proofs in the main text or appendix. The numerical sections report the model classes, target systems, reference quantities, and evaluation statistics used in each comparison; the appendix provides the tabulated Ising and tree-structured results underlying the corresponding figures. The molecular experiments use the public MD17 data and the cited OPLS-AA and rotational-isomeric-state models.

REFERENCES

Akihiro Abe, Robert L. Jernigan, and Paul J. Flory. Conformational energies of n-alkanes and the random configuration of higher homologs including polymethylene. *Journal of the American Chemical Society*, 88(4):631–639, 1966. doi: 10.1021/ja00956a001.

M S Albergo, G Kanwar, and P E Shanahan. Flow-based generative models for markov chain monte carlo in lattice field theory. *Physical Review D*, 100(3):034515, 2019. doi: 10.1103/PhysRevD.100.034515.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018. doi: 10.1214/17-AAP1326.

C Chow and C Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. doi: 10.1109/TIT.1968.1054142.

Gavin E Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726, 1999. doi: 10.1103/PhysRevE.60.2721.

Imre Csiszár. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Marylou Gabrié, Grant M Rotskoff, and Eric Vanden-Eijnden. Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022. doi: 10.1073/pnas.2109420119.

Louis Grenioux, Leonardo Galliano, Ludovic Berthier, Giulio Biroli, and Marylou Gabrié. Boltzmann generators for amorphous particle systems. *arXiv preprint arXiv:2512.16607*, 2025.

Michele Invernizzi, Andreas Krämer, Cecilia Clementi, and Frank Noé. Skipping the replica exchange ladder with normalizing flows. *The Journal of Physical Chemistry Letters*, 13(50):11643–11649, 2022. doi: 10.1021/acs.jpclett.2c03327.

C Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997. doi: 10.1103/PhysRevLett.78.2690.

William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996. doi: 10.1021/ja9621760.

Leon Klein and Frank Noé. Transferable boltzmann generators. *Advances in Neural Information Processing Systems*, 2024.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: Exact likelihood generative learning for symmetric densities. *International Conference on Machine Learning*, 2020.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994. doi: 10.1080/01621459.1994.10476469.

Zhifeng Kong and Kamalika Chaudhuri. The expressive power of a class of normalizing flow models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Alexander G. D. G. Matthews, Michael Arbel, Danilo J. Rezende, and Arnaud Doucet. Continual repeated annealed flow transport monte carlo. In *International Conference on Machine Learning*, 2022.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. doi: 10.1023/A:1008923215028.

Kim A Nicoli, Shinichi Nakajima, Nils Strodthoff, Wojciech Samek, Klaus-Robert Müller, and Pan Kessel. Asymptotically unbiased estimation of physical observables with neural samplers. *Physical Review E*, 101(2):023304, 2020. doi: 10.1103/PhysRevE.101.023304.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. doi: 10.1126/science.aaw1147.

Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117–149, 1944. doi: 10.1103/PhysRev.65.117.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272.

Zia Ur Rehman and Gero Friesecke. Sampling boltzmann distributions via normalizing flow approximation of transport maps. *arXiv preprint arXiv:2603.14258*, 2026.

Maximilian Schebek, Frank Noé, and Jutta Rogal. Scalable boltzmann generators for equilibrium sampling of large-scale materials. *arXiv preprint arXiv:2509.25486*, 2025.

Robert H Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in monte carlo simulations. *Physical Review Letters*, 58(2):86–88, 1987. doi: 10.1103/PhysRevLett.58.86.

Charlie B. Tan, Avishek Joey Bose, Chen Lin, Leon Klein, Michael M. Bronstein, and Alexander Tong. Scalable equilibrium sampling with sequential boltzmann generators. *arXiv preprint arXiv:2502.18462*, 2025a.

Charlie B. Tan, Majdi Hassan, Leon Klein, Saifuddin Syed, Dominique Beaini, Michael M. Bronstein, Alexander Tong, and Avishek Joey Bose. Amortized sampling with transferable normalizing flows. *arXiv preprint arXiv:2508.18175*, 2025b.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

Alexandre Verine, Benjamin Negrevergne, Fabrice Rossi, and Yann Chevaleyre. On the expressivity of bi-lipschitz normalizing flows. In *Asian Conference on Machine Learning*, 2022.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960. doi: 10.1147/rd.41.0066.

Ulli Wolff. Collective monte carlo updating for spin systems. *Physical Review Letters*, 62(4):361–364, 1989. doi: 10.1103/PhysRevLett.62.361.

Hao Wu, Jonas Köhler, and Frank Noé. Stochastic normalizing flows. *Advances in Neural Information Processing Systems*, 33, 2020.

A Proofs

A.1 Proof of Theorem 4

$\mathcal{Q}_G$ is the set of distributions whose density factorises according to G; it is a log-linear (exponential) family, and minimising $\mathrm{D}(\mu\|q)$ over an exponential family is a moment-matching problem whose solution μ_G is the I-projection of μ onto $\mathcal{Q}_G$ (Csiszár, 1975). At the optimum, $\mathrm{D}(\mu\|\mu_G) = -\mathrm{H}(\mu) - \mathbb{E}_\mu \log \mu_G$, and because μ_G matches μ on all clique marginals, $\mathbb{E}_\mu \log \mu_G = \mathbb{E}_{\mu_G} \log \mu_G = -\mathrm{H}(\mu_G)$. Hence $\inf_{q\in\mathcal{Q}_G} \mathrm{D}(\mu\|q) = \mathrm{H}(\mu_G) - \mathrm{H}(\mu)$. For a decomposable G the junction-tree factorisation gives $\mu_G = \prod_C \mu_C / \prod_S \mu_S$, so $\mathrm{H}(\mu_G) = \sum_C \mathrm{H}(\mu_C) - \sum_S \mathrm{H}(\mu_S)$. □

A.2 Proof of Corollary 5

Let $\Phi = \Phi_{\mathcal{Q}}(\mu)$ and $q \in \mathcal{Q}$. By definition, $\mathrm{D}(\mu\|q) \geq \Phi$. The importance-sampling cutoff results of Chatterjee & Diaconis (2018) place the logarithmic sample-size threshold around $\mathrm{D}(\mu\|q)$ under their stated regularity and asymptotic conditions. Replacing $\mathrm{D}(\mu\|q)$ by its classwise lower bound Φ therefore gives the claimed class-level lower bound on this logarithmic scale. Changing optimisation, parameter count, or training data changes the selected q but not the infimum defining Φ. □

A.3 Proof of Theorem 6

Let $\mathcal{S}$ be a partition into blocks and let $\partial\mathcal{S}$ be the set of bonds joining different blocks, $|\partial\mathcal{S}| = m(N)$. Writing the total correlation as a sum of conditional mutual informations along any ordering of the blocks and using the assumed lower bound ι per broken bond, $\mathrm{TC}_{\mathcal{S}}(\mu_N) \geq \iota\, m(N)$. For blocks of bounded size and a translation-invariant lattice, $m(N) = \Theta(N)$. Combining with Corollary 5 gives $n \geq \exp(\Omega(N))$. For the one-dimensional q-state Potts chain with single-site blocks, $\mathrm{TC} = L\,\mathrm{H}_1 - \mathrm{H}(\mu_L)$ with $\mathrm{H}_1 = \log q$ by symmetry at zero field, and $\mathrm{H}(\mu_L)$ is affine in L with slope the entropy rate, so TC is affine in L with slope $\mathrm{H}_1 - s_{\text{rate}}$, which is precisely the per-bond mutual information. We verify the equality numerically to 1.4×10^{-9}. □

A.4 Proof of Theorem 8

Applying the importance-sampling threshold stage by stage shows that the logarithmic sample-size scale at stage ℓ is controlled by $\mathrm{D}(\mu_{\lambda_\ell}\|\mu_{\lambda_{\ell-1}})$. Thus consecutive divergences of order $O(\log N)$ correspond to polynomial per-stage thresholds, and a polynomial number of such stages yields polynomial total work at this level of analysis. For a smooth exponential interpolation, $\lambda \mapsto \log Z_\lambda$ is convex and a second-order expansion gives $\mathrm{D}(\mu_{\lambda_\ell}\|\mu_{\lambda_{\ell-1}}) = \frac{1}{2}(\Delta\lambda_\ell)^2\,\partial_\lambda^2 \log Z_{\lambda_\ell} + O(\Delta\lambda_\ell^3)$. Hence approximately equal consecutive divergences are obtained by choosing $\Delta\lambda_\ell$ inversely proportional to the square root of the local Fisher information. □

B What the best tree removes, site by site

Table 1 is the measurement behind the tree floor. Every mutual information is estimated on the same Wolff samples and $\mathrm{H}(\mu)$ is Onsager's, so no term is fitted.

Table 1: The floor of the best tree-structured model on the critical square lattice, against the product floor.

L	N	product floor	best-tree floor	per site	% of floor removed
$L = 6$	36	13.922	1.894	0.0526	86.40
$L = 8$	64	24.748	4.667	0.0729	81.14
$L = 10$	100	38.672	7.595	0.0760	80.36
$L = 12$	144	55.677	10.502	0.0729	81.14
$L = 16$	256	98.970	20.350	0.0795	79.44

C The Ising ladder, site by site

Table 2 is the measurement behind Figure 4. $S(\mu)$ comes from thermodynamic integration over an inverse-temperature ladder with the Wolff cluster algorithm, and the last column compares the floor per site to Onsager's exact 0.38684 nats.

C.1 On the tightness of the sub-additivity step

Theorem 6 uses a per-bond lower bound on mutual information across an interface, whereas Theorems 2–4 are exact identities. For the one-dimensional chain, the interface relation is an equality, allowing the extensivity slope to be checked directly.

Table 2: The measured floor on the critical square-lattice Ising model, from thermodynamic integration with the Wolff cluster algorithm.

L	N	$S(\mu)$	TC	TC $/N$	vs Onsager
6	36	10.18	14.95	0.4153	$+7.4\%$
8	64	18.20	26.26	0.4104	$+6.1\%$
10	100	28.86	40.71	0.4071	$+5.2\%$
12	144	41.41	58.40	0.4056	$+4.9\%$
16	256	74.94	102.51	0.4004	$+3.5\%$
20	400	117.31	159.95	0.3999	$+3.4\%$