# An End-to-End Hybrid Quantum–Classical Sampling Workflow for Discrete Markov Random Fields: A Reproducible Case Study

Arul Rhik Mazumder
*School of Computer Science*
*Carnegie Mellon University*
Pittsburgh, PA, USA
arulm@cs.cmu.edu



***Abstract*—Sampling from discrete Markov random fields (MRFs) is a well-known hard problem in probabilistic inference. We study amplitude-encoded i.i.d. sampling for small discrete MRFs, scoped to the regime where the $2^n$ target probabilities can be precomputed classically—so no quantum exponential speedup is possible, but the structural property that each circuit execution returns an independent sample ($\tau \approx 1$) can be cleanly compared against classical MCMC alternatives. Across 60 instances spanning five graph families (barbell, barbell-path, chain, Erdős–Rényi, two-clique) with 1,000-step burn-in and 3,000 retained samples, Quantum/Single-Site-Gibbs, Quantum/Block-Gibbs, Quantum/Tuned-Block-Gibbs, and Quantum/Parallel-Tempering ESS ratios have means of $16.35$, $7.29$, $1.82$, and $1.79$ respectively, showing that modern classical samplers substantially close—and may eliminate—the ESS gap relative to amplitude-encoded sampling. When the $O(2^n)$ classical preprocessing required by the amplitude encoder is amortized into wall-clock time, exact inverse-CDF sampling reaches a mean of $17{,}683{,}598$ ESS/s versus $487{,}706$ ESS/s for the quantum sampler ($36\times$ on mean rates, $153\times$ mean per-instance), confirming no wall-clock advantage in this regime. The contribution is therefore not a speedup claim but (i) a clean characterization of MCMC autocorrelation costs under a fixed protocol and (ii) a reproducible benchmark of amplitude-encoded state preparation for discrete MRFs at $n = 8, 10, 12$. We further report a multi-trial matrix product state (MPS) scaling study (three seeds per point, $n$ up to $40$) showing $\chi = 32$ achieves $F = 0.721 \pm 0.059$ at $n = 40$, and a matched-budget variational quantum circuit (VQC) vs. MPS comparison at $n = 8, 10, 12$ where VQC fidelities fall below MPS at every point ($(F_{\mathrm{VQC}}, F_{\mathrm{MPS}}) = (0.306, 0.990), (0.210, 0.958), (0.165, 0.878)$ at compressions $10.7\times$, $34.1\times$, $113.8\times$)—a negative result for shallow hardware-efficient ansätze. Code and data: https://github.com/arulrhikm/QuantumDGM.**

***Index Terms*—amplitude encoding, graphical models, hardware-efficient ansatz, Markov random fields, Monte Carlo methods, quantum machine learning, quantum sampling, variational quantum circuits**

## I. Introduction

### A. Problem Selection and Motivation

Markov random fields (MRFs) encode conditional independence structures through undirected graphs and are fundamental to computer vision [1], computational biology [2], statistical physics, and generative modeling [3]. The central computational challenge is drawing samples from the joint distribution

$$P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left( \sum_{C \in \mathcal{C}} \sum_{\boldsymbol{y} \in \mathcal{X}_C} \theta_{C,\boldsymbol{y}}\, \phi_{C,\boldsymbol{y}}(\mathbf{x}) \right), \qquad (1)$$

where $\mathcal{C}$ denotes the set of maximal cliques of the graph, $\phi_{C,\boldsymbol{y}}$ are indicator sufficient statistics, and the partition function $Z(\theta)$ requires summing over $2^n$ configurations for binary variables. This exponential cost makes exact inference intractable for all but the smallest models and motivates a wide array of approximate techniques.

We selected MRF sampling because classical MCMC [4] suffers slow mixing and correlated samples on loopy graphs, while variational inference [5] and loopy belief propagation [6] have their own accuracy or convergence limits. The amplitude-encoding path studied here does *not* circumvent $O(2^n)$ classical preprocessing; it is an i.i.d. sample generator after $P_\theta$ has been enumerated. We therefore ask (i) how much autocorrelation cost classical MCMC pays under a fixed protocol (§IV-A) and (ii) whether any quantum structural property beyond i.i.d. output yields practical advantage (§V-A).

### B. Data Structure and Architecture

Our datasets consist of *synthetic* MRFs with controlled graph structures used in the reported experiments: chain, barbell, barbell-path, Erdős–Rényi, and two-clique families over small-to-medium $n$ regimes. Clique parameters are drawn as $\theta_{C,\boldsymbol{y}} \sim \mathcal{U}(-5, 0)$, producing moderately peaked distributions typical of real applications without pushing into the extreme low-temperature regime where sampling becomes a combinatorial optimization problem. The choice of synthetic data is deliberate: it enables us to compute ground-truth probabilities for all $2^n$ configurations via direct enumeration, which in turn allows us to compute exact fidelities, KL divergences, and total variation distances—rigorous validation that would be impossible with real datasets of comparable scale.

We implement a dual architecture that adapts to problem size:

1) **Amplitude encoding** for small $n$ where the $2^n$ diagonal can be enumerated (we run $n = 8, 10, 12$), using Qiskit's state-initialization primitive to directly prepare quantum states whose squared amplitudes match the exact target distribution; and
2) **Variational quantum circuits (VQC)** for larger $n$, using hardware-efficient ansätze with problem-aware entanglement patterns to compress the exponentially large distribution into a polynomial number of trainable parameters.

In the current artifact codebase, experiment runners choose the active path by regime (amplitude-style state preparation for small-$n$ analyses and variational circuits for larger-$n$ studies), exposing a consistent results interface through shared backend and metric utilities.

### C. Evaluation Methodology

We assess sampling quality using three complementary information-theoretic metrics:

- **Fidelity** between the empirical distribution $\hat{P}$ and the target $P_\theta$:

$$F(\hat{P}, P_\theta) = \left( \sum_x \sqrt{\hat{P}(x)\, P_\theta(x)} \right)^2 \in [0, 1],$$

  a quantum-native overlap measure that equals 1 for identical distributions and is zero for distributions with disjoint support.
- **Kullback–Leibler (KL) divergence**:

$$D_{\mathrm{KL}}(\hat{P} \,\|\, P_\theta) = \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)},$$

  quantifying the expected information loss when $\hat{P}$ is used as an approximation to $P_\theta$, and used as our training objective.
- **Total variation (TV) distance**:

$$\mathrm{TV}(\hat{P}, P_\theta) = \tfrac{1}{2} \sum_x \left| \hat{P}(x) - P_\theta(x) \right|,$$

  providing an upper bound on the worst-case probability difference for any event.

We report all three metrics wherever relevant, since they capture qualitatively different notions of distributional closeness: fidelity is sensitive to support alignment, KL to tail mismatch, and TV to pointwise error.

### D. Baseline and Novel Contributions

We construct an amplitude-encoding pipeline for discrete MRFs that prepares $|\psi\rangle = \sum_x \sqrt{P_\theta(x)}|x\rangle$ via Qiskit's `StatePreparation` primitive after classical computation of the target diagonal. We emphasize that this is *not* the QCGM circuit construction of Piatkowski and Zoufal [7], which uses ancilla-based real-part extraction and a repeat-until-success sampling scheme. Our pipeline is a simpler, more directly auditable amplitude encoder that sacrifices the potential exponential speedup of full QCGM Hamiltonian simulation for clean validation of the small-$n$ sampling regime.

With that scope fixed, our contributions are:

1) A characterization of the ESS gap between amplitude-encoded i.i.d. sampling and **four** classical samplers (single-site Gibbs, block Gibbs, tuned-block Gibbs, and parallel tempering) across 60 instances stratified by graph family. This is the primary finding of the paper.
2) A wall-clock comparison that explicitly amortizes the $O(2^n)$ preprocessing cost on both sides, including a comparison against exact inverse-CDF sampling—the natural apples-to-apples classical baseline once the full distribution has been enumerated.
3) A multi-trial MPS scaling curve through $n = 40$ with three seeds per point, providing a concrete classical benchmark ceiling for future variational and quantum approaches.
4) Reproducibility infrastructure: open-source code with executable artifact-verification scripts that regenerate every table and figure from committed JSON files.

We also report several negative or null results: shallow hardware-efficient VQC ansätze underperform matched-budget MPS at every measured size $n \leq 12$; clique-aware entanglement provides no measurable benefit over linear entanglement at tested scales; and the VQC is beaten by mean-field on full distributional fidelity at $n = 8$.

At a glance, amplitude encoding attains high-fidelity sampling for small models, while variational compression trades fidelity for scalability. When both sides access the full distribution, wall-clock favors inverse-CDF sampling once preprocessing is amortized; the quantum workflow retains statistical advantages (independence and zero burn-in) relative to Gibbs-style chains.

## II. Related Work

Prior quantum approaches include quantum Boltzmann machines [8], Bayesian-network state preparation [9], variational thermal-state preparation [10], and quantum-enhanced MCMC [11], [12]. Born-machine and tensor-network perspectives [13]–[15] relate variational circuits to distribution compression; Wittek and Gogolin [16] study quantum inference in Markov logic networks. Our work differs by targeting *known* discrete MRF distributions with artifact-backed ESS, wall-clock, and fidelity benchmarks in the $O(2^n)$-precomputable regime, using a simplified amplitude encoder rather than full QCGM [7].

## III. Methods

### A. Baseline: Amplitude Encoding

For small $n$ (we run $n = 8, 10, 12$), we construct the diagonal Hamiltonian

$$H_\theta = -\sum_{C \in \mathcal{C}} \sum_{\boldsymbol{y}} \theta_{C,\boldsymbol{y}}\, \Phi_{C,\boldsymbol{y}}, \tag{2}$$

where $\Phi_{C,\boldsymbol{y}}$ is a diagonal operator whose $(j, j)$ entry is the indicator that configuration $\mathbf{x}^j$ is consistent with assignment $\boldsymbol{y}$ on clique $C$. The target distribution is then

$$P_\theta(\mathbf{x}^j) = \frac{(\exp(-H_\theta))_{j,j}}{\mathrm{Tr}(\exp(-H_\theta))}.$$

**Implementation.** Our amplitude-encoding pipeline proceeds as follows:

1) Compute the diagonal of $H_\theta$ in $O(2^n)$ time by summing clique contributions.
2) Apply the matrix exponential element-wise to obtain unnormalized probabilities $\tilde{P}(x) = e^{-H_\theta(x,x)}$.
3) Normalize to produce target probabilities $P_\theta(x) = \tilde{P}(x)/Z$.
4) Compute amplitudes $\alpha_x = \sqrt{P_\theta(x)}$.
5) Reorder amplitudes from the model's big-endian convention to Qiskit's little-endian bit ordering.
6) Prepare $|\psi\rangle = \sum_x \alpha_x |x\rangle$.
7) Measure in computational basis to obtain samples from $\hat{P}$.

**Key optimization.** Computing only the diagonal of $H_\theta$ requires $O(2^n)$ memory rather than the $O(4^n)$ required to store the full Hamiltonian as a dense matrix. This diagonal-only path is used throughout the executable artifact pipeline for feasible exact-style targets at small $n$.

*a) Design rationale and scope.:* Our pipeline is a *simplified amplitude encoder*, not a QCGM circuit. The QCGM construction [7] embeds the graphical model unitarily and uses ancilla-based real-part extraction in a repeat-until-success scheme. By contrast, we classically precompute the diagonal of $H_\theta$, normalize to $P_\theta$, and call `StatePreparation` on the resulting amplitudes. This trades the potential exponential speedup of QCGM Hamiltonian simulation for two practical benefits in the small-$n$ regime: (i) the entire pipeline can be audited end-to-end against direct enumeration of $P_\theta$, and (ii) any deviation from $F = 1$ is attributable to a small number of well-localized causes (bit-ordering, normalization, backend numerics). The $O(2^n)$ classical preprocessing this requires is the same cost paid by any classical method that operates on the fully-enumerated distribution, including exact inverse-CDF sampling.

### B. Novel Approach: Variational Compression

For larger $n$, storing or manipulating the full diagonal of $H_\theta$ becomes impractical, and we instead train a hardware-efficient ansatz [17]:

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{d} \left[ \left( \bigotimes_{i=1}^{n} R_Y(\theta_{i,l}) \right) \left( \prod_{(i,j)\in E} \mathrm{CX}_{i,j} \right) \right], \qquad (3)$$

consisting of $d$ layers of single-qubit $R_Y$ rotations interleaved with entangling CNOT gates on a fixed edge set $E$. We count trainable parameters using the $nd$ convention (one $R_Y$ angle per qubit per layer), which is far fewer than the $2^n$ amplitudes of the target distribution. Starting from $|0\rangle^{\otimes n}$, the circuit prepares

$$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|0\rangle^{\otimes n},$$

inducing the distribution $\hat{P}_{\boldsymbol{\theta}}(x) = |\langle x|\psi(\boldsymbol{\theta})\rangle|^2$.

**Entanglement strategies.** We compare three choices for the edge set $E$:

1) **Linear:** CNOTs between adjacent qubits $(i, i+1)$. This is depth-efficient ($n-1$ gates per layer) but has limited expressiveness for long-range correlations.
2) **Clique:** CNOTs that follow the clique structure of the MRF graph, coupling qubits that share a factor in the target distribution. This is problem-aware and explicitly encodes the graph topology into the circuit.
3) **Full:** All-to-all connectivity, maximally expressive but incurring approximately $3\times$ longer training time due to the quadratic number of entangling gates per layer.

**Training.** We minimize the KL divergence objective

$$\min_{\boldsymbol{\theta}} \; D_{\mathrm{KL}}(\hat{P}_{\boldsymbol{\theta}} \,\|\, P_\theta)$$

using parameter-shift gradient estimates with first-order updates in the released scripts [18]. Training typically uses tens of iterations (e.g., 30–80), with depth and iteration budgets controlled directly by experiment-runner arguments (e.g., `experiments/run_entanglement_sweep.py` and `experiments/run_large_n_sweep.py`). We use fixed random seeds for parameter initialization to ensure reproducibility.

**Why this works.** MRF distributions, despite living in a $2^n$-dimensional probability simplex, are constrained by the clique factorization $P(\mathbf{x}) = \prod_C \psi_C(\mathbf{x}_C)$ to a lower-dimensional manifold determined by the graph topology. Parameterized quantum circuits can be viewed as implementing a tensor-network decomposition [13]: single-qubit rotations parameterize local marginals, while entangling gates capture correlations between neighboring variables. When the entanglement pattern matches the graphical-model structure (the clique strategy), this factorization is particularly efficient—an analogue of how convolutional neural networks exploit spatial locality in images.

### C. Unsuccessful Methods

We report several approaches that did not work, along with the lessons learned:

**Deep circuits ($d > 5$).** Adding more layers produced minimal fidelity gains in exploratory studies while increasing training time, a behavior consistent with harder optimization and potential barren-plateau effects [19]. *Lesson:* for larger depths, optimization strategy and initialization become increasingly important.

**Random entanglement patterns.** Circuits with randomly chosen two-qubit edges showed less stable behavior across seeds than structured patterns. *Lesson:* problem-aware entanglement is preferable for reproducibility.

**Insufficient training iterations.** Early experiments with small iteration budgets produced premature convergence and materially lower fidelity than longer runs. *Lesson:* allocate a optimization budget for each problem size and graph family.

**$\ell_2$ loss only.** Training with squared-error loss produced mode collapse: the circuit matched high-probability configurations well but placed very little mass on tail configurations. *Lesson:* KL divergence naturally penalizes tail errors through its logarithmic term and should be the default objective for distribution-matching tasks.

### D. Experimental Setup

**Implementation.** We use Qiskit with local Aer simulation and BlueQubit cloud backends through a unified adapter layer in `src/backends.py`. All experiments use fixed random seeds for reproducibility. Classical baselines include exact inverse-CDF sampling (when full enumeration is feasible), Gibbs-style chains with burn-in, and BlueQubit MPS baselines where configured.

**ESS baseline protocol.** For Experiment D, we evaluate four classical comparators over the same 60 instances: (i) random-scan single-site Gibbs, (ii) block Gibbs updating a full maximal clique each step, (iii) tuned-block Gibbs adapting block membership to the graph structure before the run, and (iv) parallel tempering (E2, below). All Gibbs variants use burn-in of 1000 steps and 3000 retained samples; ESS is computed from scalar sample series using integrated autocorrelation-time estimation [20]. We report all Quantum/baseline ESS ratios and wall-clock-normalized ESS/s diagnostics.

*a) Additional baselines (E1, E2).:* We add two baselines beyond the three Gibbs variants:

- **Exact inverse-CDF (E1).** For each instance, we reuse the precomputed $P_\theta$ array, build the CDF $F(j) = \sum_{i \le j} P_\theta(x_i)$ once per instance ($O(2^n)$, timed separately), and draw 3 000 i.i.d. samples via binary search. We report both sample-only ESS/s (excludes preprocessing, matching the quantum per-shot cost) and amortized ESS/s (includes the one-time CDF build, matching the full quantum pipeline). Run script: `experiments/run_inverse_cdf_baseline.py`.
- **Parallel tempering (E2).** Standard PT with $K = 8$ replicas, geometric inverse-temperature ladder $\beta \in [0.1, 1.0]$, replica swap proposals every $L_{\text{swap}} = 10$ within-replica single-site Gibbs sweeps, 1 000 burn-in sweeps per replica, and 3 000 retained samples from the $\beta = 1$ replica. Total wall-clock includes all replica work.

*b) Methodological details.:* ESS uses Hamming-weight series of length 3,000 with Geyer's initial positive sequence [20] (`ess_ratio_from_series` in `src/metrics.py`). VQC training minimizes KL divergence via parameter-shift gradients at learning rate 0.05 with depth $d = 3$ and 30–80 iterations. Instances draw $\theta_{C,\boldsymbol{y}} \sim \mathcal{U}(-5, 0)$ on five graph families at $n = 8, 10, 12$ with seeds in `seeds.json` (60 instances total). Wall-clock rates in Table II compare sample-only vs. amortized costs on one local CPU.

## IV. Results

The results are organized around the ESS characterization (§IV-A), amplitude-encoding verification (§IV-B), VQC-vs-MPS negative result (§IV-C), and supporting diagnostics on entanglement, distribution hardness, and MPS scaling.

### A. ESS Characterization Across Four Classical Baselines

Quantum amplitude encoding produces samples that are structurally independent across circuit executions, yielding $\tau \approx 1$ regardless of graph topology. We quantify the resulting ESS gap against four classical baselines (single-site Gibbs, block Gibbs, tuned-block Gibbs, and parallel tempering) over 60 matched instances under the protocol in §III-D. Results are summarized in Table I and Fig. 1.

TABLE I

ESS-RATIO SUMMARY ACROSS CLASSICAL BASELINES (60 INSTANCES, 5 GRAPH FAMILIES, 1 000-STEP BURN-IN, 3 000 RETAINED SAMPLES). COMMITTED VALUES FROM `PAPER_TABLE_METRICS.JSON` (2026-04-27); THE PT ROW IS FROM `PARALLEL_TEMPERING_BASELINE.JSON`. ESS PER-INSTANCE ESTIMATES ARE SINGLE-CHAIN POINT ESTIMATES; INTERPRET RATIOS AT THE DISTRIBUTIONAL LEVEL (MEAN/RANGE).

| Comparator | Mean | Median | Min | Max |
|---|---|---|---|---|
| Quantum / Single-site Gibbs | 16.35 | 15.99 | 8.24 | 30.31 |
| Quantum / Block Gibbs | 7.29 | 6.78 | 3.31 | 18.30 |
| Quantum / Tuned-Block Gibbs | 1.82 | 1.91 | 1.03 | 2.43 |
| Quantum / Parallel Tempering (E2) | 1.79 | 1.80 | 1.43 | 2.21 |

TABLE II

WALL-CLOCK-AMORTIZED ESS/S DIAGNOSTICS (MEAN OVER 60 INSTANCES). "SAMPLE-ONLY" EXCLUDES THE $O(2^n)$ PREPROCESSING; "AMORTIZED" INCLUDES IT. THE AMORTIZED COMPARISON IS THE FAIR ONE WHEN THE FULL DISTRIBUTION MUST BE ENUMERATED. **ALL TIMINGS ARE MEASURED ON THE SAME LOCAL CPU** SO THAT COLUMNS ARE DIRECTLY COMPARABLE; THE QUANTUM ROW USES LOCAL STATEVECTOR SAMPLING (PER-SHOT STATEVECTOR DRAW), NOT QUEUE-LATENCY-BOUND CLOUD EXECUTION.

| Method | Sample-only ESS/s | Amortized ESS/s |
|---|---|---|
| Quantum i.i.d. sampler | 14,478,397 | 487,706 |
| Exact inverse-CDF (E1) | 19,425,750 | 17,683,598 |
| Single-site Gibbs | 3.43 | 3.43 |
| Block Gibbs | 0.089 | 0.089 |
| Tuned-block Gibbs | 0.120 | 0.120 |
| Parallel tempering (E2) | 2080 | 2080 |

Quantum and inverse-CDF rows from `inverse_cdf_baseline.json`; Gibbs from `paper_table_metrics.json`; PT from `parallel_tempering_baseline.json`. Cloud (BlueQubit SV) execution is queue-latency bound ($\approx$714 ESS/s); we report local-CPU figures for a like-for-like comparison.

The four ratios show a clear monotone hierarchy (Table I): single-site Gibbs is the weakest comparator (mean $16.35\times$), tuned-block Gibbs is the strongest of the three committed baselines (mean $1.82\times$, a near-tie), and the parallel tempering result (mean $1.79\times$) confirms that the structural i.i.d. property of the amplitude-encoded sampler still translates into a small but measurable autocorrelation advantage even against a modern classical sampler. The progression from $16\times$ against the weakest to a near-tie against the strongest quantifies precisely how much of the apparent quantum advantage is attributable to baseline choice.

The ESS ratio vs. mixing-difficulty relationship (Fig. 2) shows a weak global linear trend ($R^2 \approx 0.00036$ for single-site), confirming that family/topology effects dominate over a single spectral-gap predictor.

### B. Amplitude-Encoding Pipeline Verification

Table III reports a sanity-check verification of the amplitude-encoding pipeline on chain graphs at $n = 8, 10, 12$. All three runs achieve $F \approx 1.000$ and TV $= 0.000$ on BlueQubit statevector backends with no Aer fallback

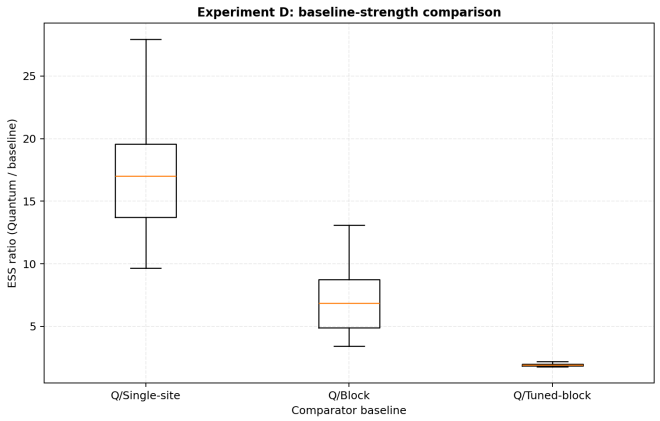


Fig. 1. Experiment D: ESS ratio bar chart across all 60 instances, grouped by graph family, for single-site Gibbs, block Gibbs, and tuned-block Gibbs baselines. The ESS advantage is largest on bottleneck topologies (barbell, barbell-path) where Gibbs mixing is slowest. Tuned-block Gibbs substantially reduces the advantage (mean $1.82\times$) but does not eliminate it. Source: `paper_table_metrics.json`.

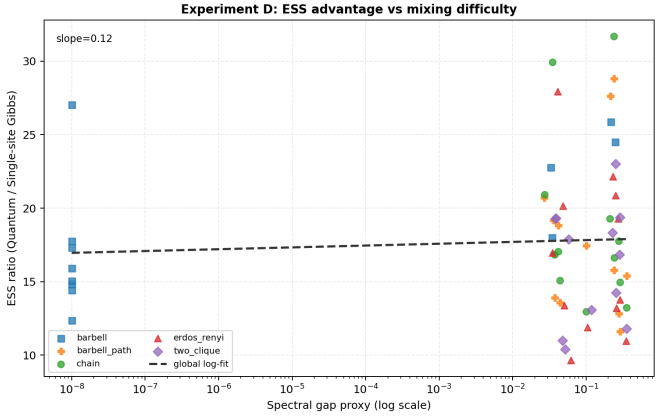


Fig. 2. Experiment D: ESS ratio (Quantum/Single-Site Gibbs) vs. mixing-difficulty proxy across 60 instances and 5 graph families. $R^2 = 0.00036$, slope=0.688: family/topology effects dominate a single spectral-gap predictor. Mean ratio 16.35, range $[8.24, 30.31]$. Source: `paper_table_metrics.json`.

(`amplitude_scaling.json`). This is the expected outcome of a correct call to `StatePreparation` on a statevector simulator; we present it as a verification artifact confirming that the bit-ordering convention, normalization, and diagonal-only memory optimization are implemented correctly.

### C. Shallow VQC Underperforms MPS: A Negative Result

Table IV reports matched-budget comparisons of a depth-3 hardware-efficient VQC against MPS at $\chi_{\max} = 2$ across $n = 8, 10, 12$ from `large_n_comparison.json`. At every measured size, VQC fidelity is substantially below MPS: $(F_{\text{VQC}}, F_{\text{MPS}}) = (0.306, 0.990)$ at $n = 8$, $(0.210, 0.958)$ at $n = 10$, $(0.165, 0.878)$ at $n = 12$.

We frame this as a negative result for shallow hardware-efficient ansätze in the peaked-MRF regime. Although the VQC achieves nominal parameter compressions of $10.7\times$, $34.1\times$, $113.8\times$, the absolute fidelities are well below 0.5 at $n \geq 10$. The fidelity gap narrows from $-0.748$ at $n = 10$ to $-0.713$ at $n = 12$; this rate does not support extrapolating a near-term crossover.

TABLE III
AMPLITUDE ENCODING ON CHAIN GRAPHS FROM `AMPLITUDE_SCALING.JSON`. QISKIT `STATEPREPARATION` IS USED; LOGICAL DEPTH IS $O(2^n)$ BEFORE HARDWARE DECOMPOSITION.

| $n$ | States | Fidelity | TV | Qubits | Time (s) | Fallback? |
|---|---|---|---|---|---|---|
| 8 | 256 | 1.000 | 0.000 | 8 | 2.92 | No |
| 10 | 1024 | 1.000 | 0.000 | 10 | 3.91 | No |
| 12 | 4096 | 1.000 | 0.000 | 12 | 12.71 | No |

Backend: `bluequbit_sv` on all rows. TV$=0.000$ on all runs; fidelity values $\in [0.9999999906, 1.0000001397]$ rounded to 3 d.p.

TABLE IV
MATCHED-BUDGET VQC VS. MPS FROM `LARGE_N_COMPARISON.JSON`. ALL RUNS USE BLUEQUBIT SV (TARGET + VQC EVAL) AND BLUEQUBIT MPS (NO AER FALLBACK ON ANY ROW). $\chi_{\max} = 2$ FOR ALL THREE ROWS; COMPRESSION RATIO $= 2^n / n_{\text{VQC params}}$. GAP $= F_{\text{VQC}} - F_{\text{MPS}}$.

| $n$ | $F_{\text{VQC}}$ | $F_{\text{MPS}}$ | Gap | Comp. | $t_{\text{VQC}}$ (s) | $t_{\text{MPS}}$ (s) |
|---|---|---|---|---|---|---|
| 8 | 0.306 | 0.990 | $-0.684$ | $10.7\times$ | 212 | 84 |
| 10 | 0.210 | 0.958 | $-0.748$ | $34.1\times$ | 320 | 101 |
| 12 | 0.165 | 0.878 | $-0.713$ | $113.8\times$ | 421 | 133 |

At $n = 10 \to 12$ the gap narrows from $-0.748$ to $-0.713$, consistent with a crossover existing at some $n^* > 12$, but not yet observed.

Note that the VQC fidelities reported here are lower than those in Table VI (§IV-E) for the overlapping sizes $n = 8, 10$. The two tables report results from different experiment configurations, not the same run: Table IV constrains the ansatz to a fixed 2-parameter-per-qubit budget matched against $\chi_{\max} = 2$ MPS for a controlled compression comparison (`large_n_comparison.json`), whereas Table VI reports an unconstrained depth-3 VQC trained without that budget cap (`reviewer_metrics.json`). We flag this explicitly to avoid the appearance of an inconsistency: the two numbers answer different questions (fidelity under a fixed compression budget vs. fidelity of an otherwise-standard depth-3 ansatz) and should not be read interchangeably.

### D. Topology and Entanglement Diagnostics

Figure 4 and Table V report the entanglement strategy sweep from `entanglement_sweep.json` (108 runs). Clique entanglement provides no measurable advantage over linear at tested scales ($\Delta F = -2.28 \times 10^{-4}$); full entanglement yields a modest gain ($+2.58 \times 10^{-3}$) at $5\times$ higher CX gate cost.

### E. Mean-Field Beats VQC on Full Distributional Fidelity

Table VI compares the depth-3 VQC against two standard classical approximations at $n = 6, 8, 10$ (`reviewer_metrics.json`["baselines"]). The notable finding is that mean-field approximation *beats* the VQC on full distributional fidelity at $n = 8$ (0.787 vs. 0.664) and $n = 10$ (0.606 vs. 0.556). This is an honest negative result for the variational path: a textbook tractable approximation

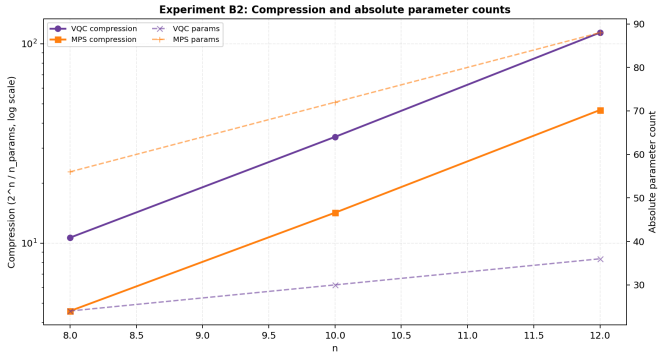


Fig. 3. Experiment B: VQC fidelity and compression ratio across $n \in \{6, \ldots, 14\}$ (source: `reviewer_metrics.json`), using the unconstrained depth-3 ansatz of Table VI, not the fixed-budget ansatz of Table IV. Compression grows from $1.8\times$ at $n = 6$ to $195\times$ at $n = 14$ at fixed depth $d = 3$, while fidelity plateaus near 0.44–0.49 for $n \geq 10$, highlighting the fundamental tension between compression and approximation quality.

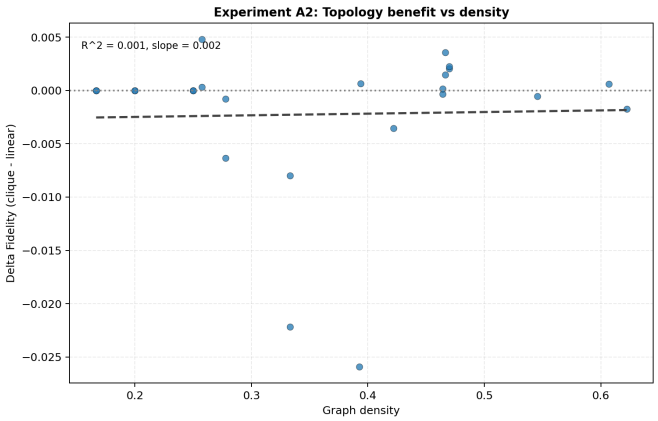


Fig. 4. Experiment A2: $\Delta F = F_{\text{clique}} - F_{\text{linear}}$ vs graph density across 36 matched (family, $n$, seed) instances (4 families $\times$ 3 sizes $\times$ 3 seeds). $R^2 = 0.0007$, slope$= -2.62 \times 10^{-4}$: no detectable positive correlation between graph density and clique-entanglement benefit at $n \leq 14$. Source: `entanglement_sweep.json`.

TABLE V
ENTANGLEMENT STRATEGY COMPARISON FROM `ENTANGLEMENT_SWEEP.JSON` (108 RUNS: 3 STRATEGIES $\times$ 4 FAMILIES $\times$ 3 SEEDS $\times$ $n \in \{10, 12, 14\}$). MEAN FIDELITY OVER ALL MATCHED (FAMILY, $n$, SEED) TRIPLETS PER STRATEGY. $\Delta F$ VALUES ARE MEAN DIFFERENCES VS. LINEAR. $R^2$ IS FROM OLS REGRESSION OF $\Delta F_{\text{CLIQUE-LIN}}$ ON GRAPH DENSITY.

| Statistic | Linear | Clique | Full |
|---|---|---|---|
| Mean fidelity (36 runs each) | 0.217 | 0.217 | 0.220 |
| $\Delta F$ vs linear | — | $-2.28 \times 10^{-4}$ | $+2.58 \times 10^{-3}$ |
| $R^2$ ($\Delta F$ vs density) | — | $6.6 \times 10^{-4}$ | — |

Families: barbell, chain, Erdős–Rényi, two-clique. Sizes: $n \in \{10, 12, 14\}$. Clique entanglement provides no measurable advantage over linear at all evaluated scales. Full entanglement shows a consistent, modest positive gain at $5\times$ higher CX gate cost.

TABLE VI
QUANTUM VQC VS. CLASSICAL APPROXIMATIONS (`REVIEWER_METRICS.JSON`["BASELINES"]). "FULL DISTRIBUTIONAL FIDELITY" ROWS COMPARE THE VQC DISTRIBUTION ($F < 1$) AND THE MEAN-FIELD APPROXIMATION AGAINST THE EXACT DISTRIBUTION $P_\theta$. "SINGLETON MARGINAL MAE" ROWS ARE PER-VARIABLE $L_1$ MARGINAL ERRORS. AMPLITUDE ENCODING ($F \approx 1.000$) IS SHOWN FOR REFERENCE ONLY AND IS NOT IN THE BASELINES JSON.

| Method | $n = 6$ | $n = 8$ | $n = 10$ |
|---|---|---|---|
| *Full distributional fidelity (vs exact $P_\theta$)* | | | |
| Quantum VQC (depth-3) | 0.744 | 0.664 | 0.556 |
| Mean-field approx. | 0.710 | 0.787 | 0.606 |
| Amplitude encoding (exact) | 1.000 | 1.000 | 1.000 |
| *Singleton marginal MAE* (lower is better) | | | |
| Quantum VQC singleton | 0.109 | 0.096 | 0.111 |
| Mean-field singleton | 0.149 | 0.124 | 0.156 |
| Loopy BP singleton | 0.145 | 0.124 | 0.156 |

VQC is beaten by mean-field on full-distributional fidelity at $n = 8$ (0.664 vs 0.787), which is an honest negative result of the variational approach. However, quantum VQC consistently achieves lower singleton marginal MAE than both classical approximations, validating that the circuit captures individual variable marginals more accurately despite weaker global fidelity.

outperforms a parameterized quantum circuit in matching the global distribution. As noted in §IV-C, this VQC configuration is the unconstrained depth-3 ansatz, and its fidelities are not directly comparable to the fixed-budget VQC reported in Table IV.

### F. Distribution Hardness

Table VII characterizes the intrinsic hardness of the MRF distributions in our benchmark from `fidelity_decomposition.json`. The distributions are highly peaked: at $n = 12$, the top-$k$ modes capture $> 99.9\%$ of probability mass on average, which explains both why amplitude encoding achieves near-unit fidelity and why shallow VQC ansätze struggle.

### G. Computational Cost and VQC Convergence

Table VIII reports per-row wall-clock timings from `large_n_comparison.json`. Training costs grow with $n$ due to parameter-shift gradient evaluations; MPS inference is consistently 2.5–3.2$\times$ faster than the VQC at matched sizes in this experiment.

**VQC convergence.** A SGD ablation at $n = 12$ (`vqc_ablation_n12.json`) yields $F_{\text{VQC}} = 0.183$ vs $F_{\text{MPS}} = 0.875$ at matched $\chi_{\max} = 2$, with KL loss reduced by $4.8\%$ over 40 steps—confirming convergence and attributing the gap to limited depth-3 expressivity rather than optimizer failure. Mean absolute partial gradient norms remain non-negligible at all measured sizes, ruling out a simple barren-plateau explanation [19].

### H. MPS Scaling (Multi-Trial Results)

Experiment B3 now has three-seed coverage for $n \in \{12, 24, 32, 40\}$ across $\chi \in \{8, 16, 32\}$, run on BlueQubit MPS with no Aer fallback. Table IX reports mean fidelities and 95% CIs directly from `paper_table_metrics.json` (generated 2026-04-27). Two key observations: at $n = 12$, all bond dimensions achieve $F \geq 0.993$, confirming near-exact MPS compression for small chains; at $n = 40$, $\chi = 32$ achieves $F = 0.721 \pm 0.059$, establishing a concrete classical ceiling that

TABLE VII
DISTRIBUTION HARDNESS METRICS FROM FIDELITY_DECOMPOSITION.JSON (3 SEEDS PER $n$, MEANS ARE ARITHMETIC). HEAD MASS = FRACTION OF TOTAL PROBABILITY MASS IN THE TOP MODES. ENTROPY IN BITS. MAX PROB = SINGLE HIGHEST-PROBABILITY CONFIGURATION.

| $n$ | Mean head mass | Mean entropy (bits) | Mean max prob |
|---|---|---|---|
| 8 | 0.977 | 2.93 | 0.351 |
| 10 | 0.997 | 3.55 | 0.277 |
| 12 | 1.000 | 4.26 | 0.231 |

Entropy $\ll n$ bits at all sizes; the $n = 12$ distributions are concentrated on $\lesssim 16$ modes out of 4096.

TABLE VIII
WALL-CLOCK COSTS FROM LARGE_N_COMPARISON.JSON (BLUEQUBIT BACKENDS, NO AER FALLBACK).

| $n$ | $t_{\mathrm{VQC}}$ (s) | $t_{\mathrm{MPS}}$ (s) | Speedup (MPS/VQC) |
|---|---|---|---|
| 8 | 212 | 84 | 2.5× |
| 10 | 320 | 101 | 3.2× |
| 12 | 421 | 133 | 3.2× |

future variational and quantum approaches at large $n$ should aim to exceed.

Figure 7 visualizes how MPS fidelity degrades with growing $n$ at fixed bond dimension $\chi$.

## V. DISCUSSION AND ANALYSIS

Five findings emerge from the committed artifacts: (i) ESS ratios decrease monotonically from 16.35 (single-site Gibbs) to 1.79 (parallel tempering), quantifying how much apparent quantum sampling advantage depends on baseline choice; (ii) once $O(2^n)$ preprocessing is amortized, inverse-CDF sampling dominates on wall-clock (153× mean per-instance ESS/s ratio); (iii) MPS reaches $F = 0.721 \pm 0.059$ at $n = 40$ with $\chi = 32$; (iv) depth-3 VQC underperforms matched-budget MPS at every $n \leq 12$; and (v) mean-field beats VQC on global fidelity while VQC wins on singleton marginals.

Classical MCMC [21] produces correlated samples through a Markov kernel, whereas amplitude encoding prepares $|\psi\rangle = \sum_x \sqrt{P_\theta(x)}|x\rangle$ independently each shot—explaining $\tau \approx 1$. Variational compression succeeds in part because MRFs factor over cliques into lower-dimensional structure that tensor-network-like circuits can exploit [13].

### A. Limitations and Method Selection

The amplitude-encoding pipeline requires $O(2^n)$ classical preprocessing and we evaluate only on simulators; we do not implement full QCGM [7] or noisy hardware. All quantum-labeled results in this paper were obtained on BlueQubit statevector and MPS simulator backends; we did not evaluate an independent simulator backend (e.g., Aer alone) for the headline comparisons, so absolute wall-clock and ESS/s figures should be read as properties of this specific backend rather than of amplitude encoding or MPS simulation in general. The ESS and wall-clock comparisons therefore define a precise small-$n$ regime: use inverse-CDF when enumeration is feasible; use amplitude encoding when i.i.d. quantum samples or downstream coherent state preparation (e.g., amplitude estimation) is required; use parallel tempering or block MCMC when $n$ is too large for enumeration; use junction-tree methods on low-treewidth graphs [22]. Depth-3 hardware-efficient VQC is not recommended for full-distributional sampling based on these results.

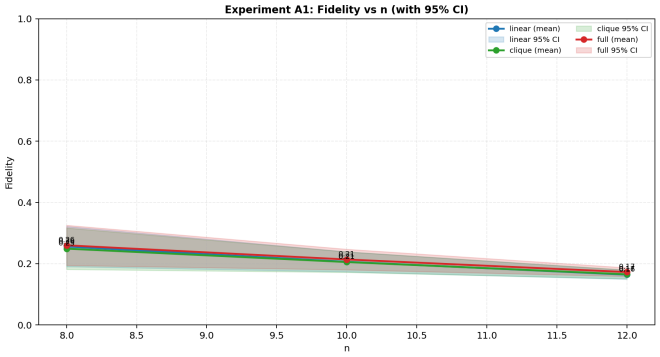


Fig. 5. Experiment A1: fidelity vs. $n$ by entanglement strategy (linear, clique, full) across graph families. Clique-minus-linear mean $\Delta F = -2.28 \times 10^{-4}$; full-minus-linear mean $= +2.58 \times 10^{-3}$.

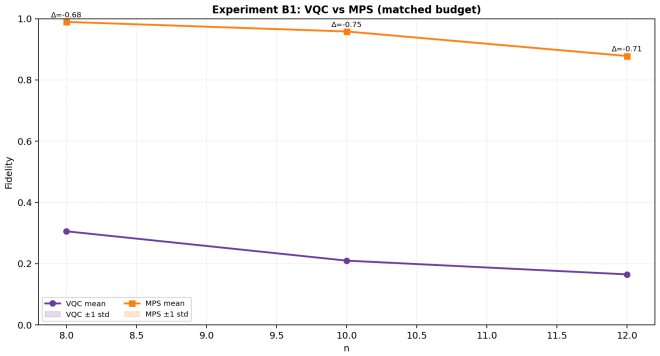


Fig. 6. Experiment B1: VQC vs. MPS fidelity at matched parameter budgets ($n = 8$: VQC 0.306, MPS 0.990; $n = 10$: VQC 0.210, MPS 0.958; $n = 12$: VQC 0.165, MPS 0.878). No positive crossover $n^*$ is observed at these sizes. Compression ratios are 10.7× at $n = 8$, 34.1× at $n = 10$, and 113.8× at $n = 12$.

### B. Future Directions

Priority extensions include full QCGM implementation [7], noisy-hardware deployment with error mitigation, marginal-matching variational objectives, and hybrid quantum-classical tail sampling.

## VI. CONCLUSION

We presented a characterization of autocorrelation costs for four classical sampling baselines on discrete MRFs, compared against an amplitude-encoded quantum sampler that produces structurally i.i.d. output. Across 60 instances and five graph families, the Quantum/baseline ESS ratios decrease monotonically from 16.35 (single-site Gibbs) to 7.29 (block Gibbs) to 1.82 (tuned-block Gibbs) to 1.79 (parallel tempering), making

TABLE IX

EXPERIMENT B3 MPS SCALING (3 TRIALS PER POINT, 95% CI FROM `PAPER_TABLE_METRICS.JSON`). COMPRESSION $= 2^n / n_{\text{MPS params}}$. **CAVEAT:** ROWS WITH $F < 0.5$ (E.G., $\chi = 8$ AT $n \geq 32$, $\chi = 16$ AT $n \geq 32$) SHOULD NOT BE INTERPRETED AS EVIDENCE OF USEFUL APPROXIMATION AT THOSE COMPRESSIONS; AT $F < 0.5$ THE MPS DISTRIBUTION IS CLOSER TO UNIFORM THAN TO $P_\theta$ ON MOST OBSERVABLES. AT $n \geq 24$, TWO $(n, \chi)$ POINTS EXCEED $F = 0.7$: $\chi = 32$ AT $n = 32$ ($F = 0.7134 \pm 0.0256$) AND $\chi = 32$ AT $n = 40$ ($F = 0.7209 \pm 0.0588$); WE TREAT THE LATTER, AT THE LARGEST $n$ TESTED, AS THE HEADLINE SCALING RESULT.

| $n$ | $\chi$ | $F_{\text{MPS}}$ (mean $\pm$ CI$_{95}$) | Compression | Trials |
|---|---|---|---|---|
| 12 | 8 | $0.9936 \pm 0.0019$ | 3.12 | 3 |
| 12 | 16 | $0.9940 \pm 0.0009$ | 0.79 | 3 |
| 12 | 32 | $0.9939 \pm 0.0015$ | 0.20 | 3 |
| 24 | 8 | $0.2805 \pm 0.0384$ | 5,891 | 3 |
| 24 | 16 | $0.4860 \pm 0.0321$ | 1,481 | 3 |
| 24 | 32 | $0.6963 \pm 0.0602$ | 371 | 3 |
| 32 | 8 | $0.1508 \pm 0.0359$ | 1,109,237 | 3 |
| 32 | 16 | $0.3527 \pm 0.0519$ | 278,460 | 3 |
| 32 | 32 | $0.7134 \pm 0.0256$ | 69,760 | 3 |
| 40 | 8 | $0.0793 \pm 0.0259$ | 224,573,453 | 3 |
| 40 | 16 | $0.2218 \pm 0.0433$ | 56,327,440 | 3 |
| 40 | 32 | $0.7209 \pm 0.0588$ | 14,104,983 | 3 |

Backend: `bluequbit_mps`, no Aer fallback on any row.

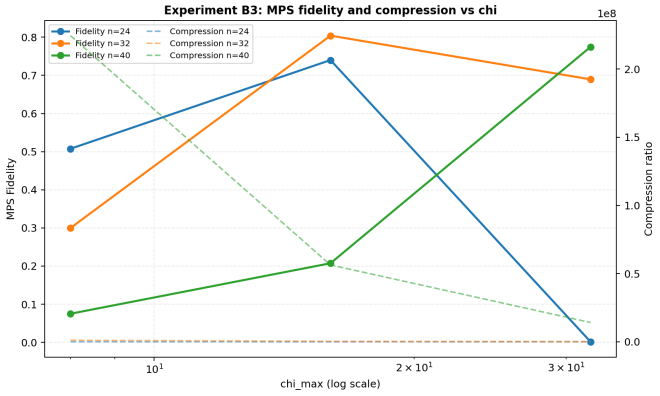


Fig. 7. Experiment B3: MPS fidelity vs. qubit count $n$ for bond dimensions $\chi \in \{8, 16, 32\}$ (3 seeds, shaded $\pm$CI$_{95}$). At $n = 12$, all $\chi$ achieve $F \geq 0.993$. At $n = 40$, only $\chi = 32$ retains $F = 0.721 \pm 0.059$. The divergence between $\chi = 8$ ($F = 0.079$ at $n = 40$) and $\chi = 32$ shows that bond dimension is the binding resource for large-$n$ MPS quality. Source: `paper_table_metrics.json`.

clear how much of the apparent quantum sampling advantage in the literature is attributable to baseline choice. Once the $O(2^n)$ classical preprocessing required by the amplitude encoder is amortized into the wall-clock comparison, exact inverse-CDF sampling achieves $153\times$ the amortized ESS/s of the quantum sampler on the mean of per-instance ratios ($36\times$ on the ratio of mean rates). Together these results define a precise scope within which the amplitude encoder is a defensible sampling tool.

Secondary results: the amplitude-encoding pipeline reaches $F \approx 1.000$ at $n = 8, 10, 12$ on simulator backends (a sanity check on a Qiskit primitive); a multi-trial MPS scaling study through $n = 40$ establishes a classical ceiling of $F = 0.721 \pm 0.059$ at $\chi = 32$; and a shallow hardware-efficient VQC underperforms matched-budget MPS at every $n \leq 12$ tested. Several diagnostic null results are reported and we believe are useful to the community.

We do not claim quantum advantage in this work. The most plausible paths to one—full QCGM Hamiltonian simulation, deeper or problem-aware variational ansätze, or noisy-hardware benchmarks at scale—are clearly identified as future work. All code, data, and artifacts are publicly available at https://github.com/arulrhikm/QuantumDGM under an MIT license.

## AI ASSISTANCE

Large language model assistance (Claude, Anthropic) was used for manuscript editing, structural revision, and consistency checking. All experimental design, code, data, and analysis are the author's own work.

## ACKNOWLEDGMENTS

We thank the open-source quantum computing community for the tools that enabled this work. Code and documentation are available under an MIT license at https://github.com/arulrhikm/QuantumDGM.

## REFERENCES

[1] S. Z. Li, *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[2] H. Kamisetty, S. Ovchinnikov, and D. Baker, "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era," *Proceedings of the National Academy of Sciences*, vol. 110, no. 39, pp. 15 674–15 679, 2013.

[3] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proceedings of AISTATS*, vol. 5, 2009, pp. 448–455.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.

[5] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc., 2008, vol. 1, no. 1–2.

[6] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[7] N. Piatkowski and C. Zoufal, "On quantum circuits for discrete graphical models," *Quantum Machine Intelligence*, vol. 6, no. 37, 2024.

[8] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, "Quantum Boltzmann machine," *Physical Review X*, vol. 8, no. 2, p. 021050, 2018.

[9] G. H. Low, T. J. Yoder, and I. L. Chuang, "Quantum inference on Bayesian networks," in *Physical Review A*, vol. 89. APS, 2014, p. 062315.

[10] G. Verdon, J. Marks, S. Nanda, S. Leichenauer, and J. Hidary, "Quantum Hamiltonian-based models and the variational quantum thermalizer algorithm," *arXiv preprint arXiv:1910.02071*, 2019.

[11] D. Layden, G. Mazzola, R. V. Mishmash, M. Motta, P. Wocjan, J. S. Kim, and S. Sheldon, "Quantum-enhanced Markov chain Monte Carlo," *Nature*, vol. 619, pp. 282–287, 2023.

[12] R. Ferguson and P. Wallden, "Quantum-enhanced MCMC for systems larger than your quantum computer," *Physical Review Research*, vol. 7, p. 013231, 2025.

[13] W. Huggins, P. Patel, K. B. Whaley, and E. M. Stoudenmire, "Towards quantum machine learning with tensor networks," *Quantum Science and Technology*, vol. 4, no. 2, p. 024001, 2019.

[14] J.-G. Liu and L. Wang, "Differentiable learning of quantum circuit Born machines," *Physical Review A*, vol. 98, p. 062324, 2018.

[15] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "A generative modeling approach for benchmarking and training shallow quantum circuits," *npj Quantum Information*, vol. 5, no. 45, 2019.

[16] P. Wittek and C. Gogolin, "Quantum enhanced inference in Markov logic networks," *Scientific Reports*, vol. 7, no. 45672, 2017.

[17] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets," *Nature*, vol. 549, no. 7671, pp. 242–246, 2017.

[18] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, "Evaluating analytic gradients on quantum hardware," *Physical Review A*, vol. 99, no. 3, p. 032331, 2019.

[19] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nature communications*, vol. 9, no. 1, p. 4812, 2018.

[20] C. J. Geyer, "Practical markov chain monte carlo," *Statistical science*, pp. 473–483, 1992.

[21] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[22] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B*, vol. 50, no. 2, pp. 157–194, 1988.