Early Fault-Tolerant Quantum Circuits Probabilistic Graphical Models

Arul Rhik Mazumder (arulm), Bryan Zhang (bryanzha)

1 Background and Motivation

Probabilistic graphical models (PGMs) are a cornerstone of modern machine learning, providing a unifying framework for expressing conditional dependencies among variables. Markov Random Fields (MRFs), in particular, define energy-based distributions of the form

$$P_\theta(x) = \frac{1}{Z(\theta)} \exp\left(\sum_{C \in \mathcal{C}} \theta_C \, \phi_C(x_C) \right),$$

where $\mathcal{C}$ denotes the set of maximal cliques in the graph and $\phi_C(x_C)$ encodes local potentials. MRFs power applications ranging from image reconstruction and protein folding to generative models such as Boltzmann machines. However, inference and sampling in MRFs remain NP-hard in general, as the partition function $Z(\theta)$ and the associated marginals scale exponentially with the number of variables. Classical approximate methods such as Markov Chain Monte Carlo (MCMC), Gibbs sampling, or perturb-and-MAP (PAM) often suffer from slow mixing, parameter sensitivity, or biased samples.

Recent progress in quantum information science offers new tools for probabilistic inference. Because the amplitudes of an n-qubit state intrinsically represent a 2^n-dimensional probability distribution, quantum computers can, in principle, perform exponentially large linear-algebraic transformations in parallel in a single shot without costly burn-in times. Unfortunately most existing quantum approaches to probabilistic modeling like quantum Boltzmann machines [2], quantum Bayesian networks [3], or annealing-based Gibbs samplers [4] either require fully fault-tolerant hardware or rely on variational training heuristics that lack theoretical guarantees.

Piatkowski and Zoufal (2022) bridged this gap by introducing *Quantum Circuits for Graphical Models* (QCGM), the first exact quantum sampling algorithm for discrete factor models [7]. QCGM constructs a unitary embedding of the MRF Hamiltonian

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

where each sufficient statistic $\Phi_{C,y}$ is encoded as a product of Pauli-Z projectors following a "Pauli-Markov" construction. The resulting circuit C_θ generates unbiased and independent samples whose distribution matches $P_\theta(x)$. Remarkably, QCGM satisfies a unitary Hammersley-Clifford theorem: the circuit factorizes over the cliques of the underlying graph, mirroring the structure of classical MRFs. Experimental validation on IBM Falcon processors confirmed the method's correctness and demonstrated parameter learning via hybrid optimization.

Despite these advances, QCGM faces three major limitations for early fault-tolerant (EFT) quantum devices:

(i) **Ancilla overhead:** the construction requires $1 + |\mathcal{C}|$ auxiliary qubits, one per clique, for real-part extraction, which restricts scalability beyond small graphs.

(ii) **Exponential success decay:** the probability of obtaining a valid sample, $\delta_* = \prod_{C \in \mathcal{C}} \delta_C$, decreases exponentially with the number of cliques.

(iii) **Circuit depth:** the embeddings $U_{C,y}(\theta_{C,y})$ entail sequential real-part extractions; implementing these naively exceeds EFT coherence times.

This proposal aims to extend QCGM toward a resource-efficient, hybrid framework suitable for EFT platforms (50–200 logical qubits, low-depth gates, minimal error-correction). The goal is to preserve the model's exact representational fidelity while reducing ancilla use and runtime via classical post-processing and adaptive control.

2 Research Objectives and Methodology

This project aims to design low-ancilla, shallow-depth quantum circuits that can faithfully model and sample from Markov Random Fields (MRFs) while leveraging hybrid classical post-processing to handle stochastic failures. The research addresses three central goals: (i) reducing ancilla requirements through circuit compression, (ii) improving sampling efficiency via hybrid quantum-classical correction, and (iii) approximating exponential maps through variational compression.

Circuit Compression and Representation

Following QCGM, each binary variable x_v corresponds to one qubit. To avoid the $1+|\mathcal{C}|$ ancilla overhead required for real-part extraction, the proposed design employs a shared ancilla line with dynamic measurement and reset. After processing clique C_i, the ancilla is measured and, if successful ($|0\rangle$), reused for subsequent cliques. This "repeat-until-success" strategy maintains unbiased sampling while reducing ancilla scaling from $O(|\mathcal{C}|)$ to $O(1)$ [8, 9].

Each clique-factor operation will be implemented through a controlled rotation

$$U_{C,y}(\theta_{C,y}) = e^{i\theta_{C,y} P_{C,y}},$$

where $P_{C,y}$ is a sparse tensor product of Pauli-Z operators. Commuting factors on disjoint cliques can be parallelized, yielding total circuit depth $O(\chi(G))$, where $\chi(G)$ is the graph's chromatic number. This provides an architecture compatible with early fault-tolerant (EFT) devices having tens to hundreds of qubits.

Hybrid Sampling and Post-Processing

Because QCGM's success probability δ_* decreases exponentially with the number of cliques, naive repetition quickly becomes inefficient. To address this, we employ hybrid post-processing techniques that extract unbiased statistics from imperfect samples:

- *Conditional Rejection Sampling:* retain both successful and failed outcomes, reweighting them according to estimated success probabilities.

Additionally, error mitigation strategies such as zero-noise extrapolation and readout correction [5, 10, 11] help stabilize the estimates without requiring full fault-tolerant operation.

Variational Compression of Exponential Maps

The exact operator $\exp(-H_\theta)$ induces large Trotter depths. We will replace it with a parameterized quantum circuit $V(\gamma)$ of fixed depth L, trained to match low-order marginals of P_θ. The loss

$$\mathcal{L}(\gamma) = \sum_{(C,y)} \|\langle \phi_{C,y} \rangle_{V(\gamma)} - \langle \phi_{C,y} \rangle_{P_\theta} \|^2$$

optimizes local correlation fidelity rather than global partition functions, allowing shallow, noise-resilient approximations. This forms a compressed, variational analogue of QCGM suited to EFT coherence times.

3 Expected Results and Scientific Impact

The project is expected to produce:

R1. Low-Ancilla Circuit Designs: Efficient QCGM-style encodings using $O(n+1)$ qubits via ancilla reuse and local factor parallelization.

R2. Hybrid Inference Methods: Classical correction schemes transforming imperfect quantum samples into unbiased estimators.

R3. Variationally Compressed Models: Fixed-depth PQCs preserving key statistical moments while achieving exponential depth reduction.

R4. Quantitative Benchmarks: Systematic comparisons between hybrid quantum-classical inference and classical MCMC baselines.

By demonstrating scalable, interpretable quantum modeling of MRFs, this work advances the integration of quantum computing with probabilistic machine learning.

4 Current Progress and Next Steps

Our current work has focused on implementing the modeling and sampling framework of Piatkowski and Zoufal [7], reproducing their Quantum Circuits for Graphical Models (QCGM) as a foundation for this project. We have verified that the circuit correctly represents discrete Markov Random Fields via a unitary embedding of the energy function, establishing a baseline for investigating extensions that improve resource efficiency on early fault-tolerant devices. This implementation confirms that QCGM can generate correct samples on small instances and provides the platform for testing new ancilla-reduction and hybrid sampling methods.

The next phase will extend this baseline to more resource-efficient designs. We plan to implement shared-ancilla architectures with mid-circuit measurement and reset to sequentially reuse qubits, and to develop hybrid post-processing schemes that combine quantum sampling with classical reweighting to counteract the exponential decay in success probability δ_*. A further direction will explore variational compression of $\exp(-H_\theta)$ using parameterized quantum circuits trained to reproduce low-order marginals, targeting shallow-depth, noise-tolerant approximations. Together, these steps will transition our current QCGM replication into a scalable, resource-aware framework for probabilistic inference on early fault-tolerant quantum hardware.

References

[1] H. Abraham et al. *Qiskit: An Open-Source Framework for Quantum Computing*, 2019.

[2] M. H. Amin et al. *Quantum Boltzmann Machine*. *Phys. Rev. X*, 8 (2018).

[3] G. H. Low et al. *Quantum Inference on Bayesian Networks*. *Phys. Rev. A*, 89 (2014).

[4] J. Nelson et al. *High-Quality Thermal Gibbs Sampling with Quantum Annealing Hardware*, 2021.

[5] S. Bravyi et al. *Mitigating Measurement Errors in Multiqubit Experiments*. *Phys. Rev. A*, 103(4), 2021.

[6] M. Niepert, P. Minervini, L. Franceschi. *Implicit MLE: Backpropagating Through Discrete Exponential Families*. *NeurIPS 34*, 2021.

[7] N. Piatkowski and C. Zoufal. *On Quantum Circuits for Discrete Graphical Models*. arXiv:2206.00398, 2022.

[8] A. Paetznick and K. M. Svore. *Repeat-Until-Success: Non-deterministic decomposition of single-qubit unitaries*. Quantum Inf. Comput., 14 (2014).

[9] A. Bocharov, M. Roetteler, K. M. Svore. *Efficient Synthesis of Universal Repeat-Until-Success Quantum Circuits*. Phys. Rev. Lett. 114, 080502 (2015).

[10] Z. Cai et al. *Quantum error mitigation*. Rev. Mod. Phys. 95, 045005 (2023).

[11] B. Nachman et al. *Unfolding quantum computer readout noise*. npj Quantum Inf. 6, 200 (2020).

[12] X. Yuan, S. Endo, Q. Zhao, Y. Li, S. C. Benjamin. *Theory of Variational Quantum Simulation*. *Quantum*, 3 (2019).