In-Detail Paper Representation: Comparing Quantum and Classical Generative Models on Molecular Data

Devansh Shah (202411032)
Kathan Vora (202411073)

Dhirubhai Ambani University (DAU)

Presented to M.V. Joshi & Bhavin Makwana
May 10, 2025

Motivation & Background

- **Generative models** enable the discovery of novel molecules and designs:
  - Drug discovery: propose candidate compounds
  - Materials design: search for optimal properties
- **Classical approaches** (VAEs, GANs, Transformers, RNNs) have advanced this field
- **Quantum sampling** holds promise:
  - Potential for exploring complex distributions
  - Intrinsic randomness may yield diverse, low-cost solutions
- **Key question:** Can quantum generative models outperform classical ones under realistic resource budgets?

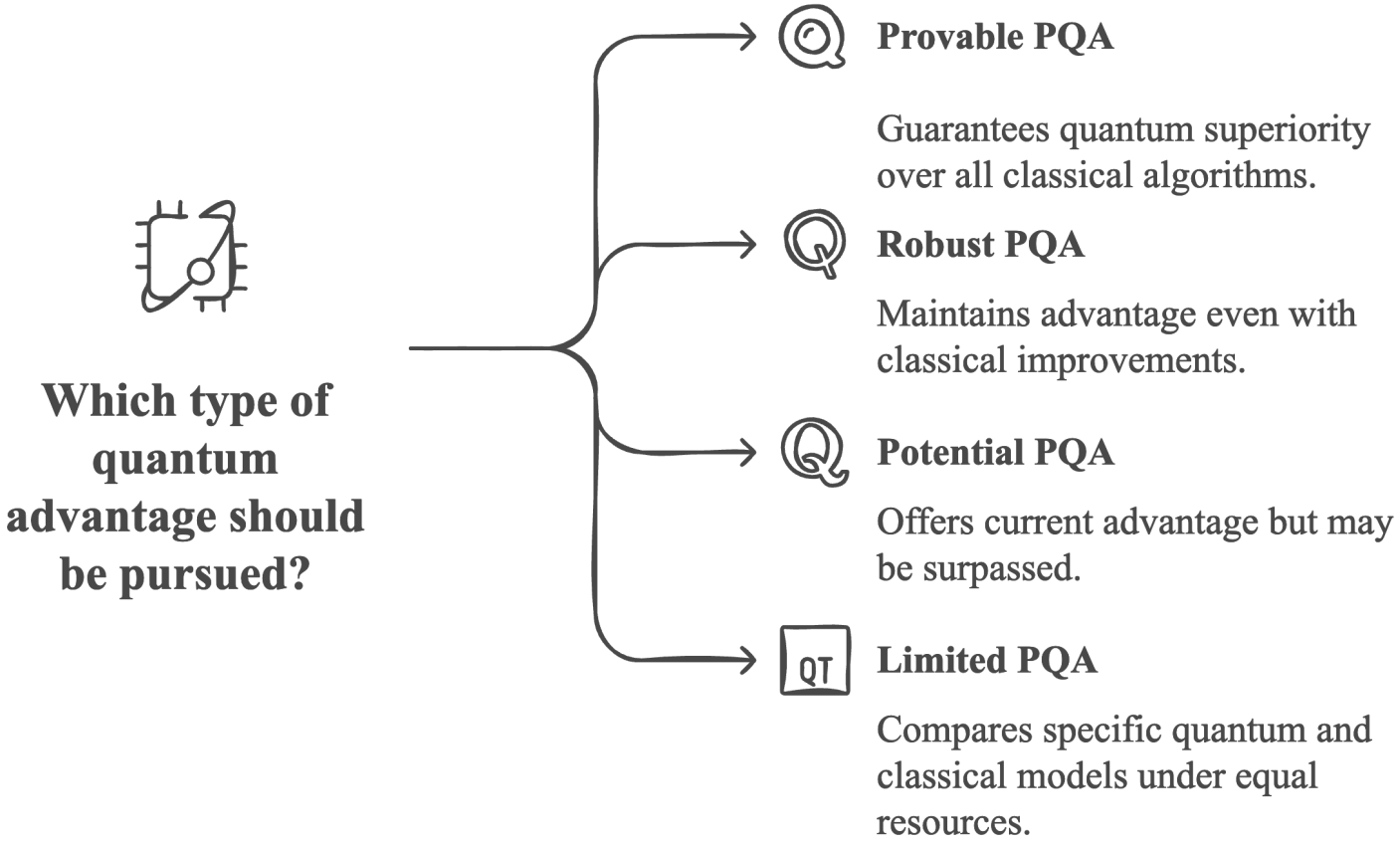
Practical Quantum Advantage (PQA)

Definition

Practical Quantum Advantage (PQA) measures when a quantum generative model outperforms or has the potential to outperform the best classical alternatives under realistic resource constraints.

- Focuses on *useful* advantage, not just asymptotic speedups
- Compares *sampling cost*, *model expressivity*, and *quality of generated solutions*
- Enables a transparent "race" between quantum and classical runners on defined *tracks*

Types of Practical Quantum Advantage



Race Analogy & Benchmark Tracks

Race Analogy

- Models as "runners" in a race
- **Strengths**: expressivity, sample diversity
- **Weaknesses**: sampling cost, training time
- Fair competition under resource constraints

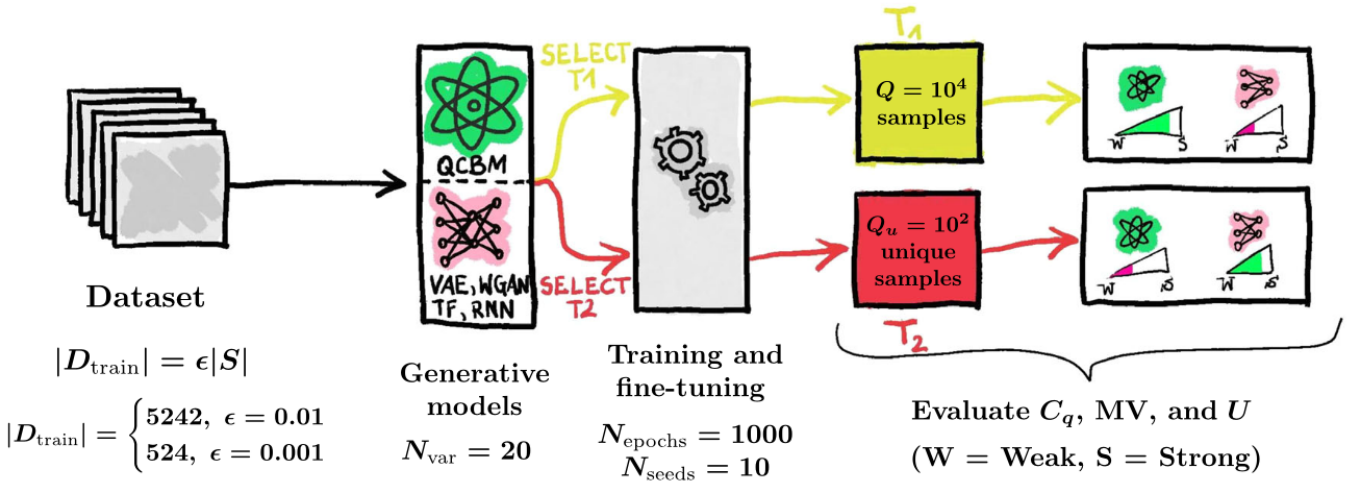
Benchmark Tracks

Track 1 (T1) Fixed total samples $Q = 10^4$ (high sampling cost scenario)

Track 2 (T2) Fixed unique valid samples $Q_u = 10^2$ (high evaluation cost scenario)

Reference: M. Hibat-Allah, M. Mauri, J. Carrasquilla & A. Perdomo-Ortiz, *A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models.*

Quality-Based Generalization Scheme



Reference: M. Hibat-Allah, M. Mauri, J. Carrasquilla & A. Perdomo-Ortiz, *A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models.*

Dataset Used in the Paper

- **Synthetic "Evens" Distribution**
  - 20-bit binary strings with an even-parity constraint
  - Cost = "negative separation" score on bit patterns
- **Data Regimes**
  - $\epsilon = 0.01$: $\approx 5{,}242$ training samples
  - $\epsilon = 0.001$: ≈ 524 training samples
- **Training Setup**
  - 1,000 epochs per model
  - 10 random seeds for averaging
  - Hyperparameter tuning via Optuna

Reference: M. Hibat-Allah, M. Mauri, J. Carrasquilla & A. Perdomo-Ortiz, *A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models.*

Key Results

- **Data-Rich Regime ($\epsilon = 0.01$)**
  - *Track 1 (T1)*: VAE achieves best utility; QCBM close second
  - *Track 2 (T2)*: WGAN leads on MV, Utility, and Coverage
- **Scarce-Data Regime ($\epsilon = 0.001$)**
  - *Track 1 (T1)*: QCBM outperforms all classical in Utility; matches or exceeds in MV and Coverage
  - *Track 2 (T2)*: QCBM VAE tie on MV/Utility; QCBM achieves highest Coverage

Reference: M. Hibat-Allah, M. Mauri, J. Carrasquilla & A. Perdomo-Ortiz, *A framework for demonstrating practical quantum advantage: comparing quantum against classical generative models.*

Molecular Dataset and Training Approach

1. **Our Dataset**
   - Molecules represented as SMILES strings (e.g., `CCCS(=O)c1ccc2[nH]c(=NC(=O)OC)[nH]c2c1`).
   - SMILES format encodes molecular structure in a text-based notation.
2. **Fingerprint Conversion**
   - Used RDKit to convert SMILES into molecular fingerprints.
   - Fingerprints are binary vectors (0s and 1s) representing molecular features.

Molecular Dataset and Training Approach (Contd.)

3. **Dimensionality Reduction**
   - Original fingerprints had high dimensionality (e.g., 1024 bits).
   - Applied PCA to reduce to 16 key features for efficient processing.
   - Ensures compatibility with quantum models (e.g., 16 qubits for QCBM).
4. **Data Normalization**
   - Scaled the 16 features between 0 and 1.
   - Standardized data for better training and model convergence.

QCBM: Circuit Structure – Initialization & First Layer

1. Initialization and Parameterized Rotation Layer:

- All 16 qubits are initialized in the ground state:

$$0^{\otimes 16} = 000\ldots0$$

- The circuit begins by applying parameterized single-qubit gates (RX and RZ) to each qubit:
  - RX: Rotation around the X-axis.
  - RZ: Rotation around the Z-axis (affects phase).
- These gates introduce tunable parameters that define a quantum superposition across all 2^{16} basis states.
- At this stage, the system is in a rich, parameter-dependent quantum state ready for entanglement.

QCBM: Circuit Structure – Entanglement & Rotation

2. **Entanglement Layer:**
   - CNOT gates are applied between qubits in a circular pattern: $q_0 \rightarrow q_1 \rightarrow \cdots \rightarrow q_{15} \rightarrow q_0$
   - This introduces **quantum correlations** across features.
   - Helps model non-linear dependencies in data.

3. **Final Rotation Layer:**
   - Another set of RX gates are applied to all qubits.
   - These gates are **trainable parameters** that shape the final output distribution.
   - Prepares qubit states for measurement.

QCBM: Training Process

Training Objective:

- The quantum circuit outputs a probability distribution over 2^{16} possible bitstrings.
- Goal: make the generated distribution similar to the real molecular fingerprint distribution.

Loss Function – KL Divergence:

- Measures difference between real distribution P and generated distribution Q
- Lower KL divergence = more accurate generation.

QCBM: Optimization & Sampling

Parameter Optimization:

- RX angles in the final layer are updated via **Adam optimizer**.
- **Parameter-shift rules** compute gradients compatible with quantum simulators.

Sampling from Trained Circuit:

- After training, the circuit is measured.
- Each qubit collapses to a classical bit (0 or 1).
- A sample is a 16-bit string — interpreted as a new molecular fingerprint.

VAE: Architecture and Structure

VAE Structure:

- Consists of an encoder and decoder.
- The encoder compresses fingerprints into a lower-dimensional latent space.

Latent Space:

- Captures essential molecular features.
- Enables smooth interpolation between different molecules.

VAE: Training and Data Generation

Training Process:

- Data flows through the encoder, producing latent variables.
- The decoder reconstructs the fingerprint from latent space.
- Loss comprises Binary Cross-Entropy (for reconstruction) and KL Divergence (for latent space regularization).

Data Generation:

- New samples are generated by decoding random noise from the latent space.
- Output fingerprints are post-processed (rounded) to yield valid binary vectors.

WGAN: Model Structure

WGAN Structure:

- Two networks: Generator and Critic.
- The Generator maps random noise to molecular fingerprints.

Critic:

- Evaluates how realistic the generated fingerprints are by assigning continuous scores.
- Unlike a binary classifier, it provides a smooth gradient for training.

WGAN: Training and Data Generation

Training Process:

- The critic is updated multiple times per generator update using Wasserstein loss.
- Weight clipping is applied to ensure training stability.

Data Generation:

- The trained Generator converts random noise into new molecular fingerprints.
- Outputs are processed (rounded) to form valid binary vectors.

How Do We Measure Model Performance?

Evaluation Metrics:

- **MV (Minimum Value):** The lowest cost among all valid, previously unseen generated samples, reflecting the single best solution found by the model.
- **U (Utility):** The average cost of the top 5% lowest-cost generated samples, capturing the quality of the best fraction of the output distribution.
- **Cq (Quality Coverage):** The fraction of generated samples whose cost is strictly below the minimum cost observed in the training set, measuring how extensively the model uncovers new high-quality solutions.

Performance Results

Frac = 0.10%

Model	MV	U	Cq
QCBM	-13.00	-12.40	0.99
VAE	-9.00	-7.70	0.91
WGAN	-13.00	-12.10	1.00

Frac = 1.00%

Model	MV	U	Cq
QCBM	-14.00	-11.90	0.99
VAE	-9.00	-7.20	0.93
WGAN	-13.00	-11.80	1.00

Conclusions

- **MV (Best Sample):** QCBM achieves the lowest MV in both scenarios (-13 and -14), matching or exceeding WGAN's best-case solutions.
- **Utility (Top 5%):** QCBM slightly outperforms WGAN in the tighter budget (Frac 0.10%), and maintains strong utility at higher sample sizes.
- **Quality Coverage (Cq):** WGAN attains perfect coverage (1.00) in both experiments; QCBM closely follows (0.99), indicating both uncover nearly all high-quality regions.
- **Comparative Insight:** VAE underperforms on all metrics, especially MV, highlighting classical VAEs' limitations in extreme low-data regimes.
- **Implication for PPQA:** QCBM demonstrates robust performance under tight sampling/data constraints, supporting its potential quantum advantage in realistic molecular design tasks.

Thank You

Presented by:
Devansh Shah (202411032)
Kathan Vora (202411073)

Dhirubhai Ambani University (DAU)