

# Quantum-Enhanced Data Mining for Subgroup Discovery

Arpita Howlader<sup>1</sup> and Dr. Chi-Ren Shyu<sup>1,2</sup>

<sup>1</sup>MU Institute for Data Science and Informatics. <sup>2</sup>Department of Electrical Engineering and Computer Science, University of Missouri

## Motivation & Research Objectives

**Subgroup discovery (SD)** identifies meaningful patient groups with shared characteristics and unusually different outcomes. A **contrast pattern** is a combination of measurements that is common in one subgroup but uncommon in another. Traditional SD struggles to explore large combinatorial search as **subgroup rules and measurement combinations increase** and overlook subgroup structures due to heuristic pruning and computational limit.

**Why Quantum:** Classical exhaustive search evaluates patterns individually, where as Quantum Approximate Optimization Algorithm (QAOA) represents feasible measurement combinations in superposition and uses interference to concentrate sampling on high-scoring candidates.

**Central question:** Does constrained QAOA explore **high-quality** biomedical contrast-pattern to explain subgroup and recover high-ranking patterns that perform better than random selection?

- Design a **hybrid quantum–classical** subgroup-discovery framework.
- Identify **meaningful contrast patterns** that distinguish subgroup pairs.
- Formulate contrast mining** as a quantum combinatorial optimization problem.
- Validate and evaluate** quantum-generated candidates against classical ground truth.

## Methods QCMSD Framework

**Quantum contrast Mining (QCM)** places measurement-pattern selection on the quantum circuit; patient-level records remain classical.

**Data preprocessing & Subgroup :** Data cleaned and **P** defining subgroups, **M** searched for contrast patterns, **Y** for screening/validation. Multivariable rules (length 2–5) built from **P** under feasibility constraints;

**Classical contrast-pattern mining [1]:** Exhaustive enumeration provides ground truth; beam search and equal-budget random search use the same candidate budget as QAOA

**QAOA formulation for QCM:** For each subgroup pair, measurement conjunctions of length  $k = 2, 3, 4, 5$ . QAOA searches for a combinatorial space of  $\binom{M}{k}$  possible measurement combinations.

The cost Hamiltonian encodes this score, while the XY mixer preserves the fixed- $k$  constraint. The QAOA state is optimized over classical parameters  $\gamma, \beta$ , with cost Hamiltonian  $H_C$  and mixer  $H_M$ :

$$H_M = \frac{1}{2} \sum_{(i,j)} (X_i X_j + Y_i Y_j) \quad | \psi(\gamma, \beta) \rangle = \prod_{\ell=1}^p e^{-i\beta_{\ell} H_M} e^{-i\gamma_{\ell} H_C} | \psi_0 \rangle$$

$$\text{Support}_i(cp) = \frac{|\{x \in SG_i : x \text{ satisfies } cp\}|}{|SG_i|}$$

$$\text{support} \geq 0.20 \text{ minimum support } (\alpha)$$

$$\text{Growth}(cp) = \frac{\max(\text{Support}_1(cp), \text{Support}_2(cp)) + \epsilon}{\min(\text{Support}_1(cp), \text{Support}_2(cp)) + \epsilon}$$

$$\text{growth} \geq 2.0 \text{ minimum growth } (\beta)$$

QAOA-generated bitstrings are decoded into candidate patterns and exact support, growth, and significance validation.

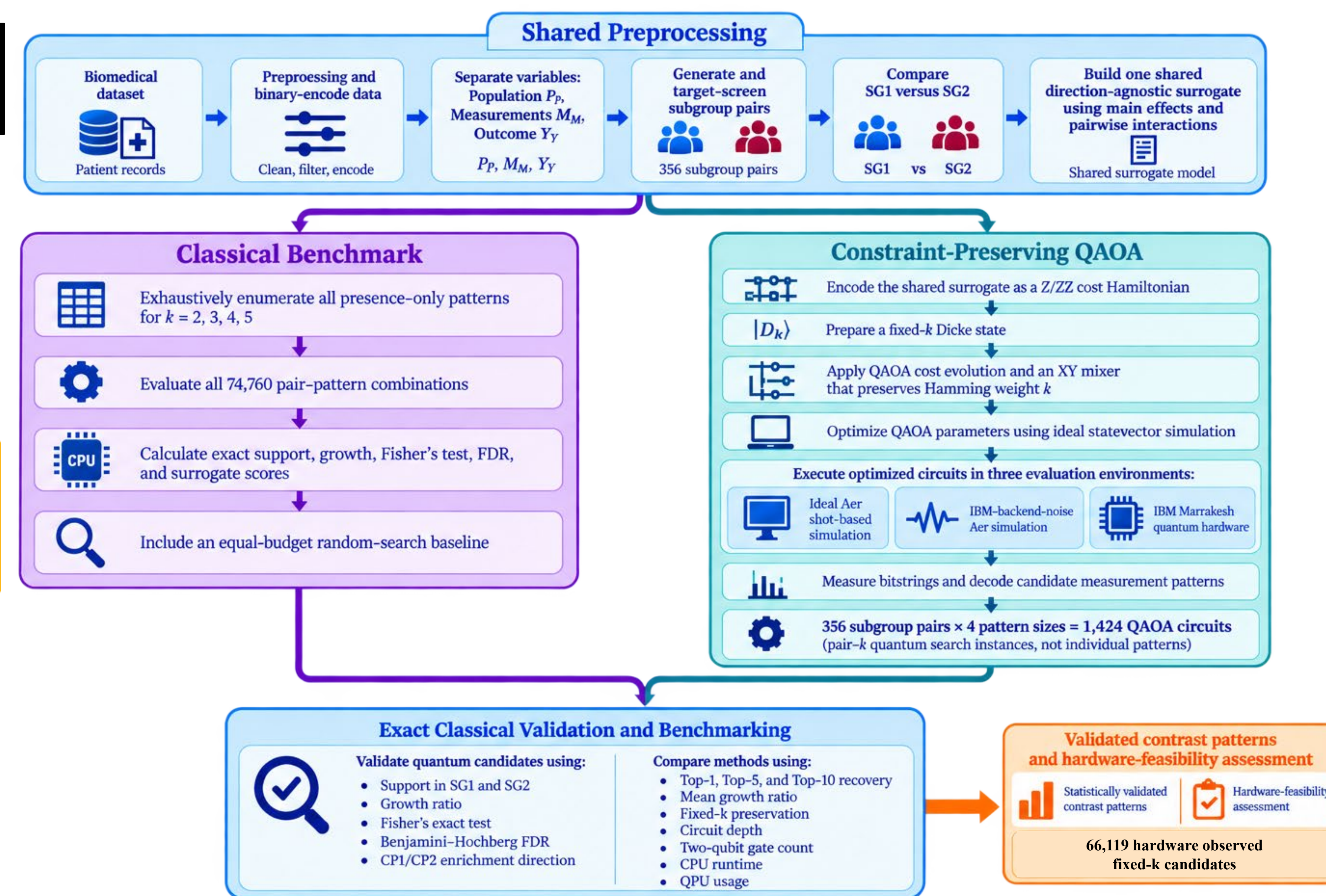


Figure 1. End-to-end QCM (Quantum Contrast Mining for Subgroup Discovery) workflow. Patient-level data remain classical; the quantum component searches the measurement-selection space, and all sampled candidates return to the classical dataset for exact validation.

## Results & Outcomes

**Lung-cancer prototype (309 records) total 931 feasible subgroup pairs generated, 356 retained** after outcome-based screening.

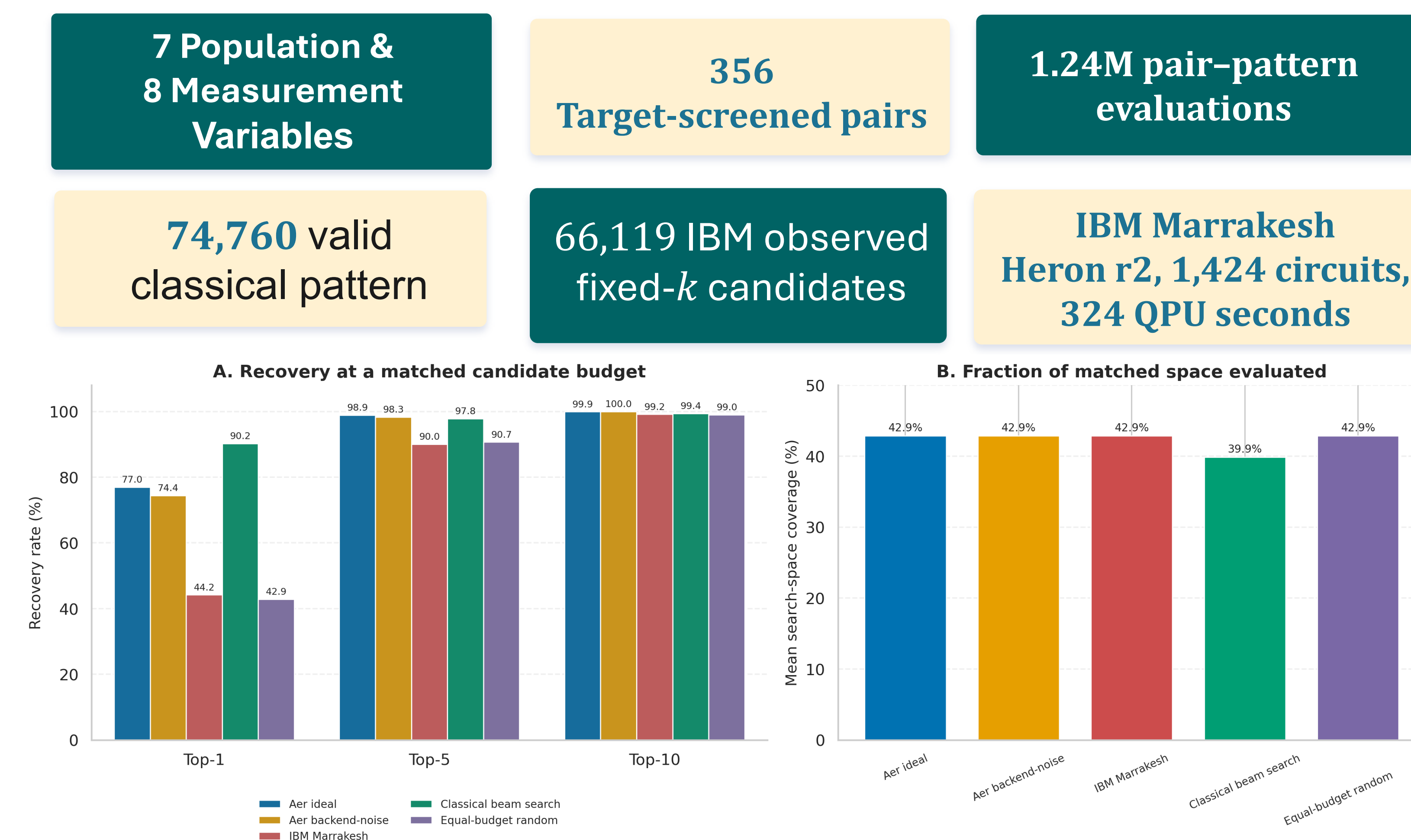


Figure 3. (A). Top-r recovery is the share of pair- $k$  instances returning at least one exhaustive top-r pattern; (B) coverage is the fraction of the matched candidate space evaluated.



Quantum Innovation Center  
University of Missouri



Institute for  
Data Science & Informatics  
University of Missouri

## QAOA Formulation for QCM

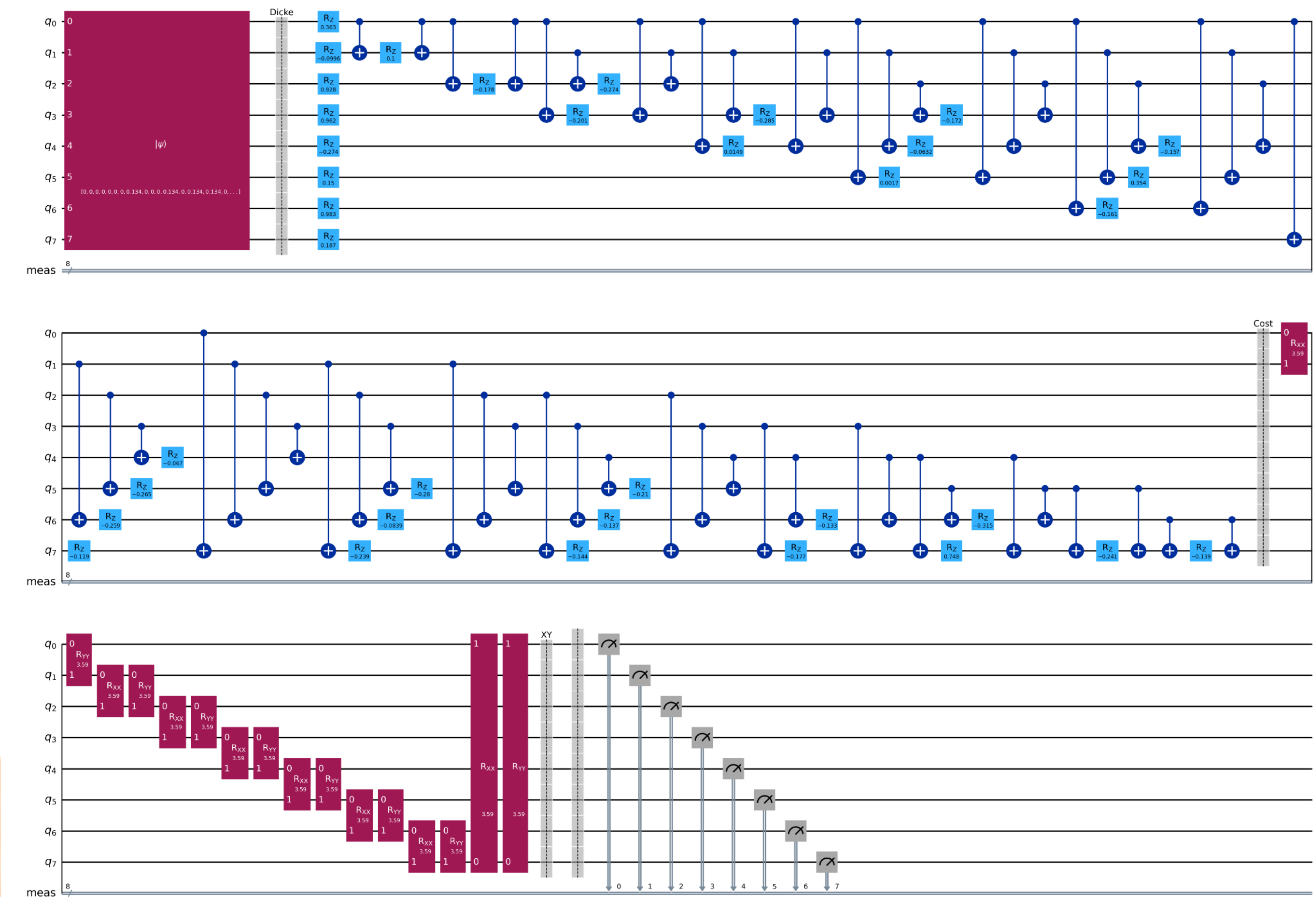


Figure 2. Representative eight-qubit circuit ( $p=1, k=3$ ): Dicke-state initialization prepares feasible subsets, the Z/ZZ cost layer encodes the contrast surrogate, and the XY mixer preserves Hamming weight before measurement. Backend: IBM Marrakesh (Heron r2), 1,424 circuits, jobs: 37, QPU: 324s.

## Impact, Challenges & Future work

- ✓ QCMSD demonstrates how quantum optimization can be integrated into a hybrid workflow for exploring large combinatorial measurement-pattern spaces
- ✓ Results establish hardware feasibility and high-ranking pattern recovery

- Translating contrast-pattern mining into a meaningful quantum objective.
- Managing deep Dicke-state preparation and numerous two-qubit interactions.
- Ensuring fair comparisons with classical and random-search baselines.

- Develop quantum-assisted generation and pruning of subgroup pairs.
- Implement a hardware-efficient Dicke-state preparation circuit.
- Test deeper QAOA, shot-based parameter optimization, and error-mitigation.

### Acknowledgements

The author acknowledges the Mizzou Quantum Innovation Center (QIC) and the University of Missouri Institute for Data Science and Informatics for supporting this research, and thanks the faculty mentor and QIC / IBM Quantum collaborators for their guidance, training, and IBM access.

Abbreviations: QPU – Quantum Processing Unit; FDR – False Discovery Rate; **P/M/Y** – population / measurement / outcome variables.

### References:

1. Liu, D., Baskett, W., Beversdorf, D., & Shyu, C. R. (2020). Exploratory data mining for subgroup cohort discoveries and prioritization. *IEEE Journal of Biomedical and Health Informatics*, 24(5), 1456–1468.
2. Farhi, Edward & Goldstone, Jeffrey & Gutmann, Sam. (2014). A Quantum Approximate Optimization Algorithm.
3. S. Hadfield et al., (2019) "From the Quantum Approximate Optimization Algorithm to a Quantum Alternating Operator Ansatz," *Algorithms*, vol. 12, no. 2, p. 34.