

Towards Maximum Independent Set Inspired Dataset Reductions with Neutral Atom Quantum Computers

Zirui Zhang
Department of Computer Science
Purdue University
West Lafayette, USA
zhan4192@purdue.edu

Pravin Mahendran
Department of Physics
The Ohio State University
Columbus, USA
mahendran.12@buckeyemail.osu.edu

Kevin Zhang
Department of Computer Science
Purdue University
West Lafayette, USA
zhan4196@purdue.edu

David E. Bernal Neira
Davidson School of Chemical Engineering
Purdue University
West Lafayette, USA
dbernaln@purdue.edu

Abstract—Quantum computers have gained popularity in recent years due to their potential to solve problems currently difficult for classical computers to tackle. This paper presents a low-dimensional proof-of-concept for dataset reduction using neutral-atom hardware. Our method constructs a graph from data geometry, estimates MIS-inspired inclusion scores from hardware measurements, and uses a classical postprocessing heuristic to build reduced training subsets. We compare our approach with classical dataset reduction techniques—k-means clustering and classical MIS—on various datasets and models. Across six synthetic 2D datasets and two larger low-dimensional benchmarks, the method is competitive with k-means and classical MIS on several tasks, with the clearest gains appearing at smaller subset sizes in selected nonlinear cases. These results suggest feasibility of hardware-assisted subset scoring on neutral-atom platforms, rather than a general performance advantage.

Index Terms—quantum computing, neutral atom, maximum independent set, dataset reduction, machine learning.

I. INTRODUCTION

Quantum computing has emerged as a new field of study in recent years with the potential to solve complex problems that are currently difficult for classical computers [1]–[3]. Despite this potential, practical applications of quantum computers remain under development. In the current Noisy Intermediate-Scale Quantum (NISQ) era, different quantum architectures have been explored for their practicality and efficiency [4]. One such architecture, the neutral atom quantum computer, has displayed notable capabilities. Using Rydberg atom arrays, this architecture has demonstrated superlinear speedup performance relative to simulated annealing in solving complex classical graph-based problems [5], [6].

Machine-learning applications of neutral-atom hardware remain relatively unexplored [7]. Our work demonstrates the potential of a novel application for neutral atom quantum computers in dataset reduction in machine learning. We utilize the

Rydberg atom platform to compute an approximate Maximum Independent Set (MIS) of graph problems constructed from datasets, which we refer to as QMIS. MIS is traditionally considered an NP-complete problem, but in this context, it corresponds to selecting a representative and diverse subset of data points used in the training of the machine learning model [8]. Our work aims to study a low-dimensional proof-of-concept for hardware-assisted subset scoring rather than a general-purpose dataset-reduction framework.

As part of this algorithm, we present a postprocessing step that aims to extract approximate MIS solutions from the probability distributions measured by the shots of the measurements, keeping the classical postprocessing lightweight relative to the subset-construction stage. It should be noted however, end-to-end execution on current cloud-accessed hardware remains slower than classical MIS solvers. The postprocessing algorithm uses a classical sort and sweep method to identify a configuration of various data points, allowing the raw distribution data to be used as an approximate MIS graph construction for the selection of important data.

Our experiments suggest that QMIS is competitive with clustering-based reduction on several low-dimensional tasks, with the clearest gains appearing at smaller subset sizes in selected cases; the advantage is not uniform across datasets or classifiers. Notably, in complex datasets where classical data reduction methods can cause a drastic skew in decision boundaries, our method preserves representative data points, leading to better model accuracy and generalization.

While prior works have explored adiabatic evolution for weak classifiers, results remain mostly theoretical [9]. However, recent works on quantum preconditioning have demonstrated promising results that make classical problems easier [10]. Our work builds on the quantum preprocessing idea, presenting a new potential use of neutral atom quantum computers

in classical task applications, showcasing a potential method in machine learning workflow through quantum dataset reduction and preprocessing.

Our contributions in this work include:

- Proof-of-concept encoding of low-dimensional dataset reduction as a unit-disk-graph problem on neutral-atom hardware
- Classical postprocessing heuristic that converts hardware measurement statistics into size-(k) training subsets.
- Empirical comparison against k-means and classical MIS on eight datasets.

The complete code can be found within our GitHub repository <https://github.com/kevin-ys-zhang/quantum-enhanced-classification>.

II. BACKGROUND

This section introduces the neutral atom quantum computing platform and highlights its potential in solving the MIS problem. Additionally, we discuss the motivation behind our work and present the importance of dataset reduction in machine learning, focusing on current classical methods.

A. Neutral Atom Platform

Analog quantum computers define their logic in a way entirely different from those of gate-based quantum computers. One example is neutral atom platforms, which utilize arrays of neutral atoms manipulated via optical tweezers and laser fields [11], [12]. In the case of QuEra's Aquila, rubidium is the neutral atom of choice [13], [14]. The architecture offers several potential advantages over others, particularly in terms of scalability and the ability to form highly controllable large-scale qubit arrays [15].

On neutral atom platforms, qubits are encoded in each atom's internal energy state. The atoms in the Rydberg state interact via long-range dipole-dipole interaction, known as a Rydberg blockade. This blockade effect prevents nearby neutral atoms from being simultaneously excited to Rydberg states, enabling entanglement and multi-qubit interactions [16].

We represent our problem with a Hamiltonian as in Eq. 1. Then, by tuning the interaction and energy level of the qubits, these systems naturally evolve towards the lowest energy ground state, representing the solution to our problem. For our work, we utilized QuEra's Aquila, a 256-qubit neutral atom quantum computer, which defines the dynamic control Hamiltonian as:

$$H = \sum_i \frac{\Omega(t)}{2} (e^{i\phi(t)} |g_i\rangle \langle r_i| + e^{-i\phi(t)} |r_i\rangle \langle g_i|) - \sum_i \Delta(t) n_i + \sum_{i < j} V_{ij} n_i n_j, \quad (1)$$

which describes a system interaction between ground $|g_i\rangle$ and excited $|r_i\rangle$ state between each particle indexed by i [17]. $\Omega(t)$ defines the time-dependent Rabi frequency, and $\Delta(t)$ defines the time-dependent detuning, forming the two hyperparameters that control the Rydberg blockade radius for MIS, similar to how classical MIS utilizes Euclidean distance to define a set

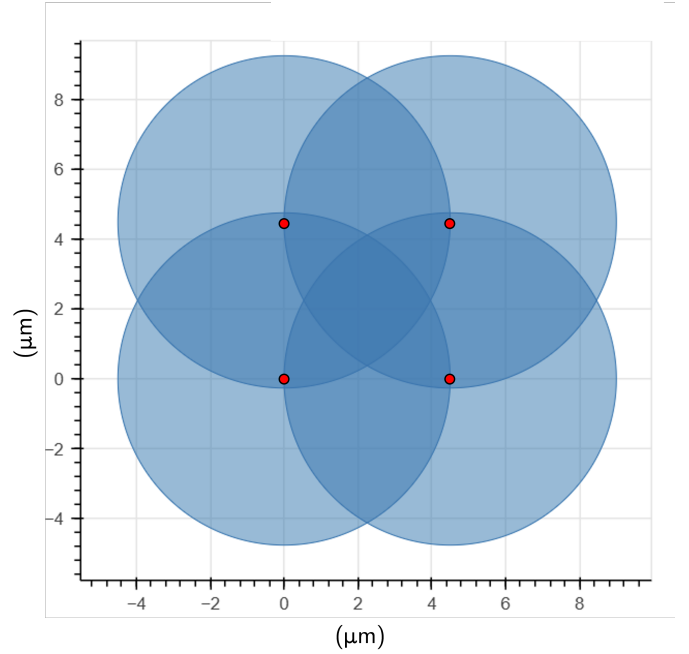


Fig. 1: Square lattice with four nodes, each with $4 \mu\text{m}$ lattice spacing, with blockade radius of $4.5 \mu\text{m}$ shown in blue and node center in red.

[17]. After measurement, a probability distribution of excited atoms is returned.

B. Maximum Independent Set

The Maximum Independent Set (MIS) problem is a form of combinatorial optimization problem in graph theory [8]. Traditionally, it has relevance in network design, resource allocation, and various optimization scenarios where mutual exclusivity is a constraint. Given an undirected graph $G = (V, E)$, an independent set is a subset of vertices $S \subseteq V$ such that no two vertices in S are connected by an edge in E . The MIS is the largest possible independent set within G . MIS is known to be an NP-hard problem, making it exponentially expensive in the worst case to solve for large graphs [18], [19].

C. MIS on Neutral Atom Platform

Many optimization-based implementations have been proposed for solving graph-based problems, though with limited practicality [20]–[23]. Neutral atom quantum computers, on the other hand, compute approximate solutions to MIS by mapping the problem onto the physical interactions of Rydberg atoms. This approach utilizes the Rydberg blockade effect to enforce the local independence constraint of the MIS directly through atom interactions [24], [25].

We encode each vertex as an atom and each graph edge as a pair of atoms within the blockade radius. Under this convention, a measured Rydberg excitation corresponds to a selected independent-set vertex, since adjacent vertices cannot be simultaneously excited. We therefore estimate per-vertex inclusion likelihoods from excitation probabilities measured over

repeated shots. For *Aquila*, we use the effective blockade-based adjacency threshold defined:

$$R_b = \left(\frac{C_6}{\sqrt{\Omega^2 + \Delta^2}} \right)^{1/6}, \quad (2)$$

where Ω and Δ are the time-dependent Rabi frequency and detuning, respectively, and C_6 is the van der Waals interaction coefficient. In this work, we use $C_6 = 2\pi \times 862690$ (in units consistent with Ω , Δ , and R_b). The system's Hamiltonian H ground state, demonstrated by Eq. 1, can be used to correspond to the MIS of a given graph, where adjacent atoms simultaneously being in the excited state will be penalized, enforcing the independence condition. In practice, we use Eq. 2 as the effective adjacency threshold for graph construction under the chosen *Aquila* control settings.

The lattice is initialized in a superposition of all possible atomic states and evolves toward the ground state of H adiabatically [17], [26], similar to adiabatic optimization models [27], [28]. Afterward, measurements are performed to determine the state of each atom. An atom found in the excited state will be included in the final set to form the maximum independent set. Due to the probabilistic nature of the result and the heuristic nature of the solution algorithm, multiple shots are performed to obtain the final probability distribution.

D. Dataset Reduction

As datasets grow in size and complexity, training machine learning models has become increasingly resource-intensive. Dataset reduction aims to offer a potential solution by selecting or generating a representative subset of the original dataset while maintaining usability. Unlike feature reduction, in dataset reduction, the data does not decrease in dimensionality.

Traditionally, clustering-based techniques, such as k -mean clustering, have been used to reduce the size of the dataset. This technique partitions the data into k clusters based on feature similarity, using cluster centroids as representative data points. Other cluster techniques, such as hierarchical clustering, build nested clusters by merging or splitting existing ones. All of these techniques aim to compute for effective dataset reduction, reducing model training time and resource consumption. Furthermore, dataset reduction helps prevent overfitting by eliminating redundant or noisy data [29], [30].

E. MIS for Dataset Reduction

MIS can be interpreted as a proxy for graph reduction, specifically graph sparsification [31]. Similar to many existing sparsification techniques, given a graph, the objective of the algorithm is to create a subset using existing nodes from the original set [32]–[35]. Specifically, we use MIS as a graph-based diversity heuristic for subset selection: given a graph whose edges connect nearby data points, an independent set yields a subset of mutually separated points. In this work we test whether that separation signal helps preserve classifier performance after reduction.

Our work utilizes a partial MIS distribution sample for dataset reduction, presenting a novel approach that seeks to select maximally independent data points while preserving classification performance. Points with high similarity, and thus low distance, are connected by an edge. By ensuring that no two selected data points are similar, the approximate MIS distribution provides a specialized subset that eliminates as many closely related data points as possible. Highly unique data points are retained as they will exist farther away from other data points.

Analytically, MIS yields a subset of mutually separated points under the chosen graph construction. We hypothesize that this separation signal can help preserve classifier performance after reduction, especially when decision boundaries depend on retaining diverse support points. Whereas clustering focuses on similarity, outliers, or unique samples that can be merged into clustered points, MIS focuses on independence, prioritizing the selection of dissimilar data points that capture a broader range of features. The sampled excitation probabilities provide a hardware-derived ranking signal that we use as a heuristic proxy for node importance in subset construction.

III. METHODOLOGY

In this section, we outline the methodology used to create the data reduction algorithm for computing the MIS using neutral atom quantum computing. First, the dataset is mapped onto a 2D lattice structure, defining adjacency between data points based on a chosen distance limit. This mapping allows the dataset to be transformed into a graph suitable for the selected hardware lattice mapping, where each data point corresponds to a lattice site. Next, the radius of the Rydberg blockade is set using the previously mentioned Eq. (2), as demonstrated in [17]. Adjusting the blockade radius effectively controls the adjacency in the quantum system, mirroring the constraints of a classical MIS algorithm. Following the results returned by the quantum hardware, a postprocessing algorithm extracts potential MIS solutions, which we describe in detail in the Postprocessing subsection.

A. Lattice Mapping

The first step involves generating two-dimensional binary data as the neutral atom quantum computing platform arranges Rydberg atoms in a two-dimensional plane. Each data point in our dataset contains two features, each representing a continuous value within a specified range. This two-dimensional representation allows us to map each data point directly onto the physical 2D lattice of atoms, with their positions reflecting the data point's feature values. The two-dimensional features of each data point are represented as (x, y) coordinates on a plane and directly mapped to the corresponding positions in the two-dimensional lattice of the neutral atom platform. For each dataset, we min-max scale the two selected features into hardware coordinates, enforce the $4-(\mu)m$ minimum spacing by global upsampling when needed, and invoke partitioning only when the qubit-count or spatial constraints are exceeded. This lattice represents the physical arrangement of atoms in

the quantum system. By placing each data point as an atom at a specific location in the lattice, a spatial representation of the dataset is created, which is usable for quantum hardware.

The blockade radius represents the range over which atoms interact through the Rydberg blockade effect. This effect is demonstrated by excited atoms that prevent nearby atoms from being excited due to strong interactions. For our purpose, the blockade radius serves as the distance threshold for defining adjacency in the graph, determining the maximum distance at which two atoms, representing data points, are considered adjacent. Similarly to classical MIS algorithms, adjacency is often defined based on a distance metric or similarity measure encoded into constraints.

In this case, two atoms are considered adjacent if the ℓ_2 Euclidean distance between them is less than a certain threshold, determined by Eq. (2). Conversely, if the distance is greater than the threshold, the atoms are not considered adjacent.

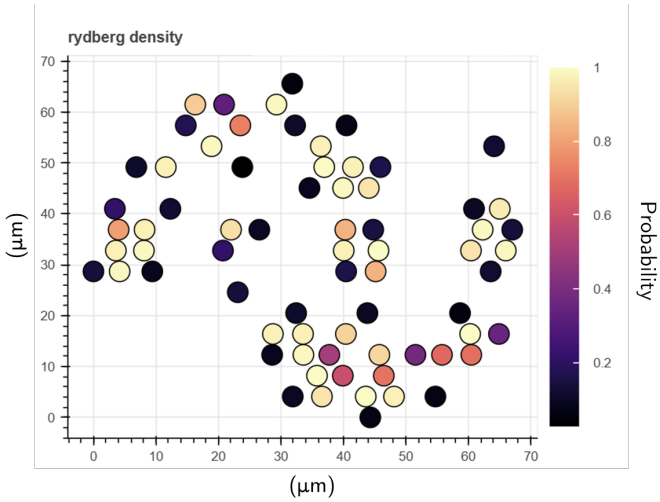


Fig. 2: Moon dataset, sampled as lattice points and processed on Aquila, with probability distribution returned by hardware showing the probability of measuring an excited state for each node.

If two data points are within the respective blockade radius of the other point, the corresponding atoms are adjacent in the graph. The Rydberg blockade effect ensures that both cannot be excited simultaneously, enforcing the MIS’s independence condition. As introduced in Section II, we take the excited atoms as the maximum independent set. Modifying the blockade radius as a hyperparameter allows the graph’s adjacency level to be adjusted. By mapping the dataset to a lattice and using the blockade radius to define adjacency, the dataset reduction problem effectively translates into a format suitable for the neutral atom platform.

B. Post-Processing

Given a complete dataset of size N , our goal is to construct a reduced dataset of size k . To guide this selection, we use the output of the quantum system, which assigns to each

lattice point i an empirical excitation probability π_i estimated from repeated hardware shots. Intuitively, points with larger excitation probability are more likely to be selected by the quantum MIS process and are therefore natural candidates for inclusion in the reduced dataset.

At the same time, a useful reduced dataset should not only favor high-probability points, but also preserve diversity in order to support downstream generalization. To balance these two objectives, we use a greedy post-processing heuristic that combines excitation probability with geometric spread over the underlying 2D manifold.

We first restrict the search space to a candidate pool consisting of the top- αk points ranked by excitation probability π_i , where $\alpha \geq 1$ is a fixed scaling factor. In all experiments, we use $\alpha = 5$. We then construct the reduced set S iteratively. The initialization is given by

$$S = \{\arg \max_i \pi_i\},$$

that is, the first selected point is the one with largest excitation probability.

For each subsequent step, and for each candidate point $i \notin S$, we define the score

$$\omega_i = \lambda \cdot \frac{1}{|S|} \sum_{s \in S} \|x_s - x_i\|_2 + \pi_i, \quad (3)$$

where $x_i \in \mathbb{R}^2$ denotes the geometric position of point i , π_i is its empirical excitation probability, and λ is a weighting factor that controls the tradeoff between diversity and quantum preference. At each iteration, we add to S the candidate maximizing ω_i . This process is repeated until $|S| = k$. The pseudocode implementation can be seen in Alg. 1.

Algorithm 1 Greedy probability-distance post-processing

Require: probabilities $\{\pi_i\}_{i=1}^N$, positions $\{x_i\}_{i=1}^N \subset \mathbb{R}^2$, target size k , scale factor α , weight λ

- 1: Let C be the top αk points ranked by π_i
- 2: Initialize $S \leftarrow \{\arg \max_{i \in C} \pi_i\}$
- 3: **while** $|S| < k$ **do**
- 4: Select

$$i^* \leftarrow \arg \max_{i \in C \setminus S} \left(\pi_i + \lambda \cdot \frac{1}{|S|} \sum_{s \in S} \|x_s - x_i\|_2 \right)$$

- 5: Update $S \leftarrow S \cup \{i^*\}$
 - 6: **end while**
 - 7: **return** S
-

In our experiments, we set $\lambda = 10^4$ so that the distance term, with ℓ_2 distance measured in meters on the atom lattice, is approximately of the same order of magnitude as the probability term. In this way, the heuristic favors points that are both highly ranked by the quantum device and well separated from those already selected.

This approach requires $O(N \log k)$ time first to do our greedy prune, making use of a size k max heap and linear sweep, then $O(\alpha k^2)$ heuristic-based search. Taking α as a

constant factor and recognizing that most tasks will require $k \in \Theta(N)$, our data reduction runs in $O(N^2)$ time, with a suitable scaling factor on the reduction to smaller samples.

IV. EVALUATION

To evaluate the effectiveness of the hybrid MIS data reduction approach, eight different classification datasets were tested against commonly used ML classification models, each selected to present varying levels of complexity and structural characteristics. This section provides an overview of performance on the selection of commonly used ML datasets, ranging from simple linear relationships to more intricate nonlinear patterns that may be challenging for standard classification models. Three specific ML models have been selected to assess the performance of our dataset reduction method. These models combine instance-based learning with models that create explicit decision boundaries, allowing for evaluation of how well the reduced dataset subsets represent the original.

A. Dataset Selection

Six synthetic binary classification datasets have been selected to evaluate the effectiveness of the hybrid MIS dataset reduction method, shown in Fig. 3. These datasets have been chosen to represent commonly used synthetic datasets for machine learning evaluations and include:

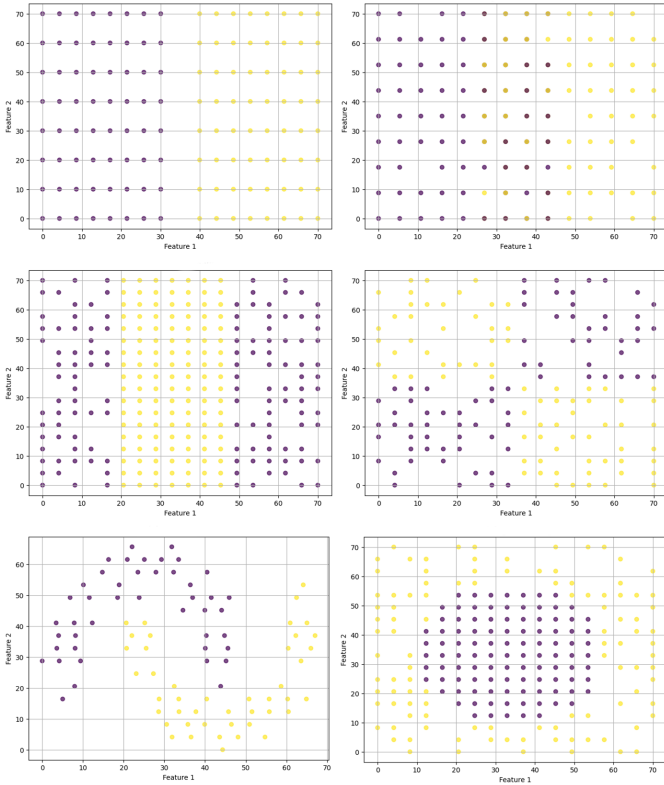


Fig. 3: Visualization of six different synthetic datasets selected for evaluation of MIS dataset reduction.

- **Linearly separable dataset** contains data points from two classes distributed in a way that they can be perfectly separated using a linear line or hyperplane.
- **Linearly non-separable dataset** uses data points from different classes overlapping or intertwined, making it impossible to be separated using a linear line.
- **Bands dataset** is a special subset of the linearly separable dataset, where one class is in between two layers of the other class, forming parallel bands. Unlike the linearly separable dataset, it will require more than one linear line or hyperplane as a decision boundary.
- **XOR dataset** creates data points placed in alternating quadrants of a Cartesian plane, such that the adjacent quadrant belongs to a different class.
- **Ring dataset** consists of concentric circles or rings, with one class forming an inner ring and the other forming the outer ring surrounding the other class.
- **Moon dataset** comprises two interleaving half-circle shapes resembling crescent moons. Due to the curved structure, the dataset is also not linearly separable.

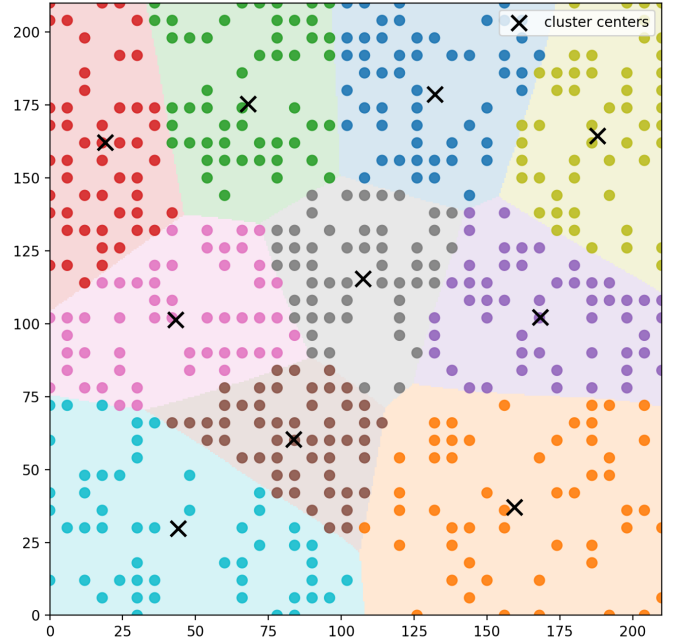


Fig. 4: Visualization of artificial 2D dataset consisting of 10 different classes with 512 data points.

B. Lattice Expansion

It is important to note that, due to hardware constraints imposed by Aquila, constraints have been placed on dataset generation, limiting value ranges for each of the data points. On the Aquila platform, the Euclidean distance between two sites must be no less than $4\mu m$, which, combined with the size of the focusing optics, constrains the total area to $75\mu m$ by $76\mu m$ rectangle. Additionally, the Aquila hardware supports up to 256 qubits, each used to represent a potential data point in the dataset [17], which means that the generated dataset

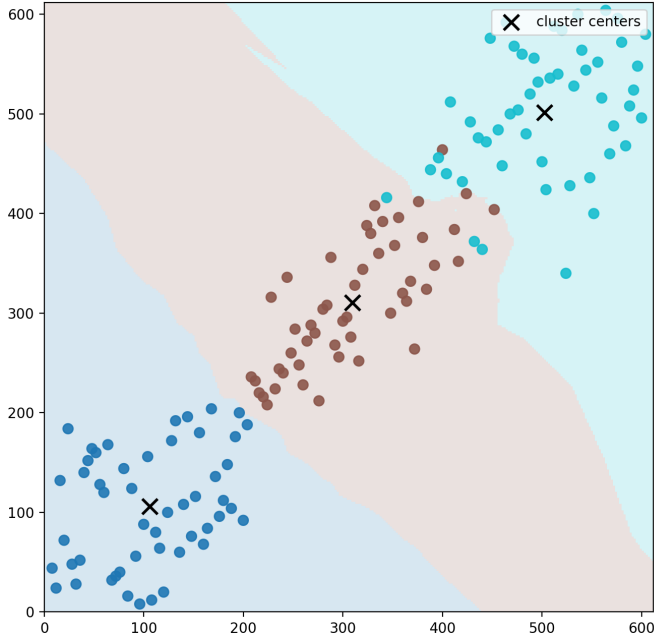


Fig. 5: Visualization of Iris dataset with petal length and petal width as feature, with 150 data points.

must contain fewer than 256 data points, each at least 4 units apart, with each of the feature values less than 75 units.

To bypass this limitation imposed by the hardware, we utilized the divide and conquer based partitioning scheme introduced by Xu et. al [36]. Following a similar partitioning idea with separator merging on the boundaries, our method can be used to reduce total data point on graphs beyond 75 by 76 units with support for more than 256 individual data points. For any data points with distance less than 4 units apart, we perform a global normalization to ensure the minimum distance is compatible with hardware. As such, we also include two additional testing datasets, one artificial dataset with 10 classes and 512 data points, as seen in Fig. 4, as well as the real world Iris dataset [37], as seen in Fig. 5. It should be noted that for Iris, we evaluate only the two-feature petal-length/petal-width projection to remain compatible with the current 2D hardware-mapping pipeline. To isolate the potential loss in dimensionality reduction, we evaluated the projection loss, defined as the paired difference between native-dimensional and projected-data performance on identical held-out folds, for each of the models. We concluded the projection is nearly lossless with an average MSE of 1.37%.

We divide the full dataset into local patches fitting inside hardware requirements. Each such leaf of the partitioning tree is passed into the hardware as its own instance, and later merged using the same greedy approach defined by Xu et. al, as demonstrated by Fig. 6. For this work, we chose a more conservative subdivision based on QuEra’s hardware specification, with a max lattice grid size of 70 units by 70 units, with each data point being at least 4 units apart in both axis, such that all datasets can be tested on physical hardware.

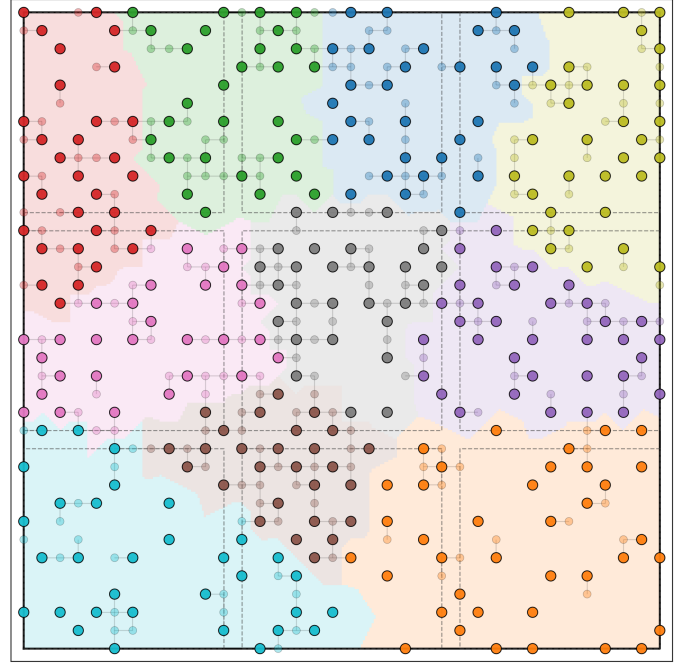


Fig. 6: Sample merged hardware output, with points selected as the MIS subset in full opacity and points removed with reduced opacity. Gray lines denote separator regions.

C. Classification Evaluation Models

Three basic machine learning models, k-nearest neighbors (KNN), logistic regression, and support vector machines (SVM), have been used to assess the impact of hybrid MIS-reduced datasets [38], [39]. These models were selected due to their varying characteristics depending on the selected dataset. By testing the reduced datasets across these models, the goal is to assess how well our approach performs under different learning approaches with different sensitivity to data distribution and structure. Unless stated otherwise, we use weighted-F1, kNN with $k = 5$, logistic regression with $C = 1$, and SVM with RBF kernel. Exact training/testing split can be found in Tab. I. All other hyperparameters are unchanged from sklearn default [40], with no explicit random seed or custom initialization policy.

1) *k-Nearest Neighbors*: kNN is an instance-based learning algorithm that classifies data points based on the majority class among their k nearest neighbors in the feature space. This model was selected due to its reliance on distance metrics and training instances, where the quality and representativeness of the reduced training subset are crucial.

2) *Logistic Regression*: Logistic regression is a statistical model commonly used for binary classification tasks. It uses the logistic function to interpret the output of a linear model as the probability that a given input belongs to a particular class. The model performs well when the classes are linearly separable, making it suitable for evaluating datasets such as the linearly separable and bands dataset. It was selected to test how well the reduced dataset maintains the linear relationship

between features.

3) *Support Vector Machines*: SVM is a form of supervised learning model used for classification tasks. Its goal is to find the optimal hyperplane that maximally separates classes in the feature space. Its ability to handle nonlinear relationships, under the Gaussian kernel, makes it suitable for evaluating the nonlinear datasets in our experiment. It has been selected as only the support vectors influence the model, which requires that these critical points be retained for training to ensure that the decision boundaries are accurate for classification.

V. RESULTS

We evaluate the method on six synthetic 2D classification datasets chosen to stress different decision-boundary geometries, plus two larger low-dimensional benchmarks mentioned in the previous section, using the three discussed classifiers: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. For each dataset, 100 total shots are used to calculate excitation probability. For datasets that are too large for a single lattice, multiple instances of 100 shots have been used per partition leaf. We compare the classifiers' performance trained with the set produced by data reduction through QMIS, clustering with k-means, as well as the classical optimal MIS. This is done quantitatively by means of the F-1 score, which is given by:

$$\text{F-1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

For each classifier, we applied our approach iteratively for varying data point sizes k ranging from 10 up to the entire dataset size [41], [42]. For each subset size, we trained the classifier and evaluated it on the test set, and the resulting F-1 score and node size were recorded. Similarly, we applied k -means clustering to create reduced datasets at various cluster sizes, training the classifier on each and recording the F-1 scores for the respective node sizes. We also trained the classifier on the full dataset to obtain its F-1 score. The F-1 scores for different node sizes were recorded, and the average F-1 score is shown in Tab. I.

To compare the QMIS with the classical MIS, we utilized the solver Gurobi to obtain the optimal MIS [43]. The

classifier was then trained on this subset, and the F-1 score and the number of data points in the subset were recorded. This results in a fixed size subset, which we show in comparison to our QMIS method at same output size in Tab I. We further generate plots of F-1 against node size for all the mentioned methods. The difference in F-1 scores between the QMIS and other methods is also calculated. In the graphs, the vertical line indicates the number of points in the classical MIS, and the horizontal lines are the values of the classical MIS and the F-1 score of a model trained over the full original dataset.

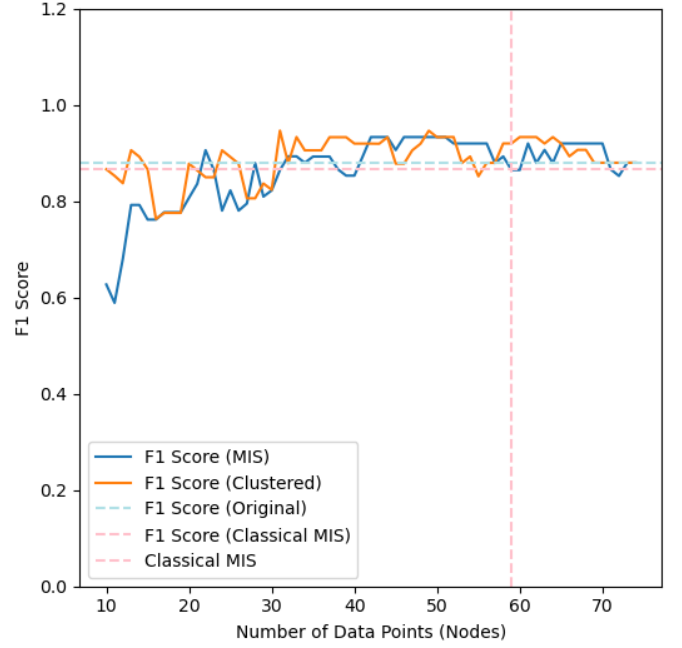


Fig. 8: F-1 score for SVM on the XOR dataset with varying output dataset sizes.

A. Comparing QMIS and Clustering

On average, based on Fig. 7, our implementation of QMIS is competitive with k-means across several datasets, and in selected cases it achieves similar F-1 at smaller subset sizes. It should be noted however, the advantage is not uniform across models or datasets. Additionally, the clustering approach exhibits greater fluctuations at lower node counts, indicating that our QMIS is often competitive at smaller subset sizes, though the effect is not uniform across datasets and classifiers, and should be interpreted together with uncertainty across splits. Reported averages and standard deviations are computed over the same hardware run with random training/testing split assignment of respective subset sizes. For the XOR and Linearly separable dataset using kNN, our QMIS achieves an average score that is higher than k-means from Tab. I. However, k-means achieves a higher average F-1 score than QMIS on datasets like XOR with the SVM classifier, though the individual F-1 scores remain comparable to k-means across both lower and higher node sizes, as shown in Fig. 8.

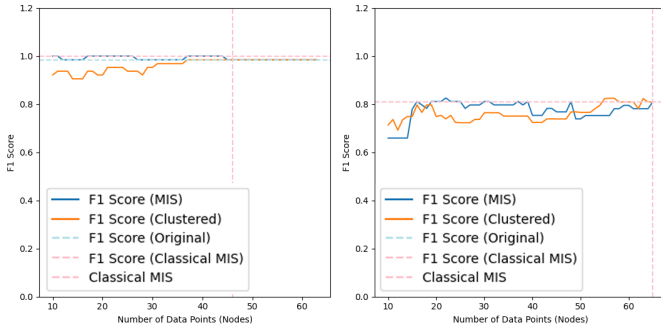


Fig. 7: F-1 score comparison linear (left) and nonlinear (right) datasets on the SVM classifier

Dataset	Dataset Size			Model Performance Results															Classical Comparison (SVM)		
	Total Data Points	Dataset Split		K-Nearest Neighbor				Logistic Regression				Support Vector Machine				Training set size	QMIS F-1 at Dataset Size	Classical MIS F-1			
		Training Points	Testing Points	Original	K-means		QMIS		Original	K-means		QMIS		Original	K-means				QMIS		
					Avg F-1	STD Dev	Avg F-1	STD Dev		Avg F-1	STD Dev	Avg F-1	STD Dev		Avg F-1				STD Dev	Avg F-1	STD Dev
Linearly Separable	128	64	64	1.0000	0.8906	0.0303	0.9290	0.0232	1.0000	0.9867	0.0403	0.9994	0.0000	0.9844	0.9631	0.0313	0.9902	0.0184	46	1.0000	1.0000
Linearly Inseparable	150	66	84	0.7962	0.731	0.0299	0.6939	0.0218	0.8111	0.7721	0.0299	0.7840	0.0250	0.8093	0.7592	0.0289	0.7716	0.0210	65	0.8093	0.8093
Bands	252	210	51	0.9410	0.9046	0.0762	0.8802	0.0715	0.6069	0.5069	0.1237	0.5655	0.0687	1.0000	0.9457	0.0295	0.9504	0.0519	109	0.9412	0.9412
XOR	150	75	75	0.8925	0.7644	0.0594	0.7784	0.0583	0.4309	0.4282	0.09848	0.3517	0.0684	0.8800	0.8877	0.0634	0.8632	0.0424	59	0.8667	0.8662
Ring	208	166	42	0.9039	0.8379	0.0714	0.8197	0.1368	0.3568	0.3531	0.0899	0.4131	0.0409	0.9523	0.9509	0.0429	0.9100	0.0455	95	0.9523	0.9039
Moon	88	70	18	0.8286	0.7652	0.0958	0.6669	0.0928	0.8286	0.7778	0.0227	0.6918	0.0268	0.7662	0.8168	0.03304	0.7805	0.0255	49	0.8875	0.8875

TABLE I: Model performance evaluation with F-1 score and standard deviation for each model and each dataset and reduced QMIS subset along with a comparison with classical MIS F-1 score for SVM classifier. Average F-1 is evaluated using the summation of all weighted F-1 score over the entire subset sweep. Training points are subsets of the original dataset or centroids created with k-means, used for training the classifier models.

B. Comparing QMIS and CMIS

The rightmost section of Tab. I shows a comparison of our approach with the optimal MIS, which is obtained classically. As described in our approach, we set a radius for connectivity in the lattice graph. Therefore, we re-use this radius to determine connectivity for the graph, for which we find a classical MIS solution. With a set graph, a single-sized optimal solution exists. As such, we directly compare our approach with the classical MIS approach at this reduction size.

Based on Tab. I, it can be seen that with an SVM model, the model trained on data reduced via our approach matches the performance of a model trained on data reduced via the optimal classical MIS approach. Furthermore, Fig. 8 demonstrates that the model trained on data obtained by reduction from our approach maintains this performance down to approximately 40% of the classical MIS size.

A similar pattern can be seen in the two complex datasets. While direct F-1 comparison only demonstrates comparable accuracy performance, as shown in Tab. II, our approach shows less variation over the subset-size sweep at low point counts, as seen in Fig. 9, suggesting a higher boundary stability when it comes to data point selection. Although, the source of this variability should be interpreted according to the averaging protocol defined above.

It should be noted that, although logistic regression is sufficient for several binary classification tasks, it inherently assumes a linear relationship between the input feature and the log-odds of the outcome. As such, results on strongly nonlinear datasets are included only as a negative control and are not central to our conclusions.

C. Decision Boundary Impact

While F-1 scores provide a quantitative assessment of classifier accuracy on dataset and dataset subsets, they may not fully capture the impact of dataset reduction on the decision boundary across various subset sizes. Visual inspection suggests that QMIS-reduced datasets often preserve the original decision boundary more closely than clustering-based reductions, particularly in the examples shown in Fig. 10.

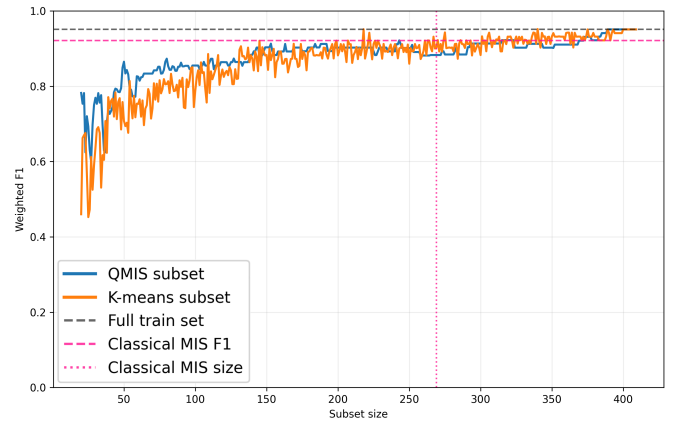


Fig. 9: F-1 score for SVM on the 10 classes cluster dataset with varying output dataset size.

Dataset	Total Data Points	Dataset Split		SVM Performance				
		Training Points	Testing Points	Full Set F-1	K-means		QMIS	
					Avg F-1	STD Dev	Avg F-1	STD Dev
Clusters	512	409	103	0.9230	0.9046	0.0855	0.9294	0.0546
Iris	150	120	30	0.9123	0.9144	0.0293	0.9002	0.0231

TABLE II: Performance comparison of F-1 score between K-means and QMIS approach on the two complex dataset.

A visual comparison indicates that models trained on the QMIS-reduced datasets show decision boundaries that align closer to the original relative to clustering, indicating the preservation of the dataset's structural and distribution properties. This is especially noticeable in nonlinear classifiers, as clustering tends to sway the decision boundary drastically. This trend can also be seen in the more complex 10 class dataset where our approach creates less variance in decision boundary, as seen in Fig. 11.

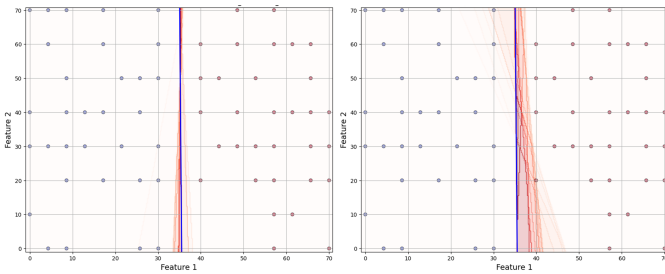


Fig. 10: Decision boundary comparison between QMIS (left) and k-means clustering (right) of logistic regression classifier on the linearly separable dataset, with original decision boundary in blue and reduced dataset decision boundary in red. The difference is highlighted in red. Reduced dataset size ranges from 10 to 50 data points

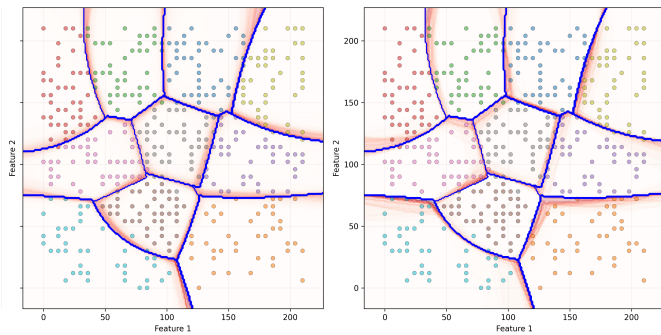


Fig. 11: Decision boundary comparison between QMIS (left) and k-means clustering (right) of SVM on 10 class clusters, with original decision boundary in blue and reduced dataset decision boundary in red. The difference is highlighted in red. Reduced dataset size ranges from 20 to 409 data points

VI. FUTURE WORK AND LIMITATION

While our work demonstrates the potential usage of neutral atom quantum computing for dataset reduction via a maximum independent set computation, several limitations still exist due to the inherent nature of current hardware. Addressing these limitations presents opportunities for future research and improvements.

A. Mapping Limitation

One such limitation is the specific mapping method. While we are able to bypass the physical lattice configuration with a partitioning scheme [36], it doesn't fully address the dimensionality limitation. Due to the neutral atom platform encoding in 2D lattice, higher dimensional data would require some form of translation or mapping down to a lower dimension, such as with PCA, for it to be compatible with the physical hardware. For simple low dimensionality datasets, it wouldn't cause significant loss in fidelity. But for modern ML tasks, often with much higher feature count, a better encoding scheme would be required to accommodate those datasets.

While outside the scope of this work, we argue that a reasonable 3D projection using the same separator decomposition idea is possible. For a basic 2D case, the hardware embedding is naive. In 3D or n-dimensional projection, recursively split into leaf nodes such that it either projects to 2D or re-embeds the leaf node as a 2D approximate unit disk graph.

Furthermore, it could also be argued that another form of encoding could be used for data point mapping. Rather than encoding feature value as a direct mapping to lattice space, some distance metric between each data point could be used as feature encoding. An improved data mapping algorithm, as opposed to direct feature mapping, could potentially provide greater flexibility.

B. Ablation Study

It's important to note that due to the limited amount of dataset and models evaluated in this study, it is not possible to draw a definitive conclusion at this time about the exact performance of this approach. To investigate whether the hardware-derived selection probabilities contain sample-specific information, we performed a preliminary random- π ablation using the 512 point clusters dataset. The observed probabilities were permuted among training samples within each class, preserving their class-conditional distributions while removing their association with individual samples. All other components—including the train-test split, class quotas, geometric diversity term, subset sizes, and downstream SVM—were held fixed. Across six representative subset sizes and 50 permutations, the original hardware probabilities produced a mean weighted F-1 score of 0.858, compared to 0.844 ± 0.019 for the permuted controls. However, this result should not be interpreted as definitive, as the clusters dataset was randomly created for the purpose of this study. As such, Future work should also repeat the ablation across a much larger set of datasets with various classification models.

C. Execution Time

Quantum hardware used in this study was accessed using a cloud-based platform. In our experiments, total wall-clock time from submission to result was within 2 minutes. This time is significantly longer than that it takes by classical methods to find the provably optimal solution to the MIS in current classical hardware. However it should be noted that the reported wall-clock time includes cloud submission and retrieval overhead, not just on-device execution, and therefore is not directly comparable to classical solver runtime. It is important to recognize that current quantum hardware will fall short of execution speed. Accessing quantum hardware through cloud-based services often involves queuing delays due to high demand and limited availability of quantum processors. Users may experience extended wait time before the job is executed, further increasing the total time of job submission and resulting in retrieval. This limits the direct comparison in runtime to classical methods.

In the current approach, the reduced dataset subset selects data points for training using marginal per-node probabilities estimated from repeated hardware shots, calculated using the greedy heuristic. However, the quantum hardware outputs a probability distribution over possible MIS configurations, which contains additional information that can be utilized. Unlike directly using the greedy heuristic to select data points, interpreting the measurement probabilities as weights allows for assigning importance measurement to each data point based on its likelihood of being included in the MIS. This weighted dataset can enhance machine learning models by emphasizing more representative or critical data points. We leave full-configuration decoding and weighting strategies as future work; the present study uses only marginal per-node probabilities. A useful future control would compare this probability-informed ranking against purely geometric greedy selection to isolate the contribution of the hardware-derived scores.

VII. CONCLUSION

In this paper, we present a proof-of-concept use of neutral-atom hardware for dataset reduction via MIS-inspired subset scoring in low-dimensional machine-learning tasks. Our approach addresses the problem of reducing low-dimensional datasets while preserving structure relevant to downstream classification, which is critical for training accurate machine learning models with reduced training time. We demonstrate the feasibility of integrating neutral-atom measurement statistics into a hybrid subset-selection pipeline for low-dimensional tasks, without claiming a runtime advantage over classical dataset-reduction methods.

Overall, the results support the feasibility of integrating neutral-atom measurement statistics into a hybrid subset-selection pipeline for low-dimensional tasks. The present evidence supports competitiveness on selected tasks rather than a general accuracy or runtime advantage over classical reduction methods.

While our work aims to showcase the potential usage of neutral atom quantum computing for dataset reduction, addressing the limitation related to qubit count and spatial constraint remains essential for practical scalability in hybrid systems. Additionally, further utilization of probabilistic output can potentially offer additional improvements for enhancing machine learning models. Our work aims to inspire future research on these areas, as they remain critical in integrating quantum computation into a hybrid machine learning pipeline.

ACKNOWLEDGMENT

We would like to thank Professor Mahdi Hosseini for supporting the Quantum Game Club. Special thanks to other members of team Plankton: Allen Baranov, Ishika Tulsian, Jeffrey Tejada, and Kevin Rapp.

- [1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, 10th ed. Cambridge: Cambridge university press, 2010.
- [2] P. W. Shor, "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1484–1509, Oct. 1997. [Online]. Available: <http://epubs.siam.org/doi/10.1137/S0097539795293172>
- [3] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing - STOC '96*. Philadelphia, Pennsylvania, United States: ACM Press, 1996, pp. 212–219. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=237814.237866>
- [4] J. Preskill, "Quantum Computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018. [Online]. Available: <https://quantum-journal.org/papers/q-2018-08-06-79/>
- [5] S. Ebadi, A. Keesling, M. Cain, T. T. Wang, H. Levine, D. Bluvstein, G. Semeghini, A. Omran, J.-G. Liu, R. Samajdar, X.-Z. Luo, B. Nash, X. Gao, B. Barak, E. Farhi, S. Sachdev, N. Gemelke, L. Zhou, S. Choi, H. Pichler, S.-T. Wang, M. Greiner, V. Vuletić, and M. D. Lukin, "Quantum optimization of maximum independent set using Rydberg atom arrays," *Science*, vol. 376, no. 6598, pp. 1209–1215, Jun. 2022. [Online]. Available: <https://www.science.org/doi/10.1126/science.abo6587>
- [6] I. H. Deutsch, G. K. Brennen, and P. S. Jessen, "Quantum Computing with Neutral Atoms in An Optical Lattice," in *Scalable Quantum Computers*, 1st ed., S. L. Braunstein, H. Lo, and P. Kok, Eds. Wiley, Dec. 2000, pp. 155–173. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1002/3527603182.ch10>
- [7] J. Wurtz, P. L. S. Lopes, C. Gorgulla, N. Gemelke, A. Keesling, and S. Wang, "Industry applications of neutral-atom quantum computing solving independent set problems," 2022, version Number: 2. [Online]. Available: <https://arxiv.org/abs/2205.08500>
- [8] R. E. Tarjan and A. E. Trojanowski, "Finding a Maximum Independent Set," *SIAM Journal on Computing*, vol. 6, no. 3, pp. 537–546, Sep. 1977. [Online]. Available: <http://epubs.siam.org/doi/10.1137/0206038>
- [9] K. L. Pudenz and D. A. Lidar, "Quantum adiabatic machine learning," *Quantum Information Processing*, vol. 12, no. 5, pp. 2027–2070, May 2013. [Online]. Available: <http://link.springer.com/10.1007/s11128-012-0506-4>
- [10] M. Dupont, T. Oberoi, and B. Sundar, "Optimization via Quantum Preconditioning," 2025, version Number: 1. [Online]. Available: <https://arxiv.org/abs/2502.18570>
- [11] C. Weitenberg, S. Kuhr, K. Mølmer, and J. F. Sherson, "Quantum computation architecture using optical tweezers," *Physical Review A*, vol. 84, no. 3, p. 032322, Sep. 2011. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevA.84.032322>
- [12] A. M. Kaufman and K.-K. Ni, "Quantum science with optical tweezer arrays of ultracold atoms and molecules," *Nature Physics*, vol. 17, no. 12, pp. 1324–1333, Dec. 2021. [Online]. Available: <https://www.nature.com/articles/s41567-021-01357-2>
- [13] M. Saffman, T. G. Walker, and K. Mølmer, "Quantum information with Rydberg atoms," *Reviews of Modern Physics*, vol. 82, no. 3, pp. 2313–2363, Aug. 2010. [Online]. Available: <https://link.aps.org/doi/10.1103/RevModPhys.82.2313>
- [14] M. Saffman, "Quantum computing with atomic qubits and Rydberg interactions: progress and challenges," *Journal of Physics B: Atomic, Molecular and Optical Physics*, vol. 49, no. 20, p. 202001, Oct. 2016. [Online]. Available: <https://iopscience.iop.org/article/10.1088/0953-4075/49/20/202001>
- [15] K. Wintersperger, F. Dommert, T. Ehmer, A. Hoursanov, J. Klepsch, W. Maurer, G. Reuber, T. Strohm, M. Yin, and S. Luber, "Neutral atom quantum computing hardware: performance and end-user perspective," *EPJ Quantum Technology*, vol. 10, no. 1, p. 32, Dec. 2023. [Online]. Available: <https://epjquantumtechnology.springeropen.com/articles/10.1140/epjqt/s40507-023-00190-1>
- [16] E. Urban, T. A. Johnson, T. Henage, L. Isenhower, D. D. Yavuz, T. G. Walker, and M. Saffman, "Observation of Rydberg blockade between two atoms," *Nature Physics*, vol. 5, no. 2, pp. 110–114, Feb. 2009. [Online]. Available: <https://www.nature.com/articles/nphys1178>
- [17] J. Wurtz, A. Bylinskii, B. Braverman, J. Amato-Grill, S. H. Cantu, F. Huber, A. Lukin, F. Liu, P. Weinberg, J. Long, S.-T. Wang, N. Gemelke, and A. Keesling, "Aquila: QuEra's 256-qubit neutral-atom

- quantum computer,” Jun. 2023, arXiv:2306.11727 [quant-ph]. [Online]. Available: <http://arxiv.org/abs/2306.11727>
- [18] E. L. Lawler, J. K. Lenstra, and A. H. G. Rinnooy Kan, “Generating All Maximal Independent Sets: NP-Hardness and Polynomial-Time Algorithms,” *SIAM Journal on Computing*, vol. 9, no. 3, pp. 558–565, Aug. 1980. [Online]. Available: <http://epubs.siam.org/doi/10.1137/0209042>
 - [19] R. M. Karp, “Reducibility among Combinatorial Problems,” in *Complexity of Computer Computations*, R. E. Miller, J. W. Thatcher, and J. D. Bohlinger, Eds. Boston, MA: Springer US, 1972, pp. 85–103. [Online]. Available: http://link.springer.com/10.1007/978-1-4684-2001-2_9
 - [20] E. Farhi, J. Goldstone, and S. Gutmann, “A Quantum Approximate Optimization Algorithm,” Nov. 2014, arXiv:1411.4028 [quant-ph]. [Online]. Available: <http://arxiv.org/abs/1411.4028>
 - [21] B. Fuller, C. Hadfield, J. R. Glick, T. Imamichi, T. Itoko, R. J. Thompson, Y. Jiao, M. M. Kagele, A. W. Blom-Schieber, R. Raymond, and A. Mezzacapo, “Approximate Solutions of Combinatorial Problems via Quantum Relaxations,” *IEEE Transactions on Quantum Engineering*, vol. 5, pp. 1–15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10586788/>
 - [22] E. Farhi, D. Gamarnik, and S. Gutmann, “The Quantum Approximate Optimization Algorithm Needs to See the Whole Graph: A Typical Case,” Apr. 2020, arXiv:2004.09002 [quant-ph]. [Online]. Available: <http://arxiv.org/abs/2004.09002>
 - [23] T. N. Bui and E. Paul H, “A hybrid genetic algorithm for the maximum clique problem,” in *Proceedings of the 6th international conference on genetic algorithms*, 1995, pp. 478–484.
 - [24] H. Pichler, S.-T. Wang, L. Zhou, S. Choi, and M. D. Lukin, “Quantum Optimization for Maximum Independent Set Using Rydberg Atom Arrays,” 2018, version Number: 1. [Online]. Available: <https://arxiv.org/abs/1808.10816>
 - [25] K. Kim, M. Kim, J. Park, A. Byun, and J. Ahn, “Quantum computing dataset of maximum independent set problem on king lattice of over hundred Rydberg atoms,” *Scientific Data*, vol. 11, no. 1, p. 111, Jan. 2024. [Online]. Available: <https://www.nature.com/articles/s41597-024-02926-9>
 - [26] T. Kato, “On the Adiabatic Theorem of Quantum Mechanics,” *Journal of the Physical Society of Japan*, vol. 5, no. 6, pp. 435–439, Nov. 1950. [Online]. Available: <http://journals.jps.jp/doi/10.1143/JPSJ.5.435>
 - [27] T. Albash and D. A. Lidar, “Adiabatic quantum computation,” *Reviews of Modern Physics*, vol. 90, no. 1, p. 015002, Jan. 2018. [Online]. Available: <https://link.aps.org/doi/10.1103/RevModPhys.90.015002>
 - [28] D. An and L. Lin, “Quantum Linear System Solver Based on Time-optimal Adiabatic Quantum Computing and Quantum Approximate Optimization Algorithm,” *ACM Transactions on Quantum Computing*, vol. 3, no. 2, pp. 1–28, Jun. 2022. [Online]. Available: <https://dl.acm.org/doi/10.1145/3498331>
 - [29] D. Napoleon and S. Pavalakodi, “A New Method for Dimensionality Reduction Using KMeans Clustering Algorithm for High Dimensional Data Set,” *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, Jan. 2011. [Online]. Available: <http://www.ijcaonline.org/volume13/number7/pxc3872471.pdf>
 - [30] S. Ougiaroglou and G. Evangelidis, “Efficient dataset size reduction by finding homogeneous clusters,” in *Proceedings of the Fifth Balkan Conference in Informatics*. Novi Sad Serbia: ACM, Sep. 2012, pp. 168–173. [Online]. Available: <https://dl.acm.org/doi/10.1145/2371316.2371349>
 - [31] M. Hashemi, S. Gong, J. Ni, W. Fan, B. A. Prakash, and W. Jin, “A Comprehensive Survey on Graph Reduction: Sparsification, Coarsening, and Condensation,” 2024, version Number: 4. [Online]. Available: <https://arxiv.org/abs/2402.03358>
 - [32] R. Wickman, X. Zhang, and W. Li, “A Generic Graph Sparsification Framework using Deep Reinforcement Learning,” 2021, version Number: 2. [Online]. Available: <https://arxiv.org/abs/2112.01565>
 - [33] J. D. Batson, D. A. Spielman, and N. Srivastava, “Twice-ramanujan sparsifiers,” in *Proceedings of the forty-first annual ACM symposium on Theory of computing*. Bethesda MD USA: ACM, May 2009, pp. 255–262. [Online]. Available: <https://dl.acm.org/doi/10.1145/1536414.1536451>
 - [34] J. Batson, D. A. Spielman, N. Srivastava, and S.-H. Teng, “Spectral sparsification of graphs: theory and algorithms,” *Communications of the ACM*, vol. 56, no. 8, pp. 87–94, Aug. 2013. [Online]. Available: <https://dl.acm.org/doi/10.1145/2492007.2492029>
 - [35] I. Althöfer, G. Das, D. Dobkin, D. Joseph, and J. Soares, “On sparse spanners of weighted graphs,” *Discrete & Computational Geometry*, vol. 9, no. 1, pp. 81–100, Jan. 1993. [Online]. Available: <http://link.springer.com/10.1007/BF02189308>
 - [36] H. Xu and A. Pothen, “Divide and Conquer-based Quantum Algorithms for Maximum Independent Set on Large Separable Graphs,” in *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*. Montreal, QC, Canada: IEEE, Sep. 2024, pp. 87–97. [Online]. Available: <https://ieeexplore.ieee.org/document/10821339/>
 - [37] R. A. Fisher, “Iris,” UCI Machine Learning Repository, 1936, DOI: <https://doi.org/10.24432/C56C76>.
 - [38] E. Patrick and F. Fischer, “A generalized k-nearest neighbor rule,” *Information and Control*, vol. 16, no. 2, pp. 128–152, Apr. 1970. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0019955870900811>
 - [39] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY: Springer New York, 2000. [Online]. Available: <http://link.springer.com/10.1007/978-1-4757-3264-1>
 - [40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, “Scikit-learn: Machine learning in Python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011, publisher: JMLR. org.
 - [41] A. Jain and R. Dubes, *Algorithms for Clustering Data*, ser. Prentice Hall advanced reference series. Prentice Hall, 1988. [Online]. Available: <https://books.google.com/books?id=7eBQAAAAMAAJ>
 - [42] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5. University of California press, 1967, pp. 281–298.
 - [43] “Gurobi.” [Online]. Available: <https://www.gurobi.com>