Quantum Data Sketches

Qin Zhang
Indiana University Bloomington, IN, USA

Mohsen Heidari
Indiana University Bloomington, IN, USA

Abstract

Recent advancements in quantum technologies, particularly in quantum sensing and simulation, have facilitated the generation and analysis of inherently quantum data. This progress underscores the necessity for developing efficient and scalable quantum data management strategies. This goal faces immense challenges due to the exponential dimensionality of quantum data and its unique quantum properties such as no-cloning and measurement stochasticity. Specifically, classical storage and manipulation of an arbitrary n-qubit quantum state requires exponential space and time. Hence, there is a critical need to revisit foundational data management concepts and algorithms for quantum data. In this paper, we propose succinct quantum data sketches to support basic database operations such as search and selection. We view our work as an initial step towards the development of quantum data management model, opening up many possibilities for future research in this direction.

2012 ACM Subject Classification Theory of computation → Quantum query complexity; Information systems → Query operators; Information systems → Query optimization

Keywords and phrases quantum data representation, data sketching, query execution

Digital Object Identifier 10.4230/LIPIcs.ICDT.2025.16

Related Version *Full Version*: `https://arxiv.org/abs/2501.06705` [56]

Funding *Qin Zhang*: supported by NSF CCF-1844234 and IU Luddy Faculty Fellowship.
Mohsen Heidari: supported by NSF CCF-2211423.

1 Introduction

Quantum information and computing, rooted in the principles of quantum mechanics, have emerged as an important field of study with far-reaching effects across a broad spectrum of disciplines. Central to the concept of quantum computing are quantum bits (or qubits), which set themselves apart from classical bits due to their ability to exist in a superposition of states, allowing a quantum computer to offer the potential computational advantage against classical computing.

Although significant advancements have been made in the development of quantum algorithms after several decades of research, only a handful provably outperform their classical counterparts. Notable examples include Shor's algorithm for factorization [48], Grover's algorithm for search [20], and linear system solvers [24]. These quantum algorithms typically start by encoding classical input data into quantum states, execute a series of quantum operations, and then measure the resulting quantum states and carry out specific post-processing on the measurement outcomes. The reasons for the difficulties in the design of quantum algorithms that can outperform classical counterparts on *classical input data* remain elusive.

In this paper, we take a different perspective, directing our attention towards quantum data themselves. The nature, along with scientific experiments spanning physics, chemistry, material science, biology, and other fields, generates massive quantities of quantum data every day. Sources include Hawking Radiation, Cosmic Microwave Background, quantum effects in neutron stars, quantum states in ultra-cold atoms, quantum information in DNA

replication, etc. In many scenarios, there is a need for us to preserve quantum data that has been collected from nature or generated in labs for future analysis. For example, scientists often use photons collected from remote stars to study the properties of those astronomical objects. It would be beneficial to store those photons as quantum states in a database, since it may not be feasible to collect fresh photons from those astronomical objects at the time of data analysis. In the case that the quantum states are prepared in the labs, generating fresh copies of quantum states on demand is often time-consuming. Let us use quantum simulation as an example. Quantum simulation is a prominent advantage of quantum computers, with significant implications for numerous areas of scientific research, including computer-aided drug design [44], high-energy physics [36], quantum chemistry [52] and many-body physics [49]. Quantum simulation typically relies on solving the Schrödinger equation for the underlying Hamiltonian. The Hamiltonian is implemented by a quantum circuit, which is applied to an initial quantum state to generate target quantum states. The construction of the Hamiltonians and the preparation of the target states can be rather time-consuming.[1] Storing the generated molecular quantum states in a database would eliminate the need to repeat the state preparation procedures during data analysis.

Once the quantum states are stored in a database, and assuming each state is associated with additional information such as the nature sources recorded at the time of collection or parameters of the experimental setup used to produce them, numerous applications can be envisioned. For example, if scientists receive photons from an unknown remote star, they can search a photon database to find a matching quantum state. Upon finding a match, they can retrieve its associated properties and other information, such as the time and method of its previous observation. They may also want to sort the states using a local observable (see Definition 9 in Section 3) with respect to certain properties (such as energy or momentum) to get an order of the photons in the database, aiding in the understanding of the spectrum of the corresponding stars in the universe. In quantum simulation, if we want to produce molecular states with average energy levels above a certain threshold relative to a specific local observable, we can perform a selection operation in our database to identify those states, and then use the associated parameters for the experimental setup to produce more of such quantum states.

Nevertheless, quantum data management remains in its infant stage. Some of the previously mentioned motivating examples are more like anticipated future problems. There has been research that leverages quantum data for learning or optimization, such as quantum machine learning [29, 22, 3], quantum variational optimization algorithm [26, 15]. and quantum neural network [46, 42, 16, 28, 40, 18]. However, their primary focus is on the sample complexity (namely, the number of copies of the quantum state needed for the task) and the convergence to optimal points, rather than on developing methods for the efficient representation and storage of quantum data for subsequent analysis.

In this paper, we introduce several quantum data sketches to support basic database operations in a *sustainable* and *efficient* manner. This paper does not aim to formulate a comprehensive quantum data management model. Rather, we view our work as an initial step towards developing a sustainable model for representing, querying, and analyzing quantum data at scale.

[1] For instance, the Hamiltonian of the two-dimensional Fermi-Hubbard model on an 8×8 lattice already requires approximately 10^7 Toffoli gates [38], which directly contribute to the query time if states need to be generated from scratch at query.

Unique Challenges in the Quantum World. The quantum world possesses several unique properties, such as superposition and entanglement, that can be leveraged to reduce resource usage in computing and information exchange. However, some of these features also post significant challenges to quantum data management. We highlight a few below.

Post-Measurement State Disturbance. The only way to extract information from a quantum state is to perform quantum measurements and observe probabilistic outcomes. However, each measurement has the effect of perturbing the quantum state. This characteristic implies that a quantum state might not be reusable post-measurement. In other words, we may need to consume many identical copies of a quantum state in order to derive enough useful information about it. This phenomenon is in stark contrast with the classical setting, in which we can consistently access the same data element for a number of times, always yielding the same result.

No-cloning. A natural thought to resolve the issue caused by state disturbance is to clone the quantum state before the operations. Unfortunately, the *no-cloning theorem* (see, e.g., [41]) in quantum mechanics asserts that it is impossible to create an exact copy of an arbitrary unknown quantum state.

Lack of Large-Scale Quantum Storage Systems. At the time of writing this paper, we are not aware of any reliable large-scale quantum storage systems. One reason for this is that qubits are highly susceptible to environmental disruptions such as temperature variations, electromagnetic radiation, or particle interactions. These disruptions lead to what is known as decoherence [35], resulting in the loss of quantum information.

Moreover, due to the quantum state disturbance and the no-cloning principle, even if we successfully build viable large-scale quantum storage systems in the future, we still need many identical copies of the quantum state for any nontrivial database operation. This implies that in order to accommodate an unlimited number of database operations (i.e., to be sustainable), we must prepare an unlimited number of copies for each quantum state in the storage, which is certainly *not* practical.

An alternative approach is to first learn the classical description of each quantum state and store it in a classical memory for future operations. Indeed, we believe that for the purpose of quantum data management, we have to store quantum states in the classical format. However, learning and storing the full information of a quantum state as a classical object is both time and space expensive, as the dimensionality of a quantum state is exponential in terms of the number of qubits.

We thus propose to design *succinct classical representations* (or, sketches) of quantum states that can be used to perform database operations efficiently. Based on the particular database operation it is intended to support, each sketch preserves only *partial* information of a quantum state. This is also the reason why we may be able to make the size of the sketch to be $o(d)$, where d is the dimension of the quantum state. We also note that the sample complexity for constructing data sketches is a secondary consideration for database management systems, as it is just a one-time preprocessing step in the database design. This is where our work departs from the quantum state learning/tomography literature, which we will discuss in Section 1.1.

Our Contribution. We give the first systematic approach to designing *space-efficient* sketches for quantum states. These sketches can then be used to develop *time-efficient* algorithms for basic database operations. In particular:

1. In Section 3, we have formalized a set of basic database operations for quantum data, including search, selection, sorting, and join. These operations differ from those for classical data as they inherently incorporate approximation in their definitions.

2. Our main technical results are the first set of classical vector sketches that preserve, up to a distortion of $(1+\iota)$ for an arbitrarily small $\iota > 0$, the trace distance of the quantum states with probability $(1-\delta)$. Our sketches have sizes $O(\log(1/\delta)/\iota^2)$, which is *independent* of the dimension of the states. Coupled with efficient nearest neighbor search via locality sensitive hashing, they can be used to support the search and join operations with time sublinear in the database size and independent of the state dimension. See Section 4.1.
3. We make use of classical shadow seeds of quantum states [32] to approximate the expectation value of any given k-local observable (to be defined in Section 3.3) using time and space *independent* of the dimension of the state. We also present a new hybrid quantum-classical algorithm to accelerate the query time. This sketch can be used for selection and sorting operations. See Section 4.2.

Paper Outline. In Section 2, we review some background on quantum information and computing as well as tools for classical data management. In Section 3, we define a set of basic database operations for quantum data. After these preparations, in Section 4, we present our classical sketches of quantum states and illustrate how to perform various database operations using these sketches. We review works that are most relevant to this paper in Section 1.1 and propose several directions for future research in Section 5.

1.1 Related Work

We are not aware of any prior work on designing classical sketches of quantum data, except for the paper [32] discussed in Section 4.2. There have been effort aiming to introduce quantum computing, quantum algorithms and quantum machine learning to the database community [13, 55, 39, 51, 9, 53]. We refer the readers to the recent tutorial [23] for an overview of these works. However, these initiatives either attempt to design and perform database operations directly on quantum data (i.e., assuming database elements are stored as quantum states) or focused on speeding up databases query optimization and transactions on classical data, setting them apart from the objectives pursued in this paper.

There are works [54, 33] focusing on applying classical data compression techniques (such as quantization) to the quantum state vector during quantum simulation. We note that our approach with sketches is quite different, as we aim to extract relevant information (often independent of the quantum states' dimension) for various database operations.

Quantum State Learning. Many studies have explored the task of characterizing and learning properties of a quantum state using multiple copies of the state, including *approximate state discrimination* [12], *quantum state discrimination* [27], *quantum state tomography* [21, 43], *quantum state property testing* [25], *quantum state certification* [7], *shadow tomography* [2, 6], and *pretty good tomography* [1].

In the problem of approximate state discrimination, we are promised that a query quantum state ϕ belongs to a set S of quantum states. The algorithm's task is to return a state $\psi \in S$ such that $D(\phi, \psi) \leq \epsilon$. The algorithm for approximate state discrimination proposed in [12] can be used together with the equality testing to handle the search operation when the available number of copies of the query state is limited, at the cost of larger time and space complexities. However, the need of fresh copies of database states for equality testing would undermine the long-term sustainability of the database system.

The problem of quantum state discrimination is very similar: We are again promised that the query state ϕ belongs to a set S, but now the algorithm needs to return the *exact* ϕ. Harrow and Winter [27] gave an algorithm for this problem where the sample complexity of the query state depends on a parameter F, which is the maximum pairwise fidelity of states in the set S.

In the quantum state tomography, we wanted to learn an unknown quantum state up to a trace distance ϵ. Optimal sample complexity $\tilde{\Theta}(d/\epsilon^2)$ has been identified [21, 43].

Quantum state property testing [25] and quantum state certification [7] can be seen as relaxations of aforementioned problems. In the former, we are given a query state ϕ and a set S of quantum states, and asked to test whether $\phi \in S$ or ϕ is ϵ-far from S (that is, for any state $\psi \in S$, we have $D(\phi, \psi) > \epsilon$), and in the latter, we are given a query state ϕ and a known state ψ, and asked to test whether $\phi = \psi$ or $D(\phi, \psi) > \epsilon$. The main issue with property testing and certification in the setting of data management is that the decision can be *arbitrary* even if the query state is very close to (but not the same as) a database state.

Both shadow tomography and pretty good tomography focus on approximating $\phi^\dagger M_i \phi$ for a query state ϕ and a set of *known* binary measurements $\{M_i\}$ [2, 6], or a distribution on them [1]. However, these algorithms cannot be used for the (η, ε)-selection for an arbitrary observable M given at the time of query. Their running time is also polynomial in terms of the state dimension d. Recently, Gong and Aaronson [19] generalized shadow tomography to a *fixed* set of measurements with multiple outcomes.

To the best of our knowledge, all the previous work on quantum state learning focuses on the sample complexity, but *not* on the space complexity for representing the quantum states for various data management operations.

2 Preliminaries

We start by giving a gentle introduction of the basics of quantum information and computing, particularly for readers who are not in the field yet. For a comprehensive treatment on this topic, we refer the readers to standard textbooks in the field, such as [41].

Quantum States and Qubits. The first axiom of quantum mechanics is concerned with *quantum state* as a way to describe a quantum system, such as a qubit. For accessibility of the paper we focus on pure state that are represented by complex-valued vectors. Moreover, we assume that each quantum data point is stored in n-qubits. Therefore, the dimentionality of the space is $d = 2^n$. In that case, the quantum stats are unit-norm vectors in $\mathbb{C}^d$. Following the Dirac bra-ket notation, a vector $u \in \mathbb{C}^d$ is simply denoted by the ket $|u\rangle$. As an example, a *qubit* is a 2-dimensional vector represented as $|\phi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle$, where $|\alpha_0|^2 + |\alpha_1|^2 = 1$. This decomposition is typically called a *superposition*. A well-known superposition is the state $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Similarly, an n-qubit state is represented by a superposition as $|u\rangle = \sum_{x_1 \cdots x_n \in \{0,1\}^n} \alpha_{x_1 \cdots x_n} |x_1 \cdots x_n\rangle$, where $\sum_{x_1 \cdots x_n \in \{0,1\}^n} \alpha^2_{x_1 \cdots x_n} = 1$. For compactness, we use $|i\rangle$ to represent each $|x_1 \cdots x_n\rangle$, where i is the decimal representation of the binary string $x_1 \cdots x_n$.

Quantum Operations. The second axiom of quantum mechanics states that the evolution of quantum states are described via unitary transformation. A unitary transformation is represented by a unitary matrix U such that $U^\dagger U = U U^\dagger = I$. If the initial state is $|\phi\rangle$, then the evolved state is $U|\phi\rangle$. In quantum computing U is typically implemented in terms of elementary quantum logical gates. In this perspective, one can study the gate complexity of

implementing U. This axiom implies a unique feature of quantum, known as the *no-cloning* principle that prohibits making copies of quantum data. As a result one needs to adopt data management procedures that abide this rule.

Quantum Measurements. The third axiom of quantum mechanics asserts that any classical information about a quantum state is obtained via *measuring* it. The act of measuring a quantum system will collapse the quantum state inevitably. The specific outcome of a measurement is probabilistic and is governed by the Born's law. These probabilities are determined by the initial state of the system and the nature of the interaction between the system and the measuring device. Measuring in an n-qubit system is typically modeled in the so-called computational basis. When the quantum state is in the superposition $|\phi\rangle = \sum_i \alpha_i |i\rangle$, the outcome of the measurement in the computational basis is going to be $i \in [2^n]$ with probability $p_i = |\alpha_i|^2$. For instance, measuring the state $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ produces a random uniform binary output. The stochasticity of quantum measurements is another feature that calls for probabilistic data management frameworks. Moreover, the state collapse phenomenon significantly complicates the tasks, as the quantum state cannot be entirely "recycled" following a measurement.

One may attempt to think of a quantum state $|\phi\rangle = \sum_i \alpha_i |i\rangle$ - as far as measurement is concerned - as a discrete probability distribution $\{p_1, \ldots, p_d\}$, but there are two fundamental differences. First, the coefficients (called *amplitudes*) α_i's are complex numbers that make superposition and interference possible. Second, the probability of an outcome in quantum mechanics is found by taking the *absolute square* of the amplitude, that is, $p_i = |\alpha_i|^2$.

In general, a certain measurement $\mathcal{M}$ on a quantum state can be obtained in three stages: (i) applying an appropriate quantum operator U to the state, (ii) measuring the evolved state $U|\phi\rangle$ in the computational basis; and (iii) applying classical post processing on the measurement outcomes. This procedure is compactly modeled as a matrix M called an *observable* that is multiplied by the original state $|\phi\rangle$. The eigenvalues of M represent the possible values of the measurement outcomes. Moreover, by $\mathcal{M}(|\phi\rangle)$ we denote the probability distribution of the measurement outcomes after applying $\mathcal{M}$ on $|\phi\rangle$. Because the outcomes are probabilistic, we are often interested in their expectation values. The expectation of the outcome distributed by $\mathcal{M}(|\phi\rangle)$ is equal to $\langle\phi|M|\phi\rangle$, where $\langle\phi|$ is the complex conjugate transpose of the vector $|\phi\rangle$.

Standard Math Notations Versus Dirac Notations. As this paper is intended for an audience within the database community, we recognize that the Dirac bra-ket notation might appear unfamiliar to database researchers without a background in quantum information and computing. To simplify, in the main text we express a pure quantum state as a column vector with dimensions denoted as d, and use ϕ and $\phi^\dagger$ to denote $|\phi\rangle$ and $\langle\phi|$, respectively. We use $\phi^\dagger M \phi$ to denote the expectation value $\langle\phi|M|\phi\rangle$ of an observable M. Throughout the paper, we reserve the notations ϕ and ψ for quantum states.

We have included a more formal (but still gentle) introduction of quantum information and computing using Dirac bra-ket notations in the full version of this paper [56].

Trace Distance. Given two quantum states ϕ and ψ, we define their trace distance to be $D(\phi, \psi) = \sqrt{1 - |\psi^\dagger \phi|^2}$. The trace distance is the most widely used distance measure for quantum states in the literature.

2.1 Performance Metrics

In the context of quantum data, similar to classical database design, the efficiency of space and time is crucial during database initialization, indexing, and querying. Minimizing the number of quantum state copies used for constructing sketches is also important, as obtaining state copies can be costly and they cannot be fully recycled due to post-measurement disturbance. However, as we mentioned earlier, sample complexity is a secondary consideration in the data management setting, since the sketch-building/initialization is a one-time process.

A unit-time quantum operation comprises standard single-qubit gates like the Hadamard gate, Pauli gates, phase gate, and T gate, as well as a two-qubit gate, such as the Controlled-NOT (CNOT) gate, that enables entangling operations.[2] The combination of these gates is sufficient to approximate any unitary operation to arbitrary accuracy. We call these gates *unit gates*, and define the *size* of a circuit (for representing a unitary operation) to be the number of unit gates in the circuit.

As mentioned, a typical quantum measurement $\mathcal{M}$ on n qubit systems consists of a unitary operator $U_{\mathcal{M}}$ followed by measurement in computational basis and classical post processing. Assuming that the classical post processing is polynomial, the overall time cost is typically dominated by the gate complexity of $U_{\mathcal{M}}$. It has been shown in [50] that a circuit depth of $\Theta(2^n/n)$ (i.e., $\Theta(d/\log d)$) is needed for constructing an arbitrary unitary operator U. To simplify matters, we assume that both executing an arbitrary d-dimensional quantum measurement and preparing an arbitrary d-dimensional state require $O(d)$ quantum time.

2.2 Nearest Neighbor in High Dimensions

As quantum states are inherently high dimensional, even after effective sketching and summarization that we will illustrate in the subsequent sections, we will thus use *Approximate Nearest Neighbor (ANN)* via *Locality Sensitive Hashing (LSH)* to further speed up some database operations. This subsection will take a brief detour from our discussion of quantum data management.

▶ **Definition 1** ((r,β)-ANN-search). *Let X be a database containing a set of vectors in $\mathbb{R}^d$ and $q \in \mathbb{R}^d$ be a query vector. Let $dist(\cdot,\cdot)$ be a distance function. If there is at least one vector $p \in X$ with $dist(p,q) \le r$, return any $p' \in X$ with $dist(p',q) \le \beta r$. Otherwise, either return a $p' \in X$ with $dist(p',q) \le \beta r$ or return $\emptyset$.*

Let us focus on the case that the distance function $dist(\cdot,\cdot)$ is ℓ_1 or ℓ_2. Indyk and Motwani [34] showed that (r,β)-ANN can be solved efficiently via LSH. The idea is that we first apply multiple hash functions to each vector in X; this part can be pre-computed and stored as an indexing. At the time of query, we apply the same set of hash functions to the query vector q. We then run over all vectors $p \in X$ such that p and q collide (i.e., fall into the same bin) on at least one hash function, and return the first vector p if $dist(p,q) \le \beta r$. If no such p found after traversing a certain number of vectors in X, we return $\emptyset$.

We will use $\texttt{ANN}(q,X,r,\beta)$ to denote the (r,β)-ANN search for a query vector q in database X. The following is a summary of results on LSH-based ANN for ℓ_1/ℓ_2 distances.

▶ **Theorem 2** ([34, 14, 5]). *For $dist(\cdot,\cdot)$ being ℓ_1 or ℓ_2, a database X of m vectors, and a d-dimensional vector q, there is an algorithm that solves $\texttt{ANN}(q,X,r,\beta)$ using $O(dm+m^{1+\gamma})$ space and $O(dm^{\gamma})$ classical time, where $\gamma \approx 1/\beta$ for ℓ_1 distance and $\gamma \approx 1/\beta^2$ for ℓ_2 distance.*

[2] We refer the readers to [41] for a detailed introduction of these gates.

▶ Remark 3. We note that if we do not terminate the algorithm after encounter the first $p \in X$ such that $dist(p,q) \leq \beta r$, then the same algorithm can return a subset $Y \subseteq X$ including *all* vectors p such that $dist(p,q) \leq r$, and excluding all vectors p such that $dist(p,q) \geq \beta r$.

▶ Remark 4. We can also use LSH to find a set J of pairs of vectors such that J includes all pairs (p,q) such that $dist(p,q) \leq r$, and excludes all pairs (p,q) such that $dist(p,q) \geq \beta r$. To this end, we first hash all vectors, and then check the distances of all pairs of vectors that collide on at least one hash function.

3 Basic Operations on Quantum Data

The characteristics of quantum information dictate that we can only obtain an *approximation* of a quantum state ϕ with a finite number of quantum state copies. A celebrated result in quantum state tomography states that to learn an unknown n-qubit quantum state ϕ up to a trace distance ϵ, we already need $\Omega\left(d/\epsilon^2\right)$ copies of the quantum state, where $d = 2^n$ is the dimension of ϕ [21, 43]. We thus consider two quantum states ϕ, ψ with $D(\phi,\psi) \leq \varepsilon$ the same state. Consequently, all the operations that we support in a quantum database also need to be approximate. The precise definition of "approximation" varies for different operations.

In this section, we formulate basic quantum data operations that we aim to support using our proposed sketches. When we say the return of a quantum state ϕ, we are referring to its identifier.

3.1 Equality Test

In the classical data setting, the equality test on two data objects returns 1 if $p = q$, and returns 0 otherwise. In the quantum setting, since we cannot distinguish two quantum states using $o(d/\varepsilon^2)$ copies of the states if their trace distance is at most ε, we need to introduce the approximation version of the equality test:

▶ **Definition 5** ((ε,β)-equality-test)**.** *Given two quantum states ϕ and ψ, output* 1 *if $D(\phi,\psi) \leq \varepsilon$, and* 0 *if $D(\phi,\psi) > \beta\varepsilon$. The output can be arbitrary if $\varepsilon < D(\phi,\psi) \leq \beta\varepsilon$.*

In words, we consider two quantum states the same if their trace distance is at most ε, and different if their trace distance is more than $\beta\varepsilon$. If the distance falls between the two values, then the decision can be arbitrary. The gap between *yes* and *no* is inevitable for quantum data.

Given two quantum states ϕ and ψ, which may be unknown, the standard method for estimating their trace distance is the swap test [8]. The algorithm uses a controlled-SWAP gate (can be implemented using $O(n) = O(\log d)$ unit gates) and two single-qubit Hadamard gates. The test outputs 1 with probability $\frac{1+\left|\phi^\dagger\psi\right|^2}{2} = 1 - \frac{D(\phi,\psi)^2}{2}$, and 0 otherwise. Therefore, using $O_\beta\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ such tests (the constant hidden in the big-O depends on the constant β), we can differentiate the case $D(\phi,\psi) > \beta\varepsilon$ from $D(\phi,\psi) \leq \varepsilon$ with a probability $1-\delta$.

The main issue with this algorithm is that we have to consume fresh copies of database states for each equality test, which is *unsustainable* for a database system that is designed to answer an unlimited number of queries.

3.2 Search and Join

In the classical data setting, given a set of objects $X = \{p_1, \ldots, p_n\}$ and a query object q, the search operation returns some $p_i \in X$ such that $p_i = q$ if such p_i exists, and $\emptyset$ otherwise. In the quantum setting, again due to the difficulty of distinguishing two quantum states within a distance of ϵ, we propose the following approximation version.

▶ **Definition 6** ((ε, β)-search). *Given a query state ϕ and a database X, if there exist a state $\psi \in X$ such that $D(\phi, \psi) \leq \varepsilon$, return a state $\psi' \in X$ with $D(\phi, \psi') \leq \beta\varepsilon$. Otherwise, either return a state $\psi' \in X$ with $D(\phi, \psi') \leq \beta\varepsilon$ or return $\emptyset$.*

In other words, if there exists a state in the database which has a trace distance no more than ε from the query state ϕ, we return a state in X whose distance is no more than $\beta\varepsilon$ from ϕ (similar to the ANN search). Else if all states in the database have distances larger than $\beta\varepsilon$ from the query state, we return $\emptyset$. In other cases, we either return a database state with distance no more than $\beta\varepsilon$ from the query state or return $\emptyset$.

The most straightforward way is to perform the (ε, β)-equality-test for each database state $\psi \in X$ with the query state ϕ. By the above algorithm for equality test (setting $\delta = 1/m^2$), we can determine with probability $(1 - m\delta) = (1 - o(1))$ whether there exists a state $\psi \in X$ such that the (ε, β)-equality-test on ϕ and ψ returns 1. The above procedure takes $O\left(m \log d \log m/\varepsilon^2\right)$ quantum time, which is linear in terms of the number of states in the database. Another significant limitation of this method is the necessity of using fresh copies of the database states for each search operation because of the equality test, making the database system unsustainable.

A closely related operation to search is join, which is one of the most important operations in relational database systems. We introduce the quantum version of *natural join* as follows.

▶ **Definition 7** ((ε, β)-natural-join). *Given two databases X and Y of quantum states, we want to output a set that includes all pairs of states (ϕ, ψ) $(\phi \in X, \psi \in Y)$ such that $D(\phi, \psi) \leq \epsilon$, and excludes all pairs (ϕ, ψ) such that $D(\phi, \psi) > \beta\epsilon$. The decisions for other pairs can be arbitrary.*

3.3 Selection and Sorting

In relational databases for classical data, selection is typically denoted by $\sigma_\theta(R)$, where R is a relation and θ is a propositional formula that involves an attribute, a comparison operator in the set $\{<, >, \leq, \geq, =, \neq\}$, and a constant value for comparison (e.g., age ≥ 8). However, in the quantum data setting, quantum states cannot be directly compared. We can only apply a measurement $\mathcal{M}$ on the state ϕ and get a random outcome according to the distribution $\mathcal{M}(\phi)$. As a classical analog, we would say a person's age is 5 with probability 0.6 and 10 with probability 0.4.[3] We thus look at the expectation value $\phi^\dagger M \phi$ for the observable M corresponding to $\mathcal{M}$.

The quantity $\phi^\dagger M \phi$ holds significant importance in quantum mechanics (see, e.g., the textbook [45]). It can be used to provide an estimate of the system's average energy in a particular state, describe the level of non-classical correlations between entangled particles, quantify quantum information such as entropy, coherence, and entanglement, etc.

We define the ε-approximate "$\geq$" selection operation for quantum data as follows.

[3] This assembles probabilistic databases, but in the quantum data setting the probability distribution is not given explicitly, and the support size of the distribution is exponential in terms of the number of qubits of each quantum state.

▶ **Definition 8** ((η, ε)-selection). *Given a database X, an observable M, a threshold η, and an error parameter ε, return a set of states $S \subseteq X$ such that S includes all database states ϕ such that $\phi^\dagger M \phi \geq \eta$, but excludes all ϕ such that $\phi^\dagger M \phi \leq \eta - \varepsilon$.*

Note that the ε-approximate equality selection can be implemented by taking the difference between $(\eta - \varepsilon, \varepsilon)$-selection and $(\eta + 2\varepsilon, \varepsilon)$-selection, which includes all ϕ with $\eta - \varepsilon \leq \phi^\dagger M \phi \leq \eta + \varepsilon$ and excludes all ϕ with $\phi^\dagger M \phi \leq \eta - 2\varepsilon$ or $\phi^\dagger M \phi \geq \eta + 2\varepsilon$. In the context of approximation, we can consider "$<$" and "$>$" the same as "$\leq$" and "$\geq$", respectively.

We also note that (ε, β)-search can also be handled by looking at $\phi^\dagger M \phi$ for a specific observable M, although this solution is not as efficient as that using the particular sketches that we shall design for the search operation. We have included a reduction from (ϵ, β)-search to (η, ϵ)-selection in the full version of this paper [56].

In the context of databases, we are particularly interested in the following type of observables.

▶ **Definition 9** (k-local observable). *An observable O of a system with n qubits is called k-local if it can be written as a sum of a constant number of terms, each acting on at most k qubits. For instance, a 2-local observable in a 3-qubit system might look like:*

$$O = O_{12} \otimes I_3 + I_1 \otimes O_{23},$$

Where O_{12} and O_{23} are operators acting on the pairs of qubits (1,2) and (2,3) respectively, while I_3 and I_1 are the identity operators acting on the remaining qubits.

k-local observables have been well studied in the literature (see [11, 37] and references therein). They are interesting because, in most practical scenarios, our goal is to identify specific properties of a quantum state (e.g., the energy, momentum, or spin of a photon) that rely on a small subset of qubits of the state. This is similar to the classical setting where most queries depend on a few attributes of a relational database table. For example, suppose we want to retrieve all records in a table containing patient information for individuals aged 80 years or older with systolic blood pressure at least 140, we only need to look at two attributes in the table: age and blood pressure. If we view each qubit of a quantum state as an attribute (e.g., spin, position, momentum, polarization, etc.), then a k-local observable performs selection on at most k attributes of the quantum state.

A related problem of selection is sorting. As a motivation, we would like to sort a set of given quantum states according to their average energy with respect to an observable determined by a particular application. Note that there is no natural order between the quantum states themselves. Therefore, introducing an observable and computing the expectation value is somewhat necessary to establish a total order between the quantum states.

We define the sorting operation with respect to an observable M as follows. Similar to the selection operation, we introduce an additive approximation ε in the sorted order.

▶ **Definition 10** (ε-sorting). *Given a database X of m states, an observable M, and an error parameter ε, return an order $(\phi_1, \phi_2, \ldots, \phi_m)$ of the states in X such that for all $i = 1, \ldots, m-1$, we have ${\phi_i}^\dagger M \phi_i \leq {\phi_{i+1}}^\dagger M \phi_{i+1} + \varepsilon$.*

4 Sketches for Quantum Data Operations

In this section, we introduce two quantum data sketches, *vector sketches* and *shadow seeds*, which are summaries of the original states for efficiently handling previously mentioned database operations.

Before delving into the details, let us use metaphors to provide some very high-level intuition of the two data summarizing methods. The vector sketches can be seen as capturing snapshots of the state from different angles, while each shadow seed can be seen as a piece of information gleaned from the state. Using multiple shadow seeds, we can reconstruct the original state at varying levels of resolution.

4.1 Vector Sketches for Equality-Test, Search, and Join

The concept of vector sketch is to represent a quantum state ϕ as a vector in $\mathbb{R}^t$ with $t \ll d$ instead of a vector in $\mathbb{C}^d$, while preserve certain distance properties. In this section, we design vector sketches for quantum states and then use them to conduct equality test, search, and join.

A natural way to construct the sketch is to take a number of random measurements on ϕ, and write down the measurement outcomes as a vector. The following result is due to Sen [47], rewritten for pure quantum states.

▶ **Theorem 11** ([47]). *Let ϕ and ψ be two pure quantum states in $\mathbb{C}^d$. With probability at least $\left(1 - e^{-\Omega(d)}\right)$ over the choice of a random measurement basis $\mathcal{M}_d = \{M_1, \ldots, M_d\}$, there exists a universal constant $c \in (0,1)$ such that*

$$c \cdot D(\phi, \psi) \leq \|\mathcal{M}_d(\phi) - \mathcal{M}_d(\psi)\|_1 \leq D(\phi, \psi). \tag{1}$$

Theorem 11 connects the trace distance of two quantum states to the ℓ_1 distance of their measurement outcome distributions. We note that the *distortion* in (1), $D(\phi,\psi)/(cD(\phi,\psi)) = 1/c$, is a big constant whose value left unspecified in [47].

Vectors $\mathcal{M}_d(\phi)$ and $\mathcal{M}_d(\psi)$ are discrete distributions with outcomes $\{1, 2, \ldots, d\}$. It is well-known that for a discrete distribution μ over a domain of size d, using $\Theta\left((d + \log(1/\delta))/\epsilon^2\right)$ samples we can obtain an empirical distribution $\widetilde{\mu}$ such that $\|\mu - \widetilde{\mu}\|_1 \leq \epsilon$ with probability $1 - \delta$ (see, e.g., [10]).

▶ **Corollary 12.** *Let $\widetilde{\mathcal{M}_d}(\phi)$ and $\widetilde{\mathcal{M}_d}(\psi)$ be the empirical distributions of measurement outcomes by applying $\mathcal{M}_d$ in Theorem 11 to $c_s(d + \log(1/\delta))/\epsilon^2$ (for a sufficiently large constant c_s) copies of ϕ and ψ, respectively. With probability $1 - \delta - e^{-\Omega(d)}$, we have*

$$c \cdot D(\phi, \psi) - \epsilon \leq \left\| \widetilde{\mathcal{M}_d}(\phi) - \widetilde{\mathcal{M}_d}(\psi) \right\|_1 \leq D(\phi, \psi) + \epsilon,$$

where $c \in (0,1)$ is a universal constant.

We can view $\widetilde{\mathcal{M}_d}(\phi)$ and $\widetilde{\mathcal{M}_d}(\psi)$ as two empirical probability vectors. However, since $d = 2^n$ for a n-qubit state, it is both space-expensive to store $\widetilde{\mathcal{M}_d}(\phi)$ and time-expensive to use it for database operations.

Embedding to L_1-space. We aim to address the issue of efficiency in both time and space by showing that there is another distribution of measurements whose number of outcomes is *independent* of the state dimension d, for which a similar connection exists between the trace distance of two quantum states and the ℓ_1 distance of the corresponding measurement outcome distributions. Moreover, the distortion of our sketching can be made arbitrarily close to 1 (compared with $1/c$ in (1)). It is worth noting that this distortion will significantly impact the efficiency of the search and join operations, as we will discuss shortly.

Our result is summarized in the following theorem.

▶ **Theorem 13.** *Let ϕ and ψ be two pure d-dimensional quantum states. For any $\iota > 0$, there is a distribution π of measurements with $k = c\log(1/\delta)/\iota^2$ outcomes for a sufficiently large constant c, such that a measurement $\mathcal{M}_k$ sampled randomly from π satisfies*

$$(1-\iota)D(\phi,\psi) \le \sqrt{\frac{d}{k}}c_\tau \|\mathcal{M}_k(\phi) - \mathcal{M}_k(\psi)\|_1 \le (1+\iota)D(\phi,\psi)$$

with probability at least $(1-\delta)$, where $c_\tau \in [0.48, \sqrt{2}]$ is a universal computable constant. Additionally, the measurement sampling can be completed in $O(\log^8 d)$ time, and the sampled measurement can be represented as a quantum circuit with a gate complexity of $O(\log^2 d)$.

Proof Overview. At a high level, our approach leverages form dimension reduction through quantum measurements. We make use of a technique called *pretty good measurement* [31] to generate random projective quantum measurements $\mathcal{M}$ with k outcomes. The output of these measurements are random vectors serving as the embedding of the state ϕ into $\mathbb{R}^k$.

We start by picking a random basis for $\mathbb{C}^d$ based on the Haar measure [30]. Let x_t, y_t $(t = 1, \ldots, d)$ be independent Gaussian random variables with mean zero and variance $\sigma^2 = \frac{1}{2d}$, and let $g \triangleq (c_1, \ldots, c_d) \in \mathbb{C}^d$ be a random vector where $c_t = x_t + iy_t$. We repeat this process and generate d *complex* Gaussian random vectors $g_1, \ldots, g_d$. These vectors are linearly independent with probability one; but they are *not* necessarily orthonormal. We make use of pretty good measurement to orthogonalize and normalize these vectors. More precisely, we construct the operator (matrix) $\Gamma \triangleq \sum_{t\in[d]} g_t^\dagger g_t$, and define the vector $\gamma_t \triangleq \Gamma^{-1/2} g_t$ for each $t \in [d]$. We can show that $\gamma_1, \ldots, \gamma_d$ are linearly independent and are orthonormal. Moreover, the distribution of γ_t is unitary invariant, and hence the Haar measure. Intuitively, γ_t is distributed uniformly over surface of the unit sphere in $\mathbb{C}^d$. Next, we randomly group γ_t's into k groups and form random projection operators as

$$\Pi_j = \sum_{\ell\in[d/k]} (\gamma_\ell^j)^\dagger \gamma_\ell^j \,. \qquad (j = 1, \ldots, k)$$

Let $\mathcal{M}_k = \{\Pi_1, \cdots, \Pi_k\}$ be the corresponding measurement. Clearly, $\mathcal{M}$ is a valid measurement with probability one. This random measurement facilitates an embedding of the quantum states in $\mathbb{C}^d$ into $\mathbb{R}^k$. We carefully analyze the distortion of the embedding (i.e., the outcome distribution by applying $\mathcal{M}_k$ to the quantum state) using tools from the concentration of measures and properties of the Haar distribution. We show that the distortion of this embedding is no more than $(1+\iota)$ with probability $(1-\delta)$ when $k = c\log(1/\delta)/\iota^2$ for a constant c. The complete proof can be found in the full version of this paper [56].

The measurement construction described above could require polynomial time in d. However, we demonstrate that it can be sampled more efficiently from the Clifford group in classical time $O(\log^8 d)$, leveraging the properties of unitary 2-designs from quantum information theory. The details can be found in the full version of this paper [56]. ◀

To approximate $\sqrt{\frac{d}{k}}c_\tau\|\mathcal{M}_k(\phi) - \mathcal{M}_k(\psi)\|_1$ up to an additive error ϵ, we have to approximate $\|\mathcal{M}_k(\phi) - \mathcal{M}_k(\psi)\|_1$ up to $\epsilon' = \frac{\epsilon}{\sqrt{d/k}\cdot c_\tau}$. We have the following immediate corollary.

▶ **Corollary 14.** *For any $\iota > 0$, let $k = c\log(1/\delta)/\iota^2$ for a sufficiently large constant c, and let $\widetilde{\mathcal{M}_k}(\phi)$ and $\widetilde{\mathcal{M}_k}(\psi)$ be the empirical distributions of the outcomes by applying independent random measurements $\mathcal{M}_k$ in Theorem 13 to $c_s d/\epsilon^2$ (for a sufficiently large constant c_s) copies of ϕ and ψ, respectively. With probability at least $1-\delta$, we have*

$$(1-\iota)D(\phi,\psi) - \epsilon \le \sqrt{\frac{d}{k}}c_\tau \left\|\widetilde{\mathcal{M}_k}(\phi) - \widetilde{\mathcal{M}_k}(\psi)\right\|_1 \le (1+\iota)D(\phi,\psi) + \epsilon,$$

where $c_\tau \in [0.48, \sqrt{2}]$ is the same constant in Theorem 13.

Embedding to L_2-space. The sketch we have constructed for the L_1-space can also be applied to the L_2-space, albeit through a different analysis. The ℓ_2 distance is interesting since we know from Theorem 2 that ℓ_2 enjoys a slightly better ANN scheme in term of time and space complexities, which will be useful for speeding up search and join operations. The proof of the following theorem can be found in the full version of this paper [56].

▶ **Theorem 15.** *Let ϕ and ψ be two pure d-dimensional quantum states. For any $\iota > 0$, there is a distribution π of measurements with $k = c\log(1/\delta)/\iota^2$ outcomes for a sufficiently large constant c, such that a measurement $\mathcal{M}_k$ sampled randomly from π satisfies*

$$(1-\iota)D(\phi,\psi) \leq \sqrt{\frac{d}{2}}\|\mathcal{M}_k(\phi) - \mathcal{M}_k(\psi)\|_2 \leq (1+\iota)D(\phi,\psi)$$

with probability at least $1-\delta$. Additionally, the measurement sampling can be completed in $O(\log^8 d)$ time, and the sampled measurement can be represented as a quantum circuit with a gate complexity of $O(\log^2 d)$.

For a discrete distribution μ over a domain of size d for any $d \geq 1$, it takes $\Theta\left(\log(1/\delta)/\epsilon^2\right)$ samples to obtain an empirical distribution $\widetilde{\mu}$ such that $\|\mu - \widetilde{\mu}\|_2 \leq \epsilon$ with probability $1-\delta$ (see, e.g., [10]). We have the following corollary.

▶ **Corollary 16.** *For any $\iota > 0$, let $k = c\log(1/\delta)/\iota^2$ for a sufficiently large constant c, and let $\widetilde{\mathcal{M}_k}(\phi)$ and $\widetilde{\mathcal{M}_k}(\psi)$ be the empirical distributions of the outcomes by applying independent random measurements $\mathcal{M}_k$ in Theorem 15 to $c_s d\log(1/\delta)/\epsilon^2$ (for a sufficiently large constant c_s) copies of ϕ and ψ, respectively. With probability $1-\delta$, we have*

$$(1-\iota)D(\phi,\psi) - \epsilon \leq \sqrt{\frac{d}{2}}\left\|\widetilde{\mathcal{M}_k}(\phi) - \widetilde{\mathcal{M}_k}(\psi)\right\|_2 \leq (1+\iota)D(\phi,\psi) + \epsilon.$$

Johnson-Lindenstrauss Lemma in Our Context. It is natural to ask whether existing dimension reduction techniques, such as the Johnson–Lindenstrauss (JL) lemma, can be applied directly to the d-dimensional vector representation $\alpha(\phi) = (\alpha_1, \ldots, \alpha_d) \in \mathbb{C}^d$ of a quantum state ϕ, or the outcome distribution $p(\phi) = (p_1, \ldots, p_d) \in \mathbb{R}^d$ ($p_i = |\alpha_i|^2$) when measured in the computational basis. After all, we can use quantum tomography to learn the representation $(\alpha_1, \ldots, \alpha_d)$ approximately. We would like to first point out that a direct application will not work, since we can construct simple examples demonstrating inherent distortions between the trace distance of quantum states and the ℓ_1/ℓ_2 distances of their d-dimensional vector representations ($\alpha(\phi)$ or $p(\phi)$), even when all the coordinates are real-valued and before any dimension reduction step. We leave the detailed examples and calculation to the full version of this paper [56]. In our examples, for the $\alpha(\phi)$ vector representation, the distortions between the trace distance of quantum states and the ℓ_1 and ℓ_2 distances of the two corresponding vectors are at least $\sqrt{d/6}$ and $\sqrt{1.5}$, respectively. And for the $p(\phi)$ vector representation, the distortions between the trace distance of quantum states and the ℓ_1 and ℓ_2 distances of the two corresponding vectors are at least $\sqrt{3}$ and $\sqrt{3d/4}$, respectively. Moreover, the JL lemma only takes real vectors.

We also note that there exists a near-linear lower bound for dimension reduction in the L_1 space [4], indicating that, unlike the JL lemma for L_2 space, dimension reduction in the L_1 space is not generally possible.

We note that there is a way to circumvent the issues for embedding quantum states into the L_2 space: for each state ϕ, we write its density matrix $\phi\phi^\dagger$ as a real-valued $2d^2$ dimensional vector v_ϕ. By some calculation, we can show that the ℓ_2 distance of v_ϕ and

v_ψ preserves the trace distance of the two original pure states ϕ and ψ. We then perform dimension reduction on the vectors v_ϕ using the JL lemma. Our sketching algorithm has the following advantages compared with this "full tomography plus JL lemma" approach (setting the error probability $\delta = 0.01$):

1. The memory usage of our sketch construction is *independent* of d, while the memory needed for storing the classical vector representation of the quantum state ϕ is $O(d)$ and that for the density matrix $\phi\phi^\dagger$ is $O(d^2)$.
2. Our sketch construction takes $\tilde{O}(d/\epsilon^2)$ time, while the full (pure) quantum state tomography takes $O(d^2/\epsilon^5)$ [17] time and the dimension reduction using the JL lemma needs another $O(d^2/\epsilon^2)$ time.

These comparisons demonstrate that our sketch construction using direct quantum measurements significantly outperforms the method of first converting the quantum state to its classical description followed by dimension reduction, both in terms of time and space, which are the main focus of this paper.

We now apply our embedding results to database operations.

The Equality-Test Operation. We observe that Corollary 14 and Corollary 16 directly provide a way for solving (ϵ, β)-equality-test. We just set $\iota = \epsilon = \frac{\varepsilon}{2}$, and use the ℓ_1 or ℓ_2 distances between the two vector sketches $\widetilde{\mathcal{M}_k}(\phi)$ and $\widetilde{\mathcal{M}_k}(\psi)$ to estimate $D(\phi, \psi)$ up to an additive error ε with probability $1 - \delta$. The running time is bounded by $O(k) = O(\log(1/\delta)/\varepsilon^2)$.

The Search Operation. We now illustrate how to use vector sketches and approximate nearest neighbor (ANN) to perform (ϵ, β)-search on quantum states.

Let ϵ and $(1+\iota)$ be the additive error and multiplicative error in Corollary 14/Corollary 16 for building $\left\{ \widetilde{\mathcal{M}_k}(\phi) \;\middle|\; \phi \in X \right\}$, respectively. We assume that an LSH indexing structure has already been built on top of $\widetilde{\mathcal{M}_k}(\phi)$'s to achieve the time and space usages stated in Theorem 2. To handle (ε, β)-search, we call $\texttt{ANN}\left(\widetilde{\mathcal{M}_k}(\phi), \left\{ \widetilde{\mathcal{M}_k}(\psi) \;\middle|\; \psi \in X \right\}, (1+\iota)\varepsilon, \beta_{nn}\right)$, where $\beta_{nn} = \beta/(1 + \iota + \epsilon/\varepsilon)$ is the parameter for the tradeoff between the distortion and the time/space complexity in ANN. By Corollary 14/Corollary 16 and Theorem 2, if there exists a state $\psi \in X$ such that $D(\phi, \psi) \leq \varepsilon$, then ANN returns a state $\psi' \in D$ such that $D(\phi, \psi') \leq \beta\varepsilon$. On the other hand, ANN either returns a state $\psi' \in D$ with $D(\phi, \psi') \leq \beta\varepsilon$, or returns $\emptyset$.

By Theorem 2, it takes $O(km^\gamma) = O(m^\gamma \log m/\varepsilon^2)$ classical time to perform the search. The space for storing the LHS index is $O(km + m^{1+\gamma}) = O(m \log m/\varepsilon^2 + m^{1+\gamma})$, where $\gamma \approx 1/\beta_{nn}$ for ℓ_1 and $\gamma \approx 1/\beta_{nn}^2$ for ℓ_2.

We note that in the above approach, we have to make sure that $\beta_{nn} \geq 1$. In other words, we can only handle (ϵ, β)-search with $\beta \geq (1 + \iota + \epsilon/\varepsilon)$. However, since ϵ and ι can be positive constants arbitrarily close to 0, we can essentially handle all constants $\beta > 1$. Certainly, the higher the value of β, the larger β_{nn} that we can pick for reducing the query time and space usage in the ANN search. In practice, a reasonably large constant β may be okay, as the trace distance between two quantum states that are generated by separate entities or experiments is typically much larger than that between two states originating from the same entity or experiment (due to quantum noise or preparation errors).

Setting $\delta = 1/m^2$, $\iota = 0.01$ and $\epsilon = 0.01\varepsilon$, we have $\beta_{nn} \geq 0.98\beta$, and consequently $\gamma \leq 1.05/\beta^2$. Applying our vector sketch with respect to the ℓ_2 distance and the corresponding ANN search, we have the following theorem.

▶ **Theorem 17.** *There is an index of size* $O\left(\frac{m\log m}{\varepsilon^2} + m^{1+\frac{1.05}{\beta^2}}\right)$*, using which we can solve* (ϵ,β)*-search on a database of* m *quantum states with success probability* $1-o(1)$ *and classical time* $O\left(m^{\frac{1.05}{\beta^2}}\cdot\frac{\log m}{\varepsilon^2}\right)$.

Note that the index space cost is *independent* of d, and the query time is *sublinear* in m (for $\beta > \sqrt{1.05}$) and *independent* of the state dimension d.

The Join Operation. The sketch-based approach can also be used for join. Given a set of sketch vectors $\left\{\widetilde{\mathcal{M}_k}(\phi) \;\middle|\; \phi\in X\right\}$, we can apply the same hashing process as that for the ANN search, and then verify (by computing the actual distance) all pairs of vectors that collide on at least one hash function. The space cost is the same as that of the search. The query time is dependent on the size of the join output, but it is still independent of the state dimension d.

4.2 Shadow Seeds for Selection and Sorting

In this section, we develop a classical data summarization that can be used to estimate the expectation value $\phi^\dagger M\phi$ for an arbitrary k-local observable M. We make use of the classical shadow tomography (CST), introduced in [32], to approximate $\phi^\dagger M\phi$ up to a small additive error. CST tries to extract minimal information about the quantum state, without performing complete tomography, to estimate certain properties of the state described by observables.

For completeness, let us briefly describe the CST procedure using Pauli measurements. For each of the N copies of ϕ, we select n unitary operators, $U_1,\ldots,U_n$, randomly and independently from the set $\{I,H,S^\dagger H\}$, where H is the Hadamard gate and $S=\sqrt{Z}$ is the square root of the Pauli-Z gate; Their matrix representations can be found in the full version of this paper [56]. We then apply U_j to the j-th qubit of ϕ and measure the state on the computational basis. The result is a binary string $b_1,\ldots,b_n\in\{0,1\}$. The n pairs $\{b_j,\texttt{index}(U_j))\}_{j=1}^n$ form a row vector, where $\texttt{index}(U_j)$ is the index of U_j in the set $\{I,H,S^\dagger H\}$. We then repeat this process for N times, getting N rows, forming the seed matrix $A(\phi)=\{b_{i,j},\texttt{index}(U_{i,j})\}_{i\in[N],j\in[n]}$. We call $A(\phi)$ the *shadow seeds*. Clearly, $A(\phi)$ can be stored using $O(nN)$ classical bits, since each entry of $A(\phi)$ belongs to $\{0,1\}\times\{0,1,2\}$.

At the time of query, given a k-local observable M, we first construct k-local classical shadows $\tilde{\phi}_i$ of the database state ϕ from each row $i\in[N]$ of its seed matrix $A(\phi)$ with respect to the k-local observable M. Suppose M depends non-trivially on the k qubits indexed by $Q\triangleq\{q_1,\ldots,q_k\}$. Let $e_0=(0,1)^T, e_1=(1,0)^T$ be the standard basis vectors in the two dimensional plane. For each row $i\in[N]$ and column $j\in Q$, we first construct a vector $v_{i,j}=U_{i,j}e_{b_{i,j}}$. Next, we construct the i-th shadow as a $2^k\times 2^k$ matrix $\hat{\rho}_i=\bigotimes_{j\in Q}\left(3v_{i,j}v_{i,j}^\dagger - I\right)$, where I is the 2×2 identity matrix. Finally, the estimator for $\phi^\dagger M\phi$ is given by $T=\frac{1}{N}\sum_{i\in[N]}\mathrm{tr}\{M\hat{\rho}_i\}$. The following theorem states that T is a good approximation of the expectation value $\phi^\dagger M\phi$.

▶ **Theorem 18** (Based on [32])**.** *The above procedure prepares an* $N\times n$ *shadow seed matrix* $A(\phi)$ *given* N *copies of an* n*-qubit quantum state* ϕ*, such that for any given* k*-local observable* M*, if* $N\geq 4^k\|M\|_\infty^2\log(1/\delta)/\varepsilon^2$*, the estimator* T *approximates* $\phi^\dagger M\phi$ *up to an additive error* ε *with probability* $(1-\delta)$ *using* $A(\phi)$*. Moreover, the time for computing* $\phi^\dagger M\phi$ *using* $A(\phi)$ *is bounded by* $O\left(2^{2k}N\right)(\propto 16^k)$*, and the space for storing* $A(\phi)$ *is* $O(Nn)$ *classical bits.*

Note that the space cost and query time are both independent of the state dimension d.

Typically, the k-local observable M can be expressed as a quantum circuit with poly(k) gate complexity. In this case, we propose a new estimation algorithm to further improve the total query time from $O(16^k)$ to $O(9^k)$ (omitting other less critical factors) by an approach we call QCQC (quantum→classical→quantum→classical). We have the following theorem, whose proof can be found in the full version of this paper [56].

▶ **Theorem 19.** *There is a procedure for preparing an $N \times n$ shadow seed matrix $A(\phi)$ given N copies of an n-qubit quantum state ϕ, such that for any given k-local observable M with poly(k) gate complexity, if $N \geq 9^k \|M\|_\infty^2 \log(1/\delta)/\varepsilon^2$, we can approximate $\phi^\dagger M \phi$ up to an additive error ε with probability $(1-\delta)$ using $A(\phi)$. Moreover, the quantum time for computing $\phi^\dagger M \phi$ using $A(\phi)$ is bounded by $O\left(N \text{poly}(k)\right) (\propto 9^k)$, and the space for storing $A(\phi)$ is $O(Nn)$ classical bits.*

The Selection Operation. It is easy to see that Theorem 19 directly implies an algorithm for handling (η, ε)-selection: Setting $\delta = 1/m^2$, we can estimate $\phi^\dagger M \phi$ up to an additive error ε with probability $(1 - 1/m^2)$ for each n-qubit database state ϕ using an $N \times n$ shadow seed matrix, where $N \geq 9^k \|M\|_\infty^2 \cdot 2 \log m/\varepsilon^2$. By a union bound over m database states, we can solve the (η, ε)-selection problem with probability $(1 - 1/m)$. The query time is bounded by $Nm \cdot \text{poly}(k) = 9^k m \log m \cdot \text{poly}(k) \|M\|_\infty^2 / \varepsilon^2$.

▶ **Theorem 20.** *There is an index of size $O\left(9^k n W^2 \log m/\varepsilon^2\right)$, using which we can solve for any k-local observable M $(\|M\|_\infty \leq W)$ the (η, ϵ)-selection on a database of m n-qubit quantum states with success probability $(1 - o(1))$ and quantum time $9^k m \log m W^2 \text{poly}(k)/\varepsilon^2$.*

The Sorting Operation. Since the shadow seed matrix can be used for estimating the expectation value $\phi^\dagger M \phi$ up to an additive error ε, we can use it for ε-sorting with the same space and time complexity as that for the selection operation.

5 Conclusion and Future Work

In this paper, we have defined basic database queries for quantum data and proposed several classical sketches of quantum states to facilitate these queries. We consider our work a preliminary step towards a comprehensive quantum data management system. Numerous questions and directions remain open following this work. We list a few below.

Support More Data Operations. This paper primarily focuses on two basic database operations: search and selection, along with several related operations. We would like to expand the support to more complex operations for data analytics, such as *clustering* and *classification*, for which we may need to develop new classical summaries of the quantum states for the sake of efficiency.

Mixed States. In various scenarios, such as when the description of a quantum system is unknown due to quantum noise, the use of a density operator (or, density matrix) for describing *mixed* quantum states becomes more convenient. Suppose the quantum system is in one of a collection of d-dimensional pure states $\{\phi_1, \ldots, \phi_k\}$, we can represent a mixed quantum state as $\rho = \sum_{i=1}^k p_i \phi_i {\phi_i}^\dagger$, where $p_1, \ldots, p_k \geq 0$ and $\sum_{i=i}^k p_i = 1$. We can view ρ as a convex combination of outer products of pure states ϕ_i, where each $\phi_i {\phi_i}^\dagger$ is associated with a probability p_i. We anticipate that results presented in this paper can be extended to mixed states, although the technical aspects of this generalization require further investigation.

The Integration with the Theory of Relational Databases. A key feature of our proposed model is that quantum data is represented entirely in the classical format. This unique aspect enables us to integrate our model with established theories related to indexing, query execution, and query optimization in relational databases designed for classical data. However, the integration process will likely require the redesign of multiple components to accommodate the inherent differences stemming from the distinct definitions of database operations for quantum data.

References

1 Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007.

2 Scott Aaronson. Shadow tomography of quantum states. In Ilias Diakonikolas, David Kempe, and Monika Henzinger, editors, *STOC*, pages 325–338. ACM, 2018. doi:10.1145/3188745.3188802.

3 Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A*, 70(5):052328, November 2004. doi:10.1103/physreva.70.052328.

4 Alexandr Andoni, Moses Charikar, Ofer Neiman, and Huy L. Nguyen. Near linear lower bound for dimension reduction in L1. In Rafail Ostrovsky, editor, *FOCS*, pages 315–323. IEEE Computer Society, 2011. doi:10.1109/FOCS.2011.87.

5 Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*, pages 459–468. IEEE Computer Society, 2006. doi:10.1109/FOCS.2006.49.

6 Costin Badescu and Ryan O'Donnell. Improved quantum data analysis. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC*, pages 1398–1411. ACM, 2021. doi:10.1145/3406325.3451109.

7 Costin Badescu, Ryan O'Donnell, and John Wright. Quantum state certification. *CoRR*, abs/1708.06002, 2017. arXiv:1708.06002.

8 Harry Buhrman, Richard Cleve, John Watrous, and Ronald De Wolf. Quantum fingerprinting. *Physical Review Letters*, 87(16):167902, 2001.

9 Umut Çalikyilmaz, Sven Groppe, Jinghua Groppe, Tobias Winker, Stefan Prestel, Farida Shagieva, Daanish Arya, Florian Preis, and Le Gruenwald. Opportunities for quantum acceleration of databases: Optimization of queries and transaction schedules. *Proc. VLDB Endow.*, 16(9):2344–2353, 2023. doi:10.14778/3598581.3598603.

10 Clément L. Canonne. A short note on learning discrete distributions, 2020. arXiv:2002.11457.

11 Thomas Chen, Shivam Nadimpalli, and Henry Yuen. Testing and learning quantum juntas nearly optimally. In Nikhil Bansal and Viswanath Nagarajan, editors, *SODA*, pages 1163–1185, 2023.

12 Kai-Min Chung and Han-Hsuan Lin. Sample efficient algorithms for learning quantum channels in PAC model and the approximate state discrimination problem. In Min-Hsiu Hsieh, editor, *TQC*, volume 197 of *LIPIcs*, pages 3:1–3:22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICS.TQC.2021.3.

13 Paul Cockshott. Quantum relational databases, 1997. arXiv:quant-ph/9712025.

14 Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In Jack Snoeyink and Jean-Daniel Boissonnat, editors, *SOCG*, pages 253–262. ACM, 2004. doi:10.1145/997817.997857.

15 Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm, 2014. doi:10.48550/arXiv.1411.4028.

16 Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors, February 2018. arXiv:1802.06002.

17 Daniel Stilck França, Fernando G. S. L. Brandão, and Richard Kueng. Fast and robust quantum state tomography from few basis measurements. In Min-Hsiu Hsieh, editor, *16th Conference on the Theory of Quantum Computation, Communication and Cryptography, TQC 2021, July 5-8, 2021, Virtual Conference*, volume 197 of *LIPIcs*, pages 7:1–7:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICS.TQC.2021.7.

18 Siddhant Garg and Goutham Ramakrishnan. Advances in quantum deep learning: An overview. *arXiv:2005.04316*, May 2020. arXiv:2005.04316.

19 Weiyuan Gong and Scott Aaronson. Learning distributions over quantum measurement outcomes. *CoRR*, abs/2209.03007, 2022. doi:10.48550/arXiv.2209.03007.

20 Lov K. Grover. A fast quantum mechanical algorithm for database search. In Gary L. Miller, editor, *STOC*, pages 212–219. ACM, 1996. doi:10.1145/237814.237866.

21 Jeongwan Haah, Aram W. Harrow, Zheng-Feng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In Daniel Wichs and Yishay Mansour, editors, *STOC*, pages 913–925. ACM, 2016. doi:10.1145/2897518.2897585.

22 Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, pages 1–1, 2017. doi:10.1109/tit.2017.2719044.

23 Rihan Hai, Shih-Han Hung, and Sebastian Feld. Quantum data management: From theory to opportunities. In *ICDE*, pages 5376–5381. IEEE, 2024. doi:10.1109/ICDE60146.2024.00410.

24 Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009.

25 Aram W. Harrow, Cedric Yen-Yu Lin, and Ashley Montanaro. Sequential measurements, disturbance and property testing. In Philip N. Klein, editor, *SODA*, pages 1598–1611. SIAM, 2017. doi:10.1137/1.9781611974782.105.

26 Aram W. Harrow and John C. Napp. Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms. *Physical Review Letters*, 126(14):140502, April 2021. doi:10.1103/physrevlett.126.140502.

27 Aram W. Harrow and Andreas J. Winter. How many copies are needed for state discrimination? *IEEE Trans. Inf. Theory*, 58(1):1–2, 2012. doi:10.1109/TIT.2011.2169544.

28 Mohsen Heidari, Ananth Y. Grama, and Wojciech Szpankowski. Toward physically realizable quantum neural networks. *Association for the Advancement of Articial Intelligence (AAAI)*, 2022.

29 Mohsen Heidari, Arun Padakandla, and Wojciech Szpankowski. A theoretical framework for learning from quantum data. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

30 Alexander S. Holevo. *A Mathematical Introduction*. De Gruyter, Berlin, Boston, 2013. doi:doi:10.1515/9783110273403.

31 Alexander Semenovich Holevo. On asymptotically optimal hypotheses testing in quantum statistics. *Teoriya Veroyatnostei i ee Primeneniya*, 23(2):429–432, 1978.

32 Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics 16, 1050–1057 (2020)*, February 2020.

33 Noah Huffman, Dmitri Pavlichin, and Tsachy Weissman. Lossy compression for schrödinger-style quantum simulations, 2024. arXiv:2401.11088.

34 Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In Jeffrey Scott Vitter, editor, *STOC*, pages 604–613. ACM, 1998. doi:10.1145/276698.276876.

35 E. Joos, H. D. Zeh, C. Kiefer, D. J. W. Giulini, J. Kupsch, and I. O. Stamatescu. *Decoherence and the Appearance of a Classical World in Quantum Theory*. Springer, 2003.

36 Stephen P. Jordan, Keith S. M. Lee, and John Preskill. Quantum algorithms for quantum field theories. *Science*, 336(6085):1130–1133, June 2012. doi:10.1126/science.1217069.

37 Julia Kempe, Alexei Y. Kitaev, and Oded Regev. The complexity of the local hamiltonian problem. *SIAM J. Comput.*, 35(5):1070–1097, 2006. doi:10.1137/S0097539704445226.

38 Ian D. Kivlichan, Craig Gidney, Dominic W. Berry, Nathan Wiebe, Jarrod McClean, Wei Sun, Zhang Jiang, Nicholas Rubin, Austin Fowler, Alán Aspuru-Guzik, Hartmut Neven, and Ryan Babbush. Improved fault-tolerant quantum simulation of condensed-phase correlated electrons via trotterization. *Quantum*, 4:296, July 2020. doi:10.22331/q-2020-07-16-296.

39 Yang Liu and Gui Lu Long. Deleting a marked item from an unsorted database with a single query, 2007. arXiv:0710.3301.

40 Fabio Valerio Massoli, Lucia Vadicamo, Giuseppe Amato, and Fabrizio Falchi. A leap among entanglement and neural networks: A quantum survey. *arXiv:2107.03313*, July 2021. arXiv:2107.03313.

41 Isaac L. Chuang Michael A. Nielsen. *Quantum Computation and Quantum Information*. Cambridge University Pr., December 2010.

42 K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, September 2018. doi:10.1103/physreva.98.032309.

43 Ryan O'Donnell and John Wright. Efficient quantum tomography. In Daniel Wichs and Yishay Mansour, editors, *STOC*, pages 899–912. ACM, 2016. doi:10.1145/2897518.2897544.

44 Alexey Pyrkov, Alex Aliper, Dmitry Bezrukov, Yen-Chu Lin, Daniil Polykovskiy, Petrina Kamya, Feng Ren, and Alex Zhavoronkov. Quantum computing for near-term applications in generative chemistry and drug discovery. *Drug Discovery Today*, page 103675, 2023.

45 Jun John Sakurai and Eugene D Commins. Modern quantum mechanics, revised edition, 1995.

46 Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. The quest for a quantum neural network. *Quantum Information Processing*, 13(11):2567–2586, August 2014. doi:10.1007/s11128-014-0809-8.

47 Pranab Sen. Random measurement bases, quantum state distinction and applications to the hidden subgroup problem. In *CCC*, pages 274–287. IEEE Computer Society, 2006. doi:10.1109/CCC.2006.37.

48 Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Comput.*, 26(5):1484–1509, 1997. doi:10.1137/S0097539795293172.

49 Adam Smith, MS Kim, Frank Pollmann, and Johannes Knolle. Simulating quantum many-body dynamics on a current digital quantum computer. *npj Quantum Information*, 5(1):106, 2019.

50 Xiaoming Sun, Guojing Tian, Shuai Yang, Pei Yuan, and Shengyu Zhang. Asymptotically optimal circuit depth for quantum state preparation and general unitary synthesis, 2023. arXiv:2108.06150.

51 Immanuel Trummer and Christoph Koch. Multiple query optimization on the d-wave 2x adiabatic quantum computer. *Proc. VLDB Endow.*, 9(9):648–659, 2016. doi:10.14778/2947618.2947621.

52 James D. Whitfield, Jacob Biamonte, and Alán Aspuru-Guzik. Simulation of electronic structure hamiltonians using quantum computers. *Molecular Physics*, 109(5):735–750, March 2011. doi:10.1080/00268976.2011.552441.

53 Tobias Winker, Sven Groppe, Valter Uotila, Zhengtong Yan, Jiaheng Lu, Maja Franz, and Wolfgang Mauerer. Quantum machine learning: Foundation, new techniques, and opportunities for database research. In Sudipto Das, Ippokratis Pandis, K. Selçuk Candan, and Sihem Amer-Yahia, editors, *SIGMOD*, pages 45–52. ACM, 2023. doi:10.1145/3555041.3589404.

54 Xin-Chuan Wu, Sheng Di, Emma Maitreyee Dasgupta, Franck Cappello, Hal Finkel, Yuri Alexeev, and Frederic T. Chong. Full-state quantum circuit simulation by using data compression. In Michela Taufer, Pavan Balaji, and Antonio J. Peña, editors, *SC*, pages 80:1–80:24. ACM, 2019. doi:10.1145/3295500.3356155.

55 Ahmed Younes. Database manipulation on quantum computers, 2007. arXiv:0705.4303.

56 Qin Zhang and Mohsen Heidari. Quantum data sketches, 2025. arXiv:2501.06705.