Quantum Physical Unclonable Functions: Possibilities and Impossibilities

Myrto Arapinis[1], Mahshid Delavar[1], Mina Doosti[1], and Elham Kashefi[1,2]

[1] School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, UK
[2] Departement Informatique et Reseaux, CNRS, Sorbonne Université,
4 Place Jussieu 75252 Paris CEDEX 05, France

Keywords: Hardware Security, Physical Unclonable Functions, Quantum Cryptography, Quantum Cryptanalysis, Quantum Emulation, Unforgeability

Abstract. Physical Unclonable Functions (PUFs) are physical devices with unique behavior that are hard to clone. A variety of PUF schemes have been considered in theoretical studies as well as practical implementations of several security primitives such as identification and key generation. Recently, the inherent unclonability of quantum states has been exploited for defining (a partial) quantum analogue to classical PUFs (against limited adversaries). There are also a few proposals for quantum implementations of classical optical PUFs. However, none of these attempts provides a comprehensive study of Quantum Physical Unclonable Functions (QPUFs) with quantum cryptographic tools as we present in this paper. We formally define QPUFs, encapsulating all requirements of classical PUFs as well as introducing new ones inherent to the quantum setting such as testability. We develop a quantum game-based security framework for our analysis and define a new class of quantum attacks, called General Quantum Emulation Attack. This class of attacks exploits previously captured valid challenge-response pairs to emulate the action of an unknown quantum transformation on new input. We devise a concrete attack based on an existing quantum emulation algorithm and use it to show that a family of quantum cryptographic primitives that rely on unknown unitary transformations do not provide existential unforgeability while they provide selective unforgeability. Then, we express our results in the case of QPUF as an unknown unitary transformation.

1 Introduction

Canetti and Fischlin's result on the impossibility of achieving secure cryptographic protocols without any setup assumptions [1] has motivated a rich line of research investigating the advantages of making hardware assumptions in protocol design. The idea was first introduced by Katz in [2], and attracted the attention of researchers and developers as it eliminates the need to trust a designated party or to rely on computational assumptions. Instead, one adopts physical assumptions. Physical Unclonable Functions (PUFs) are one of the hardware assumptions that have greatly impacted the field [3].

PUFs are hardware structures designed to utilize the random physical disorders which appear in any physical device during the manufacturing process. Because of the uncontrollable nature of these random disorders, building a clone of the device is considered impractical. The behaviour of a PUF is usually equivalent to a set of Challenge-Response Pairs (CRPs) which are extracted through physically querying the PUF and measuring its responses. The PUF's responses depend on its physical features and are assumed to be unpredictable, i.e. even the manufacturer of the PUF, with access to many CRPs, cannot predict the response to a new challenge [4]. This property makes PUFs different from other hardware tokens in the sense that the manufacturer of a hardware token is completely aware of the behaviour of the token they have built [5].

So far, the cryptographic literature has only considered what we will call classical PUFs (or CPUFs) restricted to classical CRPs limiting the adversary to classical interactions with the PUF, even though this might be too weak of an adversary. For instance, many implementations of PUFs rely on optical systems and can thus be challenged with quantum states (photons). Restricting the security analysis of such structures to classical CRPs with a purely classical adversary might hide important and exploitable security vulnerabilities as we indicate in this paper. Considering the importance of CPUFs as a hardware security primitive in several real-world applications[3], it is crucial to investigate their security in the quantum regime as well. This is the main question that we address in this paper, demonstrating how a quantum adversary could break the security in some setting as well as how quantum devices could also provide quantum-secure PUF in some other settings. In particular, we identify the requirements a QPUF needs to meet to provide the main security property required for most of the QPUF-based applications, that is *unforgeability*[4]. All prior works on QPUFs [6–9] (see related work paragraph below) analysed their security only against adversaries with limited capabilities while the formal definition of QPUFs and their security properties remain missing from the literature.

Our Contributions. We choose the standard game-based security framework for our analysis of QPUFs and adapt it to the quantum setting by defining the *quantum game-based security framework* in the most general form in Section 4 in the spirit of [10–12]. We emphasize that the framework can be used to study any cryptographic primitive that can be represented as a unitary transformation. We formalize a new quantum attack called *General Quantum Emulation Attack*, and relying on the concept of *unknown unitary* we establish general results on the security of such cryptographic primitives in Section 4. Our new attack (defined in Section 4) is a generalisation of previously studied superposition and entanglement attacks [13–15].

We then present the formal definitions of QPUFs and their main security property that is unforgeability in Section 5.

[3] Recently SAMSUNG announced that in their new processor Exynox 9820 they have integrated SRAM based PUF to store and manage personal data in perfect isolation.

[4] *Unpredictability* and *unclonability* are other equivalent terms for this notion used often in the literature.

We are guided by the classical counterparts to establish the requirements that QPUFs should satisfy to enable their usage as a cryptographic primitive. Using, the general security framework and our proposed attack models from previous sections we establish the precise setting where the security of QPUFs could be obtained in Section 5. In more details, our contributions are as follows:

Quantum Game-based Security Framework - To analyse the security of QPUFs, we need a formal security framework that captures the limitations imposed by the quantum nature of transmitted messages between honest parties and corrupted ones. To this end, we adopt the standard game-based security framework and define a quantum analogue applicable for any cryptographic primitive modelled by a unitary transformation. Compared to the quantum unforgeability definitions introduced in [10–12, 16], our framework is a unified framework for defining the notion of *unforgeability* for any cryptographic primitive in the quantum regime. Importantly, it does not restrict the adversary to choosing quantumly-encoded classical bit strings as challenges; the challenge can be any quantum state in the corresponding Hilbert space, provided it has some minimum distance from the previously learnt states to avoid trivial attacks.

Moreover, we define a new feature within our quantum framework that has not been considered in [10–12, 16], namely enabling the challenger to test the equality between two arbitrary unknown quantum states. Note that previous frameworks for defining quantum security of cryptographic primitives do not necessarily require such a quantum testing feature due to the classical nature of messages that the adversary needs to output to win the game. However, this is particularly desired when QPUFs are used in device authentication (or identification) protocols. We define a general test algorithm abstracting the existing test algorithms in the literature such as the SWAP test and its generalised version. We use this test algorithm for establishing our results in the rest of the paper. Also, to eliminate the cases that the adversary may cheat and sends the learned pair as a new input-output pair, we introduce the notion of distinguishability of the challenge state from previously queried states.

Quantum Emulation Attacks - We define a new class of quantum attacks, named quantum emulation (QE) attack, that covers all the adversarial strategies for breaking the security of a cryptographic primitive by having quantum access to it. Then, inspired by the universal quantum emulator algorithm introduced by Marvian and Lloyd in [17], we devise successful concrete QE attacks against unforgeability-type properties of unitary primitives based on the full parameter estimation of the proposed algorithm presented in Section 3. We analyse the security of cryptographic primitives that can be modelled by *unknown unitary transformations*, against Quantum Polynomial-Time adversaries. Using our quantum emulation attack we prove that:

- No such primitive provides security against adversaries that choose the target challenge themselves;
- No primitive can provide security against adaptive adversaries that can query the primitive after receiving the challenge.

Note that these no-go results are uniquely due to the quantum setting and our QE attack. Indeed such strong security features could be obtained in the classical setting due to limitations of a classical emulator/attacker while as the quantum security is not a straightforward translation of the classical results.

– However, any such cryptographic primitive is secure under some limited assumptions, provided the adversary is static and does not choose the challenge.

While the above setting is weaker however it matches the practical setting where most of such primitives will be utilised (such as the QPUFs discussed next).

Quantum Physical Unclonable Functions - We first define QPUFs as quantum channels and formalize the standard requirements of robustness, uniqueness and collision-resistance for QPUFs. We show that the collision-resistance requirement imposes the underlying transformation of any QPUF to be unitary. To emphasize this fact we refer to such class of quantum PUFs as *unitary QPUFs* or UQPUFs. Furthermore, any realistic UQPUF has to be equipped with an efficient test algorithm to check the equality between the reference and response quantum states. Since to the best of our knowledge, there is no efficient test algorithm for mixed quantum states, we restrict our study to the UQPUFs that are operating on pure quantum states. We then focus on the notion of unforgeability for UQPUFs and give a formal definition for quantum unconditional, existential, adaptive selective and static selective unforgeability of UQPUFs. We show that any UQPUF can be viewed as an unknown unitary transformation matching the criterion of our quantum emulation attack as the main toolkit to prove:

– *No UQPUF provides Quantum Unconditional Unforgeability.* The attack presented in this no-go result is the correct analogue of the brute-force attack for CPUFs.
– *No UQPUF provides Quantum Existential Unforgeability.* The attack is a particular instance of the quantum emulation attack introduced in this paper. The adversary's goal is to generate a valid input-output pair of the target UQPUF such that it successfully passes the test algorithm. We show how the adversary breaks this security property of the UQPUFs using the quantum emulation algorithm.
– *No UQPUF provides Quantum Adaptive Selective Unforgeability.* We consider adaptive quantum adversaries who can query the target UQPUF after receiving the challenge. We show a QPT adversary that wins the corresponding security game by entangling a local ancillary state with the challenge.
– *Any UQPUF provides Quantum Selective Unforgeability.* In other words, no QPT adversary can, on average, generate the response of a UQPUF to random challenges. We prove, for any QPT adversary, the target challenge with high probability will lie in the orthogonal subspace. Since the adversary can query the UQPUF only before getting the target challenge, he cannot adapt the samples accordingly. It is important to mention this is the scenario that many UQPUF applications correspond to. Hence our main positive result provides the first formal treatment of a practical family of UQPUFs.

Related Works The concept of Physical Unclonable Functions was first introduced by Pappu *et al.* [18] in 2001, devising the first implementation of an Optical PUF. Optical PUFs were subsequently improved as to generating an independent number of CRPs [19]. Several structures of Physical Unclonable Functions were further introduced including Arbiter PUFs [20], Ring-Oscillator based PUFs [21, 22] and SRAM PUFs [23]. For a comprehensive overview of existing PUF structures, we refer the reader to [24, 25]. There are also a few works that have formally defined secure PUFs and PUF-based protocols investigating their security in the game-based model or in the Universal Composability (UC) framework [26, 5, 27, 3].

Recently, the concept of "quantum read-out of PUF (QR-PUF)" was introduced in [6] to exploit the no-cloning feature of quantum states to potentially solve the spoofing problem in the remote device identification. In this line of research, QR-PUFs are assumed to be unitary transformations that map pure challenge quantum states to pure response quantum states. The identification protocol based on QR-PUF proposed in [6, 7] has been implemented in [28]. Continuous variable encoding has been also exploited to define practical QR-PUF based identification protocol [8]. However, the security of all these schemes has been analysed in non-standard security models by imposing the attacker to efficiently estimate or clone the state of a single copy of an unknown challenge quantum state, [6, 7, 29–31]. In contrary, in this work, we formalise the notion of QPUF and analyse its main property, i.e. unforgeability, in a standard model against general quantum adversaries including unbounded and Quantum Polynomial-Time (QPT) adversaries. Our framework would enable one to revisit these prior proposals to obtain practical yet secure QPUFs with currently emerging quantum devices and quantum communication networks.

2 Preliminaries

2.1 Physical Unclonable Functions

In this section, we briefly present the formal definition of Physical Unclonable Functions (PUFs) as found in the classical literature [26, 4, 5]. Let a $\mathcal{D}$-family be a set of physical devices generated through the same manufacturing process. Due to unavoidable variations during manufacturing, each device has some unique features that are not easily clonable. A Physical Unclonable Function (PUF) is an operation making these features observable and measurable by the holder of the device.

As in [26, 5], we formalize the manufacturing process of a PUF by defining the Gen algorithm that takes the security parameter λ as input and generates a PUF with an identifier **id**. Note that each time the Gen algorithm is run, a new PUF with new **id** is built. So, we have:

$$\mathrm{PUF}_{\mathbf{id}} \leftarrow \mathrm{Gen}(\lambda). \tag{1}$$

Also, we define the Eval algorithm that takes a challenge x and $\mathrm{PUF}_{\mathbf{id}}$ as inputs and generates the corresponding response $y_{\mathbf{id}}$ as output:

$$y_{\mathbf{id}} \leftarrow \mathrm{Eval}(\mathrm{PUF}_{\mathbf{id}}, x). \tag{2}$$

Due to variations in the environmental conditions, for any given $\mathrm{PUF}_{\mathbf{id}}$, the Eval algorithm may generate a different response to the same challenge x. It is required that this noise be bounded as follows; if $\mathrm{Eval}(\mathrm{PUF}_{\mathbf{id}}, x)$ is run several times, the maximum distance between the corresponding responses should at most be δ_r. This requirement is termed the *robustness requirement*.

Consider a family of $\mathrm{PUF}_{\mathbf{id}}$ created by the same Gen algorithm, and assume the algorithm Eval is run on all of them with a single challenge x. To be able to distinguish each PUF, it is required that the minimum distance between the corresponding responses be at least δ_u. This requirement is termed the *uniqueness requirement*.

The other requirement considered in [26] is *collision-resistance*. This requires that whenever the Eval algorithm is run on $\mathrm{PUF}_{\mathbf{id}}$ with different challenges, the minimum distance between the different responses must be at least δ_c. The parameters δ_r, δ_u, δ_c are determined by the security parameter λ. Robustness, uniqueness and collision-resistance are crucial for correctness of cryptographic schemes built on top of PUFs. The conditions $\delta_r \leq \delta_u$ and $\delta_r \leq \delta_c$ must be satisfied to allow for distinguishing different challenges and PUFs [26].

According to the above, a $(\lambda, \delta_r, \delta_u, \delta_c)$-PUF is defined as a pair of algorithms: Gen and Eval that provides the robustness, uniqueness and collision-resistance requirements. We call a $(\lambda, \delta_r, \delta_u, \delta_c)$-PUF a Classical PUF (CPUF), if the Eval algorithm runs on classical information such as bit strings.

2.2 Quantum tools

We use the term quantum bit or qubit [32] to denote a simple two-level physical system with quantum behaviour which is the quantum analogue to classical bits. Quantum states are denoted as unit vectors in a Hilbert space $\mathcal{H}$. Any D-dimensional Hilbert space is equipped with a set of D orthonormal vectors called a basis. In the case of a single qubit where D=2, the following set of vectors are a complete basis referred to as the computational bases:

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and any qubit state can be written as $|x\rangle = \alpha |0\rangle + \beta |1\rangle$ where $|\alpha|^2 + |\beta|^2 = 1$ for some $\alpha, \beta \in \mathbb{C}$. The above form is called a superposition of two quantum states. We say a quantum state is pure if it deterministically describes a vector in Hilbert space. On the other hand, a mixed quantum state is described as a probability distribution over different pure quantum states:

$$\rho = \sum_s p_s |\psi_s\rangle\langle\psi_s|$$

represented as a density matrix. If a quantum state can be written as the tensor product of all its subsystems, we say that the state is separable, eg.

$$|\psi_{AB}\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$$

otherwise, it is referred to as entangled state.

Expectation value or classical information of a quantum system is obtained by measurements. A measurement operator is defined as a family of linear operators $\{M_j\}$ acting on the state where the index j refers to each measurement outcome. If $|x\rangle$ is the quantum state before the measurement, then the probability of obtaining result j is:

$$Pr(j) = \langle x| M_j^\dagger M_j |x\rangle .$$

Transformations between pure quantum states is usually described by unitary operators which are reversible and preserve the inner product. General quantum transformations are Completely Positive Trace Preserving (CPTP or CPT) maps which include also unitary matrices. If a CPT map does not preserve the trace, we call it a Completely Positive (CP) map. In Appendix A the list of all unitary operators used in this paper is given.

An important difference between quantum and classical bits is the impossibility of creating perfect copies of general unknown quantum states, known as the *no-cloning theorem*. This is an important limitation imposed by quantum mechanics which is particularly relevant for cryptography. A variation of the same feature states that it is impossible to obtain the exact classical description of quantum states by having a single copy of it. Therefore, there exists a bound on how well one can derive the classical description of quantum states depending on their dimension and the number of available copies. Hence, distinguishing between unknown quantum states can be achieved only probabilistically. A useful and relevant notion of quantum distance that we exploit in this paper is *fidelity*. The fidelity of two pure quantum states $|\psi\rangle$ and $|\phi\rangle$ is defined as

$$F(|\psi\rangle , |\phi\rangle) = | \langle\psi| \phi\rangle|^2.$$

Generally the fidelity of mixed states ρ and σ is defined by the Uhlmann fidelity:

$$F(\rho, \sigma) = [Tr(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})]^2.$$

Now based on this quantum distance we can introduce the distinguishability and indistinguishability of two quantum states.

Definition 1 (μ-distinguishability and ν-indistinguishability). *Let $F(\cdot, \cdot)$ denote the fidelity distance, and μ and ν the distinguishability and indistinguishability threshold parameters respectively such that $0 \le \mu, \nu \le 1$. We say two quantum states ρ and σ are μ-distinguishable if $0 \le F(\rho, \sigma) \le 1 - \mu$ and ν-indistinguishable if $\nu \le F(\rho, \sigma) \le 1$.*

Note that two quantum states, ρ and σ, are *completely distinguishable* or 1-distinguishable ($\mu = 1$), if $F(\rho, \sigma) = 0$ and they are *completely indistinguishable* or 1-indistinguishable ($\nu = 1$) if $F(\rho, \sigma) = 1$.

Due to the impossibility of perfectly distinguishing between all quantum states according to the above definition, checking equality of two completely unknown states is a non-trivial task. This is one major difference between classical bits and qubits. Nevertheless, a probabilistic comparison of unknown quantum states can be achieved through the simple quantum SWAP test algorithm [33]. The SWAP test and its generalisation to multiple copies introduced recently in [34] have been discussed in more details in the Appendix B. Here we abstract for specific tests and define necessary conditions for a general quantum test.

Definition 2 (Quantum Testing Algorithm). *Let $\rho^{\otimes\kappa_1}$ and $\sigma^{\otimes\kappa_2}$ be κ_1 and κ_2 copies of two quantum states ρ and σ, respectively. A Quantum Testing algorithm $\mathcal{T}$ is a quantum algorithm that takes as input the tuple ($\rho^{\otimes\kappa_1}$,$\sigma^{\otimes\kappa_2}$) and accepts ρ and σ as equal (outputs 1) with the following probability*

$$\Pr[1 \leftarrow \mathcal{T}(\rho^{\otimes\kappa_1}, \sigma^{\otimes\kappa_2})] = 1 - \Pr[0 \leftarrow \mathcal{T}(\rho^{\otimes\kappa_1}, \sigma^{\otimes\kappa_2})] = f(\kappa_1, \kappa_2, F(\rho, \sigma))$$

where $F(\rho, \sigma)$ is the fidelity of the two states and $f(\kappa_1, \kappa_2, F(\rho, \sigma))$ satisfies the following limits:

$$\begin{cases} \lim_{F(\rho,\sigma)\to 1} f(\kappa_1, \kappa_2, F(\rho, \sigma)) = 1 & \forall\, (\kappa_1, \kappa_2) \\ \lim_{\kappa_1,\kappa_2\to\infty} f(\kappa_1, \kappa_2, F(\rho, \sigma)) = F(\rho, \sigma) \\ \lim_{F(\rho,\sigma)\to 0} f(\kappa_1, \kappa_2, F(\rho, \sigma)) = Err(\kappa_1, \kappa_2) \end{cases}$$

with $Err(\kappa_1, \kappa_2)$ characterising the error of the test algorithm.

We will use the Diamond Norm as the the most common distance measure for quantum operators Q_E and Q_F, and is defined in terms of l_1 trace norm

$$\| Q_E - Q_F \|_\diamond \equiv max_\rho(\| (Q_E \otimes \mathbb{I})[\rho] - (Q_F \otimes \mathbb{I})[\rho]) \|_1 .$$

Finally, we will let λ denote the security parameter. A non-negative function $\epsilon(\lambda)$ is negligible if, for any constant c, $\epsilon(\lambda) \leq \frac{1}{\lambda^c}$ for all sufficiently large λ.

3 Quantum Emulation Algorithm

In this section, we describe the Quantum Emulation (QE) algorithm presented in [17] as a quantum process learning tool that can outperform the existing approaches based on quantum tomography [35]. The main idea behind quantum emulation comes from the question on the possibility of emulating the action of an unknown unitary transformation on an unknown input quantum state by having some of the input-output samples of the unitary. An emulator is not trying to completely recreate the transformation or simulate the same dynamics. Instead, it outputs the action of the transformation on a quantum state. Although the original intended application for this algorithm is in the field of quantum process tomography and quantum phase estimation, here we repurpose it and provide a new analysis as a practical and implementable algorithm for cryptanalysis of cryptographic primitives. We formalize it as a general quantum emulation attack in the next section. In this section, however, we focus on the circuit of the algorithm in [17] and introduce a new fidelity analysis for our purpose.

3.1 The Circuit and Description

The circuit of the quantum emulation algorithm has been depicted in Figure 5 in Appendix C [17] and works as follows.

Let U be a unitary transformation on a D-dimensional Hilbert space $\mathcal{H}^D$, $S_{in} = \{|\phi_i\rangle ; i = 1, ..., K\}$ be a sample of input states and $S_{out} = \{|\phi_i^{out}\rangle ; i = 1, ..., K\}$ the set of corresponding outputs, i.e $|\phi_i^{out}\rangle = \mathrm{U} |\phi_i\rangle$. Also, let d be the dimension of the Hilbert space $\mathcal{H}^d$ spanned by S_{in} and $|\psi\rangle$, a challenge state. The goal of the algorithm is to find the corresponding output of U, that is $\mathrm{U} |\psi\rangle$. The main building block of the algorithm are controlled-reflection gates described as:

$$R_c(\phi) = |0\rangle \langle 0| \otimes \mathbb{I} + |1\rangle \langle 1| \otimes e^{i\pi |\phi\rangle \langle \phi|} \tag{3}$$

The gate acts as the identity ($\mathbb{I}$) if the control qubit is $|0\rangle$, and as $R(\phi) = e^{i\pi |\phi\rangle \langle \phi|} = \mathbb{I} - 2 |\phi\rangle \langle \phi|$ if the control qubit is $|1\rangle$. The circuit also uses Hadamard and SWAP gates, and consists of four stages.

Stage 1 $K + 1$ number of sample states is chosen as well as number of ancillary qubits used through the algorithm. We assume the algorithm uses all of the states in S_{in}. The ancillary systems are all qubits prepared at $|-\rangle$. Let $|\phi_r\rangle \in S_{in}$ be considered as the reference state. This state can be chosen at random or from a special distribution. The first step consists of K blocks where in each block, the following gates run on the state of the system and an ancilla:

$$W(i) = R_c(|\phi_i\rangle) H R_c(|\phi_r\rangle). \tag{4}$$

According to equation 4, a controlled-reflection around the reference state $|\phi_r\rangle$ is performed on $|\psi\rangle$ with the control qubit being on the $|-\rangle$ ancillary state. Then a Hadamard gate runs on the ancilla followed by another controlled-reflection around the sample state $|\phi_i\rangle$. This is repeated for each of the K states in S_{in} such that the input state is being entangled with the ancillas and also it is being projected into the subspace $\mathcal{H}^d$ in a way that the information of $|\psi\rangle$ is encoded in the coefficients of the general entangled state. This information is the overlap of $|\psi\rangle$ with all the sample inputs. By reflecting around the reference state in each block, the main state is pushed to $|\phi_r\rangle$ and the probability of finding the system at the reference state increases. The overall state of the circuit after Stage 1 is:

$$[W(K)...W(1)] |\psi\rangle |-\rangle^{\otimes K} \approx |\phi_r\rangle |\Omega(anc)\rangle \tag{5}$$

where $|\Omega(anc)\rangle$ is the entangled state of K ancillary qubits. The approximation comes from the fact that the state is not only projected on the reference quantum state but it is also projected on other sample quantum states with some probability. We present a more precise formula in the next subsection.

Stage 2 In this stage, first a reflection around $|\phi_r\rangle$ is performed and after applying a Hadamard gate on an extra ancilla, that ancilla is measured in the computational basis $\{|0\rangle , |1\rangle\}$. Based on the output of the measurement, one

can decide whether the first step was successful (when output of the measurement is 0) or not. If the first step is successful, the main state has been pushed to the reference state. In this case, the algorithm proceeds with Stage 3. If the output is 1, the projection was unsuccessful and the input state remains almost unchanged. In this case, either the algorithm aborts or it goes back to the first stage and picks a new state as the reference. This stage has a post-selection role which can be skipped to output a mixed state of two possible outputs.
Stage 3 The main state is swapped with $|\phi_r^{out}\rangle = \mathrm{U}\,|\phi_r\rangle$ that is the output of the reference state. This is done by means of a SWAP gate. At this point, the overall state of the system is:

$$(\mathrm{SWAP} \otimes I^{\otimes K})\,|\phi_r^{out}\rangle\,|\phi_r\rangle\,|\Omega(anc)\rangle = |\phi_r\rangle\,|\phi_r^{out}\rangle\,|\Omega(anc)\rangle\,. \tag{6}$$

By tracing out the first qubit, the state of the system becomes $|\phi_r^{out}\rangle\,|\Omega(anc)\rangle$.

Stage 4 The last stage is very similar to the first one except that all blocks are run in reverse order and the reflection gates are made from corresponding output quantum states. The action of stage 4 is equivalent to:

$$W^{out}(i) = R_c(|\phi_i^{out}\rangle)HR_c(|\phi_r^{out}\rangle) = (\mathrm{U} \otimes I)W(i)(\mathrm{U}^\dagger \otimes \mathbb{I}). \tag{7}$$

After repeating this gate for all the output samples, U is acted on the projected components of $|\psi\rangle$ and by restoring back the information of $|\psi\rangle$ from the ancilla, the input state approaches $\mathrm{U}\,|\psi\rangle$. The overall output state of the circuit at the end of this stage is (equality whenever the success probability of Stage 2 is 1):

$$[W^{out}(1)...W^{out}(K)]\,|\phi_r^{out}\rangle\,|\Omega(anc)\rangle \approx \mathrm{U}\,|\psi\rangle\,|-\rangle^{\otimes K}\,. \tag{8}$$

3.2 Output fidelity analysis

We are interested in the fidelity of the output state $|\psi_{QE}\rangle$ of the algorithm and the intended output $\mathrm{U}\,|\psi\rangle$ to estimate the success. In the original paper, the fidelity analysis is first provided for ideal controlled-reflection gates and later a protocol is presented to implement them efficiently. In this paper, as we are more interested in the theoretical bounds for the fidelity, all the gates including the controlled-reflection gates are assumed to be ideal keeping in mind that the implementation is possible [17, 36]. We recall the main theorem of [17]:

Theorem 1. *[17] Let $\mathcal{E}_\mathrm{U}$ be the quantum channel that describes the overall effect of the algorithm presented above. Then for any input state ρ, the Uhlmann fidelity of $\mathcal{E}_\mathrm{U}(\rho)$ and the desired state $\mathrm{U}\rho\mathrm{U}^\dagger$ satisfies:*

$$F(\rho_{QE}, \mathrm{U}\rho\mathrm{U}^\dagger) \geq F(\mathcal{E}_\mathrm{U}(\rho), \mathrm{U}\rho\mathrm{U}^\dagger) \geq \sqrt{P_{succ-stage1}} \tag{9}$$

where $\rho_{QE} = |\psi_{QE}\rangle\langle\psi_{QE}|$ is the main output state(tracing out the ancillas) when the post-selection in Stage 2 has been performed. $\mathcal{E}_\mathrm{U}(\rho)$ is the output of the whole circuit without the post-selection measurement in Stage 2 and $P_{succ-stage1}$ is the success probability of Stage 1.

For the purpose of this paper, we need a more precise and concrete expression for the output fidelity not covered in [17]. From the proof of Theorem 1 in [17], it can be seen that the success probability of Stage 1 is calculated as follows:

$$P_{succ-stage1} = |\langle\phi_r| Tr_{anc}(|\chi_f\rangle\langle\chi_f|)|\phi_r\rangle|^2 \tag{10}$$

where $|\chi_f\rangle$ is the final state of the circuit after Stage 1 and $Tr_{anc}(\cdot)$ computes the reduced density matrix by tracing out the ancillas. The overlap of the resulting state and the reference state equals the success probability of Stage 1. Now relying on Theorem 1, we only use equation (10) for our analysis henceforward.

The fidelity of the output state of the circuit highly depends on the choice of the reference state (equation (10)) such that it may increase or decrease the success probability of the adversary in different security models as we will discuss in the Section 4.2. We establish the following recursive relation for the state of the circuit after the i-th block of Stage 1, in terms of the previous state:

$$|\chi_i\rangle = \frac{1}{2}[(I - R(\phi_r))|\chi_{i-1}\rangle|0\rangle + R(\phi_i)(\mathbb{I} + R(\phi_r))|\chi_{i-1}\rangle|1\rangle]. \tag{11}$$

Now by using this relation, we can prove the following theorem:

Theorem 2. *Let $|\chi_f\rangle$ be the final overall state of the circuit (Figure 5). Let K be the number of blocks in the circuit. Let $|\psi\rangle$ be the input state of the circuit, $|\phi_r\rangle$ the reference state and $|\phi_i\rangle$ other sample states. The final state is:*

$$\begin{aligned}|\chi_f\rangle &= \langle\phi_r|\psi\rangle|\phi_r\rangle|0\rangle^{\otimes K} + |\psi\rangle|1\rangle^{\otimes K} - \langle\phi_r|\psi\rangle|\phi_r\rangle|1\rangle^{\otimes K}\\ &+ \sum_{i=1}^{K}\sum_{j=0}^{i}[f_{ij}2^{l_{ij}}|\langle\phi_r|\psi\rangle|^{x_{ij}}|\langle\phi_i|\psi\rangle|^{y_{ij}}|\langle\phi_r|\phi_i\rangle|^{z_{ij}}]|\phi_r\rangle|q_{anc}(i,j)\rangle\\ &+ \sum_{i=1}^{K}\sum_{j=0}^{i}[g_{ij}2^{l'_{ij}}|\langle\phi_r|\psi\rangle|^{x'_{ij}}|\langle\phi_i|\psi\rangle|^{y'_{ij}}|\langle\phi_r|\phi_i\rangle|^{z'_{ij}}]|\phi_i\rangle|q'_{anc}(i,j)\rangle\end{aligned} \tag{12}$$

where l_{ij}, x_{ij}, y_{ij}, z_{ij}, l_{ij}, x'_{ij}, y'_{ij} and z'_{ij} are integer values indicating the power of the terms of the coefficient. Note that f_{ij} and g_{ij} can be 0, 1 or -1 and $q_{anc}(i,j)$ and $q'_{anc}(i,j)$ output a computational basis of K qubits (other than $|0\rangle^{\otimes K}$).

Proof. We prove this by induction. The proof can be found in Appendix D. □

Having a precise expression for $|\chi_f\rangle$ from Theorem 2, one can calculate $P_{succ-step1}$ of equation (10) by tracing out all the ancillary systems from the density matrix of $|\chi_f\rangle\langle\chi_f|$. Also, now it is clear that if $|\psi\rangle$ is orthogonal to the $\mathcal{H}^d$, the only term remaining in equation (12) is $|\psi\rangle|1\rangle^{\otimes K}$. So, the input state remains unchanged after the first stage and $P_{succ-step1} = 0$.

For states projected in the subspace spanned by S_{in}, the overall channel describing the quantum emulation algorithm has always a fixed point inside the subspace [17]. Hence, Stage 1 is successful with probability close to 1 by assuming the gates to be ideal.

For the cases that the classical description of the samples are known, the controlled-reflection gates are perfect and no further assumption is needed. Finally, to also capture the effect of the noise and imperfection of the gates we have a brief discussion in Appendix E.

4 Quantum Game-based security

In this section, we introduce a quantum game-based framework for analysing the security of quantum cryptographic primitives. Generalising the idea of quantum emulation on unknown quantum processes presented in Section 3, we introduce a new class of quantum attacks termed Quantum Emulation Attacks. This notion generalises previously considered quantum attacks, *e.g.* superposition attacks [13–15]. We then use this formal framework to establish general results as to the security of quantum crypto primitives under different threat models.

4.1 Quantum Game-based Security Framework

The game-based security framework is a standard model for defining security properties of cryptographic primitives such as encryption algorithms, digital signatures schemes and CPUFs [26, 10, 37, 12, 16]. Also, security analysis of the classical cryptographic primitives in a quantum game-based framework, where parties are Quantum Turing Machines (QTM), has been widely studied in [10–12, 16]. Inspired by these works, we introduce a similar but more generalised framework, in the sense that it unifies the previous unforgeability definitions for any classic/quantum primitive that concerns this security property.

Moreover, to have a precise framework which captures non-trivial quantum attacks on quantum primitives, we have introduced the notion of *μ-distinguishability* to characterise the level of quantum security and also the notion of *test algorithm* for verifying the output of unknown quantum states.

Now, we introduce our quantum game-based security framework for a typical cryptographic primitive $\mathcal{F} = (\mathcal{S}, \mathcal{E}, \mathcal{T})$ where $\mathcal{S}$, $\mathcal{E}$ and $\mathcal{T}$ are setup, evaluation and test algorithm, respectively, where the test algorithm satisfies Definition 2. We say $\mathcal{F}$ is a *unitary cryptographic primitive* if $\mathcal{E}$ can be modelled as a unitary transformation $\mathrm{U}_{\mathcal{E}}$ over a D-dimensional Hilbert space $\mathcal{H}^D$.

Similar to the classical setting, the security of $\mathcal{F}$ is captured by a game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$. The challenger models the honest parties, while the adversary captures the corrupted parties. The adversary's goal is to *closely approximate* the output of the evaluation algorithm $\mathcal{E}$ on a quantum challenge $|\psi\rangle$. The games considered here have 5 phases. First, $\mathcal{C}$ runs the setup algorithm $\mathcal{S}$ to generate the parameters required throughout the game. The game begins with a first learning phase and is followed by a challenge phase. Then, a second learning phase is run, and finally, $\mathcal{A}$ has to return his response to the state of the challenge phase. The learning phases define the threat model, and the challenge phase determines the security notion captured by the game. The formal description of our quantum games is shown in Figure 1.

Setup - In the setup phase, $\mathcal{C}$ generates the parameters required in subsequent phases by running the setup algorithm of the primitive $\mathcal{F}$ on input λ.

Learning phases - In the learning phases, we grant different levels of oracle access on $\mathcal{E}$ to $\mathcal{A}$. We analyse two main types which we call quantum Unknown Input (qUI) and quantum Chosen Input (qCI), depending on the adversary's capabilities. In a qUI learning phase, the adversary has no control over the inputs, i.e. $S_{in} = \{|\phi_{j,1}\rangle, \ldots, |\phi_{j,k_j}\rangle\}_{j=1,2}$ where j denotes the index of the learning phase, is a set of unknown quantum states. The set may include t copies of each unknown quantum state where t is determined by the concrete protocol that the primitive is used for. Considering t to be polynomial in $\log(D)$, the state of the quantum inputs will be still unknown for the adversary. While in a qCI learning phase, the adversary *smartly* chooses the states $|\phi_{j,i}\rangle_{i=1:k_j}$ where $j \in \{1, 2\}$ to build S_{in} where he also knows the classical description of the inputs and can thus prepare perfect multiple copies of the state from their description. If the considered adversary is *adaptive*, the game has a second learning phase after the challenge phase. Otherwise, the considered adversary is *static*. Whenever ℓ_1 and/or ℓ_2 is equal to null, we drop it from $\mathcal{G}^{\mathcal{F}}_{\ell_1,c,\ell_2}$. Also, in this type of learning phase, the adversary prepares at least two copies of each input $|\phi_{j,i}\rangle$, adds the first one to S_{in} and sends the second one to the challenger to obtain the output quantum state $|\phi^{out}_{j,i}\rangle = U|\phi_{j,i}\rangle$ which they add to S_{out}.

Challenge phase - In this phase, the challenge $|\psi\rangle$, that the adversary has to respond to, is chosen. We want to capture two different notions of security that correspond to two different types of challenge phases. More precisely, if $|\psi\rangle$ is chosen by the challenger $\mathcal{C}$, we call it a *quantum selective challenge*, denoted by qSel in Figure 1. If it is chosen by the adversary, it is called a *quantum existential challenge*, denoted by qEx. We impose different conditions on the challenge phases. These conditions prevent the adversary from mounting trivial attacks. We will discuss the design of the game and the conditions later in the discussion subsection. In the security analysis of the unitary primitive as follows, the success probability is calculated only for the events that the existential or selective challenges are at least μ-distinguishable, according to Definition 1, from all the inputs queried in the first and second learning phases, respectively. Also, for the qSel challenges against static adversaries, we relax this condition but we require that the chosen challenge $|\psi\rangle$ is picked at random from a uniform distribution over the whole Hilbert space $\mathcal{H}^D$.

Guess phase - In this phase, the adversary responds to the challenge $|\psi\rangle$ chosen during the challenge phase. The adversary wins the game if he closely approximates the effect of $\mathrm{U}_{\mathcal{E}}$ on $|\psi\rangle$. More precisely, if the output of the test algorithm $\mathcal{T}$ is 1, where $\mathcal{T}$ is an algorithm satisfying Definition 2.

Discussion on the game definition and conditions. In the introduced framework, the type of learning phase (qCI or qUI) characterises the access level of the adversary to the cryptographic primitive.

The quantum challenge phase needs to be carefully specified to avoid capturing trivial attacks such as sending one of the previously learnt states as the

The game $\mathcal{G}^{\mathcal{F}}_{\ell_1,\ell_2,c,\mu}(\lambda,\mathcal{A})$[a]

First learning phase:
- if $\ell_1 = \mathsf{qUI}$
  (a) $\mathcal{A}$ sends k_1 to $\mathcal{C}$. Then $\mathcal{C}$ proceeds as follows:
  (b) For $i = 1 : k_1$
    - $\mathcal{C}$ prepares a quantum state $|\phi_{1,i}\rangle \in \mathcal{H}^D$ according to the protocol
    - For $l = 1 : t$[b]
      - $\mathcal{C}$ prepares two copies of $|\phi_{1,i,l}\rangle = |\phi_{1,i}\rangle$ and appends one to S_{in}
      - $\mathcal{C}$ applies $\mathrm{U}_{\mathcal{E}}$ on 2^{nd} copy of $|\phi_{1,i,l}\rangle$ and obtains $|\phi^{out}_{1,i,l}\rangle = \mathrm{U}_{\mathcal{E}}\,|\phi_{1,i,l}\rangle$
      - $\mathcal{C}$ appends $|\phi^{out}_{1,i,l}\rangle$ to S_{out}
  (c) $\mathcal{C}$ sends S_{in} and S_{out} to $\mathcal{A}$ and keeps the classical description of $|\phi_{1,i}\rangle_{i=1:k_1}$
- if $\ell_1 = \mathsf{qCI}$
  (a) For $i = 1 : k_1$
    - $\mathcal{A}$ prepares two copies of a quantum state $|\phi_{1,i}\rangle \in \mathcal{H}^D$, appends one to S_{in} and sends the other to $\mathcal{C}$
    - $\mathcal{C}$ applies $\mathrm{U}_{\mathcal{E}}$ to $|\phi_{1,i}\rangle$, gets $|\phi^{out}_{1,i}\rangle = \mathrm{U}_{\mathcal{E}}\,|\phi_{1,i}\rangle$, and sends $|\phi^{out}_{1,i}\rangle$ to $\mathcal{A}$
    - $\mathcal{A}$ appends $|\phi^{out}_{1,i}\rangle$ to S_{out}
- if $\ell_1 = \mathsf{null}$, the adversary does nothing

Challenge phase:
- if $c = \mathsf{qEx}$: $\mathcal{A}$ picks a quantum state $|\psi\rangle \notin_\mu S_{in}$[c] and sends κ_1 copies of it to $\mathcal{C}$.
- if $c = \mathsf{qSel}$: $\mathcal{C}$ chooses a quantum state $|\psi\rangle$ at random from the uniform distribution over the Hilbert space $\mathcal{H}^D$. $\mathcal{C}$ keeps κ_1 copies of $|\psi\rangle$ and sends an extra copy of $|\psi\rangle$ to $\mathcal{A}$.

Second learning phase: The same as the *first learning phase* but with ℓ_2 controlling its type, and k_2 the number of queries to $\mathrm{U}_{\mathcal{E}}$. For $S_{in} = \{|\phi_{1,i}\rangle\,;\ i = 1 : k_1\} \cup \{|\phi_{2,i}\rangle\,;\ i = 1 : k_2\}$ the condition $|\psi\rangle \notin_\mu S_{in}$ must also hold.

Guess phase:
- $\mathcal{A}$ sends κ_2 copies of $|\omega\rangle$ to $\mathcal{C}$
- $\mathcal{C}$ uses the κ_1 copies of the challenge, applies $\mathrm{U}_{\mathcal{E}}$ to all of them and obtains $|\psi^{out}\rangle^{\otimes\kappa_1} = (\mathrm{U}_{\mathcal{E}}\,|\psi\rangle)^{\otimes\kappa_1}$
- $\mathcal{C}$ runs the test algorithm $b \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})$ where $b \in \{0,1\}$
- Outputs b

[a] $\ell_1, \ell_2 \in \{\mathsf{qUI}, \mathsf{qCI}, \mathsf{null}\}$, $c \in \{\mathsf{qEx}, \mathsf{qSel};\ 0 \le \mu \le 1\}$, S_{in}, S_{out}: ordered lists.
[b] t is the number of copies of each unknown challenge prescribed by the protocol.
[c] $\notin_\mu$ denotes at least μ-distinguishablity from all the states in S_{in}.

Fig. 1: Formal definition of the quantum games $\mathcal{G}^{\mathcal{F}}_{\ell_1,\ell_2,c,\mu}(\lambda,\mathcal{A})$ where $\mathcal{F} = (\mathcal{S},\mathcal{E},\mathcal{T})$ is a unitary cryptographic primitive with $\mathcal{S}$, $\mathcal{E}$ and $\mathcal{T}$ as a setup, an evaluation and a test algorithm, respectively; $\mathrm{U}_{\mathcal{E}}$ is the unitary modelling $\mathcal{E}$ and D its dimension; and λ is the security parameter which includes D, κ_1 and κ_2.

challenge of the adversary. More precisely, in the qEx challenge phase, we impose the adversary to choose a quantum state that is μ-distinguishable ($0 \leq \mu \leq 1$) from the quantum states queried in the learning phase.

Note that the case $\mu = 1$ implies the challenge quantum state is orthogonal to all the quantum states queried in the learning phase; it morally captures the standard classical unforgeability definitions where the adversary does not have quantum access to the primitive. When $\mu < 1$, we grant the adversary meaningful quantum access to the primitive. Note that we do not specify how the challenger could check whether the adversary meets the condition or not. Implementing this check is not crucial for defining security, where we only need to be able to characterise the instances that might present a security violation. We, however, believe that there are approaches that could be used for this purpose such as sending multiple copies or the classical description of the queried quantum states in the learning phase, or generating maximally entangled quantum states as proposed in [12].

For qSel challenge phases against static adversaries, it is enough to make sure that the adversary does not have any information about the challenge that will be picked by the challenger later. This is due to the fact that the learning phase is run before the challenge phase. Thus the μ-distinguishability condition can be dropped. However, we do need to ensure that the adversary has no knowledge of the subspace or distribution of the challenge space, which could lead to other trivial attacks. To this end, we impose the challenge to be picked uniformly at random from the whole Hilbert space.[5]

On the other hand, when considering adaptive adversaries, because the adversary runs the second learning phase after getting the challenge, the quantum states queried should again be μ-distinguishable from the challenge. We clarify that we do not bound the adversary to separable states. Although to keep the notations as simple as possible we do not use entangled states in the formal definition of the game, while as we allow them in the attacks.

In comparison to quantum unforgeability definitions found in [13, 16] where the challenge is a classical message, our framework captures a more general setting in the sense that the challenge can be any quantum state in the Hilbert space as long as it meets the above mentioned conditions. Also, we have characterised the distance between the challenge quantum state and the learnt quantum states that leads to more precise security analysis of the primitive compared to others.

Finally, we clarify how the challenger can hold the necessary number of copies of a quantum state to run the test algorithm $\mathcal{T}$ in the guess phase. Recall that in the qSel challenge phase, the challenger picks the challenge; thus it can prepare multiple copies of the challenge, apply $\mathrm{U}_{\mathcal{E}}$ on them and get multiple copies of the corresponding response. It then runs the test algorithm on these copies and the adversary's response to the chosen challenge. In qEx challenge phases, the adversary picks the challenge quantum state and sends multiple copies of it to the challenger which enables the challenger to run the test algorithm $\mathcal{T}$. The

[5] Protocols analysed against this security definition need to meet this condition too.

challenger prepares multiple copies of the response quantum state by querying the unitary primitive and runs the test algorithm $\mathcal{T}$.

4.2 Security analysis of unitary cryptographic primitives and quantum emulation attacks

We now focus on the security of unitary cryptographic primitives and define a new class of attacks that we call *quantum emulation attacks.* These attacks capture quantum adversaries that try to win the previously defined games. As we will see, the QE algorithm presented in Section 3 is a general attack that wins some of these games (depending on the attacker's capabilities and the security notion considered) with non-negligible probability.

Definition 3 (Security against Quantum Emulation Attacks). *We say that a unitary cryptographic primitive $\mathcal{F}$ is secure against τ-QEA (Quantum Emulation Attack) where $\tau := (\ell_1, c, \ell_2, \mu)$ if for any QTM adversary $\mathcal{A}$, the probability to win $\mathcal{G}^{\mathcal{F}}_{\ell_1,c,\ell_2,\mu}(\mathcal{A}, \lambda)$ is negligible in the security parameter λ,*

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\ell_1,c,\ell_2,\mu}(\mathcal{A}, \lambda)] = negl(\lambda).$$

If $\mathcal{A}$ is a Quantum Polynomial-Time (QPT) algorithm in λ, we call the attack a Polynomial Quantum Emulation Attack or PQEA.

We are now ready to present several general results on the security of unitary primitives. The first one establishes a general no-go result. More precisely, we show that there exists a successful existential PQEA against any such primitive.[6]

Theorem 3 (No primitive $\mathcal{F}$ is secure against $(\mathsf{qCl}, \mathsf{qEx}, \mu)$-PQEA). *For any unitary cryptographic primitive $\mathcal{F}$ and any μ such that $\mu \leq 1 - non\text{-}negl(\lambda)$, there exists a QPT adversary $\mathcal{A}$ such that*

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCl},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = non\text{-}negl(\lambda).$$

Proof. We show there is a QPT adversary $\mathcal{A}$ that wins the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCl},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$. $\mathcal{A}$ runs the algorithm defined and explained in Figure 2.

The detailed probability analysis of the above attack has been discussed in Appendix F.1. The main idea of the attack is that the adversary can create a small subspace from the states of the learning phase, with a good overlap wrt the challenge and then use the quantum emulation algorithm to emulate the output. As different distinguishability parameter μ characterises the different level of security, we discuss two different limits for the μ in this attack. We show that if $0 < \mu \leq \frac{1}{2}$, the adversary can produce completely indistinguishable output states and win the game with probability 1. In general, considering that the security parameter λ includes the number of copies used in the test algorithm

[6] Note that Theorems 3 and 4 focus on standard security notions such as existential unforgeability under chosen or random message attacks. We will focus on these notions later on in the paper when turning to the security of QPUFs.

$(\mathsf{qCI}, \mathsf{qEx}, \mu)$-*PQEA*

First learning phase:

choose $\lvert\phi_1\rangle$ query $\lvert\phi_1\rangle$ and receive $\lvert\phi_1^{out}\rangle$ **if** $0 \leq \mu \leq \frac{1}{2}$: choose $\lvert\phi_2\rangle = \frac{1}{\sqrt{2}}(\lvert\phi_1\rangle + \lvert\phi_3\rangle)$ **else if** $\frac{1}{2} \leq \mu \leq 1-$non-$negl(\lambda)$: choose $\lvert\phi_2\rangle = \sqrt{\mu}\lvert\phi_1\rangle + \sqrt{1-\mu}\lvert\phi_3\rangle$ query $\lvert\phi_2\rangle$ and receive $\lvert\phi_2^{out}\rangle$ set $S_{in} = \{\lvert\phi_1\rangle, \lvert\phi_2\rangle\}$	set $S_{out} = \{\lvert\phi_1^{out}\rangle, \lvert\phi_2^{out}\rangle\}$ Without loss of the generality, we assume $\mathcal{A}$ chooses one of the computational basis of $\mathcal{H}^D$ as $\lvert\phi_1\rangle$. Then, $\mathcal{A}$ chooses an orthogonal state to $\lvert\phi_1\rangle$ as $\lvert\phi_3\rangle$ and sets $\lvert\phi_2\rangle$ the superposition of these two states.

Challenge phase:

set the challenge $\lvert\psi\rangle \leftarrow \lvert\phi_3\rangle$	$\lvert\phi_3\rangle$ satisfies condition $\lvert\phi_3\rangle \notin_{\mu} S_{in}$.

Guess phase:

set the reference $\lvert\phi_r\rangle \leftarrow \lvert\phi_2\rangle$ $\lvert\omega\rangle \leftarrow QE(\lvert\psi\rangle, \lvert\phi_r\rangle, S_{in}, S_{out})$ output $\lvert\omega\rangle$	$QE(\lvert\psi\rangle, S_{in}, S_{out}, \lvert\phi_r\rangle)$ is the quantum emulation algorithm.

Fig. 2: $(\mathsf{qCI}, \mathsf{qEx}, \mu)$-PQEA: adversary's algorithm against game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qEx},\mu}$

(κ_1 and κ_2), by increasing them the probability of accepting will converge to the fidelity which will lead to the following result according to Appendix F.1:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(\lvert\psi^{out}\rangle^{\otimes\kappa_1}, \lvert\omega\rangle^{\otimes\kappa_2})] = non\text{-}negl(\lambda). \qquad \square$$

This theorem implies that the adversary can always generate the correct response to his chosen challenge provided that he can query it in superposition with other quantum states during the learning phase. This is a generic quantum attack to all unitary cryptographic primitives and it is independent of their construction as long as the adversary has quantum access to the primitive. The theorem shows that with reasonable quantum access to the primitive in terms of the parameter μ, a non-trivial attack exists. Note that since output quantum states in the learning phase are unknown to the adversary, the more straightforward strategy of superposing the learnt output quantum states cannot be implemented by the adversary. More precisely, the adversary cannot prepare the precise target superposition of the output states which are completely unknown [38].

We now show that even weaker adversaries with qUI access to the primitive, break the existential security of the primitive.

Theorem 4 (No primitive $\mathcal{F}$ is secure against a $(\mathsf{qUI}, \mathsf{qEx}, \mu)$-PQEA). *For any unitary cryptographic primitive $\mathcal{F}$, and for any $0 \leq \mu \leq \frac{1}{2}$ there exists a QPT adversary $\mathcal{A}$ st.*

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = \textit{non-negl}(\lambda).$$

Proof. Let $\mathcal{A}$ be the QPT adversary playing game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ and running the algorithm defined and explained in Figure 3.

<u>$(\mathsf{qUI}, \mathsf{qEx}, \mu)$-*PQEA*</u>

First learning phase:

choose $k_1 = 2, \forall t_1 \geq 2$ receive: $S_{in} = \{\|\phi_{1,1}\rangle, ..., \|\phi_{1,t}\rangle, \|\phi_{2,1}\rangle, ... \|\phi_{2,t}\rangle\}$ $S_{out} = \{\|\phi^{out}_{1,1}\rangle, ... \|\phi^{out}_{1,t}\rangle, ..., \|\phi^{out}_{2,1}\rangle, ..., \|\phi^{out}_{2,t}\rangle\}$	At qUI learning phase, $\mathcal{A}$ receives S_{in} and S_{out} consisting of k_1 unknown quantum states μ-distinguishable.

Challenge phase:

$\|\psi_{sup}\rangle \leftarrow \text{Superpose}(\|\phi_{1,1}\rangle, \|\phi_{2,1}\rangle)$	$\mathcal{A}$ creates the unknown superposition using Superpose(), a subroutine defined in Appendix F.2.

Guess phase:

$b \xleftarrow{\$} \{0, 1\}$ set the reference $\|\phi_r\rangle \leftarrow \|\phi_{b,t}\rangle$ $\|\omega\rangle \leftarrow QE(\|\psi_{sup}\rangle, \|\phi_r\rangle, S_{in}, S_{out})$ output $\|\omega\rangle$	 $QE(\|\psi\rangle, S_{in}, S_{out}, \|\phi_r\rangle)$ is the quantum emulation algorithm.

Fig. 3: $(\mathsf{qUI}, \mathsf{qEx}, \mu)$-PQEA: adversary's algorithm against game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}$

The probability analysis of the above attack has been discussed in detail in Appendix F.2. The main idea of the attack is similar to the proof of Theorem 3. But since the learning phase states are not being chosen by the adversary, to win the game the adversary needs to create superposition of any two unknown states randomly chosen from S_{in}. But as mentioned before, there is no algorithm to create such an exact target superpositions. Although inspired by probabilistic algorithms such as [38, 39], we present an algorithm for preparing an unknown superposition of unknown states that we denote Superpose$(\cdot, \cdot)$. Then we use the quantum emulation algorithm on the new superposed state which satisfies the indistinguishability condition, and we use the other two states as reference and sample states for the emulation. We show the fidelity of the algorithm is bounded by $\frac{1}{2}$. For any $0 < \mu \leq \frac{1}{2}$ the success probability will be as follows in the security parameters including D, κ_1 and κ_2.

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qUI,qEx},\mu}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes \kappa_1}, |\omega\rangle^{\otimes \kappa_2})] \geq \frac{1}{2}$$

Thus, $\mathcal{A}$ wins the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qUI,qEx},\mu}(\lambda, \mathcal{A})$ with non-negligible probability. □

In a qSel challenge phase, the challenge quantum state is picked uniformly at random by the challenger. Thus, the adversary receives a completely unknown and random quantum state as a challenge. This unknown challenge is an important point of difference between classical and quantum security frameworks, as it is known that a quantum state cannot be perfectly determined from a single copy and also a quantum state will be disturbed and collapsed to another state once measured to extract information from it [32]. These limitations can be exploited to enhance the security of a cryptographic primitive. Here, we identify the two main assumptions for a unitary cryptographic primitive to be secure and withstand our newly introduced quantum emulation attacks. Firstly, the unitary primitive should initially be unknown to the adversary as formalised in Definition 4, i.e. the adversary should have no prior information about $\mathrm{U}_{\mathcal{E}}$ before starting the game. Secondly, the challenger needs to pick the challenge uniformly at random from the Hilbert space $\mathcal{H}^D$. In the literature, picking a uniform set of quantum states is known as picking states distributed according to the Haar measure [40, 41]. Since the challenges are unknown and picked from a uniform distribution from $\mathcal{H}^D$, no degree of μ will lead to a trivial attack, so, we drop the condition of μ-distinguishability for this challenge type.

Definition 4 (Unknown Unitary Transformation). *We call a unitary transformation* U *with dimension D Unknown, if for any QPT algorithm $\mathcal{A}$ before making any query to* U*, the probability of outputting a response corresponding to any state in $\mathcal{H}^D$ with non-negligible fidelity with the correct output is negligible:*

$$Pr[\forall |\psi\rangle \in \mathcal{H}^D : F(\mathcal{A}(|\psi\rangle), \mathrm{U}|\psi\rangle) \geq \textit{non-negl}(\log(D))] = \textit{negl}(\log(D))$$

In Theorem 6, we investigate the security of the cryptographic primitive $\mathcal{F}$ whose underlying evaluation algorithm is unknown to the adversary when the challenge is chosen uniformly at random by the challenger as captured by game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda, \mathcal{A})$. But to prove Theorem 6 we need to establish the following lemma.

Lemma 1. *Let $\mathcal{H}^D$ be a D-dimensional Hilbert space and $\mathcal{H}^d$ a subspace of $\mathcal{H}^D$ with dimension d. Also, let Π_d be any operator projecting any quantum state in $\mathcal{H}^D$ into $\mathcal{H}^d$. The average probability that any state $|\psi\rangle \in \mathcal{H}^D$ is projected into $\mathcal{H}^d$ is equal to $\frac{d}{D}$*

$$\underset{|\psi\rangle, \Pi_d}{Pr}[\langle\psi| \Pi_d |\psi\rangle] = \frac{d}{D}$$

Proof. The proof can be found in the Appendix G.1. The proof is mainly based on the symmetry of the Hilbert space and the fact that the probability of falling into each subspace is equal for a state uniformly picked at random. □

To establish our go result, we first present a primary theorem which demonstrates the security of the unitary primitives irrespective to the power of the test algorithm, or more precisely, with an ideal test algorithm which asymptotically satisfies the notion of distance. We formalize the $\mathcal{T}_{\delta}^{ideal}$ test as follows:

Definition 5 ($\mathcal{T}_{\delta}^{ideal}$ Test Algorithm). *We call a test algorithm according to Definition 2, a $\mathcal{T}_{\delta}^{ideal}$ Test Algorithm when for any two state $|\psi\rangle$ and $|\phi\rangle$ with fidelity $F(|\psi\rangle, |\phi\rangle)$ the test responds as follows:*

$$\mathcal{T}_{\delta}^{ideal} = \begin{cases} 1 & F(|\psi\rangle, |\phi\rangle) \geq \delta \\ 0 & otherwise \end{cases}$$

Theorem 5. *For any unknown unitary cryptographic primitive $\mathcal{F} = (\mathcal{S}, \mathcal{E}, \mathcal{T}_{\delta}^{ideal})$, where $\mathcal{T}_{\delta}^{ideal}$ is defined according to Definition 5, for any non-zero δ, the success probability of any adversary $\mathcal{A}$ in the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda, \mathcal{A})$ is bounded as follows:*

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda, \mathcal{A})] \leq \frac{d+1}{D}$$

where D is the dimension of the Hilbert space that the challenge quantum state is picked from, and $0 \leq d \leq D-1$ is the dimension of the largest subspace of $\mathcal{H}^D$ that the adversary can span during the first learning phase of $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda, \mathcal{A})$.

Proof. The complete proof can be found in Appendix G.2. Here we only sketch the proof. We are interested in the average success probability of the adversary playing the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda, \mathcal{A})$ and spanning a d-dimensional subspace of $\mathcal{H}^D$ in the learning phase. More generally we calculate the average fidelity of the adversary's state and the correct output, over all choices of $|\psi\rangle$. We require this fidelity to be greater than a value δ imposed by the $\mathcal{T}_{\delta}^{ideal}$:

$$Pr_{success} = \underset{|\psi\rangle \in \mathcal{H}^D}{Pr}[F \geq \delta].$$

Also, we bound the probability by the success probability of a more powerful adversary with full knowledge over the learnt subspace. We then calculate the success probability of that adversary in terms of its partial probability for the states orthogonal to the learning phase subspace and the rest of the space:

$$Pr_{success} = \underset{|\psi\rangle \in \mathcal{H}^{d^{\perp}}}{Pr}[F \geq \delta] Pr[|\psi\rangle \in \mathcal{H}^{d^{\perp}}] + \underset{|\psi\rangle \notin \mathcal{H}^{d^{\perp}}}{Pr}[F \geq \delta] Pr[|\psi\rangle \notin \mathcal{H}^{d^{\perp}}].$$

The probability of projection into the orthogonal subspace and the conjugate subspace can be obtained by calling Lemma 1 and the only remaining term to calculate is the probability that the average fidelity is greater than δ in the orthogonal subspace. Using the fact that each state $|\psi\rangle$ is picked at random from a uniform distribution of states on $\mathcal{H}^D$ which asymptotically covers the whole Hilbert space uniformly, we show that the necessary condition for adversary's states to achieve the desired fidelity on average is that the distribution of adversary's output must be uniform over the Hilbert space as well. Then by calculating the average fidelity according to Haar measure, we show that the average probability for non-zero fidelity is bounded as:

$$\underset{|\psi^{out}\rangle \in \mathcal{H}_{out}^{d\perp}}{Pr}[F \neq 0] \leq \frac{1}{D-d}$$

which results to the final target probability being:

$$Pr_{success} \leq \frac{d+1}{D} \qquad \square$$

Theorem 6. *Any unknown unitary cryptographic primitive $\mathcal{F} = (\mathcal{S}, \mathcal{E}, \mathcal{T})$, is asymptotically secure against a* $(\mathsf{qCI}, \mathsf{qSel})$*-PQEA, if the error of the $\mathcal{T}$ defined according to Definition 2, satisfies $Err(\kappa_1, \kappa_2) = negl(\kappa_1, \kappa_2)$. Then the success probability of any QPT adversary $\mathcal{A}$ in the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})$ is:*

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})] = negl(\lambda).$$

Proof. The complete proof can be found in Appendix G.3, the proof sketch is to write the conditional probability of the test algorithm $T(|\psi^{out}\rangle^{\otimes \kappa_1}, |\omega\rangle^{\otimes \kappa_2})$ outputting 1 in two different cases depending on the fidelity. We show that for the cases that the fidelity is non-negligible, in the limit of the security parameters κ_1 and κ_2, the probability of the test algorithm outputting 1 converges to their non-negligible fidelity, but this event is rare (According to Theorem 5), so overall this part will be a negligible function. The probability of fidelity being negligible on the other hand will converge to 1 but the probability that the test algorithm outputs 1, in this case, will converge to the error $Err(\kappa_1, \kappa_2)$. By restricting the error to be a negligible function of the security parameters, we then will have:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}] = negl(\lambda) + Err(\kappa_1, \kappa_2) = negl(\lambda) \qquad \square$$

Corollary 1. *Any unitary cryptographic primitive $\mathcal{F}$ is secure against a* $(\mathsf{qUI}, \mathsf{qSel})$*-PQEA if the error of $\mathcal{T}$ satisfies $Err(\kappa_1, \kappa_2) = negl(\kappa_1, \kappa_2)$.*

Proof. This follows from Theorem 6 as the adversary is weaker and does not even choose the input quantum states in the learning phase. $\square$

Theorem 5 provides an upper-bound for the success probability of any adversary depending on the relative dimensions of the learnt subspace and the Hilbert space of the unknown transformation. However, we now show that adaptive adversaries that are given access to an extra learning phase after the qSel challenge phase can win the game with overwhelming probability. The algorithm exploits entanglement to break the security of the primitive.

Theorem 7 (No unitary cryptographic primitive $\mathcal{F}$ is secure against a $(\mathsf{null}, \mathsf{qSel}, \mathsf{qCI}, \mu)$-PQEA). *There exists a QPT adversary $\mathcal{A}$ such that for any $0 \leq \mu \leq 1 - non\text{-}negl(\lambda)$*

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}(\lambda, \mathcal{A})] = 1.$$

$(\mathsf{null}, \mathsf{qSel}, \mathsf{qCI}, \mu)$-*PQEA*

First learning phase: null

Challenge phase:

prepare qubit $|0\rangle_a$
receive $|\psi\rangle_c$ as a challenge.

$\mathcal{A}$ receives the unknown challenge state $|\psi\rangle = \sum_{i=1}^{D} \alpha_i |b_i\rangle$ where $\{|b_i\rangle\}_{i=1}^{D}$ are set of complete orthonormal basis for $\mathcal{H}^D$.

Second learning phase:

$|\psi\rangle_{ca} \leftarrow CNOT_{c,a}(|\psi\rangle |0\rangle)$
query state c
receive $\mathrm{U}_{\mathcal{E}}\rho_c = (\mathrm{U}_{\mathcal{E}} \otimes \mathcal{I}) |\psi\rangle_{ca}$

The sub-index c denotes the challenge and the sub-index a denotes the adversary's qubit.

$\mathcal{A}$ sends the challenge part of the entangled system as a request.

ρ_c is the challenge part of the entangled state.

Guess phase:

$|\psi^{out}\rangle \otimes |\pm\rangle \leftarrow Measure(|\psi\rangle_{ca}, \{|\pm\rangle\})$
if $|\pm\rangle = |+\rangle$
  output: $|\omega\rangle = |\psi^{out}\rangle$
else
  output: $|\omega\rangle = CZ^{\otimes n-1}(|\psi^{out}\rangle)$

$Measure(|\psi\rangle_{ca}, \{|\pm\rangle\}$ outputs the result of the measurement.

Fig. 4: $(\mathsf{null}, \mathsf{qSel}, \mathsf{qCI}, \mu)$-PQEA: adversary's algorithm against game $\mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}$

Proof. Let $\mathcal{A}$ be the QPT adversary playing the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}(\lambda, \mathcal{A})$ and running the algorithm described in Figure 4. The probability analysis and complete proof have been discussed in Appendix G.4. The main ingredient of the proof is the entanglement between $\mathcal{A}$'s local system and the challenge state $|\psi\rangle$ which allows $\mathcal{A}$ to adaptively ask for a part of this entangled state in the second learning phase. Then by performing a good measurement, $\mathcal{A}$ can extract the $U|\psi\rangle$ from the entangled state. Also, we have shown that the μ-distinguishability is satisfied on average, over all choices of $|\psi\rangle$. The proof concludes that:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI}}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes \kappa_1}, |\omega\rangle^{\otimes \kappa_2})] = 1 \qquad \square$$

5 Quantum Physical Unclonable Function (QPUF)

A PUF is a process that makes the unique features of each device observable. Many implementations of PUFs rely on optical systems and hence can be potentially queried with quantum states (encoded as photons). For instance, consider a set of optical media with a high density of scatterers that have been created by the same manufacturing process. Each such optical device responds with a quantum output when probed by light pulses. It has been shown in [18, 8, 28], that due to the unique features of each medium, the generated quantum outputs corresponding to each medium are distinguishable when these are probed with a single set of quantum inputs. This setup is the simplest practical realization of a Quantum PUF (QPUF). In this section, we consider such QPUFs, or quantum readouts of CPUFS as termed in [7]. We first formally define QPUFs and present the desired notion of security for QPUFs, that is *selective* and *existential unforgeability*, in the quantum game-based framework from Section 4.1, and then give general results as to their security against different attacker capabilities.

5.1 Definition of QPUFs

Let there be a set of quantum devices that have been created through the same manufacturing process. These devices respond with a quantum output when challenged with a quantum input. Similarly, to Section 2.1, we formalize the manufacturing process of QPUFs by defining a QGen algorithm:

$$\mathrm{QPUF}_{\mathbf{id}} \leftarrow \mathrm{QGen}(\lambda)$$

where **id** is the identifier of $\mathrm{QPUF}_{\mathbf{id}}$ and λ the security parameter.

We also need to define the QEval algorithm mapping any input quantum state $\rho_{in} \in \mathcal{H}^{d_{in}}$ to an output quantum state $\rho_{out} \in \mathcal{H}^{d_{out}}$ where $\mathcal{H}^{d_{in}}$ and $\mathcal{H}^{d_{out}}$ are the domain and range Hilbert spaces of $\mathrm{QPUF}_{\mathbf{id}}$, denoted as:

$$\rho_{out} \leftarrow \mathrm{QEval}(\mathrm{QPUF}_{\mathbf{id}}, \rho_{in}).$$

For now, we allow for the most general form of trace-preserving (CPT) quantum maps for the QEval and denote it:

$$\rho_{out} = \Lambda_{\mathbf{id}}(\rho_{in})$$

So far, we have just characterised the existing algorithms in the context of PUFs to fit in the quantum setting. However, we need an extra algorithm for QPUFs which is not required for CPUFs, that is an efficient test algorithm for testing the equality between two unknown quantum states. To this end, we consider the algorithm introduced in Definition 2 as the test algorithm of QPUFs.

Definition 6 (Quantum Physical Unclonable Function). *Let λ be the security parameter, and $\delta_r, \delta_u, \delta_c \in [0, 1]$ the robustness, uniqueness and collision resistance thresholds. A $(\lambda, \delta_r, \delta_u, \delta_c)$-QPUF includes the algorithms:* QGen, QEval *and $\mathcal{T}$ satisfying Requirements 1, 2, and 3 defined below:*

Requirement 1 (δ_r-Robustness) *For any $\mathrm{QPUF}_{\mathbf{id}}$ and any two input states ρ_{in} and σ_{in} that are δ_r-indistinguishable, the corresponding output quantum states ρ_{out} and σ_{out} are also δ_r-indistinguishable with overwhelming probability,*

$$\Pr[\delta_r \leq F(\rho_{out}, \sigma_{out}) \leq 1] = 1 - negl(\lambda).$$

Requirement 2 (δ_u-Uniqueness) *For any two* QPUF*s generated by the* QGen *algorithm, i.e.* $\text{QPUF}_{\mathbf{id}_i}$ *and* $\text{QPUF}_{\mathbf{id}_j}$*, the corresponding CPT map models, i.e.* Λ_{id_i} *and* Λ_{id_j} *are* δ_u*-distinguishable with overwhelming probability,*

$$\Pr[\ \| (\Lambda_{id_i} - \Lambda_{id_j})_{i \neq j} \|_{\diamond} \geq \delta_u\ ] = 1 - negl(\lambda).$$

Requirement 3 (δ_c-Collision-Resistance) *For any* $\text{QPUF}_{\mathbf{id}}$ *and any two input states* ρ_{in} *and* σ_{in} *that are* δ_c*-distinguishable, the corresponding output states* ρ_{out} *and* σ_{out} *are also* δ_c*-distinguishable with overwhelming probability,*

$$\Pr[0 \leq F(\rho_{out}, \sigma_{out}) \leq 1 - \delta_c] = 1 - negl(\lambda).$$

In QPUF-based applications such as device authentication (or identification), it is necessary that there is a clear distinction between different QPUFs [26]. To this end, the following conditions should be satisfied: $\delta_c \leq 1 - \delta_r$ and $\delta_u \leq 1 - \delta_r$. We have initially allowed for any CPT map as QEval algorithms. Now, we show however that random CPT maps do not provide collision-resistance.

Theorem 8. *Non-unitary trace preserving quantum operations are not* $(\lambda, \delta_r, \delta_u, \delta_c)$*-QPUFs for any* λ*,* δ_r*,* δ_u*, and* δ_c*.*

Proof. The contractive property of trace-preserving operations [32] states that for two general density operators ρ and σ we have:

$$\mathcal{D}_{tr}(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \leq \mathcal{D}_{tr}(\rho, \sigma)$$

where $\mathcal{E}$ denotes a Completely Positive Trace-preserving (CPT) map. Similarly, because of the relation between trace distance and the fidelity we have

$$F(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \geq F(\rho, \sigma)$$

Equality holds only when $\mathcal{E}$ is a unitary operation. Thus, no non-unitary trace-preserving quantum channel satisfies requirement 3. □

According to Theorem 8, the only possible candidates for such trace-preserving maps are unitary transformations, as long as one requires both robustness and collision-resistance requirements to be strongly satisfied for any δ_r and δ_c. As a result, we restrict ourselves to QPUFs with unknown unitary evaluation algorithm QEval which manipulate pure quantum states. We call this type of QPUFs, Unitary QPUFs (or simply UQPUFs) and formally define them in Definition 7.

Definition 7 (Unitary QPUF (UQPUF)). *A Unitary* QPUF (UQPUF) *is a* QPUF *such that the* QEval *algorithm can be modelled by an unknown unitary transformation* $\text{U}_{\mathbf{id}}$ *over a* D*-dimensional Hilbert space,* $\mathcal{H}^D$ *operating on pure quantum input states* $|\psi_{in}\rangle \in \mathcal{H}^D$ *and returning pure outputs* $|\psi_{out}\rangle \in \mathcal{H}^D$*,*

$$|\psi_{out}\rangle = \text{QEval}(\text{UQPUF}_{\mathbf{id}}, |\psi_{in}\rangle) = \text{U}_{\mathbf{id}} |\psi_{in}\rangle .$$

We can now prove that UQPUFs satisfy requirements 1, 2 and 3 of QPUFs.

Theorem 9. *Any UQPUF is a $(\lambda, \delta_r, \delta_u, \delta_c)$-QPUF for any δ_r, δ_u, and δ_c.*

Proof. To show any UQPUF satisfies requirements 1 and 3, it is enough to note that any unitary transformation preserves the fidelity of two quantum states. So, if two input states are δ_r-indistinguishable, then, their corresponding outputs are also δ_r-indistinguishable. Also, if two input states are δ_c-distinguishable, then, their corresponding outputs are δ_c-distinguishable, too. So, any UQPUF provides δ_r-robustness and δ_c-collision-resistance. Finally, by picking UQPUF uniformly at random from Haar family of unitaries, as the average fidelity of Haar states is in the order of $\frac{1}{D}$[42] and the diamond norm captures the distance for maximum fidelity of the channel's output states[43], Then the diamond norm between any two UQPUFs is by definition, close to one. So, a UQPUF is also δ_u-unique. □

As a result of the distance-preserving property of UQPUFs, we drop δ_r, δ_c, and δ_u from the notation and we simply characterise λ-UQPUFs. Requiring UQPUF transformations to be initially unknown (or exponentially hard to recover by tomography) is a hardware assumption which is also considered in classical PUFs mentioning the PUF behaviour is unknown even for the manufacturer [4]. Further, this assumption is a realistic one given that there exist efficient ways to sample a family of unitaries indistinguishable from the Haar family of unitary transformations [44, 45].

5.2 Security Analysis of Unitary QPUFs (UQPUFs)

The security of most PUF-based cryptographic protocols relies on the unforgeability of PUFs, that is estimating the output of a PUF on a given input should be impossible without actually being in possession of the PUF [26, 5]. We formalise this notion for Unitary QPUFs (UQPUFs) in the framework introduced in Section 4.1 considering UQPUF = $(\mathrm{QGen}, \mathrm{QEval}, \mathcal{T})$ as our unitary cryptographic primitive. We first investigate unforgeability of UQPUFs against unbounded quantum adversaries and call this property *quantum unconditional unforgeability*. We then consider *quantum existential* (and *selective*) unforgeability of UQPUFs against QPT adversaries.

Quantum Unconditional Unforgeability. In the classical setting, CPUFs can be fully described by the finite set of CRPs, and this suffices for breaking unforgeability. More precisely, an unbounded adversary can extract the entire set of CRPs by querying the target CPUF with all possible challenges [46]. If the challenges are n-bit strings, the number of possible challenges is 2^n. However, in the quantum setting, a UQPUF can generate an infinite number of quantum challenge-response pairs such that extracting all of them is hard, even for unbounded adversaries. This, combined with limitations imposed by quantum mechanics such as no-cloning [47] and the limits on state estimation [48], raise the question if UQPUFs could satisfy unforgeability against unbounded adversaries. Informally, unconditional unforgeability of UQPUFs implies no unbounded quantum adversary can generate even a single valid challenge-response pair without possessing it. We now formally define *quantum unconditional unforgeability* and prove that no UQPUF provides quantum unconditional unforgeability.

Definition 8 (Quantum Unconditional Unforgeability). *We say that a λ-UQPUF is quantum unconditionally unforgeable if the success probability of any unbounded quantum adversary $\mathcal{A}$ of winning the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})$ for any μ, is negligible in the security parameter λ,*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})] = negl(\lambda).$$

We prove a stronger version. We show that even if the challenge is chosen by the challenger (i.e. we consider a selective challenge phase), the adversary still can win the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel},\mu}(\lambda,\mathcal{A})$ with non-negligible probability.

Theorem 10. *[No UQPUF provides quantum unconditional unforgeability] For any λ-UQPUF and any μ, there exists an unbounded quantum adversary $\mathcal{A}$ such that*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel},\mu}(\lambda,\mathcal{A})] = non\text{-}negl(\lambda).$$

Proof. A complete proof of this theorem can be found in Appendix G.5. The key idea of the proof is based on complexity analysis of unitary tomography and implementation of a general unitary by single and double qubit gates, since for an unbounded quantum adversary, it will be feasible to extract the unitary matrix by tomography and then build the extracted unitary by general gate decomposition method. By using the Solovay-Kitaev theorem [32], we then show that the adversary can build the unitary matrix of the UQPUF performing on n-qubits, within an arbitrarily small distance ϵ using $O(n^2 4^n \log^c(n^2 4^n))$ gates and hence win the game with any test algorithm $\mathcal{T}$:

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel},\mu}(\lambda,\mathcal{A})] = 1 \qquad \square$$

Quantum Existential Unforgeability. We now look at bounded adversaries, i.e. Quantum Polynomial-Time (QPT) algorithms. And start by investigating *quantum existential unforgeability* of UQPUFs. Informally, a UQPUF provides quantum existential unforgeability if no QPT adversary can generate even a single valid quantum challenge-response pair. We show this notion of security cannot be achieved even under quantum random input learning phases as long as the new challenge has overlap with the learnt inputs, i.e. $\mu \neq 1$.

Definition 9 (Quantum Existential Unforgeability). *A λ-UQPUF provides quantum existential unforgeability if the success probability of any QPT adversary $\mathcal{A}$ of winning the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})$ is negligible in λ,*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})] = negl(\lambda).$$

Theorem 11 (No UQPUF provides quantum existential unforgeability). *For any λ-UQPUF, and $\mu \leq 1 - non\text{-}negl(\lambda)$, there exits a QPT adversary $\mathcal{A}$ such that*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})] = non\text{-}negl(\lambda).$$

Proof. Let $\mathcal{A}$ be a QPT adversary playing the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ and $\text{U}_{\mathbf{id}}$ an initially unknown unitary transformation over D-dimensional Hilbert space $\mathcal{H}^D$ modelling $\text{UQPUF}_{\mathbf{id}}$. According to Theorems 3 and 4, there exists a PQEA that outputs a state $|\omega\rangle$ with a non-negligible fidelity. As the probability of any test algorithm according to Definition 2, is lower bounded by the fidelity, we have

$$Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})] = non\text{-}negl(\lambda)$$

Also, as a strong result, in qCI learning phase and $\mu \leq \frac{1}{2}$ according to Theorem 3:

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = 1. \qquad \square$$

It is worth mentioning that according to Theorem 11, the QPT adversary wins the game for all values of μ except the case that $\mu = 1$ which implies the challenge state is orthogonal to all the queried states in the learning phase. In other words, the adversary has no quantum access to the QPUF. So, regardless of the quantification of μ, no λ-UQPUF can provide quantum existential unforgeability.

Quantum Selective Unforgeability. Given the previously established no-go results, we adapt the security notion and focus on selective unforgeability. This notion will be sufficient in many UQPUF-based protocols such as identification as in most of the PUF-based applications introduced in the literature [26, 49], the PUF needs to respond to a challenge chosen by the verifier. This is precisely the scenario captured by games with Selective Challenge phase. We formally define the notion of *quantum selective unforgeability* and provide a general result as to the security of UQPUFs wrt this notion.

Definition 10 (Quantum Selective Unforgeability). *A λ-UQPUF is quantum selectively unforgeable if the success probability of any QPT adversary $\mathcal{A}$ of winning the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})$ is negligible in the security parameter λ,*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})] = negl(\lambda).$$

We now prove that any UQPUF provides quantum selective unforgeability.

Theorem 12. *[Any UQPUF provides Selective Unforgeability] if there exist a Test $\mathcal{T}$ with error $Err(\kappa_1, \kappa_2) = negl(\lambda)$. Then for any λ-UQPUF and any QPT adversary $\mathcal{A}$, we have:*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})] = negl(\lambda).$$

Proof. By assumption, $\text{U}_{\mathbf{id}}$ is an unknown unitary according to Definition 4. According to Theorem 6, the average success probability of any QPT algorithm is $negl(\lambda)$. This concludes our proof. $\square$

Finally, we consider the case of adaptive adversary that runs the second learning phase after getting the challenge in the qSel challenge phase and define *adaptive quantum selective unforgeability.*

Definition 11 (Adaptive Quantum Selective Unforgeability). *A λ-UQPUF provides adaptive quantum selective unforgeability if the success probability of any QPT adversary $\mathcal{A}$ of winning the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}(\lambda, \mathcal{A})$ is negligible in λ,*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}(\lambda, \mathcal{A})] = negl(\lambda).$$

According to the Theorem 7 and the corresponding proof, no UQPUF provides adaptive quantum selective unforgeability.

Theorem 13 (No UQPUF provides Adaptive Quantum Selective Unforgeability). *For any λ-UQPUF, and $\mu \leq 1 - non\text{-}negl(\lambda)$ there exits a QPT adversary $\mathcal{A}$ such that*

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}(\lambda, \mathcal{A})] = 1.$$

6 Conclusion and Future Directions

We presented a general quantum game-based framework for analysing cryptographic primitives; as well as go and no-go results for such primitives. We then used our framework and results for providing precise definitions for secure QPUFs. In particular, we proved UQPUFs provide quantum selective unforgeability, which establishes that UQPUFs could be applicable in several applications such as device authentication.

The extension of our analysis to other adversarial settings such as bounded storage or noisy memory is our target for further QPUF feasibility analysis. Another direction is the possibility of defining QPUFs with mixed input-output quantum states if efficient solutions for testability were to be devised. Moreover, proposing QPUF-based protocols and their security analysis in the standard models will be considered in our future works.

Finally, our proposed quantum emulation attack along with our defined quantum game-based framework could be further exploited for investigating the security of classical PUFs and QPUF-based identification protocols found in the literature as well as other cryptographic primitives such as digital signatures.

Acknowledgement The authors are grateful to Niraj Kumar, Petros Walden, Thomas Zacharias and Mojtaba Jafarzadegan for useful discussions. The work was supported by EPSRC grants, ref EP/N003829/1.

References

1. R. Canetti and M. Fischlin, "Universally composable commitments," in *Proceedings of Annual International Cryptology Conference*, pp. 19–40, Springer, 2001.
2. J. Katz, "Universally composable multi-party computation using tamper-proof hardware," in *Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 115–128, Springer, 2007.

3. S. Badrinarayanan, D. Khurana, R. Ostrovsky, and I. Visconti, "Unconditional uc-secure computation with (stronger-malicious) pufs," in *Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 382–411, Springer, 2017.
4. U. Rührmair and D. E. Holcomb, "Pufs at a glance," in *Proceedings of the conference on Design, Automation & Test in Europe*, p. 347, European Design and Automation Association, 2014.
5. C. Brzuska, M. Fischlin, H. Schröder, and S. Katzenbeisser, "Physically uncloneable functions in the universal composition framework," in *Proceedings of Annual International Cryptology Conference*, pp. 51–70, Springer, 2011.
6. B. Škorić, "Quantum readout of physical unclonable functions," in *Proceedings of International Conference on Cryptology in Africa*, pp. 369–386, Springer, 2010.
7. B. ŠKORIĆ, "Quantum readout of physical unclonable functions," *International Journal of Quantum Information*, vol. 10, no. 01, p. 1250001, 2012.
8. G. M. Nikolopoulos and E. Diamanti, "Continuous-variable quantum authentication of physical unclonable keys," *Scientific reports*, vol. 7, p. 46047, 2017.
9. R. Young, U. Roedig, and J. Roberts, "Quantum physical unclonable function," 2019. US Patent App. 16/178,670.
10. D. Boneh and M. Zhandry, "Secure signatures and chosen ciphertext security in a quantum computing world," in *Proceedings of Annual International Cryptology Conference*, pp. 361–379, Springer, 2013.
11. T. Gagliardoni, "Quantum security of cryptographic primitives," *arXiv preprint arXiv:1705.02417*, 2017.
12. G. Alagic, T. Gagliardoni, and C. Majenz, "Unforgeable quantum encryption," in *Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 489–519, Springer, 2018.
13. D. Boneh, Ö. Dagdelen, M. Fischlin, A. Lehmann, C. Schaffner, and M. Zhandry, "Random oracles in a quantum world," in *Proceedings of International Conference on the Theory and Application of Cryptology and Information Security*, pp. 41–69, Springer, 2011.
14. M. Zhandry, "Secure identity-based encryption in the quantum random oracle model," *International Journal of Quantum Information*, vol. 13, no. 04, p. 1550014, 2015.
15. I. Damgård, J. Funder, J. B. Nielsen, and L. Salvail, "Superposition attacks on cryptographic protocols," in *Proceedings of International Conference on Information Theoretic Security*, pp. 142–161, Springer, 2013.
16. V. Soukharev, D. Jao, and S. Seshadri, "Post-quantum security models for authenticated encryption," in *Proceedings of 7th International Workshop on Post-Quantum Cryptography*, pp. 64–78, Springer, 2016.
17. I. Marvian and S. Lloyd, "Universal quantum emulator," *arXiv preprint arXiv:1606.02734*, 2016.
18. R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions," *Science*, vol. 297, no. 5589, pp. 2026–2030, 2002.
19. C. Mesaritakis, M. Akriotou, A. Kapsalis, E. Grivas, C. Chaintoutis, T. Nikas, and D. Syvridis, "Physical unclonable function based on a multi-mode optical waveguide," *Scientific reports*, vol. 8, no. 1, p. 9653, 2018.
20. B. Gassend, D. Clarke, M. Van Dijk, and S. Devadas, "Silicon physical random functions," in *Proceedings of the 9th ACM conference on Computer and communications security*, pp. 148–160, ACM, 2002.

21. G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation," in *Proceedings of 44th ACM/IEEE Design Automation Conference*, pp. 9–14, IEEE, 2007.
22. M. Delavar, S. Mirzakuchaki, and J. Mohajeri, "A ring oscillator-based puf with enhanced challenge-response pairs," *Canadian Journal of Electrical and Computer Engineering*, vol. 39, no. 2, pp. 174–180, 2016.
23. J. Guajardo, S. S. Kumar, G.-J. Schrijen, and P. Tuyls, "Fpga intrinsic pufs and their use for ip protection," in *Proceedings of International workshop on cryptographic hardware and embedded systems*, pp. 63–80, Springer, 2007.
24. R. Maes, *Physically Unclonable Functions: Constructions, Properties and Applications.* Springer-Verlag Berlin Heidelberg, 1st ed., 2016.
25. B. Halak, *Physically Unclonable Functions: From Basic Design Principles to Advanced Hardware Security Applications.* Springer International Publishing, 1st ed., 2019.
26. F. Armknecht, D. Moriyama, A.-R. Sadeghi, and M. Yung, "Towards a unified security model for physically unclonable functions," in *Proceedings of Cryptographers' Track at the RSA Conference*, pp. 271–287, Springer, 2016.
27. R. Ostrovsky, A. Scafuro, I. Visconti, and A. Wadia, "Universally composable secure computation with (malicious) physically uncloneable functions," in *Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 702–718, Springer, 2013.
28. S. A. Goorden, M. Horstmann, A. P. Mosk, B. Škorić, and P. W. Pinkse, "Quantum-secure authentication of a physical unclonable key," *Optica*, vol. 1, no. 6, pp. 421–424, 2014.
29. B. Škoric, "Security analysis of quantum-readout pufs in the case of challenge-estimation attacks," *Quantum Information and Computation*, vol. 16, 2016.
30. Y. Yao, M. Gao, M. Li, and J. Zhang, "Quantum cloning attacks against puf-based quantum authentication systems," *Quantum Information Processing*, vol. 15, no. 8, pp. 3311–3325, 2016.
31. G. M. Nikolopoulos, "Continuous-variable quantum authentication of physical unclonable keys: Security against an emulation attack," *Physical Review A*, vol. 97, no. 1, p. 012324, 2018.
32. M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information.* Cambridge University Press, 10th ed., 2010.
33. H. Buhrman, R. Cleve, J. Watrous, and R. De Wolf, "Quantum fingerprinting," *Physical Review Letters*, vol. 87, no. 16, p. 167902, 2001.
34. U. Chabaud, E. Diamanti, D. Markham, E. Kashefi, and A. Joux, "Optimal quantum-programmable projective measurement with linear optics," *Physical Review A*, vol. 98, no. 6, p. 062318, 2018.
35. G. D'Ariano and P. L. Presti, "Quantum tomography for measuring experimentally the matrix elements of an arbitrary quantum operation," *Physical Review Letters*, vol. 86, no. 19, p. 4195, 2001.
36. S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum principal component analysis," *Nature Physics*, vol. 10, no. 9, p. 631, 2014.
37. T. Gagliardoni, A. Hülsing, and C. Schaffner, "Semantic security and indistinguishability in the quantum world," in *Proceedings of Annual International Cryptology Conference*, pp. 60–89, Springer, 2016.
38. M. Oszmaniec, A. Grudka, M. Horodecki, and A. Wójcik, "Creating a superposition of unknown quantum states," *Physical Review Letters*, vol. 116, no. 11, 2016.
39. M. Doosti, F. Kianvash, and V. Karimipour, "Universal superposition of orthogonal states," *Physical Review A*, vol. 96, no. 5, p. 052318, 2017.

40. K. Zyczkowski and H.-J. Sommers, "Induced measures in the space of mixed quantum states," *Journal of Physics A: Mathematical and General*, vol. 34, no. 35, p. 7111, 2001.
41. A. Van Daele, "The haar measure on a compact quantum group," in *Proceedings of the American Mathematical Society*, vol. 123, pp. 3125–3128, 1995.
42. K. Życzkowski and H.-J. Sommers, "Average fidelity between random quantum states," *Physical Review A*, vol. 71, no. 3, p. 032313, 2005.
43. D. Aharonov, A. Kitaev, and N. Nisan, "Quantum circuits with mixed states," *arXiv preprint quant-ph/9806029*, 1998.
44. C. Dankert, R. Cleve, J. Emerson, and E. Livine, "Exact and approximate unitary 2-designs and their application to fidelity estimation," *Physical Review A*, vol. 80, no. 1, p. 012304, 2009.
45. A. Ambainis and J. Emerson, "Quantum t-designs: t-wise independence in the quantum world," in *Proceedings of Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pp. 129–140, IEEE, 2007.
46. C.-H. Chang, Y. Zheng, and L. Zhang, "A retrospective and a look forward: Fifteen years of physical unclonable function advancement," *IEEE Circuits and Systems Magazine*, vol. 17, no. 3, pp. 32–62, 2017.
47. W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, no. 5886, p. 802, 1982.
48. D. Bruss, A. Ekert, and C. Macchiavello, "Optimal universal quantum cloning and state estimation," *Physical Review Letters*, vol. 81, no. 12, p. 2598, 1998.
49. M. Delavar, S. Mirzakuchaki, M. H. Ameri, and J. Mohajeri, "Puf-based solutions for secure communications in advanced metering infrastructure (ami)," *International Journal of Communication Systems*, vol. 30, no. 9, p. e3195, 2017.

Supplementary materials

A Quantum Gates

We introduce the quantum gates that we have used in the paper. From the universality of the quantum computation, we know that any n-qubit unitary gate can be broken to a special set of universal gates. One of these sets is the single-qubit gates and CNOT (defined below). As an example of the single-qubit gates we introduce the Z gate or Pauli-Z gate which acts on a general qubit as follows:

$$X(\alpha |0\rangle + \beta |1\rangle) = \beta |0\rangle + \alpha |1\rangle\,, \quad \text{where} \quad X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The X-gate is one of the Pauli operators (the others being the Z and Y), which together with the identity operator $\mathbb{I}$, form a basis for the vector space of 2×2 Hermitian matrices. Z and Y gates have the following unitary matrices:

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$$

The X gates switch between $|0\rangle$ and $|1\rangle$ and the Z gates transform $|1\rangle$ to $-|1\rangle$ and keep the $|0\rangle$ unchanged. Also, $Y = iXZ$. Another important single-qubit gate is Hadamard gate, denoted as H which acts as follows:

$$H|0\rangle = |+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \quad H|1\rangle = |-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$$

As $|+\rangle$ and $|-\rangle$ are also an orthonormal basis, the Hadamard gate transforms these two bases to each other. Also, the Hadamard gate creates the symmetric superposition of computational bases. CNOT is a 2-qubit gate described by the following matrix

$$CNOT = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,1\,0 \end{bmatrix}.$$

The CNOT gate flips the second qubit (the target qubit) if and only if the first qubit (the control qubit) is $|1\rangle$. The CNOT is an entangling gate as by using CNOT one can create an entangled state from two separable qubits. Another useful gate that we will use throughout this paper is another two-qubit (or multi-qubit) gate known as the SWAP gate. The SWAP gate on two quantum states with arbitrary dimension acts as follows:

$$\text{SWAP}|\psi\rangle|\phi\rangle = |\phi\rangle|\psi\rangle.$$

This gate swaps between the Hilbert space of two quantum states. The qubit SWAP gate can be built from three CNOT gates.

B SWAP test and generalised SWAP test

The SWAP test is a quantum circuit which receives two quantum states and outputs a 0 or 1 as equality or non-equality of these two states. The swap test's circuit uses the controlled version of a swap gate known as Controlled-SWAP which performs swap gate if the control qubit is $|1\rangle$. Also, it uses two Hadamard gates and an extra qubit with state $|0\rangle$ which we call ancillary qubit or ancilla. Finally it outputs $|0\rangle$ with probability $\frac{1}{2} + \frac{1}{2}F(|\psi\rangle, |\phi\rangle)$ and it outputs $|1\rangle$ with probability $\frac{1}{2} - \frac{1}{2}F(|\psi\rangle, |\phi\rangle)$. To match it with the classical definition we say the output bit of SWAP test is 1 when the output of the measurement is $|0\rangle$. The success probability of this test depends on the overlap (or fidelity) of the states. This occurs because of the quantum nature of these states and probabilistic nature of the measurements in quantum mechanics. As a result, it is not possible to perfectly distinguish two none-orthogonal quantum states with a limited number of copies of them. This means that the SWAP test has always a one-sided error. The generalised SWAP test has been introduced recently in[34]. This SWAP test uses one copy of one state and $M-1, (M \geq 2)$ copies of other state and acts better than using a SWAP test M times which will need M copy of both states. The success probability of this test is $\frac{1}{M} + \frac{M-1}{M}F(|\psi\rangle, |\phi\rangle)$.

C The Quantum Emulation Algorithm

The following figure shows the circuit of the quantum emulation algorithm described in section 3.

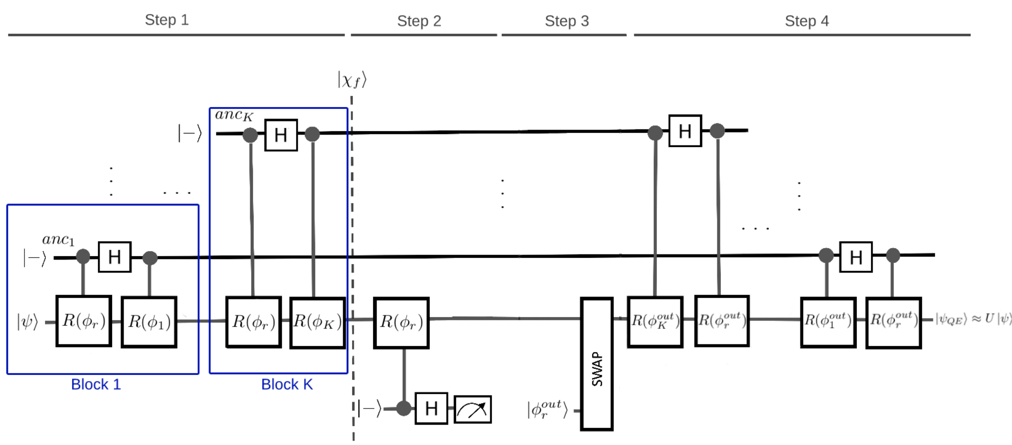

Fig. 5: The quantum emulation algorithm's circuit. $|\phi_r\rangle$ is the reference state and $|\phi_r^{out}\rangle$ is the output of the reference state. $R(*)$ gates are controlled-reflection. In each Block of Step 1, a reflection around the reference and another sample state is being performed. At Step 2 the algorithm post-select based on the success or failure of Step 1. At Step 3, the main state has been swapped with the reference output. Finally, at step 4, all the Blocks of the first step are performed in reverse order and with output samples.

D The proof of Theorem 2

We state the theorem again and we establish the proof:

Theorem 14. *Let $|\chi_f\rangle$ be the final overall state of the circuit shown in figure 5. Let K be the number of the blocks existing in the circuit. Let $|\psi\rangle$ be the input state of the circuit, $|\phi_r\rangle$ be the reference state and $|\phi_i\rangle$ other sample states. Then the final state has the following form:*

$$
\begin{aligned}
|\chi_f\rangle = {} & \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|0\rangle^{\otimes K} + |\psi\rangle\,|1\rangle^{\otimes K} - \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|1\rangle^{\otimes K} \\
& + \sum_{i=1}^{K}\sum_{j=0}^{i}[f_{ij}2^{l_{ij}}\,|\,\langle\phi_r|\,\psi\rangle|^{x_{ij}}\,|\,\langle\phi_i|\,\psi\rangle|^{y_{ij}}\,|\,\langle\phi_r|\,\phi_i\rangle|^{z_{ij}}]\,|\phi_r\rangle\,|q_{anc}(i,j)\rangle \\
& + \sum_{i=1}^{K}\sum_{j=0}^{i}[g_{ij}2^{l'_{ij}}\,|\,\langle\phi_r|\,\psi\rangle|^{x'_{ij}}\,|\,\langle\phi_i|\,\psi\rangle|^{y'_{ij}}\,|\,\langle\phi_r|\,\phi_i\rangle|^{z'_{ij}}]\,|\phi_i\rangle\,|q'_{anc}(i,j)\rangle
\end{aligned}
$$

Where l_{ij}, x_{ij}, y_{ij}, z_{ij}, l_{ij}, x'_{ij}, y'_{ij} and z'_{ij} are integer values which indicate the power of the terms of the coefficient. f_{ij} and g_{ij} can be 0, 1 or -1 and

$q_{anc}(i,j)$ and $q'_{anc}(i,j)$ are outputting a computational basis of K qubits (other than $|0\rangle^{\otimes K}$).

Proof. We prove the theorem by induction.We use equation (11) to show that for $K=1$ the form has satisfied. The term $I - R(\phi_r) = 2\,|\phi_r\rangle\langle\phi_r|$ in the equation projects the previous state to $|\phi_r\rangle$ with the coefficient $\langle\phi_r|\,\chi_{i-1}\rangle$. The operator $R(\phi_i)(I + R(\phi_r))$ has a more complicated form:

$$R(\phi_i)(I + R(\phi_r)) = 2[I - |\phi_r\rangle\langle\phi_r| - 2\,|\phi_i\rangle\langle\phi_i| + 2\,\langle\phi_i|\,\phi_r\rangle\,|\phi_i\rangle\langle\phi_r|].$$

Now for $K=1$, we have $|\chi_0\rangle = |\psi\rangle$ and the $|\chi_1\rangle$ is equal to

$$\begin{aligned}|\chi_1\rangle = {} & \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|0\rangle + |\psi\rangle\,|1\rangle - \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|1\rangle - 2\,\langle\phi_1|\,\psi\rangle\,|\phi_1\rangle\,|1\rangle \\ & + 2\,\langle\phi_r|\,\psi\rangle\,\langle\phi_r|\,\phi_1\rangle\,|\phi_1\rangle\,|1\rangle\end{aligned}$$

which satisfies the form of equation 12 where the first sum is zero and in the second sum $g_{10} = -1, g_{11} = +1$, $l_{10} = l_{11} = 1$ and $x'_{10} = z'_{10} = 0, y'_{10} = 1$ and $x'_{11} = z'_{11} = 1, y'_{11} = 0$.

Now we assume that the $|\chi_{k-1}\rangle$ satisfies equation 12, we will show that $|\chi_k\rangle$ will also satisfy 12. We use the recursive equation again. We will have

$$\begin{aligned}|\chi_k\rangle = {} & \langle\phi_r|\,\chi_{k-1}\rangle\,|\phi_r\rangle\,|0\rangle + |\chi_{k-1}\rangle\,|1\rangle - \langle\phi_r|\,\chi_{k-1}\rangle\,|\phi_r\rangle\,|1\rangle - 2\,\langle\phi_k|\,\chi_{k-1}\rangle\,|\phi_k\rangle\,|1\rangle \\ & + 2\,\langle\phi_r|\,\chi_{k-1}\rangle\,\langle\phi_r|\,\phi_k\rangle\,|\phi_k\rangle\,|1\rangle\end{aligned}$$

where $|\chi_{k-1}\rangle$ has the form of equation 12. Lets calculate each term in the above formula:

$$\begin{aligned}\langle\phi_r|\,\chi_{k-1}\rangle\,|\phi_r\rangle\,|0\rangle = {} & \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|0\rangle^{\otimes k} + \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|1\rangle^{\otimes k-1}\,|0\rangle - \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|1\rangle^{\otimes k-1}\,|0\rangle + \\ & + \sum_{i=1}^{K-1}\sum_{j=0}^{i}[f_{ij}2^{l_{ij}}\,|\,\langle\phi_r|\,\psi\rangle|^{x_{ij}}\,|\,\langle\phi_i|\,\psi\rangle|^{y_{ij}}\,|\,\langle\phi_r|\,\phi_i\rangle|^{z_{ij}}]\,|\phi_r\rangle\,|q_{anc}(i,j)\rangle\,|0\rangle \\ & + \sum_{i=1}^{K-1}\sum_{j=0}^{i}[g_{ij}2^{l'_{ij}}\,|\,\langle\phi_r|\,\psi\rangle|^{x'_{ij}}\,|\,\langle\phi_i|\,\psi\rangle|^{y'_{ij}}\,|\,\langle\phi_r|\,\phi_i\rangle|^{z'_{ij}+1}]\,|\phi_i\rangle\,|q'_{anc}(i,j)\rangle\,|0\rangle\,.\end{aligned}$$

The third term is the same only with minus sign and the ancillary states are $|0\rangle^{\otimes k-1}|1\rangle$ for the first term and $|1\rangle^{\otimes k}$ for the second and third term and $|q_{anc}(i,j)\rangle\,|1\rangle$ for the sigma terms. The second term is

$$\begin{aligned}\langle\phi_r|\,\chi_{k-1}\rangle\,|1\rangle = {} & \langle\phi_r|\,\psi\rangle\,|0\rangle^{\otimes k-1}\,|1\rangle + |\psi\rangle\,|1\rangle^{\otimes k} - \langle\phi_r|\,\psi\rangle\,|\phi_r\rangle\,|1\rangle^{\otimes k} + \\ & + \sum_{i=1}^{K-1}\sum_{j=0}^{i}[f_{ij}2^{l_{ij}}\,|\,\langle\phi_r|\,\psi\rangle|^{x_{ij}}\,|\,\langle\phi_i|\,\psi\rangle|^{y_{ij}}\,|\,\langle\phi_r|\,\phi_i\rangle|^{z_{ij}}]\,|\phi_r\rangle\,|q_{anc}(i,j)\rangle\,|1\rangle \\ & + \sum_{i=1}^{K-1}\sum_{j=0}^{i}[g_{ij}2^{l'_{ij}}\,|\,\langle\phi_r|\,\psi\rangle|^{x'_{ij}}\,|\,\langle\phi_i|\,\psi\rangle|^{y'_{ij}}\,|\,\langle\phi_r|\,\phi_i\rangle|^{z'_{ij}}]\,|\phi_i\rangle\,|q'_{anc}(i,j)\rangle\,|1\rangle\,.\end{aligned}$$

The next two terms, $-2\langle\phi_k|\chi_{k-1}\rangle|\phi_k\rangle$ and $2\langle\phi_r|\chi_{k-1}\rangle|\phi_k\rangle$ produce the same sigma terms while adding a power 1 to the functions $l_{i,j}$, $l'_{i,j}$, $x_{i,j}$, etc. Now by adding all this expression together, we will have:

$$\begin{aligned}|\chi_f\rangle &= \langle\phi_r|\psi\rangle|\phi_r\rangle|0\rangle^{\otimes K} + |\psi\rangle|1\rangle^{\otimes K} - \langle\phi_r|\psi\rangle|\phi_r\rangle|1\rangle^{\otimes K}\\ &+ \sum_{i=1}^{K}\sum_{j=0}^{i}[f_{ij}2^{l_{ij}}|\langle\phi_r|\psi\rangle|^{x_{ij}}|\langle\phi_i|\psi\rangle|^{y_{ij}}|\langle\phi_r|\phi_i\rangle|^{z_{ij}}]|\phi_r\rangle|q_{anc}(i,j)\rangle\\ &+ \sum_{i=1}^{K}\sum_{j=0}^{i}[g_{ij}2^{l'_{ij}}|\langle\phi_r|\psi\rangle|^{x'_{ij}}|\langle\phi_i|\psi\rangle|^{y'_{ij}}|\langle\phi_r|\phi_i\rangle|^{z'_{ij}}]|\phi_i\rangle|q'_{anc}(i,j)\rangle\,.\end{aligned}$$

And the theorem claim is correct by induction. □

E The quantum emulation algorithm with imperfect reflection gate

It has been discussed in [17] that there exists an algorithm for building an approximation of the gates with multiple copies of each sample. Although we do not discuss the optimality of this algorithm. The algorithm presented in [17] implements the reflection gates or controlled version of it using n copies of the unknown state with error $\mathcal{O}(\frac{1}{n})$. For this paper and to capture the effect of the noise and imperfection of the gates, we assume that the algorithm that constructs the controlled-reflection gate with m unknown copies of each sample, has an error $\epsilon(m)$ which will affect the output fidelity:

$$F_{approx}(\rho_{QE}, \mathrm{U}\rho\mathrm{U}^\dagger) \approx F_{ideal}(\rho_{QE}, \mathrm{U}\rho\mathrm{U}^\dagger) - \epsilon(m). \tag{13}$$

Although this is not a proper noise model, it will capture almost all of the imperfections of the circuit.

F Security proofs

F.1 Proof of Theorem 3: No-go result on the security of unitary primitives against $(\mathsf{qCI}, \mathsf{qEx}, \mu)$-PQEA

Proof. We show there is a QPT adversary $\mathcal{A}$ that wins the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ with non-negligible probability in terms of the security parameter λ. Let $\mathrm{U}_\mathcal{E}$ be the unitary transformation corresponding to $\mathcal{F}$. $\mathcal{A}$ runs the algorithm pictured in Figure 2. To show that $\mathcal{A}$ wins the game we need to show the test algorithm $\mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})$ returns 1 with non-negligible probability, where $|\psi^{out}\rangle = \mathrm{U}_\mathcal{E}|\psi\rangle = \mathrm{U}_\mathcal{E}|\phi_3\rangle$.

Relying on Theorem 2, the output state of Stage 1 of the QE algorithm is:

$$\begin{aligned}|\chi_f\rangle &= \langle\phi_r|\psi\rangle|\phi_r\rangle|0\rangle + |\psi\rangle|1\rangle - \langle\phi_r|\psi\rangle|\phi_r\rangle|1\rangle\\ &- 2\langle\phi_1|\psi\rangle|\phi_1\rangle|1\rangle + 2\langle\phi_r|\psi\rangle\langle\phi_r|\phi_1\rangle|\phi_1\rangle|1\rangle\,.\end{aligned}$$

Note that $\langle \phi_1 | \psi \rangle = \langle \phi_1 | \phi_3 \rangle = 0$ and we set $\langle \phi_r | \psi \rangle = \alpha$ and $\langle \phi_r | \phi_1 \rangle = \beta$ based on the choice of $|\phi_2\rangle$, the above equation can be simplified as:

$$|\chi_f\rangle = \alpha |\phi_r\rangle |0\rangle + |\psi\rangle |1\rangle - \alpha |\phi_r\rangle |1\rangle + 2\alpha\beta |\phi_1\rangle |1\rangle .$$

Now, according to Theorem 1, the final fidelity in terms of the success probability of Stage 1 can be obtained by calculating the density matrix of $|\chi_f\rangle$ and tracing out the ancillas:

$$P_{succ-stage1} = | \langle \phi_r | Tr_{anc}(|\chi_f\rangle \langle \chi_f|) |\phi_r\rangle |^2 = |\alpha^2(1 + 4\alpha^2\beta^2)|^2.$$

We have different choices for the reference state depending on the distinguishability parameter μ. For cases where the adversary is allowed to produce a new state with at least overlap half with all the states in the learning phase, by choosing the uniform superposition of the states where $\alpha = \beta = \frac{1}{\sqrt{2}}$, the output fidelity will be:

$$F(|\omega\rangle \langle \omega| , U^\dagger |\psi\rangle \langle \psi| U) \geq \sqrt{P_{succ-stage1}} = 1.$$

Thus $|\omega\rangle$ is completely indistinguishable from $\mathrm{U}_{\mathcal{E}} |\psi\rangle$ and the winning probability of $\mathcal{A}$ for any test according to Definition 2 is:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1} , |\omega\rangle^{\otimes\kappa_2})] = 1$$

which is the optimal choice of the reference. On the other hand, for the cases where the adversary is restricted to produce a challenge more than half distinguishable, we can still create a superposed state with $\alpha = \sqrt{1-\mu}$ and $\beta = \sqrt{\mu}$ and end up with the following fidelity of the emulation:

$$F(|\omega\rangle \langle \omega| , U^\dagger |\psi\rangle \langle \psi| U) \geq |\alpha^2(1+4\alpha^2\beta^2)| = |(1-\mu)(1+4\mu(1-\mu))| = \textit{non-negl}(\lambda).$$

Consequently, the probability of these states passing the test algorithm is also lower bounded by the fidelity, and is *non-negl*(λ). Recall that the security parameter λ includes the number of copies used in the test algorithm (κ_1 and κ_2), by increasing them the probability of accepting will converge to the above fidelity thus for any $\frac{1}{2} < \mu \leq 1 - \textit{non-negl}(\lambda)$:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1} , |\omega\rangle^{\otimes\kappa_2})] = \textit{non-negl}(\lambda). \quad \square$$

F.2 Proof of Theorem 4: No-go result on the security of unitary primitives against $(\mathsf{qUI}, \mathsf{qEx}, \mu)$-PQEA

Proof. Let $\mathcal{A}$ be the QPT adversary playing game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ and running the algorithm defined and explained in Figure 3. For $\mathcal{A}$ to win game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ we need to show that the test algorithm $\mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1} , |\omega\rangle^{\otimes\kappa_2})$ outputs 1 with non-negligible probability. In the qUI learning phase the states have not been chosen by the adversary and can be unknown quantum states. To provide the proof for the strongest case we assume that the states in S_{in}

are chosen at random and their classical description is unknown to $\mathcal{A}$. For this attack, $\mathcal{A}$ chooses $k_1 = 2$ and we assume at least 2 copies of each state exists. We use the fact that the success probability of the QE algorithm is non-negligible as long as there exists enough overlap between the reference quantum state and the given challenge quantum state. So, the adversary only needs to generate a suitable superposition of any two states randomly chosen from S_{in}. Although there is a no-go theorem implying the impossibility of building a precise target superposition of completely unknown quantum states [38], there are however some superposition algorithms such as [38, 39] that can be used for building a superposition of unknown quantum states with partial prior knowledge. By using the idea of these papers, we show a construction for preparing an unknown superposition of unknown quantum states that we denote Superpose$(\cdot, \cdot)$. The original circuit in [39] creates desired superposition of two unknown orthogonal qubit states. We modify it to the circuit shown in figure 6 and generalise it for n-qubit states. We simplify the notations of $|\phi_{1,1}\rangle$ and $|\phi_{2,1}\rangle$ as $|\phi_1\rangle$ and $|\phi_2\rangle$. Then for any unknown input states $|\phi_1\rangle$ and $|\phi_2\rangle$ randomly picked from S_{in}, the algorithm works as follows: The circuit creates the following superposition:

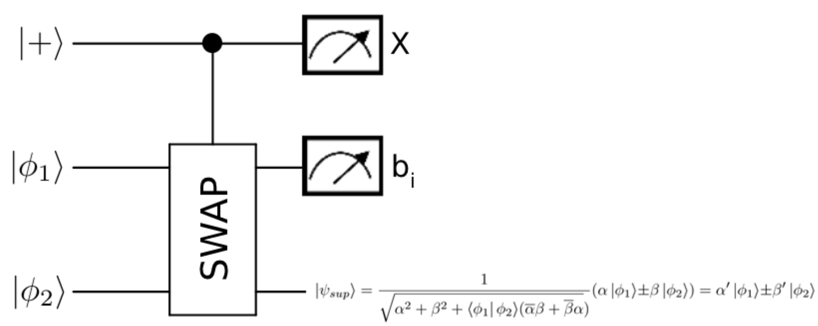

Fig. 6: The circuit for creating a superposition of two completely unknown states with unknown amplitudes. After the controlled-SWAP gate, the ancillary qubit will be measured in Pauli-X basis and the other n-qubit state is measured in one random computational basis of $\mathcal{H}^D$

$$|\psi_{sup}\rangle = \frac{1}{\sqrt{\alpha^2 + \beta^2 + \langle\phi_1|\,\phi_2\rangle(\overline{\alpha}\beta + \overline{\beta}\alpha)}}(\alpha\,|\phi_1\rangle \pm \beta\,|\phi_2\rangle) = \alpha'\,|\phi_1\rangle \pm \beta'\,|\phi_2\rangle \quad (14)$$

where α' and β' are unknown amplitudes depending on the overlap of $|\phi_1\rangle$ and $|\phi_2\rangle$ as well as the measurement basis and $\alpha'^2 + \beta'^2 = 1$. It can be seen that it is not possible to create a superposition that is more than (1/2)-distinguishable from both states $|\phi_1\rangle$ and $|\phi_2\rangle$ thus this attack holds if the acceptance threshold of a new state is at most $\frac{1}{2}$. On the other hand, if the overlap of the two states is not very small or the $\mu = 1-$non-$negl(\lambda)$. The probability that the following superposition collapses to a state with negligible overlap with one of the states, is negligible in the security parameter. Thus a non-trivial superposition of the output can be achieved by using the quantum emulation algorithm:

$$\begin{aligned}|\chi_f\rangle &= \langle\phi_r|\,\psi_{sup}\rangle\,|\phi_r\rangle\,|0\rangle + |\psi_{sup}\rangle\,|1\rangle - \langle\phi_r|\,\psi_{sup}\rangle\,|\phi_r\rangle\,|1\rangle \\ &- 2\,\langle\phi_{\bar{r}}|\,\psi_{sup}\rangle\,|\phi_{\bar{r}}\rangle\,|1\rangle + 2\,\langle\phi_{\bar{r}}|\,\psi_2\rangle\,\langle\phi_{\bar{r}}|\,\psi_{sup}\rangle\,|\phi_{\bar{r}}\rangle\,|1\rangle\,.\end{aligned}$$

To compute the fidelity of $|\psi^{out}\rangle$ wrt $|\omega\rangle = \mathrm{U}_{\mathcal{E}}\,|\psi\rangle$, first, we calculate the state of the QE algorithm after Stage 1. The reference state is picked at random between two quantum states. According to Theorem 1 we have:

$$\begin{aligned}|\chi_f\rangle &= \langle\phi_r|\,\psi_{sup}\rangle\,|\phi_r\rangle\,|0\rangle + |\psi_{sup}\rangle\,|1\rangle - \langle\phi_r|\,\psi_{sup}\rangle\,|\phi_r\rangle\,|1\rangle \\ &- 2\,\langle\phi_{\bar{r}}|\,\psi_{sup}\rangle\,|\phi_{\bar{r}}\rangle\,|1\rangle + 2\,\langle\phi_{\bar{r}}|\,\psi_2\rangle\,\langle\phi_{\bar{r}}|\,\psi_{sup}\rangle\,|\phi_{\bar{r}}\rangle\,|1\rangle\,.\end{aligned}$$

As the reference has been picked at random between $|\phi_1\rangle$ and $|\phi_2\rangle$, in the above equation $|\phi_r\rangle$ is one of the two states and $|\phi_{\bar{r}}\rangle$ is the other depending on the choice of $|\phi_r\rangle$. By calculating the reduced density matrix $|\chi_f\rangle\,\langle\chi_f|$ and tracing out the ancillary qubits, the success probability of Stage 1 of the QE algorithm is:

$$P_{succ-stage1} = |\,\langle\phi_r|\,Tr_{anc}(|\chi_f\rangle\,\langle\chi_f|)\,|\phi_r\rangle\,|^2 = |a^2 + 4c^2(ac-b)^2|^2$$

where $a = \langle\phi_r|\,\psi\rangle$, $b = \langle\psi|\,\phi_{\bar{r}}\rangle$ and $c = \langle\phi_r|\,\phi_{\bar{r}}\rangle = \langle\phi_1|\,\phi_2\rangle$. To obtain the average success probability based on the fidelity we need to calculate the average fidelity of these two cases based on the choice of the reference state that leads to

$$\begin{aligned}&F(|\omega\rangle\,\langle\omega|\,,{\mathrm{U}_{\mathcal{E}}}^{\dagger}\,|\psi\rangle\,\langle\psi|\,\mathrm{U}_{\mathcal{E}})_{avg} \geq \\ &\frac{\alpha'^2 + \beta'^2 + 4|\,\langle\phi_1|\,\phi_2\rangle|^2((\alpha'\,\langle\phi_1|\,\phi_2\rangle - \beta')^2 + (\beta'\,\langle\phi_1|\,\phi_2\rangle - \alpha')^2)}{2} = \\ &\frac{1}{2} + 2|\,\langle\phi_1|\,\phi_2\rangle|^2((\alpha'\,\langle\phi_1|\,\phi_2\rangle - \beta')^2 + (\beta'\,\langle\phi_1|\,\phi_2\rangle - \alpha')^2)\end{aligned}$$

As the second term is always positive, the minimum fidelity is $\frac{1}{2}$. So, according to Definition 2, the minimum value of the probability will be $\frac{1}{2}$ as the security parameter including κ_1, κ_2 increase:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}\,,|\omega\rangle^{\otimes\kappa_2})] \geq \frac{1}{2}$$

Thus, $\mathcal{A}$ wins the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qUI},\mathsf{qEx},\mu}(\lambda,\mathcal{A})$ with non-negligible probability. □

G ($\mathsf{qUI}, \mathsf{qEx}, \mu$) Quantum Emulation Attack with with imperfect controlled-reflection gates of the emulation algorithm

In this subsection, we show that the attack presented in Theorem 3 is successful even by assuming the imperfect controlled-reflection gate in the emulation algorithm. Here for simplicity we focus on cases where $0 < \mu \leq \frac{1}{2}$. The attack is the same as the proof in the Appendix F.1. We only discuss the effect of the imperfection of the gates on the fidelity.

Proof. Now to also capture the case with imperfect gates, we assume that the adversary can perform the gates with error $\epsilon(m)$, where m is the number of copies that exist for each sample state. As the input samples will be chosen by $\mathcal{A}$, the classical description is known and the gates build for input reflection are perfect. But the outputs will remain unknown in the general case thus we assume that the error will apply for output reflection gates. In the attack presented in Theorem3, there only exist two reflection gate thus even if the error of the reflection gates will add up, the fidelity will remain large enough to pass the test algorithm. Using equation 13, and by using the fact that there exists an algorithm where the error $\epsilon(m) \approx O(1/m)$, in the worst case, the fidelity of the imperfect attack will be

$$F_{approx}(|\omega\rangle\langle\omega|, U^{\dagger}|\psi\rangle\langle\psi|U) = 1 - \frac{2}{poly(\lambda)} = non\text{-}negl(\lambda)$$

So, according to Definition 2, as the security parameter including κ_1, κ_2 increase:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCl,qEx},\mu}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})] = non\text{-}negl(\lambda)$$

Thus, $\mathcal{A}$ wins the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCl,qEx},\mu}(\lambda, \mathcal{A})$ with non-negligible probability. □

G.1 Proof of Lemma 1

Proof. Any state $|\psi\rangle \in \mathcal{H}^D$ can be written in terms of the orthonormal basis of $\mathcal{H}^D$ denoted by $|b_i\rangle$, as follows:

$$|\psi\rangle = \sum_{i=0}^{D-1} \alpha_i |b_i\rangle \quad \text{with} \quad \sum_{i=0}^{D-1} |\alpha_i|^2 = 1$$

where α_i are complex coefficients. A projection into a smaller subspace consists of choosing d basis of $\mathcal{H}^D$ in the form of $\sum_{j=0}^{d-1} |b_j\rangle\langle b_j|$. Without loss of generality, we can assume that $D = md$ where m is an integer. This assumption is always correct for qubit spaces. This means that the larger Hilbert space can be divided into m smaller subspaces with dimension d. Let $\{|e_i\rangle\}_{i=0}^{d-1}$ be a subset of $\mathcal{H}^D$ which makes a complete set of basis for one of these subspaces. A projector will project $|\psi\rangle$ into one of these subspaces. As $|\psi\rangle$ has been picked at random and the subspaces are symmetric, the probability of falling into each subspace is equal and is equal to $\frac{1}{m}$ which is $\frac{d}{D}$. Otherwise either the sum of all probabilities would not be 1 or the assumption would not be valid. This shows that on average the probability of projecting a general state ψ is $\frac{d}{D}$. This can also be seen by the fact that the sum of all projectors in a complete set of projectors is equal to one. In this case, we have

$$\sum_{i=0}^{D-1} \Pi_i = \mathbb{I}$$

By sandwiching $|\psi\rangle$ on both sides we have:

$$\sum_{i=0}^{D-1} \langle\psi| \Pi_i |\psi\rangle = 1.$$

Clearly, each term $\langle\psi| \Pi_i |\psi\rangle$ is equal to $\sum_{j=0}^{d-1} |\langle\psi| d_j\rangle|^2$ where our projector is chosen as described above. This is equal to d number of $|\alpha_i|^2$ where each of them is in average $\frac{1}{2}$. Thus, the probability of being projected into one of the subspaces is $\frac{d}{D}$ otherwise the above equation or $\sum_{i=0}^{D-1} |\alpha_i|^2 = 1$ would be violated. □

G.2 Proof of Theorem 5

Proof. Let $\mathcal{A}$ be a quantum adversary playing the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCl,qSel}}(\lambda, \mathcal{A})$. Let the input and output database of the adversary after the qCl learning phase be S_{in} and S_{out}, both with size k_1, respectively. Also, Let $\mathcal{H}^d$ be the d-dimensional Hilbert space spanned by elements of S_{in} where $d \leq k_1$ and $\mathcal{H}^d_{out}$ be the Hilbert space spanned by elements of S_{out} with the same dimension. $\mathcal{A}$ receives an unknown quantum state $|\psi\rangle$ as a challenge in the qSel challenge phase and tries to output a state $|\omega\rangle$ as close as possible to $|\psi^{out}\rangle$. In other words, we calculate the average probability of $\mathcal{A}$'s output state $|\omega\rangle$ to have a fidelity larger or equal to δ. More generally we want to show that for any $\delta \neq 0$ the success probability will be negligible. Also, to be able to calculate the success probability of $\mathcal{A}$ in the average case, we need to clarify that the probability is over all the possible states of $|\psi\rangle$ that the Challenger picks as a challenge. Each state $|\psi\rangle$ is picked at random from a uniform distribution of states on $\mathcal{H}^D$ which asymptotically covers the whole Hilbert space uniformly. Thus we are interested in the following probability:

$$\underset{|\psi\rangle\in\mathcal{H}^D}{Pr}[F(\mathcal{A}(S_{in}, S_{out}, |\psi\rangle), U|\psi\rangle) \geq \delta] = \underset{|\psi\rangle\in\mathcal{H}^D}{Pr}[|\langle\omega| \psi^{out}\rangle|^2 \geq \delta].$$

Where $|\omega\rangle$ is the output state of the adversary and $|\psi^{out}\rangle = U|\psi\rangle$ is the correct output state. We denote the above target probability as $Pr_{success}$. According to the game definition, as the adversary selects states of the learning phase, the classical description of these states are usually known for him, but the received responses are unknown quantum states. Clearly, if the adversary receives also the classical description of the outputs, or the complete set of basis of $\mathcal{H}^d$ and $\mathcal{H}^d_{out}$, he will have a complete description of the map in the subspace and as a result have a greater success probability in general. more formally we have

$$Pr_{success}[\mathcal{A}(S_{in}, S_{out}, |\psi\rangle)] \leq Pr_{success}[\mathcal{A}(S_{in}, S^{classic}_{out}, \{|e^{in}_i\rangle, |e^{out}_i\rangle\}^d_{i=1}, |\psi\rangle)]$$

where $\{|e^{in}_i\rangle\}^d_{i=1}$ and $\{|e^{out}_i\rangle\}^d_{i=1}$ are set of orthonormal basis of the inputs and output subspaces and $S^{classic}_{out}$ denotes the set of output states with their classical description. Therefore from now on throughout the proof, we assume that the adversary has full knowledge of the subspace. Also, we mention that here using the entanglement will not enhance the adversary's knowledge on the subspace as by entangling its local system to the challenges of the learning phase, the reduced density matrix of the challenge/response entangled state will lie on the same subspace $\mathcal{H}^d$ and $\mathcal{H}^d_{out}$. Hereby upper-bounding our adversary with the adversary with full subspace knowledge we have included the entangled queries as well. Thus without loss of generality and to avoid complicated notations, we consider the adversary's state as a pure state $|\omega\rangle$. Now, we partition the set of all the challenges to two parts: the challenges that are completely orthogonal to $\mathcal{H}^d$ subspace, and the rest of the challenges that have non-zero overlap with $\mathcal{H}^d$. We denote the subspace of all the states orthogonal to $\mathcal{H}^d$ as $\mathcal{H}^{d^\perp}$. In other words, we will analyse the target probability $Pr_{success} = \underset{|\psi\rangle\in\mathcal{H}^D}{Pr}[|\langle\omega|\,\psi^{out}\rangle|^2 \geq \delta]$ in terms of the partial probabilities

$$\underset{|\psi\rangle\in\mathcal{H}^D, |\psi\rangle\in\mathcal{H}^{d^\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle|^2 \geq \delta] \text{ and } \underset{|\psi\rangle\in\mathcal{H}^D, |\psi\rangle\notin\mathcal{H}^{d^\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle|^2 \geq \delta].$$

Because the probability of $|\psi\rangle$ being in any particular subset is independent of the adversary's learning phase, the above probability can be written as:

$$Pr_{success} = \underset{|\psi\rangle\in\mathcal{H}^{d^\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle|^2 \geq \delta] \times Pr[|\psi\rangle \in \mathcal{H}^{d^\perp}] + \underset{|\psi\rangle\notin\mathcal{H}^{d^\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle|^2 \geq \delta] \times Pr[|\psi\rangle \notin \mathcal{H}^{d^\perp}]$$

where $Pr[|\psi\rangle \in \mathcal{H}^{d^\perp}] = 1 - Pr[|\psi\rangle \notin \mathcal{H}^{d^\perp}]$ denotes the probability of $|\psi\rangle$ that is picked uniformly at random from $\mathcal{H}^D$ being projected into the subspace of $\mathcal{H}^{d^\perp}$. From lemma 1, we know that this probability for any subspace, is equal to the ratio of the dimensions. As $\mathcal{H}^{d^\perp}$ is a $D-d$ dimensional subspace, $Pr[|\psi\rangle \in \mathcal{H}^{d^\perp}] = \frac{D-d}{D}$ and respectively $Pr[|\psi\rangle \notin \mathcal{H}^{d^\perp}] = \frac{d}{D}$. Also the probability is upper-bounded by the cases that the adversary can always win the game for $|\psi\rangle \notin \mathcal{H}^{d^\perp}$:

$$Pr_{success} \leq \underset{|\psi\rangle\in\mathcal{H}^{d^\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle|^2 \geq \delta] \times (\frac{D-d}{D}) + \frac{d}{D}$$

Finally, the only term of the target probability to be calculated is the first term. Any $|\psi\rangle \in \mathcal{H}^D$ can be written in any set of full basis of $\mathcal{H}^D$ as $|\psi\rangle = \sum_{i=1}^{D} c_i\,|e_i\rangle$. For any $|\psi\rangle \in \mathcal{H}^{d^\perp}$, the set of $\{|e_i\rangle\}_{i=1}^{D}$ can consist of the subspace basis $\{|e_i^{in}\rangle\}_{i=1}^{d}$ and the rest of the basis $\{|e'_i\rangle\}_{i=d+1}^{D}$ which are a set of basis for $\mathcal{H}^{d^\perp}$ and orthogonal to all the $|e_i^{in}\rangle$. Then $|\psi\rangle$ can be written as $|\psi\rangle = \sum_{i=1}^{d} c_i^{in}\,|e_i^{in}\rangle + \sum_{i=d+1}^{D} c'_i\,|e'_i\rangle$. Because $\langle\psi|\,e_i^{in}\rangle = 0$ then in this basis all the $c_i^{in} = 0$. Similarly for the output state $|\psi^{out}\rangle$, as the unitary preserves the inner product, $\langle e_i^{out}|\,\psi^{out}\rangle = \langle e_i^{in}|\,U^\dagger U\,|\psi\rangle = \langle e_i^{in}|\,\psi\rangle = 0$, and the correct output state can be written as $|\psi^{out}\rangle = \sum_{i=1}^{d} c_i^{out}\,|e_i^{out}\rangle + \sum_{i=d+1}^{D} \alpha_i\,|b_i\rangle$ where again all $c_i^{out} = 0$ and the $\{|b_i\rangle\}_{i=1}^{D-d}$ are a set of basis for $\mathcal{H}_{out}^{d^\perp}$. On $\mathcal{A}$'s side, any output of the algorithm can be written as

$$|\omega\rangle = \sum_{i=1}^{d} \beta_i\,|e_i^{out}\rangle + \sum_{i=d+1}^{D} \gamma_i\,|q_i\rangle$$

where the first sum represents part of the output state, that has been produced by $\mathcal{A}$ from the learnt output subspace and the second part has been produced in $\mathcal{H}_{out}^{d^{\perp}}$ with $\{|q_i\rangle\}_{i=1}^{D-d}$ being a set of bases for $\mathcal{H}_{out}^{d^{\perp}}$. Then because of the above argument, the fidelity of the first part is always zero as $\langle b_i| e_i^{out}\rangle = 0$. The normalization condition implies that $\sum_{i=1}^{d} |\beta_i|^2 + \sum_{i=d+1}^{D} |\gamma_i|^2 = 1$, thus for any state $|\omega\rangle$ that has a non-zero overlap with the learnt outputs, the fidelity with the correct state will decrease. So in order to make the $\mathcal{A}$'s strategies more optimal we assume that the state of all the adversaries are in the form of $\sum_{i=1}^{D-d} \gamma_i |q_i\rangle \in \mathcal{H}_{out}^{d^{\perp}}$ where the normalization condition is $\sum_{i=1}^{D-d} |\gamma_i|^2 = 1$. Now since there are infinite choices of basis orthogonal to $\{|e_i^{out}\rangle\}_{i=1}^{d}$, there is no way to uniquely choose or obtain the rest of the basis to complete the set. Also, another input of the adversary is the state $|\psi\rangle$ which according to the game definition, is an unknown state from a uniform distribution. As a result, the choice of the $|q_i\rangle$ basis are also independent of $|e'_i\rangle$ or $|b_i\rangle$. Thus knowing a matching pair of $(|q_i\rangle, |b_i\rangle)$ will increase the dimensionality of the known subspace by one. For each new challenge, $\mathcal{A}$ produces a state $|\omega\rangle = \sum_{i=1}^{D-d} \gamma_i |q_i\rangle$ with a totally independent choice of basis. Without loss of generality we can fix the basis $|q_i\rangle$ for different $|\omega\rangle$. To calculate the probability of interest, we calculate the fidelity averaging over all the possible choices of the input state. As the unitary transformation preserves the distance, it maps a uniform distribution of states to a uniform distribution and it leads to a uniform distribution of all the possible $|\psi^{out}\rangle$. Then the success probability can be written in terms of the outputs which is

$$\underset{|\psi\rangle\in\mathcal{H}^{d^{\perp}}}{Pr}[|\langle\omega|\psi^{out}\rangle|^2 \geq \delta] = \underset{|\psi^{out}\rangle\in\mathcal{H}_{out}^{d^{\perp}}}{Pr}[|\langle\omega|\psi^{out}\rangle|^2 \geq \delta].$$

Now, we show that for the adversary to win the game in the average case, $\mathcal{A}$ also needs to output $|\omega\rangle$ according to the uniform distribution. Assume that $\mathcal{A}$ outputs the states according to a probability distribution $\mathfrak{D}$ which is not uniform. Then, by repeating the experiment asymptotically many times, the correct response $|\psi^{out}\rangle$ will cover the whole $\mathcal{H}_{out}^{d^{\perp}}$ while $|\omega\rangle$ will cover a subspace of $\mathcal{H}_{out}^{d^{\perp}}$. This will decrease the average success probability of the adversary. So, generating the states $|\omega\rangle$ such that they span the whole $\mathcal{H}_{out}^{d^{\perp}}$, i.e. outputting them according to the uniform distribution, is required for the adversary to win the game. Using the above argument, and the fact that all the $|\omega\rangle$ are produced independently, we show that the average fidelity over all the $|\psi^{out}\rangle$ is equivalent to average fidelity over all the $|\omega\rangle$. There are different methods for calculating the average fidelity[42], but most commonly the average fidelity can be written as:

$$\int_{|\psi^{out}\rangle\in\mathcal{H}_{out}^{d^{\perp}}} |\langle\omega|\psi_x^{out}\rangle|^2 d\mu_x$$

where $d\mu$ is a measure based on which the reference state has been produced and parameterized. According to our uniformity assumption, the $d\mu$ here is the Haar measure. Note that $|\omega\rangle$ can be different for any new challenge. Now we rewrite the above average with the new parameters as:

$$\int_{|\psi^{out}\rangle \in \mathcal{H}_{out}^{d\perp}} |\langle \omega | \psi_x^{out} \rangle|^2 d\mu_x = \int_{|\psi^{out}\rangle \in \mathcal{H}_{out}^{d\perp}} |\sum_{i=1}^{D-d} \overline{\gamma_i} \langle q_i | \psi_x^{out} \rangle|^2 d\mu_x =$$

$$\int_{|\psi^{out}\rangle \in \mathcal{H}_{out}^{d\perp}} |\sum_{i=1}^{D-d} \overline{\gamma_{i_x}} \langle q_i | \psi^{out} \rangle|^2 d\mu_x = \int_{|\omega\rangle \in \mathcal{H}_{out}^{d\perp}} |\langle \omega_x | \psi^{out} \rangle|^2 d\mu_x$$

The above equality holds as the fidelity is a symmetric function of two states and also because the measure of integral for both cases is equal. Finally, we can use equality for averaging all the possible outputs for one $|\psi^{out}\rangle$. We want to calculate the probability of the average fidelity being greater than δ. To this end, we first calculate a more general probability that is the probability of the average fidelity to be non-zero. As we have

$$\underset{|\omega\rangle \in \mathcal{H}_{out}^{d\perp}}{Pr}[|\langle \omega | \psi^{out} \rangle|^2 \neq 0] + \underset{|\omega\rangle \in \mathcal{H}_{out}^{d\perp}}{Pr}[|\langle \omega | \psi^{out} \rangle|^2 = 0] = 1,$$

we calculate the probability of the zero fidelity for simplicity. We have:

$$\underset{|\omega\rangle \in \mathcal{H}_{out}^{d\perp}}{Pr}[|\langle \omega | \psi^{out} \rangle|^2 = 0] = Pr[(\int |\sum_{i=1}^{D-d} \overline{\gamma_{i_x}} \langle q_i | \psi^{out} \rangle|^2 d\mu_x) = 0] =$$

$$\underset{x}{Pr}[(\sum_{i,j=1}^{D-d} \overline{\gamma_{i_x}} \alpha_j \langle q_{i_x} | b_j \rangle)^2 = 0]$$

Now we use the Cauchy–Schwarz inequality to obtain the following inequality for probability

$$\underset{x}{Pr}[(\sum_{i,j=1}^{D-d} \overline{\gamma_{i_x}} \alpha_j \langle q_i | b_j \rangle)^2 = 0] \geq \underset{x}{Pr}[(\sum_{i,j=1}^{D-d} |\overline{\gamma_{i_x}} \alpha_j|^2 |\langle q_i | b_j \rangle|^2) = 0] =$$

$$\underset{x}{Pr}[(\sum_{i,j=1}^{D-d} |\overline{\gamma_{i_x}} \alpha_j|^2 |\langle q_i | b_j \rangle \langle b_j | q_i \rangle|) = 0] = \underset{x}{Pr}[(\sum_{i,j=1}^{D-d} |\overline{\gamma_{i_x}} \alpha_j|^2 |\langle q_i | \Pi_j | q_i \rangle |) = 0]$$

The last term is the probability of being projected into the orthogonal subspace averaging over all the projectors. We call again Lemma 1. As the subspace includes only one vector of the Hilbert space, the dimension of the orthogonal subspace is always one dimension less which here is equal to $D - d - 1$. Thus the last term of the probability is equal to:

$$\underset{x}{Pr}[(\sum_{i,j=1}^{D-d} |\overline{\gamma_{i_x}} \alpha_j|^2 |\langle q_i | \Pi_j | q_i \rangle |) = 0] = \frac{D-d-1}{D-d}.$$

Hence we showed that the probability of the average fidelity to be zero is greater than $\frac{D-d-1}{D-d}$ and consequently we have:

$$\underset{|\psi^{out}\rangle \in \mathcal{H}_{out}^{d\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle| \neq 0] \leq \frac{1}{D-d}$$

thus

$$\underset{|\psi\rangle \in \mathcal{H}^{d\perp}}{Pr}[|\langle\omega|\,\psi^{out}\rangle| \geq \delta] \leq \frac{1}{D-d}$$

for any non-zero δ. Substituting this into the original success probability formula we will have

$$Pr_{success} \leq \frac{1}{D-d} \times (\frac{D-d}{D}) + \frac{d}{D} = \frac{d+1}{D}$$

and the theorem has been proved. □

G.3 Proof of Theorem 6

Proof. The target probability is the probability of test algorithm outputting 1 (accept) for κ_1 copies of the correct output $|\psi^{out}\rangle = \mathrm{U}_{\mathcal{E}}\,|\psi\rangle$ and κ_2 copies of $|\omega\rangle$, $\mathcal{A}$'s output state, for the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda,\mathcal{A})$:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCI,qSel}}(\lambda,\mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})]$$

We can calculate the probability that the test algorithm outputs 1 conditioning on two cases for fidelity which is $Pr[F(|\omega\rangle, |\psi^{out}\rangle) \geq \delta]$ and $Pr[F(|\omega\rangle, |\psi^{out}\rangle) < \delta] = 1 - Pr[F(|\omega\rangle, |\psi^{out}\rangle) \geq \delta]$. We write the conditional probability denoting $Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})]$ as simply $Pr[1 \leftarrow \mathcal{T}]$ and $F(|\psi^{out}\rangle, |\omega\rangle)$ as simply F:

$$Pr[1 \leftarrow \mathcal{T}] = Pr[1 \leftarrow \mathcal{T}|F \geq \delta] \times Pr[F \geq \delta] + Pr[1 \leftarrow \mathcal{T}|F < \delta] \times Pr[F < \delta].$$

From Theorem 5 we have $Pr[F \geq \delta] \leq \frac{d+1}{D}$. Using this result and also setting the δ =non-$negl(\lambda)$ will lead to

$$\begin{aligned} Pr[1 \leftarrow \mathcal{T}] =& Pr[1 \leftarrow \mathcal{T}|F = \delta = \text{non-}negl(\lambda)] \times (\frac{d+1}{D}) \\ &+ Pr[1 \leftarrow \mathcal{T}|F = negl(\lambda)] \times (\frac{D-d-1}{D}) \end{aligned}$$

Also for any fidelity δ according to the Definition 2,
$\lim_{\kappa_1,\kappa_2\to\infty}(f(\kappa_1,\kappa_2,\delta)) = \delta$ and $\lim_{F\to 0}(f(\kappa_1,\kappa_2,F)) = Err(\kappa_1,\kappa_2)$. Thus:
$Pr[1 \leftarrow \mathcal{T}|F = \delta] \to \delta$ and $Pr[1 \leftarrow \mathcal{T}|F = negl(\lambda)] \to Err(\kappa_1,\kappa_2)$ and we will have:

$$Pr[1 \leftarrow \mathcal{T}] = \delta(\frac{d+1}{D}) + Err(\kappa_1,\kappa_2)(\frac{D-d-1}{D}).$$

Also, the input and output states of the target $\mathcal{F}$ be n-qubit states and $D = 2^n$. If $\mathcal{A}$ is a QPT adversary, the number of the learning query k_1 is $poly(n)$ and as a result the subspace dimension $d = poly(n)$. Then $\frac{d+1}{D} = \frac{poly(n)}{2^n} = negl(n)$ and $\lim_{n\to\infty} \frac{D-d-1}{D} = 1$. Consequently in the asymptotic limit of the security parameters n, κ_1, κ_2 we will have:

$$Pr[1 \leftarrow \mathcal{T}] = negl(n) + Err(\kappa_1, \kappa_2)$$

with the assumption that $Err(\kappa_1, \kappa_2) = negl(\kappa_1, \kappa_2)$, we conclude:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{qCl},\mathsf{qSel}}(\lambda, \mathcal{A})] = negl(\lambda)$$

and the proof is complete. □

G.4 Proof of Theorem 7

Proof. Let $\mathcal{A}$ be the QPT adversary playing the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCl},\mu}(\lambda, \mathcal{A})$ and running the algorithm described in Figure 4. $\mathcal{A}$ does not query the primitive $\mathcal{F}$ during the first learning phase. $\mathcal{A}$ receives an unknown challenge state $|\psi\rangle = \sum_{i=1}^{D} \alpha_i |b_i\rangle$ where $\{|b_i\rangle\}_{i=1}^{D}$ is a set of complete orthonormal bases for $\mathcal{H}^D$. Then, $\mathcal{A}$ prepares state $|0\rangle$ and performs a CNOT gate on the first qubit of the unknown challenge state and the ancillary qubit ($|0\rangle$) with the control qubit on the challenge state. We can assume the order of the bases is such that in the first half, the first qubit is $|0\rangle$ and in the second half the first qubit is $|1\rangle$. Then the output entangled state is

$$|\psi\rangle_{ca} = \sum_{i=1}^{D/2} \alpha_i |b_i\rangle_c \otimes |0\rangle_a + \sum_{i=\frac{D}{2}+1}^{D} \alpha_i |b_i\rangle_c \otimes |1\rangle_a$$

Now we can compute the final state of the two systems after the second qCl learning phase which is:

$$|\psi^{out}\rangle_{ca} = \sum_{i=1}^{D/2} \alpha_i (\mathrm{U}_{\mathcal{E}} \otimes \mathbb{I})(|b_i\rangle_c \otimes |0\rangle_a) + \sum_{i=\frac{D}{2}+1}^{D} \alpha_i (\mathrm{U}_{\mathcal{E}} \otimes \mathbb{I})(|b_i\rangle_c \otimes |1\rangle_a).$$

By rewriting the first qubit in the $|+\rangle$ basis we have

$$|\psi^{out}\rangle = [\mathrm{U}_{\mathcal{E}}(\sum_{i=1}^{D} \alpha_i |b_i\rangle_c)]\frac{|+\rangle}{\sqrt{2}} + [\mathrm{U}_{\mathcal{E}}(\sum_{i=1}^{D/2} \alpha_i |b_i\rangle_c - \sum_{i=\frac{D}{2}+1}^{D} \alpha_i |b_i\rangle_c)]\frac{|-\rangle}{\sqrt{2}}.$$

Then, the adversary measures his local qubit in the $\{|+\rangle, |-\rangle\}$ bases. If he obtains $|+\rangle$, the state collapses to $\mathrm{U}_{\mathcal{E}}(\sum_{i=1}^{D} \alpha_i |b_i\rangle_c) = \mathrm{U}_{\mathcal{E}} |\psi\rangle$ that is the desired state with fidelity 1. If the output of the measurement is $|-\rangle$, half of the terms have a minus sign. In this case, $\mathcal{A}$ applies a controlled-Z gate on the second half of the state to obtain again $\mathrm{U}_{\mathcal{E}} |\psi\rangle$. As a result, for any κ_1 and κ_2, we have:

$$Pr[1 \leftarrow \mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCl}}(\lambda, \mathcal{A})] = Pr[1 \leftarrow \mathcal{T}(|\psi^{out}\rangle^{\otimes \kappa_1}, |\omega\rangle^{\otimes \kappa_2})] = 1.$$

Now to complete the proof, we show that the μ-distinguishability is satisfied on average. We need to calculate the reduced density matrix of this state and compare it with the density matrix $\rho_\psi = |\psi\rangle\langle\psi|$ in terms of the Uhlmann's fidelity. The reduced density matrix of the challenge state can be calculated as follows:

$$\rho_c = Tr_a[|\psi\rangle\langle\psi|_{ca}] = \sum_{i=1}^{D} |\alpha_i|^2 |b_i\rangle\langle b_i| + \sum_{i=j=1}^{\frac{D}{2}} \sum_{j\neq i, j=\frac{D}{2}+1}^{D} \overline{\alpha_i}\alpha_j |b_i\rangle\langle b_j| + \sum_{i=\frac{D}{2}+1}^{D} \sum_{j\neq i, j=1}^{\frac{D}{2}} \overline{\alpha_i}\alpha_j |b_i\rangle\langle b_j|$$

where Tr_a denoted the partial trace taken over the adversary's sub-system. And the first sum shows the diagonal terms of the density matrix. As it can be seen these density matrices are different in half of the non-diagonal terms with the ρ_ψ. According to the Uhlmann's fidelity definition in the preliminary, and the fact that $|\psi\rangle$ is a pure state the fidelity reduce to:

$$F(\rho_\psi, \rho_c) = [Tr(\sqrt{\sqrt{\rho_\psi}\rho_c\sqrt{\rho_\psi}})]^2 = \langle\psi|\rho_c|\psi\rangle = \sum_{i=1}^{D} |\alpha_i|^2 \langle b_i|\rho_c|b_i\rangle .$$

By substituting the ρ_c from above, the result will be as follows:

$$F(\rho_\psi, \rho_c) = \sum_{i=1}^{D} |\alpha_i|^4 + \sum_{i=1}^{\frac{D}{2}} \sum_{j=\frac{D}{2}+1}^{D} 2|\alpha_i\alpha_j|^2 = 1 - \sum_{i=1}^{\frac{D(D-1)}{4}} 2|\gamma_i|^2$$

where $|\gamma_i|^2$ denoted the square of a quarter of the non-diagonal elements of ρ_ψ. This is a positive value and on average over all the state $|\psi\rangle$, non-negligible compared to the dimensionality of the state. Hence:

$$F(\rho_\psi, \rho_c) \leq 1 - \textit{non-negl}(\lambda)$$

and the distinguishability condition is satisfied and the proof is complete. □

G.5 Proof of Theorem 10: Impossibility of quantum unconditional unforgeability for UQPUFs

Proof. Let $\text{UQPUF}_{\textbf{id}}$ operate on n-qubit input-output pairs where $n = \log(D)$. In the qCI learning phase, $\mathcal{A}$ selects a complete set of orthonormal basis of $\mathcal{H}^D$ denoted as $\{|b_i\rangle\}_{i=1}^{2^n}$ and queries $\text{UQPUF}_{\textbf{id}}$ with each base 2^n times. So, the total number of queries in the qCI learning phase is $k_1 = 2^{2n}$.

Then, $\mathcal{A}$ runs a *unitary tomography* algorithm to extract the mathematical description of the unknown unitary transformation corresponding to the $\text{UQPUF}_{\textbf{id}}$, say $\text{U}_{\textbf{id}}$. It has been shown in [32] that the complexity of this algorithm is $\mathcal{O}(2^{2n})$ for n-qubit input-output pairs. This is feasible for an unbounded adversary. It is clear that once the mathematical description of the unitary is extracted, $\mathcal{A}$ can simply calculate the response of the unitary to a known challenge quantum state and wins the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ for any value of μ. So, we have:

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = 1.$$

More precisely, for generating the response of the unitary to an unknown challenge quantum state which is the case in the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel}}(\lambda, \mathcal{A})$, the unitary needs to be implemented. It is known that any unitary transformation over $\mathcal{H}^{2^n}$ requires $O(2^{2n})$ two-level unitary operations or $O(n^2 2^{2n})$ single qubit and CNOT gates [32] to be implemented. However, according to Solovay-Kitaev theorem [32], to implement a unitary with an accuracy ϵ using any circuit consisting of m single qubit and CNOT gates, $O(m \log^c(m/c))$ gates from the discrete set are required where c is a constant approximately equal to 2. Thus, an arbitrary unitary performing on n-qubit can be approximately implemented within an arbitrarily small distance ϵ using $O(n^2 4^n \log^c(n^2 4^n))$ gates.

So, $\mathcal{A}$ implements the unitary $\text{U}'_{\mathbf{id}}$ with error ϵ. Let $\mathcal{A}$ get the challenge state $|\psi\rangle$ in the qSel Challenge phase. The adversary queries $\text{U}'_{\mathbf{id}}$ with $|\psi\rangle$ and gets $|\omega\rangle = \text{U}'_{\mathbf{id}} |\psi\rangle$ as output. Since the ϵ can be arbitrary small, then $F(\text{U}_{\mathbf{id}} |\psi\rangle, \text{U}'_{\mathbf{id}} |\psi\rangle) \geq 1 - negl(\lambda)$. So, $\mathcal{A}$'s output $|\omega\rangle$ passes any test algorithm $\mathcal{T}(|\psi^{out}\rangle^{\otimes\kappa_1}, |\omega\rangle^{\otimes\kappa_2})$ with probability close to 1. Again, an unbounded adversary wins the game $\mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel},\mu}(\lambda, \mathcal{A})$ with probability 1:

$$Pr[1 \leftarrow \mathcal{G}^{\text{UQPUF}}_{\mathsf{qCI},\mathsf{qSel},\mu}(\lambda, \mathcal{A})] = 1 \qquad \square$$