Unforgeability in the quantum world

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: Unforgeability, Quantum Security, Quantum Cryptography, Quantum Cryptanalysis

Abstract. Defining unforgeability and designing cryptographic primitives that provide unforgeability in the quantum setting, *i.e.* where the adversary has quantum capabilities including quantum oracle access to the primitive, has proven to be a hard challenge. The classical notions and techniques do not transpose directly to the quantum setting. In this paper, we continue the line of work initiated by Boneh and Zhandry at CRYPTO 2013 and EUROCRYPT 2013 in which they formally define the notion of unforgeability against quantum adversaries specifically for Message Authentication Codes and Digital Signatures schemes. We develop a general and parameterized quantum game-based security framework for both classical and quantum primitives modelled by unitary transformations. We provide general possibility and impossibility results for such primitives. In particular, we show that no unitary primitive can provide existential unforgeability against quantum adversaries. Our main impossibility result relies on a new and generic quantum attack. We demonstrate this attack both on classical and quantum primitives to show its applicability as well as the completeness/integrity of our definitions of security. On the other hand, we show that selective unforgeability is satisfied by a specific class of unitaries that we term unknown unitaries.

1 Introduction

Recent advances in quantum technologies threaten the security of many widely-deployed cryptographic primitives. This calls for quantum-secure cryptographic schemes. In this context, cryptographers have considered two main security models when analysing the security of cryptographic primitives against quantum adversaries [1–7]: 1) the *standard security* model, often also termed post-quantum security, where the adversary only has classical access to the primitive but can locally perform quantum computations; or 2) the *quantum security* model where the adversary has further quantum access to the primitive, *i.e.* he can issue quantum queries. In the standard model, formal definitions of security are directly adapted from the classical ones. But, in the quantum model, because of the quantum nature of interaction with the primitives, a broader range of attack scenarios emerge making the task of transposing security definitions to the quantum setting highly non-trivial and subtle. For instance when considering a keyed classical function $O_k : \{0,1\}^n \rightarrow \{0,1\}^m$ with the key k in the quantum model, the adversary is given access to the unitary operator $U_{O_k} : \; |x\rangle |y\rangle \rightarrow |x\rangle |y \oplus O_k(x)\rangle$ [2–6].

One of the key elements of quantum models is the fact that the adversary can query the oracle with quantum states in superposition. Superposition queries are more likely to leak sensitive information to the adversary and lead to non-trivial attacks that are not possible in the classical regime. Another important aspect is that having access to the input-output pairs of the oracle in the form of quantum states enables the adversary to run quantum algorithms and take advantage of quantum speedup. Of course, a possible countermeasure against *superposition attacks* is to forbid any kind of quantum access to the oracle through measurements. However, in such a setting the security relies on the physical implementation of the measurement tool which itself could be potentially exploited by a quantum adversary. Thus, and as it has previously been advocated in [2–4, 6], providing security guarantees in the quantum security model is crucial.

In this paper, we pursue the line of work initiated by Boneh and Zhandry in [2, 3] as well as Alagic *et al.* in [6] on formalizing the notion of unforgeability in the quantum security model. This notion is the security property desired for many primitives such as Message Authentication Codes, Digital Signatures, or Physical Unclonable Function schemes. Informally, unforgeability ensures that the adversary cannot produce valid input-output pairs of the oracle without access to the full description of its circuit.

1.1 Our Contributions

Definition of Quantum Unforgeability We propose a general and unified definition of quantum unforgeability for both classical and quantum unitary cryptographic primitives. We present our definitions in the quantum-game based framework in the spirit of [3, 8, 9]. In particular, previous definitions [2, 3, 6] do not apply to quantum primitives such as quantum readouts of PUFs for instance. Furthermore, our definitions precisely capture the quantum capabilities of the adversary in terms of overlap between the challenge and the queried states in the learning phase. This formalizes the full spectrum of unforgeability from classical to fully quantum, revealing new attacks that previous definitions do not capture.

Informally speaking, we define μ-existential unforgeability as follows: *The unitary primitive $\mathcal{F}$ satisfies μ-existential unforgeability, if the success probability of any QPT adversary to output a "new" μ-distinguishable input-output pair is negligible in the security parameter.* The notion of μ-distinguishability captures the overlap of the challenge with the learning phase, and allows characterising "new" challenges in a fine-grained manner. This contrasts with previous definitions of unforgeability which characterise "new" challenges through counting the queries in the learning phase and the challenge phase. Such definitions essentially encode the equality testing algorithm in the counts, but as we show with our case studies, also misses some forgery attacks.

As said above, our definition also caters for the standard security model. Note that 1-existential unforgeability corresponds to the setting where the adversary has to output a valid input-out pair that is orthogonal to all the queries in the learning phase. This captures the standard definition for existential unforgeability where the adversary only has classical access to the primitive.

Quantum Emulation Attacks We define a new class of quantum attacks, termed Quantum Emulation (QE) attack, that covers all adversarial strategies for forgeries against both classical and quantum primitives. Inspired by the universal quantum emulator algorithm introduced by Marvian and Lloyd in [10], we devise concrete QE attacks against unforgeability of unitary primitives. This algorithm was developed and proposed in the context of quantum process tomography, thus the analysis did not consider any adversarial behaviour. In particular, revealing successful QE attacks relying on this universal quantum emulator algorithm required the full parameter estimation of the algorithm (as we provide in Section 3). We further design an adaptive QE attack by exploiting entanglement. These two attacks are novel and not captured by previous definitions of quantum unforgeability.

(Im)possibility Results We then question the possibility of unitary cryptographic primitives satisfying quantum unforgeability. And we establish the main following general results.

We show that quantum existential unforgeability is a strong notion to defend against quantum static adversaries and that selective quantum unforgeability is also difficult to protect against adaptive adversaries.

Theorem 1 (informal). *No unitary primitive $\mathcal{F}$ satisfies μ-existential quantum unforgeability against static adversaries with chosen input access to $\mathcal{F}$ for any $\mu \leq 1 - non\text{-}negl(\lambda)$.*

Theorem 2 (informal). *No unitary primitive $\mathcal{F}$ satisfies μ-selective quantum unforgeability against adaptive adversaries with chosen input access to $\mathcal{F}$ for any $\mu \leq 1 - non\text{-}negl(\lambda)$.*

On the other hand, we prove a weaker yet realistic setting i.e. selective quantum unforgeability for unknow unitaries could be achieved. More precisely, we define the notion of a family of unknown unitaries, as a family of unitaries such that a unitary randomly picked from such family can only be *learned* through queries, and establish that this is sufficient for achieving selective quantum unforgeability.

Theorem 3 (informal). *Any unknown unitary cryptographic primitive $\mathcal{F}$ satisfies selective quantum unforgeability against static adversaries with chosen input access to $\mathcal{F}$.*

Case Studies Finally, we turn to case studies to show the generality of our previous theoretical investigations. We show how our quantum game-based framework provides a unified definition for analysing the quantum security of both classical and quantum cryptographic primitives. We also present how our proposed quantum emulation attack technique reveals new vulnerabilities for both quantum and classical primitives with quantum oracle access in the learning phase that was not covered in previous works.

Message Authentication Codes (MACs) We show that common MAC constructions such as HMAC, PMAC, and NMAC [11] do not satisfy quantum existential unforgeability. This allows us to concretely show the limitations of previous definitions of quantum unforgeability [2, 3].

Physical Unclonable Functions (PUFs) We show that the existential unforgeability notion is too strong but that selective unforgeability can be achieved.

Symmetric Encryption The relevant security property for such primitives is indistinguishability. We show that our new quantum emulation attack is not limited to the notion of quantum unforgeability, by turning it to an attack on Symmetric Encryption with quantum oracle access.

1.2 Other Related Work

Quantum Unforgeability Another proposal for defining quantum unforgeability has been provided by Alagic *et al.* in [6]. To address some of the limitations of the Boneh-Zhandry definition, they defined the notion of *blind unforgeability* where the quantum adversary has restricted access to the domain of the signing algorithm. More precisely, the adversary is not allowed to query the oracle on a random subset of the domain termed the blinded subset. The goal of the adversary is then to forge a valid input-output pair inside the blinded subset. Their definition is provided for the one-time query scenarios, but its generalization to q-time forgery again relies on counting queries and thus suffers from the same issues as the Boneh-Zhandry definition as we demonstrate in this paper.

Quantum Adversarial Algorithms Several algorithms have been proven to break the security of common cryptographic primitives. Famously, Shor's algorithm is known to break asymmetric primitives whose security rely on the hardness of factoring and of the discrete logarithm problem [12]. And Grover's [13] and Simon's [14] algorithms are known to provide quantum speedup in key recovery and collision attacks on symmetric crypto-systems [4, 15, 16]. As we show, Quantum Emulation is yet another class of such quantum algorithms that break the security of symmetric crypto-systems. It relies on very different properties of quantum systems and can be used to target different types of attacks.

Security in the quantum model In recent years, security against adversaries with quantum capabilities has drawn a lot of attention and generated a plethora works providing formal models and theoretical results. Several cryptographic primitives have been studied in the quantum security model as defined above. Most of these works [3, 5, 7, 17, 18] focus on indistinguishability type properties, which we have not fully explored beyond our case study in Section 5.3.

2 Preliminaries

We use the term quantum bit or qubit [19] 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 are 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* [20]. 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 $0 \leq \mu, \nu \leq 1$. We say two quantum states ρ and σ are μ-distinguishable if $0 \leq F(\rho, \sigma) \leq 1 - \mu$ and ν-indistinguishable if $\nu \leq F(\rho, \sigma) \leq 1$.*

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 [21]. The SWAP test and its generalisation to multiple copies introduced recently in [22] 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

$$\parallel Q_E - Q_F \parallel_\diamond \equiv max_\rho(\parallel (Q_E \otimes \mathbb{I})[\rho] - (Q_F \otimes \mathbb{I})[\rho]) \parallel_1 .$$

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

3 Quantum Emulation Algorithm

In this section, we describe the Quantum Emulation (QE) algorithm presented in [10] as a quantum process learning tool that can outperform the existing approaches based on quantum tomography [23]. 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. The original algorithm was developed and proposed in the context of quantum process tomography, thus the analysis did not consider any adversarial behaviour. For our cryptanalysis purposes, we need to provide a new fidelity analysis for challenges not fully lying within the subspace of the learning phase. We further optimise the success probability of our attack by optimising the choice of the reference state.

3.1 The Circuit and Description

The circuit of the quantum emulation algorithm has been depicted in Figure 5 in Appendix C also in [10] and works as follows. The quantum emulation algorithm introduced in [10] 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:

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

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 the 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 wherein each block, the following gates run on the state of the system and an ancilla:

$$W(i) = R_c(|\phi_i\rangle) H R_c(|\phi_r\rangle). \tag{2}$$

According to equation (2), 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:

$$[W(K)...W(1)] |\psi\rangle |-\rangle^{\otimes K} \approx |\phi_r\rangle |\Omega(anc)\rangle \tag{3}$$

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 the 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:

$$(\mathrm{SWAP} \otimes I^{\otimes K}) |\phi_r^{out}\rangle |\phi_r\rangle |\Omega(anc)\rangle = |\phi_r\rangle |\phi_r^{out}\rangle |\Omega(anc)\rangle . \tag{4}$$

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:

$$W^{out}(i) = R_c(|\phi_i^{out}\rangle) H R_c(|\phi_r^{out}\rangle) = (\mathrm{U} \otimes I) W(i) (\mathrm{U}^\dagger \otimes \mathbb{I}). \tag{5}$$

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:

$$[W^{out}(1)...W^{out}(K)] |\phi_r^{out}\rangle |\Omega(anc)\rangle \approx \mathrm{U}|\psi\rangle |-\rangle^{\otimes K} \tag{6}$$

where equality is obtained whenever the success probability of Stage 2 is equal to 1.

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 [10, 24]. We recall the main theorem of [10]:

Theorem 4. *[10] 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:*

$$F(\rho_{QE}, \mathrm{U}\rho\mathrm{U}^\dagger) \geq F(\mathcal{E}_\mathrm{U}(\rho), \mathrm{U}\rho\mathrm{U}^\dagger) \geq \sqrt{P_{succ-stage1}} \tag{7}$$

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 [10]. From the proof of Theorem 4 in [10], it can be seen that the success probability of Stage 1 is calculated as follows:

$$P_{succ-stage1} = |\langle\phi_r| Tr_{anc}(|\chi_f\rangle\langle\chi_f|) |\phi_r\rangle|^2 \tag{8}$$

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 4, we only use equation (8) for our analysis henceforward.

The fidelity of the output state of the circuit highly depends on the choice of the reference state (equation (8)) 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:

$$\left|\chi_i\right\rangle = \frac{1}{2}[(I - R(\phi_r))\left|\chi_{i-1}\right\rangle\left|0\right\rangle + R(\phi_i)(\mathbb{I} + R(\phi_r))\left|\chi_{i-1}\right\rangle\left|1\right\rangle]. \tag{9}$$

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

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

$$\begin{aligned}\left|\chi_f\right\rangle = {} & \left\langle\phi_r\middle|\psi\right\rangle\left|\phi_r\right\rangle\left|0\right\rangle^{\otimes K} + \left|\psi\right\rangle\left|1\right\rangle^{\otimes K} - \left\langle\phi_r\middle|\psi\right\rangle\left|\phi_r\right\rangle\left|1\right\rangle^{\otimes K} \\ & + \sum_{i=1}^{K}\sum_{j=0}^{i}[f_{ij}2^{l_{ij}}|\left\langle\phi_r\middle|\psi\right\rangle|^{x_{ij}}|\left\langle\phi_i\middle|\psi\right\rangle|^{y_{ij}}|\left\langle\phi_r\middle|\phi_i\right\rangle|^{z_{ij}}]\left|\phi_r\right\rangle\left|q_{anc}(i,j)\right\rangle \\ & + \sum_{i=1}^{K}\sum_{j=0}^{i}[g_{ij}2^{l'_{ij}}|\left\langle\phi_r\middle|\psi\right\rangle|^{x'_{ij}}|\left\langle\phi_i\middle|\psi\right\rangle|^{y'_{ij}}|\left\langle\phi_r\middle|\phi_i\right\rangle|^{z'_{ij}}]\left|\phi_i\right\rangle\left|q'_{anc}(i,j)\right\rangle\end{aligned} \tag{10}$$

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 $\left|0\right\rangle^{\otimes K}$).

Proof. The proof is by induction and the details is found in Appendix D. □

Having a precise expression for $\left|\chi_f\right\rangle$ from Theorem 5, one can calculate $P_{succ-step1}$ of equation (8) by tracing out all the ancillary systems from the density matrix of $\left|\chi_f\right\rangle\left\langle\chi_f\right|$. Also, now it is clear that if $\left|\psi\right\rangle$ is orthogonal to the $\mathcal{H}^d$, the only term remaining in equation (10) is $\left|\psi\right\rangle\left|1\right\rangle^{\otimes K}$. 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 [10]. Hence, Stage 1 is successful with probability close to 1 by assuming the gates to be ideal.

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 [25–27]. 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 physical unclonable functions. [3, 5, 9, 28, 29]. 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 [3, 8, 9, 29]. 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 (qCl or qUl) 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 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 [9].

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[c] $|\psi\rangle \notin_\mu S_{in}$ 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.

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.[3]

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 we allow them in the attacks.

In comparison to quantum unforgeability definitions found in [25, 29] 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 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*

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

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.

From this generalised definition, we can formally define *Existential* and *Selective Unforgeability* of unitary primitives as instances of our game as follows:

Definition 4 (Quantum μ-Existential Unforgeability). *A unitary cryptographic primitive $\mathcal{F}$, with unitary transformation $\mathrm{U}_{\mathcal{E}}$ provides μ-quantum existential unforgeability if the success probability of any QPT adversary $\mathcal{A}$ of winning the game $\mathcal{G}^{\mathrm{U}_{\mathcal{E}}}_{\mathsf{qCl},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ is negligible in λ,*

$$Pr[1 \leftarrow \mathcal{G}^{\mathrm{U}_{\mathcal{E}}}_{\mathsf{qCl},\mathsf{qEx},\mu}(\lambda, \mathcal{A})] = negl(\lambda).$$

Definition 5 (Quantum Selective Unforgeability). *A unitary cryptographic primitive $\mathcal{F}$, with unitary transformation $\mathrm{U}_{\mathcal{E}}$ is quantum selectively unforgeable if the success probability of any QPT adversary $\mathcal{A}$ of winning the game $\mathcal{G}^{\mathrm{U}_{\mathcal{E}}}_{\mathsf{qCl},\mathsf{qSel}}(\lambda, \mathcal{A})$ is negligible in the security parameter λ,*

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

We are now ready to present several general results on the security of unitary primitives. First we prove a general impossibility result by presenting a successful existential PQEA against any such primitive.

Theorem 6 (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 E.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 (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 E.1:

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

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

First learning phase:

```
choose $|\phi_1\rangle$
query $|\phi_1\rangle$ and receive $|\phi_1^{out}\rangle$
if $0 \leq \mu \leq \frac{1}{2}$:
    choose $|\phi_2\rangle = \frac{1}{\sqrt{2}}(|\phi_1\rangle + |\phi_3\rangle)$
else if $\frac{1}{2} \leq \mu \leq 1-$non-$negl(\lambda)$:
choose $|\phi_2\rangle = \sqrt{\mu}\,|\phi_1\rangle + \sqrt{1-\mu}\,|\phi_3\rangle$
query $|\phi_2\rangle$ and receive $|\phi_2^{out}\rangle$
set $S_{in} = \{|\phi_1\rangle, |\phi_2\rangle\}$
```

set $S_{out} = \{|\phi_1^{out}\rangle, |\phi_2^{out}\rangle\}$
Without loss of the generality, we assume $\mathcal{A}$ chooses one of the computational basis of $\mathcal{H}^D$ as $|\phi_1\rangle$. Then, $\mathcal{A}$ chooses an orthogonal state to $|\phi_1\rangle$ as $|\phi_3\rangle$ and sets $|\phi_2\rangle$ the superposition of these two states.

Challenge phase:

```
set the challenge $|\psi\rangle \leftarrow |\phi_3\rangle$
```

$|\phi_3\rangle$ satisfies condition $|\phi_3\rangle \notin_\mu S_{in}$.

Guess phase:

```
set the reference $|\phi_r\rangle \leftarrow |\phi_2\rangle$
$|\omega\rangle \leftarrow QE(|\psi\rangle, |\phi_r\rangle, S_{in}, S_{out})$
output $|\omega\rangle$
```

$QE(|\psi\rangle, S_{in}, S_{out}, |\phi_r\rangle)$ is the quantum emulation algorithm.

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

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 [30].

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

Theorem 7 (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.

```
$(\mathsf{qUI}, \mathsf{qEx}, \mu)$-PQEA

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

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

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

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 E.2. The main idea of the attack is similar to the proof of Theorem 6. But since the learning phase states are not being chosen by the adversary, to win the game the adversary needs to create a superposition of any two unknown states randomly chosen from S_{in}. But as mentioned before, there is no algorithm to create such exact target superpositions. Although inspired by probabilistic algorithms such as [30, 31], 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.

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

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

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 [19]. 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 formalized in Definition 6, 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 [32, 33]. 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 6 (Unknown Unitary Transformation). *We say that a family of unitary transformations $\mathcal{U}$, over a D-dimensional Hilbert space $\mathcal{H}^D$ is a family of Unkown Unitaries, if for all QPT adversaries $\mathcal{A}$ the following holds:*

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

An example of such family of unknown unitary matrices is the Haar measure family of unitaries or t-designs [34, 35] that are the quantum analogue of t-wise independent distributions [7]. Also the definition can capture keyed family of unitaries $\mathcal{U} = \{U_k\}_{k\in\mathcal{K}}$ if they satisfy the above pre-query condition. In general, we state the above definition as a sufficient condition for satisfying the notion of selective unforgeability against any quantum adversary. This condition could be enforced through other construction or assumptions for families of unitaries used in different cryptographic constructions to allow application of our general results for such setting. In the rest of the paper, we will say that a unitary is unknown if it is sampled at random from a family of unknown unitaries $\mathcal{U}$ and all probabilistic statements for such unitaries are over $\mathcal{U}$.

In Theorem 9, 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 the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCl,qSel}}(\lambda, \mathcal{A})$. But to prove Theorem 9 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 E.3. 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 possibility 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 7 ($\mathcal{T}_\delta^{ideal}$ 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 8. *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 7, for any non-zero δ, the success probability of any adversary $\mathcal{A}$ in the game $\mathcal{G}_{\mathsf{qCI,qSel}}^{\mathcal{F}}(\lambda, \mathcal{A})$ is bounded as follows:*

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

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 $\mathcal{G}_{\mathsf{qCI,qSel}}^{\mathcal{F}}(\lambda, \mathcal{A})$.

Proof. The complete proof can be found in Appendix E.4, here we only sketch the main idea. We are interested in the average success probability of the adversary playing the game $\mathcal{G}_{\mathsf{qCI,qSel}}^{\mathcal{F}}(\lambda, \mathcal{A})$ 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 $Pr_{success}$ 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:

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

which results to the final target probability being:

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

Theorem 9. *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 E.5, the main idea 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 8), 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 obtain:

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

Corollary 1. *Any unknown 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 9 as the adversary is weaker and does not even choose the input quantum states in the learning phase. □

Theorem 8 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 10 (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-negl(λ)*

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

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 E.6. 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 the appropriate measurement given in the proof (Appendix E.6), $\mathcal{A}$ can extract the $U|\psi\rangle$ from the entangled state. Also, we have shown that the μ-distinguishability is satisfied on average, over all the

```
$(\mathsf{null}, \mathsf{qSel}, \mathsf{qCI}, \mu)$-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.
        $\rho_c$ 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 $\mathcal{G}^{\mathcal{F}}_{\mathsf{null},\mathsf{qSel},\mathsf{qCI},\mu}$

possible 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 Case Study

In this section, we demonstrate the generality of our work through several case studies. In particular, we show how our quantum game-based framework provides unified definitions for analysing the quantum security of both classical and quantum cryptographic primitives. Furthermore, we present how our proposed quantum emulation attack technique reveals new vulnerabilities for both quantum and classical primitives with quantum oracle access in the learning phase that was not covered in previous works.

First, we study quantum unforgeability of Message Authentication Codes (MACs) schemes as an example of classical primitive with quantum oracle access. We show that common MAC constructions such as HMAC, PMAC, and NMAC do not satisfy quantum existential unforgeability. We further compare our definitions of unforgeability to previous ones by Boneh and Zhandry [2, 3]. We provide a different concrete forgery attack on common MAC schemes that clearly shows the limitations of these previous definitions.

In our second example, we investigate the security of Physical Unclonable Functions (PUFs) with quantum oracle access or quantum readouts of PUFs as termed in [36]. This primitive illustrates our definitions and results on quantum primitives as their evaluation can be naturally modelled by a unitary transformation and their relevant security property is unforgeability. We show that the existential unforgeability notion is too strong to be satisfied by these schemes under quantum attacks. On the other hand, we establish that selective unforgeability can be achieved.

Finally, we show that our new quantum emulation attack is not limited to the notion of quantum unforgeability, but can also be applied to indistinguishability-type properties such as the security of encryption schemes with quantum oracle access. We demonstrate a concrete attack based on the emulation technique that breaks the quantum and classical indistinguishability of symmetric encryption with a quantum oracle access.

5.1 Quantum Existential Unforgeability of Message Authentication Codes (MACs)

A MAC system consists of a keyed MAC signing algorithm and a verification algorithm. The signing algorithm $S(k, m) : \mathcal{K} \times \mathcal{M} \rightarrow \mathcal{T}$ produces a tag t using secret key k from $\mathcal{K}$. The verification algorithm $V(k, t, m)$ verifies whether the message-tag pair (m, t) is valid under k. The security notion relevant to MACs is unforgeability. That is, no adversary should be able to produce a new valid message-tag pair without knowing the secret key k, *i.e.* solely through polynomial access to the signing oracle. In the standard security model, the adversary only has classical access to the oracle in the learning phase; while in the quantum security model quantum access to the signing oracle has been granted to the adversary and the quantum signing oracle is modelled by a unitary transformation as follows [2]:

$$U_{MAC} : \sum_{m,x,y} \alpha_{m,x,y} |m, x, y\rangle \rightarrow \sum_{m,x,y} \alpha_{m,x,y} |m, x \oplus S(k, m), y\rangle$$

where $m \in \mathcal{M}$ is a classical message and x, y are ancillary states making the oracle transformation a unitary matrix.

The following corollary of Theorem 6 establishes that most MAC schemes[4] are not μ-Existentially Unforgeable under quantum chosen message attacks for any $0 \leq \mu \leq 1 - \textit{non-negl}(N)$ where N is the security parameter:

Corollary 2. *MAC schemes with unitary quantum signing oracle U_{MAC} of dimension $D = 2^N$ do not satisfy μ-existential unforgeability for any $0 < \mu \leq 1 - \textit{non-negl}(N)$. That is, there exists a QPT adversary $\mathcal{A}$ that wins the game $\mathcal{G}^{U_{MAC}}_{\mathsf{qCl},\mathsf{qEx},\mu}(\lambda, \mathcal{A})$ with non-negligible probability.*

Comparison with previous work Definition of existential unforgeability under quantum chosen-message attack (EUF-qCMA) as presented in [2] does not capture our quantum emulation attack. The main difference resides in the way they enforce the challenge to be "new" with respect to the queries of the learning phase. Following the classical security model definitions, they restrict the game to new challenges by imposing that if q queries have been issued to the quantum signing oracle during the learning phase, then the adversary should be able to produce in the guess phase $q + 1$ distinct valid message-tag pairs to win the security game.

Definition 8 (EUF-qCMA [2]). *A MAC system is existentially unforgeable under a quantum chosen message attack (EUF-qCMA) if no adversary after issuing q quantum chosen message queries, can generate $q + 1$ valid classical message-tag pairs with non-negligible probability in the security parameter.*

The reason for characterising "new challenges" through the counting of queries and message-tags in the challenge phase is to avoid trivial guessing attacks through measuring. We argue however that this definition does not capture all attacks. In particular, we show below an execution that this definition does not characterise as an attack, but which is not the trivial guessing attack and whose success probability is 1.

Existential Forgery Attack Let the MAC signing oracle U_{MAC} be defined over a $D = 2^N$-dimensional Hilbert space $\mathcal{H}^D$. We consider the following adversary:

[4] Any MAC scheme that can be modelled by a unitary matrix. This includes most common constructions such HMAC, PMAC and NMAC.

Learning phase $\mathcal{A}$ issues the following queries to the signing oracle:

$$|\phi_1\rangle = |m_1, 0, 0\rangle \quad |\phi_r\rangle = \sqrt{1 - 2|\mu|^2}\, |m_1, 0, 0\rangle + \mu\, |m_2, 0, 0\rangle + \mu\, |m_3, 0, 0\rangle$$

where m_1, m_2 and m_3 are any classical messages. $\mathcal{A}$ obtains back $|\phi_1^{out}\rangle$ and $|\phi_r^{out}\rangle$. Note that $|\phi_1\rangle$ is one of the computational basis of $\mathcal{H}^D$ and so is $|\phi_1^{out}\rangle$. Thus, this query is equivalent to a classical query.

Challenge phase $\mathcal{A}$ sets m_1, m_2, and m_3 as his challenges, and he needs to produce the valid classical tag under the challenger's key for all these three messages. Note that $\mathcal{A}$ already has a classical message-tag pair for m_1 from the learning phase, thus $\mathcal{A}$ is left with forging tags for m_2 and m_3.
$\mathcal{A}$ achieves this performing the quantum emulation attack presented in the proof of Theorem 6 on challenges $|m_2, 0, 0\rangle$ and $|m_3, 0, 0\rangle$ which uses the quantum emulation algorithm with only one block.

Probability analysis. We show how the choice of the reference state $|\phi_r\rangle$ for the emulation optimizes this forgery attack. As the reference state is symmetric over the choice of $|m_2, 0, 0\rangle$ and $|m_3, 0, 0\rangle$, $\mathcal{A}$ can emulate the outputs of both $|\phi_2\rangle = |m_2, 0, 0\rangle$ and $|\phi_3\rangle = |m_3, 0, 0\rangle$ with equal fidelity. Let $\alpha^2 = |\langle\phi_r|\,\phi_2\rangle|^2 = |\langle\phi_r|\,\phi_3\rangle|^2 = |\mu|^2$ and $\beta^2 = |\langle\phi_1|\,\phi_r\rangle|^2 = 1 - 2|\mu|^2$, then the fidelity of the emulation for both states is:

$$F(|\omega\rangle\langle\omega|, U^\dagger |\psi\rangle\langle\psi| U) \geq |\alpha^2(1 + 4\beta^4)| = |\mu^2(1 + 4(1 - 2\mu^2)^2)|$$

which is non-negligible for any valid value of μ.

Also, one can see the different levels of resistance to forgeability attacks *wrt* parameter μ. If the reference state has been chosen to be a uniform superposition, i.e. $\mu = \frac{1}{\sqrt{3}}$, the output states for both $|\phi_2\rangle$ and $|\phi_3\rangle$ can be generated with fidelity $F \approx 0.48$, which means that no such MAC scheme can satisfy more than $\frac{1}{\sqrt{3}}$-unforgeability under this attack. A more interesting case of attack can be shown by optimizing the superposition overlaps of state $|\phi_r\rangle$ to get the maximum possible fidelity:

$$F = |\mu|^2(1 + 4(1 - 2|\mu|^2)^2) = 1$$

resulting in $\mu \approx 0.4831$, hence one can emulate the output of U_{MAC} with almost perfect fidelity for the two new messages m_2 and m_3.

The above example distinctly demonstrates that there are MAC schemes that are unforgeable according to the EUF-qCMA definition but which are not secure with respect to our μ-existential unforgeability. Note that in the presented attack above, the adversary is able to produce 3 classical message-tag pairs with high probability but to be able to output both classical outputs for m_2 and m_3 the emulation algorithm needs to be run twice which means that another copy of the output state $|\phi_r^{out}\rangle$ might be needed. Thus formally, the attack does not break EUF-qCMA while it is clearly a forgery on the MAC scheme.

This attack calls for new definitions of unforgeability in the quantum security model. Simply counting the quantum queries does not capture the quantum attacks where the queries are being consumed inside the quantum attack algorithm as also pointed by Alagic *et al.* in [6].

It is worth mentioning that in the above attack the output fidelity does not depend on the dimension of the Hilbert space of the oracle ($\mathcal{H}^D$). It only relies on the ability of the adversary to create enough overlap with the desired state and the learning phase subspace. Therefore while expanding the dimension of the oracle's Hilbert space exponentially reduces the adversary's success probability of gaining information (from the superposition queries through measurement), nevertheless it does not reduce the subspace dimension and overlaps needed for the described emulation attack. This shows a fundamental gap between feasible quantum attacks and current quantum security definitions.

5.2 Quantum Unforgeability of Physical Unclonable Functions with Quantum Access

Physical Unclonable Functions (PUFs) are hardware cryptographic primitives based on unique features of devices that are hard to clone [28, 37, 38]. These unique features can be observed and exploited for security through challenge-response pairs that can be extracted by physically querying the PUF and measuring its responses. Several implementations of PUFs rely on optical systems and hence can potentially be 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 [39–41], 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. In general, quantum access to a classical or quantum-readout PUF can be modelled by a unitary transformation oracle U_{PUF} 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 $|\psi_{out}\rangle \in \mathcal{H}^D$

$$U_{PUF} : |\psi_{in}\rangle \rightarrow |\psi_{out}\rangle = U_{PUF} |\psi_{in}\rangle .$$

Quantum-readout PUFs are in effect quantum primitives as their input and output states are general quantum states in the Hilbert space $\mathcal{H}^D$ and not necessarily encoded over computational (or even any other orthonormal) basis. Thus their unitary transformation can be chosen from a larger set of unitary matrices of certain dimension compared to classical primitives.

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 [28, 38]. Due to a larger set of valid challenges and the properties of Hilbert space, the security definition of such quantum primitives does not reduce to usual quantum security definitions. We investigate desired notion of security for (quantum readouts of) PUFs and similar quantum primitives, 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.

Quantum Existential Unforgeability of PUFs Using the definition of quantum existential unforgeability (Definition 4), and the impossibility result established in Theorem 6, we conclude the following impossibility result for existential unforgeability of any unitary PUF:

Corollary 3. *No unitary PUF with quantum oracle U_{PUF} of dimension $D = 2^N$ satisfies μ-existential unforgeability for any $0 < \mu \leq 1 - \textit{non-negl}(N)$. That is, there exists a QPT adversary that wins the game $\mathcal{G}^{U_{PUF}}_{\mathsf{qCI,qEx},\mu}(\lambda, \mathcal{A})$ with non-negligible probability.*

The attack is exactly the one presented in the proof of Theorem 6. It is worth noting that the QPT adversary wins the game for all interesting values of μ. Indeed, values of $\mu > 1 - \textit{non-negl}(N)$ prevent the adversary from meaningful quantum access to the unitary. This is evidently too restrictive when considering quantum primitives such as PUFs. Hence, regardless of the quantification of μ, no quantum primitive can provide quantum existential unforgeability.

Quantum Selective Unforgeability of PUFs Given the previously established impossibility result, we turn to selective unforgeability. This property will be sufficient for many PUF-based protocols such as identification. In effect, in most PUF-based applications introduced in the literature [28, 42], the PUF needs to respond to a challenge chosen by the verifier. This is precisely the scenario captured by games with Selective Challenge phase. As a direct corollary of our Theorem 9 we can state that *quantum selective unforgeability* can be satisfied by PUF or other quantum primitives as long as their unitary transformation is Unknown Unitary according to Definition 6:

Corollary 4. *Any PUF with unknown unitary transformation U_{PUF} according to Definition 6, satisfies selective unforgeability. That is, no QPT adversary $\mathcal{A}$ can win the game $\mathcal{G}^{U_{PUF}}_{\mathsf{qCI,qSel}}(\lambda, \mathcal{A})$ with non-negligible probability.*

As PUFs are usually considered to be unclonable and unknown even to the manufacturer it is reasonable to assume they are unknown unitaries. Also for other quantum primitives, if their transformation has been randomly picked from a set of unitaries which are indistinguishable from Haar measure or t-designs [35] family, they will be unknown unitaries and as a result be selectively unforgeable according to Corollary 4.

5.3 Quantum Indistinguishability of Encryption

In this section, we show that the emulation attacks presented in Section 4.1 do not only apply against unforgeability but are more general. They also hold against indistinguishability of encryption schemes with quantum oracle. A symmetric key encryption scheme is a triple of algorithms (Gen, Enc, Dec) where the key generation algorithm Gen returns a random key from key space $\mathcal{K}$ and the Enc algorithm operates on a message space $\mathcal{M} = \{0,1\}^m$. For all $k \in \mathcal{K}$, and any message $x \in \mathcal{M}$, the encryption and decryption algorithms satisfy $Pr[Dec(k, Enc_k(x)) = x] = 1$. The quantum oracle was introduced by Boneh *et al.* in [3] and later defined by Gagliardoni *et al.* in [5] as the following unitary transformation:

$$U_{Enc_k} : \sum_{x,y} \alpha_{x,y} |x, y\rangle \to \sum_{x,y} \alpha_{x,y} |x, y \oplus Enc_k(x)\rangle$$

The security of a symmetric key encryption scheme is captured by an indistinguishability game. We consider here the definition of *IND-qCPA* introduced by Boneh and Zhandry in [2, 5]. The indistinguishability is defined through the following game between a Challenger $\mathcal{C}$ which produces a legitimate key k which is used throughout the game, and an adversary $\mathcal{A}$.

qCPA learning phase $\mathcal{A}$ gets oracle access to the encryption oracle U_{Enc_k}

IND challenge phase $\mathcal{A}$ picks two challenge messages m_0 and m_1 and sends these to $\mathcal{C}$. Then $\mathcal{C}$ samples the bit $b \overset{\$}{\leftarrow} \{0,1\}$ and sends back $Enc_k(m_b)$

Guess phase $\mathcal{A}$ guesses b^*.

Definition 9 (IND-qCPA [2, 3]). *A symmetric key encryption scheme is said to be IND-qCPA secure if the success probability of any QPT adversary winning the above game is at most negligibly close (in the security parameter) to $\frac{1}{2}$.*

We show how the emulation attack can be used to win the indistinguishability game of symmetric encryption schemes.

Distinguishing attack Let the encryption oracle U_{Enc_k} be defined over a D-dimensional Hilbert space $\mathcal{H}^D$ and $\mathcal{M}$ be the set of all the classical messages in the domain of the encryption algorithm. The adversary plays the games IND-qCPA as follows:

qCPA learning phase $\mathcal{A}$ queries the following states:

$$|\phi_1\rangle = |m, 0\rangle\,, \quad |\phi_r\rangle = \frac{1}{\sqrt{2}}(|m, 0\rangle + |m', 0\rangle)$$

Where m and m' are any two classical messages.

IND challenge phase $\mathcal{A}$ pick the challenge messages as follows $m_0 = m'$ and $m_1 \neq m$ (m_1 can be any classical message other than m and m') and sends these to $\mathcal{C}$. Note that $\mathcal{A}$ cannot pick m but can pick m' as the m' has never been queried in the learning phase and it is also $\frac{1}{2}$-distinguishable from $|\phi_r\rangle$. $\mathcal{C}$ sends back $Enc_k(m_b)$.

Guess phase $\mathcal{A}$ performs the quantum emulation attack presented in Theorem 6 on $|m', 0\rangle$ and obtains $|\psi^{out}\rangle = |m', Enc_k(m')\rangle$ with fidelity 1, from which he extracts $Enc_k(m')$. Finally he outputs 0 if $Enc_k(m_b) = Enc_k(m')$, and 1 otherwise.

Probability analysis. As the output fidelity is 1, the adversary can perfectly extract the $Enc_k(m')$ and can guess b with probability 1.

The same attack strategy can be used by $\mathcal{A}$ to also win the indistinguishability game with qIND challenge phase where the adversary prepares a quantum state like $|m_0, m_1, 0\rangle$ with the same choice of m_0 and m_1 and $\mathcal{C}$ applies the following transformation:

$$|m_0, m_1, 0\rangle \rightarrow |m_0, m_1, Enc_k(m_b)\rangle$$

This shows that the above attack breaks both qIND and IND security of symmetric encryption.

Comparison with previous superposition attacks The same impossibility result for qIND security of symmetric key encryption has been demonstrated by Boneh and Zhandry in [3] through their superposition attack. Although, our attack does not only work for the two specific messages $m_0 = |0^m\rangle$ and $m_1 = H|0^m\rangle$. It holds for any two messages provided one of the two is at least μ-distinguishable from the learning-phase subspace. Furthermore as demonstrated above, the emulation attack applies to IND-qCPA which is a weaker notion. To conclude, this new class of attacks based on the subspace of the adversary's learning-phase reveal fundamental capabilities for quantum adversaries when moving towards quantum security that have not been fully explored before.

6 Conclusion and Future Directions

In this work, we presented novel formal definitions for different notions of quantum unforgeability which address some limitations of previously proposed ones. We devised novel quantum attacks on unforgeability, which allow us to establish several impossibility results. The first one inspired by the universal quantum emulator algorithm and the other one exploiting quantum entanglement. We also formalized the notion of a family of unknown unitaries and proved that this is a sufficient condition for achieving selective unforgeability. Finally, we demonstrated the applicability of our results, and in particular of our attack to MACs, quantum read-out of PUFs and symmetric encryption schemes. An interesting future direction for our work is to confront our results and techniques to indistinguishability properties such as studies in recent works [17, 18].

References

1. M. Zhandry, "How to construct quantum random functions," in *53rd Annual Symposium on Foundations of Computer Science*, pp. 679–687, Oct 2012.
2. D. Boneh and M. Zhandry, "Quantum-secure message authentication codes," in *Advances in Cryptology – EUROCRYPT 2013* (T. Johansson and P. Q. Nguyen, eds.), (Berlin, Heidelberg), pp. 592–608, Springer Berlin Heidelberg, 2013.

3. D. Boneh and M. Zhandry, “Secure signatures and chosen ciphertext security in a quantum computing world,” in *Advances in Cryptology – CRYPTO 2013* (R. Canetti and J. A. Garay, eds.), (Berlin, Heidelberg), pp. 361–379, Springer Berlin Heidelberg, 2013.
4. M. Kaplan, G. Leurent, A. Leverrier, and M. Naya-Plasencia, “Breaking symmetric cryptosystems using quantum period finding,” in *Advances in Cryptology – CRYPTO 2016* (M. Robshaw and J. Katz, eds.), (Berlin, Heidelberg), pp. 207–237, Springer Berlin Heidelberg, 2016.
5. T. Gagliardoni, A. Hülsing, and C. Schaffner, “Semantic security and indistinguishability in the quantum world,” in *Advances in Cryptology – CRYPTO 2016* (M. Robshaw and J. Katz, eds.), (Berlin, Heidelberg), pp. 60–89, Springer Berlin Heidelberg, 2016.
6. G. Alagic, C. Majenz, A. Russell, and F. Song, “Quantum-secure message authentication via blind-unforgeability,” *arXiv preprint arXiv:1803.03761*, 2018.
7. F. Song and A. Yun, “Quantum security of nmac and related constructions,” in *Advances in Cryptology – CRYPTO 2017* (J. Katz and H. Shacham, eds.), (Cham), pp. 283–309, Springer International Publishing, 2017.
8. T. Gagliardoni, “Quantum security of cryptographic primitives,” *arXiv preprint arXiv:1705.02417*, 2017.
9. G. Alagic, T. Gagliardoni, and C. Majenz, “Unforgeable quantum encryption,” in *Advances in Cryptology – EUROCRYPT 2018* (J. B. Nielsen and V. Rijmen, eds.), (Cham), pp. 489–519, Springer International Publishing, 2018.
10. I. Marvian and S. Lloyd, “Universal quantum emulator,” *arXiv preprint arXiv:1606.02734*, 2016.
11. M. Bellare, R. Canetti, and H. Krawczyk, “Keying hash functions for message authentication,” in *Advances in Cryptology — CRYPTO ’96* (N. Koblitz, ed.), (Berlin, Heidelberg), pp. 1–15, Springer Berlin Heidelberg, 1996.
12. P. W. Shor, “Algorithms for quantum computation: discrete logarithms and factoring,” in *35th Annual Symposium on Foundations of Computer Science*, pp. 124–134, Nov 1994.
13. L. K. Grover, “A fast quantum mechanical algorithm for database search,” in *28th annual ACM symposium on Theory of computing*, pp. 212–219, 1996.
14. D. R. Simon, “On the power of quantum computation,” *SIAM journal on computing*, vol. 26, no. 5, pp. 1474–1483, 1997.
15. T. Santoli and C. Schaffner, “Using simon’s algorithm to attack symmetric-key cryptographic primitives,” *arXiv preprint arXiv:1603.07856*, 2016.
16. X. Bonnetain, A. Hosoyamada, M. Naya-Plasencia, Y. Sasaki, and A. Schrottenloher, “Quantum attacks without superposition queries: The offline simon’s algorithm,” in *Advances in Cryptology – ASIACRYPT 2019* (S. D. Galbraith and S. Moriai, eds.), (Cham), pp. 552–583, Springer International Publishing, 2019.
17. J. Czajkowski, C. Majenz, C. Schaffner, and S. Zur, “Quantum lazy sampling and game-playing proofs for quantum indifferentiability,” *arXiv preprint arXiv:1904.11477*, 2019.
18. T. V. Carstens, E. E. Ebrahimi, G. N. Tabia, and D. Unruh, “On quantum indifferentiability.,” *IACR Cryptology ePrint Archive*, vol. 2018, p. 257, 2018.
19. M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 10th ed., 2010.
20. W. K. Wootters and W. H. Zurek, “A single quantum cannot be cloned,” *Nature*, vol. 299, no. 5886, p. 802, 1982.
21. H. Buhrman, R. Cleve, J. Watrous, and R. De Wolf, “Quantum fingerprinting,” *Physical Review Letters*, vol. 87, no. 16, p. 167902, 2001.

22. 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.
23. 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.
24. S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum principal component analysis," *Nature Physics*, vol. 10, no. 9, p. 631, 2014.
25. D. Boneh, Ö. Dagdelen, M. Fischlin, A. Lehmann, C. Schaffner, and M. Zhandry, "Random oracles in a quantum world," in *Advances in Cryptology – ASIACRYPT 2011* (D. H. Lee and X. Wang, eds.), (Berlin, Heidelberg), pp. 41–69, Springer Berlin Heidelberg, 2011.
26. M. Zhandry, "Secure identity-based encryption in the quantum random oracle model," *International Journal of Quantum Information*, vol. 13, no. 04, p. 1550014, 2015.
27. I. Damgård, J. Funder, J. B. Nielsen, and L. Salvail, "Superposition attacks on cryptographic protocols," in *International Conference on Information Theoretic Security*, pp. 142–161, Springer, 2013.
28. F. Armknecht, D. Moriyama, A.-R. Sadeghi, and M. Yung, "Towards a unified security model for physically unclonable functions," in *Cryptographers' Track at the RSA Conference*, pp. 271–287, Springer, 2016.
29. V. Soukharev, D. Jao, and S. Seshadri, "Post-quantum security models for authenticated encryption," in *7th International Workshop on Post-Quantum Cryptography*, pp. 64–78, Springer, 2016.
30. M. Oszmaniec, A. Grudka, M. Horodecki, and A. Wójcik, "Creating a superposition of unknown quantum states," *Physical Review Letters*, vol. 116, no. 11, 2016.
31. M. Doosti, F. Kianvash, and V. Karimipour, "Universal superposition of orthogonal states," *Physical Review A*, vol. 96, no. 5, p. 052318, 2017.
32. 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.
33. A. Van Daele, "The haar measure on a compact quantum group," in *Proceedings of the American Mathematical Society*, vol. 123, pp. 3125–3128, 1995.
34. 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.
35. A. Ambainis and J. Emerson, "Quantum t-designs: t-wise independence in the quantum world," in *22nd Annual IEEE Conference on Computational Complexity (CCC'07)*, pp. 129–140, IEEE, 2007.
36. B. ŠKORIĆ, "Quantum readout of physical unclonable functions," *International Journal of Quantum Information*, vol. 10, no. 01, p. 1250001, 2012.
37. U. Rührmair and D. E. Holcomb, "Pufs at a glance," in *the conference on Design, Automation & Test in Europe*, p. 347, European Design and Automation Association, 2014.
38. C. Brzuska, M. Fischlin, H. Schröder, and S. Katzenbeisser, "Physically uncloneable functions in the universal composition framework," in *Advances in Cryptology – CRYPTO 2011* (P. Rogaway, ed.), (Berlin, Heidelberg), pp. 51–70, Springer Berlin Heidelberg, 2011.
39. R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions," *Science*, vol. 297, no. 5589, pp. 2026–2030, 2002.

40. G. M. Nikolopoulos and E. Diamanti, “Continuous-variable quantum authentication of physical unclonable keys,” *Scientific reports*, vol. 7, p. 46047, 2017.
41. 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.
42. 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.
43. K. Życzkowski and H.-J. Sommers, “Average fidelity between random quantum states,” *Physical Review A*, vol. 71, no. 3, p. 032313, 2005.

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[22]. 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.

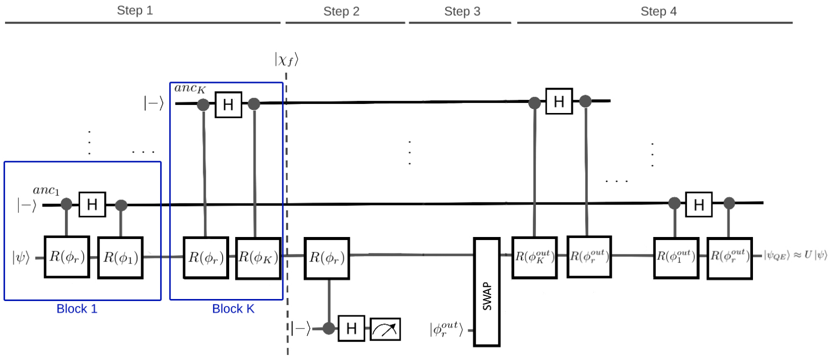

Fig. 5: The quantum emulation algorithm's circuit. $|\phi_r\rangle$ is the reference state and $|\phi_r^{out}\rangle$ is the output of the reference state. $R(*)$ gates are controlled-reflection. In each Block of Step 1, a reflection around the reference and another sample state is being performed. At 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 5

We state the theorem again and we establish the proof:

Theorem 11. *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:*

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

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 (9) 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

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

which satisfies the form of equation (10) 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 (10), we will show that $|\chi_k\rangle$ will also satisfy 10. We use the recursive equation again. We will have

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

where $|\chi_{k-1}\rangle$ has the form of equation (10). 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 Security proofs

E.1 Proof of Theorem 6: Impossibility 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 5, the output state of Stage 1 of the QE algorithm is:

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

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 4, 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 $\frac{1}{2} < \mu \leq 1 - \textit{non-negl}(\lambda)$:

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

E.2 Proof of Theorem 7: Impossibility 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 the impossibility of building a precise target superposition of completely unknown quantum states was prove in [30], there are however some superposition algorithms such as [30, 31] 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 [31] 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:

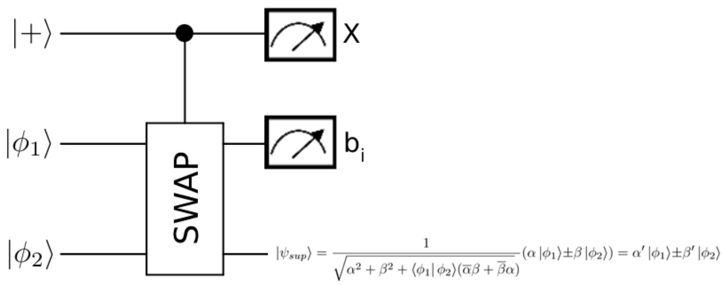
$|+\rangle$
$|\phi_1\rangle$
$|\phi_2\rangle$
SWAP
X
b_i
$|\psi_{sup}\rangle = \frac{1}{\sqrt{\alpha^2 + \beta^2 + \langle\phi_1|\,\phi_2\rangle(\overline{\alpha}\beta + \overline{\beta}\alpha)}}(\alpha\,|\phi_1\rangle \pm \beta\,|\phi_2\rangle) = \alpha'\,|\phi_1\rangle \pm \beta'\,|\phi_2\rangle$

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 $\mathcal{H}^D$

$$|\psi_{sup}\rangle = \frac{1}{\sqrt{\alpha^2 + \beta^2 + \langle\phi_1|\,\phi_2\rangle(\overline{\alpha}\beta + \overline{\beta}\alpha)}}(\alpha\,|\phi_1\rangle \pm \beta\,|\phi_2\rangle) = \alpha'\,|\phi_1\rangle \pm \beta'\,|\phi_2\rangle \quad (11)$$

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 4 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

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

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:

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

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

E.3 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

$$\sum_{i=0}^{D-1} \mathit{\Pi}_i = \mathbb{I}$$

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

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

Clearly, each term $\langle\psi| \mathit{\Pi}_i |\psi\rangle$ 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. □

E.4 Proof of Theorem 8

Proof. Let $\mathcal{A}$ be a quantum adversary playing the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{qCl},\mathsf{qSel}}(\lambda, \mathcal{A})$. 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:

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

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

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

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 $\mathcal{H}^{d^{\perp}}_{out}$ while $|\omega\rangle$ will cover a subspace of $\mathcal{H}^{d^{\perp}}_{out}$. This will decrease the average success probability of the adversary. So, generating the states $|\omega\rangle$ such that they span the whole $\mathcal{H}^{d^{\perp}}_{out}$, 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[43], but most commonly the average fidelity can be written as:

$$\int\limits_{|\psi^{out}\rangle\in\mathcal{H}^{d^{\perp}}_{out}} |\langle\omega|\,\psi^{out}_x\rangle|^2 d\mu_x$$

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:

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

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

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

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

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

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. □

E.5 Proof of Theorem 9

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 8 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:

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

and the proof is complete. □

E.6 Proof of Theorem 10

Proof. Let $\mathcal{A}$ be the QPT adversary playing the game $\mathcal{G}^{\mathcal{F}}_{\mathsf{null,qSel,qCI},\mu}(\lambda, \mathcal{A})$ 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. □