

Routing Anonymity and Identifiability of Noisy Quantum Hardware

Ben Priestley^{1,2} Mina Doosti²

¹ Centre for Quantum Information and Foundations, DAMTP, University of Cambridge

² Quantum Software Lab, School of Informatics, University of Edinburgh

Full version available at <https://arxiv.org/abs/2607.05281>.

1 Introduction

Present-day quantum computing is already predominantly cloud-based: a user submits a circuit to a service provider, who executes it on proprietary backend hardware and returns finite-shot classical data. This architecture is convenient, and arguably unavoidable in the near future, since the most capable quantum processors are expensive, delicate, and centrally operated. However, it also creates a new information channel. Even when the provider hides implementation details, scheduling choices, calibration data, or even which physical device was used, noisy finite-shot outputs may carry backend-specific “fingerprints”, information imprinted in the classical output distribution that can reveal the backend identity. Such fingerprints have been observed experimentally and studied as tools for benchmarking, verification, and hardware forensics [1–6].

In this work, we frame this phenomenon as a dual *learning–security* problem for quantum cloud services. On the learning side, the user has black-box finite-shot access to an unknown noisy backend channel, through chosen workload circuits, and tries to learn the latent backend identity from the resulting classical transcripts. On the security side, the provider asks whether this latent route can remain anonymous, even when the user is allowed to choose informative probes and apply arbitrary statistical or machine-learning classifiers to the returned data. This leads to the central question of our work:

When, and to what extent, can a quantum cloud service keep its routing choice anonymous?

This question is particularly natural for the quantum machine learning (QML) community, especially in light of recent trends that seek quantum advantage at deployment or inference time rather than only during training, and that treat noisy quantum devices as data-generating systems. In this setting, the relevant degrees of freedom available to a QML practitioner are the family of circuits being designed and the finite-shot samples that can be collected from them. Backend identifiability is precisely the *learning problem* that such data, or learned representations of it, can contain enough information to classify the hidden physics of the device. In this sense, backend fingerprints are not only benchmarking criteria, but genuinely learnable and operationally meaningful features of noisy quantum hardware. Routing anonymity is the corresponding security notion, asking when such learning should be impossible.

We formalise this notion through a formal but operationally motivated backend-identification security game. A user submits workload circuits sampled from a chosen ensemble, the provider executes them on a hidden backend, optionally post-processes the observed finite-shot data, and returns a transcript. The user wins if she correctly identifies the hidden backend. This gives rise to the security notion of *routing anonymity*, which asks whether even the best possible identifier can do substantially better than the baseline strategy of guessing the most likely route.

2 Our results

We develop, to our knowledge, the first formal framework connecting backend identifiability, statistical hypothesis testing, privacy, utility, and learnability. We summarise the main contributions below, and refer the interested reader to the full version of the paper for complete definitions, proofs, experiments, and further discussion.

Backend identifiability as hypothesis testing. For a workload ensemble μ , number of shots m , backend D_i , and post-processing map ϕ , we associate an induced transcript law,

$$P_i(\mu, m, \phi) = \phi_{\#} Q_i(\mu, m) \quad ,$$

where $Q_i(\mu, m)$ is the raw transcript law before post-processing. The backend-identification game then reduces to a hypothesis-testing problem with hypotheses $H_i : X \sim P_i$. In particular, for two equally likely routes D_i, D_j , the optimal success probability is characterized by total-variation distance between the transcript laws; $p_s^* = (1 + \text{TV}(P_i, P_j))/2$. More generally, for an arbitrary prior π over routes and $r^* \in \arg \max_i \pi_i$, we show that,

$$p_s^* \leq \pi_{r^*} + \sum_{i \neq r^*} \pi_i \text{TV}(P_i, P_{r^*}) \quad .$$

Consequently, pairwise closeness of transcript laws directly implies *routing anonymity*, in the sense that the excess advantage beyond π_{r^*} is arbitrarily small.

Repeated access amplifies identifiability. To go from the above simple example to a more realistic scenario, we also study *persistent routing*, where the same hidden backend is probed independently over T rounds. In this case the route-conditioned transcript law becomes $P_i^{\otimes T}$. For two equally likely routes, standard asymptotic hypothesis testing gives,

$$1 - p_s^*(T) = \exp\left(-T D_{\text{Ch}}(P_i, P_j) + o(T)\right) \quad ,$$

where D_{Ch} is the Chernoff information [7]. Thus, whenever two backends induce distinguishable transcript laws, repeated access exponentially amplifies identifiability. Routing anonymity is therefore a genuine operational security requirement.

A utility–anonymity trade-off theorem. A provider can try to improve anonymity by post-processing the raw output before returning it to the user. Indeed,

$$\text{TV}(\phi_{\#} Q_i, \phi_{\#} Q_j) \leq \text{TV}(Q_i, Q_j) \quad ,$$

so post-processing cannot increase route-specific information. However, the provider is also expected to preserve some useful service. We model the promised service by a utility map u on the raw observation, and say that ϕ preserves this utility exactly if there exists a decoder ψ such that $u = \psi \circ \phi$. We prove a no-free-lunch statement: for any pair of routes,

$$\begin{aligned} \text{TV}(P_i(\dots, \phi), P_j(\dots, \phi)) &\geq \text{TV}(P_i(\dots, u), P_j(\dots, u)) \quad , \\ D_{\text{Ch}}(P_i(\dots, \phi), P_j(\dots, \phi)) &\geq D_{\text{Ch}}(P_i(\dots, u), P_j(\dots, u)) \quad . \end{aligned}$$

Therefore, post-processing can only remove route-specific information that is extraneous to the promised utility (that is, only that information which is not intrinsically tied to the useful output).

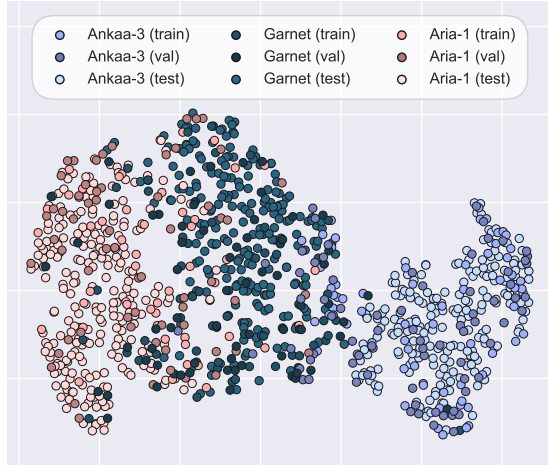
Backend identifiability is an intermediate-depth phenomenon. We further show that backend fingerprints need not grow monotonically with circuit depth. Using a Pauli-transfer-matrix model, suppose that the traceless action of backend D_i per layer is,

$$N_i = \lambda I + E_i \quad ,$$

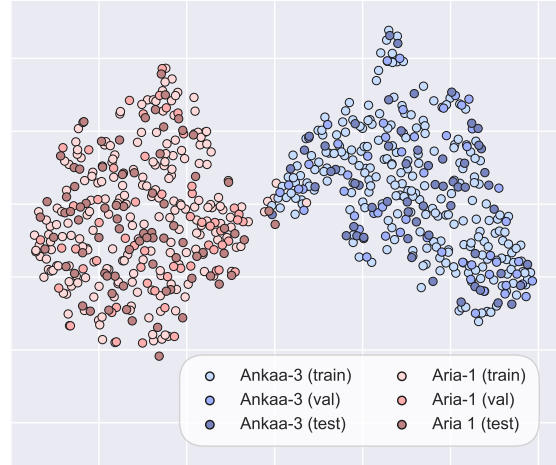
where λI is a common contractive mixing component, $\|E_i\|_{1 \rightarrow 1} \leq \epsilon$, and $\lambda + \epsilon < 1$. Under a generic non-degeneracy condition on the workload, the expected single-shot distinguishing bias satisfies,

$$\Omega(d\lambda^{d-1}) \leq \text{Adv}(d) \leq \mathcal{O}(d(\lambda + \epsilon)^{d-1}) \quad .$$

This gives a theoretical explanation for a practical observation of interest to learning-based experiments on noisy hardware: shallow circuits may not expose enough backend-specific information, while very deep circuits may be washed out by noise. The strongest fingerprints should therefore arise within some ‘intermediate depth’ window.



(a) All-type classifier embeddings.



(b) Differing-type classifier embeddings.

Experiments on cloud quantum hardware. We instantiate the identification game on Amazon Braket using Rigetti’s *Ankaa-3* and IQM’s *Garnet* superconducting QPUs, and IonQ’s *Aria-1* trapped-ion QPU. We use two 5-qubit probe workloads: depth-varied random circuits built from Haar-random two-qubit gates in a brickwork pattern, and fixed GHZ-preparation circuits repeated over time. The random-circuit experiments use 512 shots per circuit, with depths $d = 5, 10, 15, 20$ on the superconducting devices and $d = 5, 10$ on the ion-trap device; the GHZ experiments use 1024 shots to resolve leakage outside the ideal two-outcome support. For each circuit we collect finite-shot computational-basis histograms and train simple classifiers on raw and post-processed transcript representations.

The experimental results show that backend identity is highly learnable from finite-shot data. On raw random-circuit transcripts, classifiers identify the two superconducting devices with approximately 87–90% test accuracy, and distinguish superconducting from trapped-ion hardware with 96–100% test accuracy. For repeatedly probed GHZ workloads, we observe perfect classification. Importantly, identifiability survives several natural post-processing maps; for example, log-ratio features retain substantial identifying information even when the raw histogram is not exposed.

The experiments also reveal structure relevant for other applications of classical machine learning to quantum noise. Several processed features become more informative when moving from depth $d = 5$ to $d = 10$, while deeper like-type experiments show a general decay in distinguishability. This supports the intermediate-depth picture in which useful backend-specific signal first accumulates and is later washed out by mixing. In the time-varied GHZ experiments, batch-level classification reveals a learnable long-range drift signal, while within-batch early-versus-late classification is much closer to chance. This suggests that backend fingerprints are not only device-specific, but also temporally structured objects that can be learned from data.

Overall, our results and framework initiate the study of a new problem at the intersection of learning, security, and privacy, through the dual notions of backend identifiability and routing anonymity. The associated learning problem has several distinctive features: it is lighter than full noise characterisation, different in spirit from standard reconstruction tasks such as classical shadows, and sensitive to operational choices such as circuit depth, temporal drift, and post-processing. This depth dependence follows from learning noisy quantum channels, where backend-specific signal can first accumulate and later be washed out by mixing. In the full paper, we propose several applications of this framework and demonstrate its practical relevance on the AWS quantum cloud platform, addressing a concrete privacy question already present in today’s cloud-based quantum computing and quantum machine learning landscape.

References

- [1] Stefano Martina, Lorenzo Buffoni, Stefano Gherardini, and Filippo Caruso. Learning the noise fingerprint of quantum devices, 2021.
- [2] Rupshali Roy, Archisman Ghosh, and Swaroop Ghosh. Forensics of transpiled quantum circuits. In *Proceedings of the Great Lakes Symposium on VLSI 2025*, GLSVLSI 2025, pages 354–359. Association for Computing Machinery, 2025. doi: 10.1145/3716368.3735240.
- [3] Kaitlin N Smith, Joshua Vizlai, Lennart Maximilian Seifert, Jonathan M Baker, Jakub Szefer, and Frederic T Chong. Fast fingerprinting of cloud-based nisq quantum computers. In *2023 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pages 1–12. IEEE, 2023.
- [4] Allen Mi, Shuwen Deng, and Jakub Szefer. Short paper: Device-and locality-specific fingerprinting of shared nisq quantum computers. In *Proceedings of the 10th International Workshop on Hardware and Architectural Support for Security and Privacy*, pages 1–6, 2021.
- [5] Vincent Mutolo, Devon Campbell, Quinn Manning, Henri Witold Dubourg, Ruibin Lyu, Simha Sethumadhavan, Daniel Rubenstein, and Salvatore Stolfo. Quantum computer fingerprinting using error syndromes, 2025. URL <https://arxiv.org/abs/2506.16614>.
- [6] Jindi Wu, Tianjie Hu, and Qun Li. Q-id: Lightweight quantum network server identification through fingerprinting. *Netwrk. Mag. of Global Internetwkg.*, 38(5):146–152, September 2024. ISSN 0890-8044. doi: 10.1109/MNET.2024.3400893. URL <https://doi.org/10.1109/MNET.2024.3400893>.
- [7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.