

Adversarial robustness in rotationally equivariant quantum machine learning

Maureen Krümtünger¹, Martin Sevier¹, Muhammad Usman^{1,2}

¹School of Physics, University of Melbourne, Parkville, VIC 3010, Australia

²Data61, CSIRO, Clayton, VIC 3168, Australia

Abstract

Group-equivariant quantum models impose symmetry constraints that can improve trainability, but their effect on adversarial robustness remains unclear. We address this question through a feature-level analysis of equivariant quantum models under adversarial transfer attacks. For models with invariant readout, predictions depend only on the group-twirled input, separating accessible invariant information from an invisible subspace. For a rotationally equivariant quantum model, we show that the accessible feature space can be characterized as rotation-invariant image statistics resolved across symmetry sectors. Using targeted input transformations, we identify which of these statistics the model relies on across several datasets. We find that equivariance does not by itself provide transfer robustness, because the invariant feature space can still contain brittle features. We then show that suppressing the symmetry sector associated with the brittle feature improves robustness while preserving accuracy. Together, these results provide a way to diagnose brittle invariant features and improve robustness in equivariant QML models.

Keywords: Adversarial robustness; geometric quantum machine learning; feature-level analysis.

This extended abstract is based on our arXiv preprint [arXiv:2604.15552](https://arxiv.org/abs/2604.15552).

1 Introduction

Geometric quantum machine learning (QML) uses group-equivariant architectures to improve inductive bias and, in several constructions, guarantee trainability [6, 9–12, 16]. How such symmetry constraints affect adversarial robustness remains much less well understood. In classical machine learning, adversarial perturbations [4, 18] are often understood as exploiting a model’s reliance on brittle, non-robust features [5, 13]. Related observations in QML suggest that differences in robustness between quantum and classical models may likewise arise from differences in the predictive features they exploit [2, 15, 17]. Motivated by this feature-level perspective, we ask how equivariance reshapes the feature space of a quantum model, and whether the resulting symmetry constraints lead to enhanced robustness against adversarial transfer. We analyze equivariant QML models using a twirling-based framework, and apply it to the rotationally equivariant model of Ref. [16] as a representative setting. Equivariance decomposes the input space into components that are invisible to the model and rotationally invariant feature channels that can contribute to its predictions. We then use controlled input transformations to identify which invariant statistics the model relies on across datasets, showing that feature use is highly dataset dependent. Moreover, adversarial transfer attacks generated from classical surrogate models reveal that reliance on particular invariant feature channels is closely linked to adversarial vulnerability. Thus, while equivariance removes sensitivity to a large perturbation subspace, it does not automatically eliminate non-robust features. Finally, we show that suppressing symmetry channels associated with brittle predictive features substantially improves robustness while preserving clean accuracy. Overall, our results show that a feature-level understanding of equivariant quantum models can help explain adversarial vulnerability and guide targeted interventions to improve robustness.

2 Preliminaries

Rotationally equivariant quantum model. We consider supervised image classification tasks in which the label is invariant under a group $\mathcal{G}$, i.e., $y(x) = y(gx)$ for all $g \in \mathcal{G}$. As a representative example, we use the rotationally equivariant model of Ref. [16]. This model encodes an image by sampling its pixel values at vertices arranged on N_r concentric regular polygons, which we refer to as *rings*, each with N_ϕ angular positions [16]:

$$|\psi(x)\rangle = \frac{1}{\|x\|} \sum_{r,\phi} x_{r,\phi} |r, \phi\rangle. \quad (1)$$

In this setup, discrete rotations act as cyclic shifts on the angular register, realizing a representation of $\mathbb{Z}_{N_\phi}$. Applying a quantum Fourier transform to the angular register diagonalizes this action and decomposes it into Fourier modes, corresponding to the one-dimensional irreducible representations of $\mathbb{Z}_{N_\phi}$ labeled by $m = 0, \dots, N_\phi - 1$ [16]. The measurement operator is chosen as $M_j := Z_j$ ($j = 1, \dots, n_{\text{classes}}$), with $n_{\text{classes}} \leq n_{\text{rad}}$, which commutes with the group action.

Adversarial transfer attacks. Given a clean input x , an adversary constructs a perturbed input $x' = x + \delta$ with $\|\delta\|_\infty \leq \varepsilon$, intended to induce misclassification by the target quantum model. We consider a black-box setting in which the adversary has no access to the target model’s parameters, gradients, or queries. Instead, perturbations are generated using FGSM [3] and PGD [8] on a classical surrogate model trained on the same data distribution, and are then transferred to the quantum model.

3 Main results

We summarize the main feature-level analysis and numerical results on the STM [14, 16] and MNIST [7] datasets.

Characterization of features available to the model. In the rotationally equivariant model of Ref. [16], predictions depend only on the group-twirled representation of the input. This separates perturbations that are invisible to the model from rotation-invariant statistics that can affect its decision. In the symmetry-sector decomposition, the $m = 0$ sector corresponds to the ring-averaged intensity profile, while the $m \neq 0$ sectors contribute invariant angular-correlation features. Different sectors therefore define distinct feature channels whose predictive relevance and robustness can be probed independently. Figure 1(a) illustrates this probe using three controlled transformations: T_1 preserves the twirled representation and is therefore invisible to the model, T_2 preserves the ring-averaged intensity profile while perturbing angular structure, and T_3 removes the ring average while preserving angular structure.

Accessible and utilized features. The invariant feature channels used for prediction are dataset-dependent. To demonstrate this, we train the rotationally equivariant model on clean versions of the datasets introduced above and compared its accuracy on clean test inputs with its accuracy on the transformed test sets T_1 , T_2 , and T_3 , as shown in Figure 1(b). T_1 leaves the test accuracy unchanged, confirming that perturbations preserving the twirled representation do not affect the model’s decision. In contrast, the test accuracies under T_2 and T_3 vary across datasets: For STM, T_2 -transformed images remain highly classifiable, showing that the ring-averaged intensity profile already contains substantial predictive information. For MNIST, the opposite trend is closer to visual intuition: T_3 -transformed images are classified substantially better than T_2 -transformed images, indicating that angular structure is more informative than the ring average.

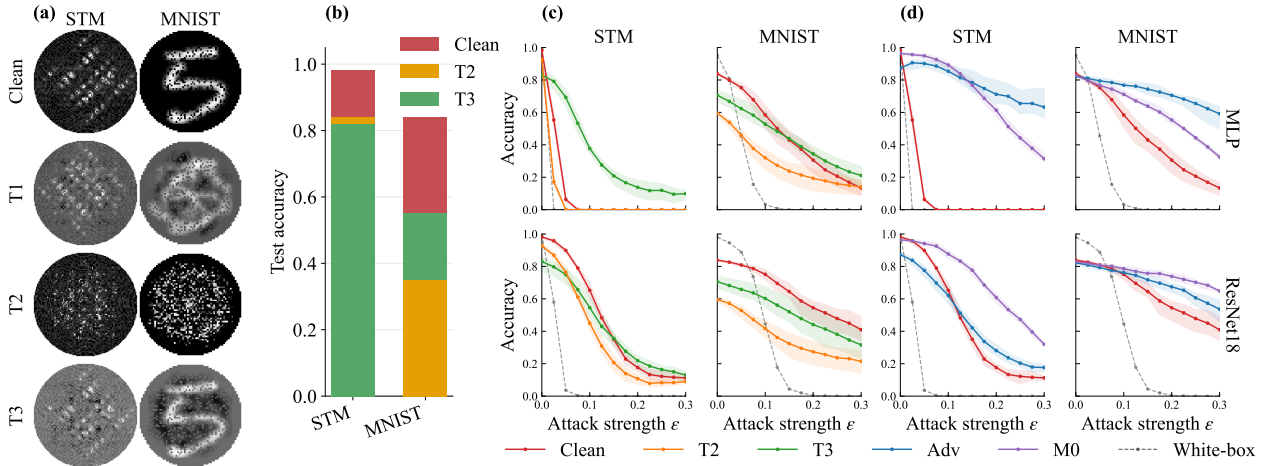


Figure 1: (a) Examples of the three controlled transformations applied to representative STM and MNIST images. (b) Test accuracy of the clean-trained rotationally equivariant model when evaluated on clean, T_2 -transformed, and T_3 -transformed test inputs. (c) Test accuracy as a function of adversarial attack strength for transfer attacks generated using classical MLP and ResNet18 surrogate models. Results compare models trained on clean data, T_2 -transformed data, and T_3 -transformed data. (d) Test accuracy as a function of adversarial attack strength for transfer attacks generated using classical MLP and ResNet18 surrogate models. Results compare the clean-trained model, the adversarially trained model, and the model in which the $m = 0$ symmetry sector is suppressed at measurement.

Brittle and robust features under transfer attacks. We next show that these invariant feature channels differ not only in predictive relevance, but also in adversarial robustness. Transfer attacks generated by classical surrogate models affect models trained on T_2 - and T_3 -transformed inputs differently, as shown in Figure 1(c). Models that rely more strongly on the feature channel preserved by T_2 are consistently less robust than models that rely on the feature channel preserved by T_3 . This identifies the ring-averaged intensity profile as a brittle predictive feature channel in our setting, and indicates that equivariance alone does not remove adversarially vulnerable features.

Improving robustness through targeted feature suppression. The feature-level diagnosis suggests a targeted intervention by suppressing the brittle $m = 0$ sector at measurement. Figure 1(d) compares this intervention with adversarial training [1, 8, 19]. In our setting, adversarial training can improve robustness beyond $m = 0$ sector suppression, but its effectiveness depends on the attacker model used during training and comes at the cost of a reduction in clean accuracy. In contrast, projecting out the $m = 0$ mode yields a more consistent robustness improvement across all surrogate attacks while preserving clean accuracy better.

4 Discussion and Outlook

Our results show that equivariance removes non-invariant perturbation directions, but does not by itself guarantee adversarial robustness. Brittle predictive features can persist within the invariant subspace. At the same time, by exploiting the irreducible decomposition of the model, we identify the symmetry sector associated with brittle features and show that suppressing this sector improves transfer robustness without a substantial loss in clean accuracy. Although the detailed interpretation is specific to the rotationally equivariant architecture and encoding studied here, the underlying twirling-based framework applies more broadly to equivariant quantum models. Thus, investigating how such feature channels can be combined, monitored, and controlled to guide the design of robust equivariant QML architectures remains an important direction for future work.

References

- [1] Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness (2021), <https://arxiv.org/abs/2102.01356>
- [2] Dowling, N., West, M.T., Southwell, A., Nakhel, A.C., Sevier, M., Usman, M., Modi, K.: Adversarial robustness guarantees for quantum classifiers. *npj Quantum Information* **12**(1) (Jan 2026). <https://doi.org/10.1038/s41534-025-01129-3>, <http://dx.doi.org/10.1038/s41534-025-01129-3>
- [3] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015), <https://arxiv.org/abs/1412.6572>
- [4] Ibitoye, O., Abou-Khamis, R., el Shehaby, M., Matrawy, A., Shafiq, M.O.: The threat of adversarial attacks on machine learning in network security – a survey (2023), <https://arxiv.org/abs/1911.02621>
- [5] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features (2019), <https://arxiv.org/abs/1905.02175>
- [6] Larocca, M., Sauvage, F., Sbahi, F.M., Verdon, G., Coles, P.J., Cerezo, M.: Group-invariant quantum machine learning. *PRX Quantum* **3**, 030341 (Sep 2022). <https://doi.org/10.1103/PRXQuantum.3.030341>, <https://link.aps.org/doi/10.1103/PRXQuantum.3.030341>
- [7] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998). <https://doi.org/10.1109/5.726791>
- [8] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations* (2018), <https://openreview.net/forum?id=rJzIBfZAb>
- [9] Meyer, J.J., Mularski, M., Gil-Fuster, E., Mele, A.A., Arzani, F., Wilms, A., Eisert, J.: Exploiting symmetry in variational quantum machine learning. *PRX Quantum* **4**, 010328 (Mar 2023). <https://doi.org/10.1103/PRXQuantum.4.010328>, <https://link.aps.org/doi/10.1103/PRXQuantum.4.010328>
- [10] Nguyen, Q.T., Schatzki, L., Braccia, P., Ragone, M., Coles, P.J., Sauvage, F., Larocca, M., Cerezo, M.: Theory for equivariant quantum neural networks. *PRX Quantum* **5**, 020328 (May 2024). <https://doi.org/10.1103/PRXQuantum.5.020328>, <https://link.aps.org/doi/10.1103/PRXQuantum.5.020328>
- [11] Schatzki, L., Larocca, M., Nguyen, Q.T., Sauvage, F., Cerezo, M.: Theoretical guarantees for permutation-equivariant quantum neural networks. *npj Quantum Information* **10**(1) (Jan 2024). <https://doi.org/10.1038/s41534-024-00804-1>, <http://dx.doi.org/10.1038/s41534-024-00804-1>
- [12] T West, M., Sevier, M., Usman, M.: Reflection equivariant quantum neural networks for enhanced image classification. *Machine Learning: Science and Technology* **4**(3), 035027 (aug 2023). <https://doi.org/10.1088/2632-2153/acf096>, <https://doi.org/10.1088/2632-2153/acf096>
- [13] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: *International Conference on Learning Representations* (2019), <https://openreview.net/forum?id=SyxAb30cY7>

- [14] Usman, M., Wong, Y.Z., Hill, C.D., et al.: Framework for atomic-level characterisation of quantum computer arrays by machine learning. *npj Computational Materials* **6**(1), 19 (2020). <https://doi.org/10.1038/s41524-020-0282-0>, <https://doi.org/10.1038/s41524-020-0282-0>
- [15] West, M.T., Erfani, S.M., Leckie, C., Sevier, M., Hollenberg, L.C.L., Usman, M.: Benchmarking adversarially robust quantum machine learning at scale. *Physical Review Research* **5**(2) (Jun 2023). <https://doi.org/10.1103/physrevresearch.5.023186>, <http://dx.doi.org/10.1103/PhysRevResearch.5.023186>
- [16] West, M.T., Heredge, J., Sevier, M., Usman, M.: Provably trainable rotationally equivariant quantum machine learning. *PRX Quantum* **5**, 030320 (Jul 2024). <https://doi.org/10.1103/PRXQuantum.5.030320>, <https://link.aps.org/doi/10.1103/PRXQuantum.5.030320>
- [17] West, M.T., Tsang, S.L., Low, J.S., Hill, C.D., Leckie, C., Hollenberg, L.C.L., Erfani, S.M., Usman, M.: Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence* **5**(6), 581–589 (May 2023). <https://doi.org/10.1038/s42256-023-00661-1>, <http://dx.doi.org/10.1038/s42256-023-00661-1>
- [18] Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning (2018), <https://arxiv.org/abs/1712.07107>
- [19] Zhao, M., Zhang, L., Ye, J., Lu, H., Yin, B., Wang, X.: Adversarial training: A survey (2024), <https://arxiv.org/abs/2410.15042>