

Spatially Informed Quantum Neural Networks for Interpretable Cancer Mutation Classification

Caitlin Jones, Hugh Warden, Sara Metwalli, Tommy Williams, Luke Boulter, Sjoerd Beentjes, Mina Doosti, Petros Wallden, Elham Kashefi, Ava Khamseh.

We introduce a new application of variational quantum models in which the structure is designed to reflect domain-specific spatial information in the data. Rather than treating an image as an unstructured collection of features, our approach represents different spatial regions as quantum subsystems and models their interactions through a parametrised Hamiltonian. The main property of our model is that it has a built-in form of interpretability: after training, the learned interaction coefficients can be extracted as region-level relevance scores, indicating which parts of the spatial environment contribute most strongly to the classification.

We apply this approach to image-based cancer mutation classification, focusing on the task of distinguishing KRAS and CTNNB1 mutations from tumour images centred on individual cancerous cells. In numerical experiments, the model achieves strong classification performance and identifies the target cell itself as the most relevant contributor to the prediction. These results suggest that Hamiltonian-structured quantum models can provide a natural way to combine spatial inductive bias with interpretable classification, without requiring additional post-hoc explanation methods or extra evaluation cost.

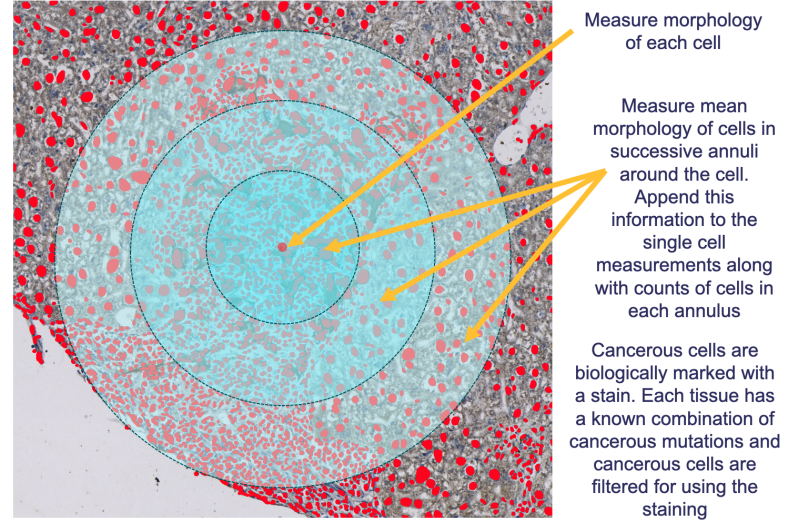


Figure 1: Example of histology image from a mouse liver tumour. The type of mutation is to be classified using cell-intrinsic features as well as features in concentric regions at different distances around the target cell.

Motivation and problem context

Different cancerous mutations alter cellular morphology and the structure of tissue in different ways. This means that cellular morphology can be used diagnostically to predict these mutations, potentially leading to faster and cheaper clinical testing for patients. The current state of the art in this area is to train large, weakly supervised, foundation models on large tissue sections [1]. Here we investigate if a quantum machine learning approach can better capture the relationship between these morphological properties and improve over pixel-based classification approaches.

We use a quantum neural network (QNN) architecture [2] to classify the cancer cell and its mutation type, defining different regions of the image using concentric rings around the cell, as shown in Figure 1, and treating these regions as separate quantum subsystems. Cellular features extracted from each region are encoded in independent quantum registers, and the relationships between the primary cell and its neighbouring regions is modelled with a parametrised interaction Hamiltonian. This approach has two main benefits. First, the interaction coefficients (λ) of the Hamiltonian can be extracted from the trained model and indicate which neighbouring regions contribute most strongly to the classification outcome, potentially revealing biologically relevant spatial patterns such as cell-to-cell communication [3]. Second, this approach does not require pixel-by-pixel encoding of the image, or significant downsampling, allowing the model to train successfully within a 19-qubit budget while avoiding known data-loading bottlenecks and training stability issues [4].

Hamiltonian-structured QNN architecture

The underlying algorithm follows the well-known hybrid QNN structure in which the classical training data $\mathbf{x}$ is encoded into a quantum circuit and then a parameterised unitary operator $U(\theta)$, where θ is a vector of initially

zeroed trainable parameters. The dimension of θ is determined by the data width and number of circuit layers. Then a measurement is performed to extract an output. Here, the output is the Z -polarisation of the zeroth qubit as the task is binary classification. The image is divided into regions, and the data for each region corresponds to a different register of the quantum circuit. Region A having n_A qubits, region B having n_B qubits, and similarly for n_C and n_D , for a total of $n = n_A + n_B + n_C + n_D$ qubits. The features of each region are encoded onto their register using angle embeddings so that the initial state is given by:

$$|\psi(\mathbf{x}_i)\rangle = R_X(\mathbf{x}_{i_A})|0\rangle_A \otimes R_X(\mathbf{x}_{i_B})|0\rangle_B \otimes R_X(\mathbf{x}_{i_C})|0\rangle_C \otimes R_X(\mathbf{x}_{i_D})|0\rangle_D, \quad (1)$$

where $\mathbf{x}_{i_A}$ is a vector of length n_A containing the features associated with region A , and the parameterised single qubit rotation gate is $R_X(\phi) = \exp(-i\sigma_x\phi/2)$. The evolution operator $U(\theta)$ is based on a parameterised Ising model designed to model the relationship between the intra and inter-region features of the image. The corresponding Hamiltonian is defined as:

$$H(\theta) = \lambda_A H_A(\theta) + \lambda_{AB} H_{AB}(\theta) + \lambda_{AC} H_{AC}(\theta) + \lambda_{AD} H_{AD}(\theta) \\ + \lambda_B H_B(\theta) + \lambda_C H_C(\theta) + \lambda_D H_D(\theta), \quad (2)$$

where the $\lambda_\alpha \in \mathbb{R}_{\geq 0}$ are non-negative, real-valued parameters, and the sub-Hamiltonians are given by

$$H_A(\theta) = \sum_{i,j \in n_A} A_{ij}(\theta) \sigma_i^z \sigma_j^z - \sum_{i \in n_A} h_i(\theta) \sigma_i^x \quad (3)$$

and

$$H_{AB}(\theta) = \sum_{\substack{i \in n_A, j \in n_B \\ i-j=n_A}} B_{ij}(\theta) \sigma_i^z \sigma_j^z, \quad (4)$$

and equivalent definitions for the other sub-Hamiltonians H_α . This Hamiltonian defines the parameterised unitary operator via $U(\theta) = e^{-itH(\theta)}$. Each layer of the circuit is a first-order Trotter expansion of $U(\theta)$ with the number of layers treated as a hyperparameter of the model. Although the system is being treated as an interacting quantum system, we set $t = 1$ as there is no actual physical time evolution.

Basing the variational quantum circuit on Eq. (2) allows for interpretability of the model. At the end of training (after the gradient of the loss function flattens below the threshold of the early stopping function or the maximum number of epochs, in our experiments 200), once the θ are learned, the values of λ_α can be extracted and, by construction, interpreted as interaction coefficients of the Hamiltonian. To ensure that the magnitude of the λ_α are comparable, we impose the following constraints:

$$\sum_{i,j \in n_A} |A_{ij}(\theta)| + \sum_{i \in n_A} |h_i| = 1, \quad \sum_{i \in n_A, j \in n_B} |B_{ij}| = 1. \quad (5)$$

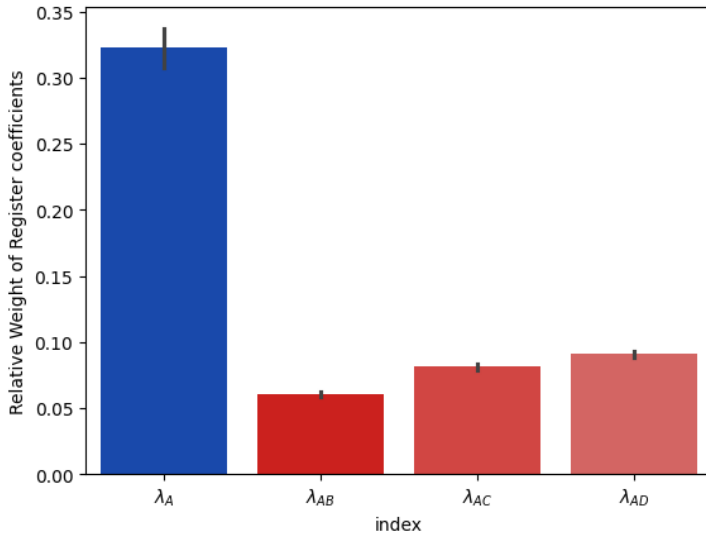


Figure 2: Relative importance of different spatial regions for classification for best found model configuration, error bars are standard deviation over 5 random seeds. Circuit layers = 2, Batch Size = 250, Learning rate = 0.01

Model	F1 Score
QNN	0.891 ± 0.015
Best Classical	0.944

Table 1: Average F1 results for both classification outcomes for the best found model configurations. The error given on the quantum model is the standard deviation over 5 random seeds.

Interpretability

There are several existing works concerned with interpretability in quantum machine learning. In [5] a model-agnostic local (offering explanations at the level of individual data points) method called Q-LIME is developed. Similarly [6] and [7] both present local methods for finding Shapley values accounting for the additional statistical uncertainty from the inherent probabilistic nature and the device noise of quantum computers. These methods are active, meaning that additional computation is required to extract insights, which the authors of [6] note can be prohibitively expensive for larger system sizes. Other approaches [8], include more directly adapting existing classical methods. For example, in [9] a QNN is used for an image classification task followed by gradient weighted class activation mapping.

In contrast, our model provides a model-specific, global, and passive form of interpretability. It is model-specific because the explanation is tied directly to the structure of the ansatz: different spatial regions are encoded into different quantum registers, and the learned coefficients λ_α control the strength of the corresponding intra- and inter-region Hamiltonian terms. It is global because these coefficients describe the relevance of spatial regions across the trained model, rather than explaining a single input image. It is passive because, once training is complete, no additional optimisation, perturbation procedure, or circuit evaluation is required; the relevance scores are read directly from the trained Hamiltonian parameters, up to the normalisation described above in Eq. (5). For the biological application considered here, this form of interpretability is well aligned with the scientific question. We are not only interested in explaining individual classifications, but in understanding which spatial regions around the target cell carry predictive information across the dataset. The resulting relevance scores should therefore be interpreted as model-level spatial relevance indicators, rather than as direct evidence of a biological mechanism. This interpretability mechanism is not generic to arbitrary variational quantum models; it relies on the combination of regional encoding, Hamiltonian interaction structure, and normalisation constraints. However, the framework is flexible, and different choices of region boundaries, feature sets, and Hamiltonian connectivity could be used to test alternative hypotheses about spatial structure in tumour images.

Numerical Results

The above method was applied to a dataset of 2,600 cells (split into 3 folds) from a mouse model of hepatocellular carcinoma (primary liver cancer). Amongst cancerous cells, two different types of gain-of-function (GOF) mutations can occur: KRAS or CTNNB1, together with a loss-of-function (LOF) mutation P53. Each image was centred on a single cancerous cell to be classified as either the KRAS or the CTNNB1 mutation. From each image, 19 features were extracted, four associated with region *A* containing the cell itself, and five with each of regions *B*, *C* and *D*. For each region the features were the average of the cell area, eccentricity, perimeter over the cells within that region and the total count. Region *A* had one fewer feature as, by construction, the cell count is always one. The quantum model was trained with the ADAM optimiser with early stopping [10] using the L^2 -norm as the loss function. To find a performant model, a small grid-search hyperparameter optimisation was performed over the learning rate, batch size and circuit depth parameters, with each of the configurations repeated five times with different random seeds. All quantum model results were obtained via noiseless expectation value simulations using PennyLane [11] and JAX [12]. Classical benchmarking was also performed, using a Super Learner [13] approach in which an ensemble of models (including linear regression, elastic net, random forest, and neural network) are all trained within a certain training budget, with the budget proportional to the number of hyperparameter degrees of freedom of that model type. The training budget was selected such that the classical and quantum models had the same number of trials. We report results for the quality of the models using the F1 metric shown in Table 1.

The values of λ for the best performing model configuration are shown in Figure 2. The error bars denote the standard deviation over the random seeds of the hyperparameter configuration. The value of λ_A is highest, implying that the features of the cell itself have the largest impact on the classification. It is not the next inner-most region *B* that is the most impactful, instead regions *C* and *D*. This proof of concept result shows how non-trivial insights can be obtained from our technique. The λ values are extracted from the model by summing over the parameters of the relevant register for each layer, and then normalising by the number of parameters. As the trainable parameters are initialised to be zero this behaviour arises purely from training and is not an artifact of the quantum circuit structure or initialisation. To demonstrate this, we compare the results for a fully trained model (shown in Figure 2) to the λ values of a model at the beginning of training (epoch = 1) with the same hyperparameter configuration, which has values $\lambda_A = 0.019$, $\lambda_B = 0.002$, $\lambda_C = 0.005$, $\lambda_D = 0.005$ showing no structure.

References

- [1] Q. Li, Z. Hu, Y. Wang, L. Li, Y. Fan, I. King, G. Jia, S. Wang, L. Song, and Y. Li, “Progress and opportunities of foundation models in bioinformatics,” *Briefings in Bioinformatics*, vol. 25, no. 6, p. bbae548, 2024.
- [2] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, “Variational quantum algorithms,” *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021.
- [3] A. Agrawal, “Decoding cell–cell communication using spatial transcriptomics,” *Nature Reviews Genetics*, vol. 26, no. 5, pp. 295–295, 2025.
- [4] M. Larocca, S. Thanasilp, S. Wang, K. Sharma, J. Biamonte, P. J. Coles, L. Cincio, J. R. McClean, and Z. Holmes, “Barren plateaus in variational quantum computing,” *Nature Reviews Physics*, vol. 7, no. 4, pp. 174–189, 2025.
- [5] L. Pira and C. Ferrie, “On the interpretability of quantum neural networks,” *Quantum Machine Intelligence*, vol. 6, no. 2, p. 52, 2024.
- [6] R. Heese, T. Gerlach, S. Mücke, S. Müller, M. Jakobs, and N. Piatkowski, “Explaining quantum circuits with shapley values: towards explainable quantum machine learning,” *Quantum Machine Intelligence*, vol. 7, Feb. 2025.
- [7] P. Steinmüller, T. Schulz, F. Graf, and D. Herr, “Explainable ai for quantum machine learning,” *arXiv preprint arXiv:2211.01441*, 2022.
- [8] L. Power and K. Guha, “Feature importance and explainability in quantum machine learning,” *arXiv preprint arXiv:2405.08917*, 2024.
- [9] F. Mercaldo, G. Ciaramella, G. Iadarola, M. Storto, F. Martinelli, and A. Santone, “Towards explainable quantum machine learning for mobile malware detection and classification,” *Applied Sciences*, vol. 12, no. 23, 2022.
- [10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [11] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. Akash-Narayanan, A. Asadi, *et al.*, “PennyLane: Automatic differentiation of hybrid quantum-classical computations,” *arXiv preprint arXiv:1811.04968*, 2018.
- [12] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, Y. Katariya, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, “JAX: composable transformations of Python+NumPy programs,” 2018.
- [13] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, “Super learner,” *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1.