Benchmarking Attention-based Quantum State Tomography

Abhishek Abhishek
Dept. of Electrical and Computer Engineering
University of British Columbia
Student ID : 29706215
abhiabhi@ece.ubc.ca

Abstract

Quantum tomography is the process of characterizing a quantum system through a series of measurements. Recent experimental realizations of increasingly large and complex quantum information processing devices have led to a need for resource efficient and accurate tomography techniques. An important problem in quantum tomography is the exponential scaling of resources with system size. In this project, we explore a recent work which applies transformer neural networks to improve resource efficiency of quantum state tomography by modelling correlations in measurement outcomes. We perform empirical evaluation of the framework on different quantum states of interest in quantum information processing, and benchmark its performance against standard techniques in quantum state tomography.

1 Introduction

Quantum information processing is experiencing a resurgence of interest in a very similar manner to deep learning. Originally proposed by Richard Feynman in 1982 to simulate physical quantum systems [1], quantum computation is currently being explored to solve challenging problems in many disciplines including biology [2], chemistry [3], and finance [4]. On the experimental frontier, increasingly large and complex quantum devices with longer coherence times and higher fidelities are being developed [5, 6]. Accurate and resource efficient techniques for characterization and validation of these devices is crucial to the development of large-scale quantum computers.

Quantum tomography is the process of characterizing a quantum system by performing a set of different measurements. The aim of quantum state tomography (QST) is to reconstruct the quantum state of a system by analyzing the outcomes of a set of measurements on identically prepared copies of the system. The number of degrees of freedom of a quantum system grows exponentially with system size. This leads to an exponential growth in the number of measurements and the amount of computing resources required to performing exact QST, rendering it intractable for large quantum systems [7].

2 Background

Density matrix A density matrix ρ is a positive semi-definite matrix with $\text{Tr}[\rho] = 1$. It provides the most general description of a many-body quantum system. For a system consisting of N_q qubits, it is a $d \times d$ matrix, where d is the size of the Hilbert space, $d = 2^{N_q}$.

POVM A positive-operator valued measurement (POVM) is a set of $d \times d$ positive semi-definite operators $\{E_x\}, x \in \{1, \ldots, N_m\}$ where each x is a possible measurement outcome and N_m is the no. of possible measurement outcomes associated with the set. Born's rule in quantum mechanics states

Preprint. Under review.

that for a system described by a density matrix ρ, the probability to measure and observe an outcome x is given by $p_\rho(x) = \text{Tr}[E_x\rho]$. For multi-qubit systems with N_q qubits, we often use a tensor-product POVM where measurement outcomes are vectors, $\mathbf{x} = [x^1, x^2, ..., x^{N_q}], \mathbf{x} \in \{1, \ldots, N_m\}^{\otimes N_q}$ and POVM elements are tensor products of single-qubit POVM elements $E_\mathbf{x} = \{E_x\}^{\otimes N_q}, x \in \{1, \ldots, N_m\}$.

Measurement Dataset The measurement outcome dataset $X = \{\mathbf{x}_i\}_{i=1}^N$ consists of N one-shot local measurements where each $\mathbf{x}_i \in \{1, \ldots, N_m\}^{\otimes N_q}$ is a vector sampled from the distribution:

$$\mathbf{p}_\rho(\mathbf{x}) = \text{Tr}[E_\mathbf{x}\rho] \tag{1}$$

where $E_\mathbf{x}$ are elements of the tensor-product POVM. A key thing to note here is that since the measurement outcomes are discrete, $\mathbf{p}_\rho(\mathbf{x})$ is a probability mass function and can be represented as a probability vector of size $N_m^{N_q}$.

Linear inversion Given such dataset X, linear inversion is the simplest procedure to reconstruct the density matrix from measurement outcomes. It requires constructing a probability vector $\tilde{\mathbf{p}}(\mathbf{x})$ which is a frequency-based estimate to $\mathbf{p}_\rho(\mathbf{x})$. This is then inverted using the following equation to obtain an estimate for the density matrix $\tilde{\rho}$:

$$\tilde{\rho} = \sum_{\mathbf{x}',\mathbf{x}''} \tilde{\mathbf{p}}(\mathbf{x}) T^{-1}_{(\mathbf{x}',\mathbf{x}'')} E_{\mathbf{x}''} \tag{2}$$

where $T^{-1}_{(\mathbf{x}',\mathbf{x}'')}$ is an element of the inverse of the POVM T-matrix where each element $T_{(\mathbf{x}',\mathbf{x}'')} = \text{Tr}[E_{\mathbf{x}'} E_{\mathbf{x}''}]$. It should be noted that since $\tilde{\mathbf{p}}(\mathbf{x})$ is a frequency-based approximation to $\mathbf{p}_\rho(\mathbf{x})$, the accuracy of the reconstruction $\tilde{\rho}$ is highly dependent on the no. of measurements performed. This is especially problematic if N is small where the reconstructed density matrix $\tilde{\rho}$ has been observed to not satisfy the positivity and trace conditions. Although $\tilde{\mathbf{p}}(\mathbf{x})$ approaches $\mathbf{p}_\rho(\mathbf{x})$ in the limit $N \to \infty$, in practice only a finite no. of measurements can be performed.

Maximum Likelihood Estimation [8, 9] MLE is another standard procedure to obtain an estimate of the density matrix from measurement data. This is achieved by finding an estimate $\tilde{\rho}$ which maximizes the likelihood

$$\mathcal{L}(\rho) = \prod_j \langle \mathbf{x}_j | \rho | \mathbf{x}_j \rangle^{f_j} \tag{3}$$

where each $\mathbf{x}_j$ is a different possible outcomes, and f_j is its observed relative frequency in the dataset, $j \in \{1, ..., N_m^{N_q}\}$. The MLE estimate for the density matrix is then

$$\tilde{\rho} = \underset{\rho}{\text{argmax}}\ \mathcal{L}(\rho) \tag{4}$$

This procedure is preferred over linear inversion since it easily allow for constraints on positive semi-definiteness and trace of the density matrix to be incorporated in the optimization. Gaussian MLE is a popular variant of standard MLE that considers measurement outcomes as being subjected to an additive Gaussian noise, and has been shown to work well in practice [10].

3 Related work

In recent years, there has been a growing interest in applying ML tools to characterize quantum many-body systems [11]. This is achieved by either reconstructing the quantum wavefunction $|\psi\rangle$ or the full density matrix ρ. In most works, generative models based on Restricted Boltzmann Machines (RBMs) [12–14] and Recurrent Neural Networks (RNNs) [15, 16] are trained on measurement datasets. The goal is to obtain a generative model $\mathbf{p}_\theta(\mathbf{x})$ which represents the many-body wavefunction and provides tractable sampling for efficient calculation of physical properties such as ground-state energies, correlation functions and entanglement entropies.

In one of the earliest works in this area [12], the authors applied RBMs with complex-valued weights to represent many-body quantum wavefunctions. The main idea was that once trained, the RBM can output both the amplitude and the phase of the many-body wavefunction for any configuration of the system. The authors showed the model can be used to find ground states and simulate time-evolution of several quantum many-body systems of interest in physics. The above approach was extended to quantum state tomography (QST) in [13] with two RBM networks $p_\lambda(\mathbf{x})$ and $\phi_\mu(\mathbf{x})$ representing the

amplitude and phase of the wavefunction. The author demonstrated the ability to perform QST of a W state [17] and the transverse-field Ising model. It was shown that unsupervised machine learning approaches are able to reconstruct complex many-body quantum systems with only a limited number of measurements.

Other works in this area include explorations of different neural-network architectures such as convolutional neural networks (CNNs) [18], as well as explorations of deep generative models including variational autoencoders (VAEs) [19] and generative adversarial networks (GANs) [20]. Attention-based Quantum Tomography (AQT) [21, 22] is a recent development which applies transformer neural networks as generative models of quantum many-body systems. The framework is motivated by the success of RNNs for QST, and aims to address their limitations such as restricted ability to capture long-range correlations.

4 Methodology

Attention-based Quantum Tomography (AQT) In this project, we studied AQT which applies transformers [23] to construct a probability vector $\mathbf{p}_T(\mathbf{x})$ as an approximation to $\mathbf{p}_\rho(\mathbf{x})$ (see eq. 1). A key idea behind the model is that for entangled many-body quantum systems, the measurement outcomes $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^{N_q}]$ exhibit short and long-range correlations which is quite similar to the short and long-term correlations exhibited by words in sentences in natural language processing (NLP). Additionally, similar to words, the measurement outcomes for individual qubits $\{x_i^j\}$ come from a dictionary of fixed size, i.e. $x_i^j \in \{1, \ldots, N_m\}$.

Dataset We train on a dataset X consisting of samples from $\mathbf{p}_\rho(\mathbf{x})$ (see Measurement Dataset 2) where each sample is a POVM outcome sequence $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^{N_q}]$ of length N_q. In order to train the transformer, each measurement sequence $\mathbf{x}_i$ is pre-processed by appending start and end tokens i.e. $[1]$ and $[2]$ respectively and adding an offset of 3 to each x_i^j to obtain $\tilde{x}_i^j$. (see figure 1)

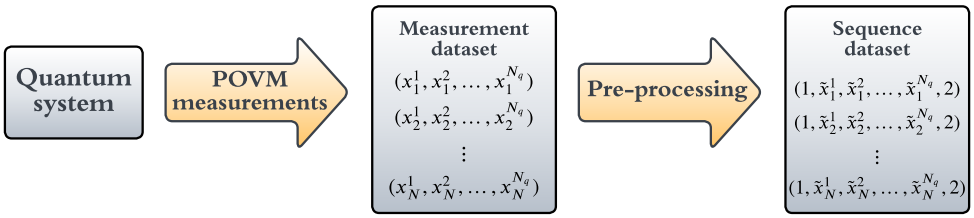

Figure 1: The workflow to obtain the measurement outcome dataset X. A set of POVM measurements (each multiple times) are performed on identically prepared copies of a quantum system.

Model In order to capture correlations in the measurement outcomes of different qubits, the AQT model factorizes $p_T(\mathbf{x})$ as a fully auto-regressive model:

$$\mathbf{p}_T(\mathbf{x}) = \prod_{j=1}^{N_q} \mathbf{p}_\theta(x^j | x^{i<j}) \tag{5}$$

In the framework, a standard transformer decoder consisting of N_l layers of two successive masked multi-headed self attention layers followed by a position-wise feed-forward network is used to parameterize $\mathbf{p}_\theta$. The standard positional encoding scheme from [23] is used to take into account qubit indices. The output of the transformer decoder is passed through a linear projection layer followed by a softmax to yield $\mathbf{p}_\theta(x^j | x^{i<j})$. The model is trained using label smoothing loss [24] which regularizes the softmax prediction of the next token in the sequence. (see figure 2)

POVM Inversion Once trained, the AQT model is used to build a probability table $\mathbf{p}_T(\mathbf{s})$ by passing all possible POVM outcome sequences $\mathbf{s}_i, i = \{1, \ldots, N_m^{N_q}\}$ and computing $\mathbf{p}_T(\mathbf{s})$ under the trained model. Standard linear inversion (Eq. 2) can then be performed in order to obtain a reconstruction of the density matrix ρ. (see figure 3)

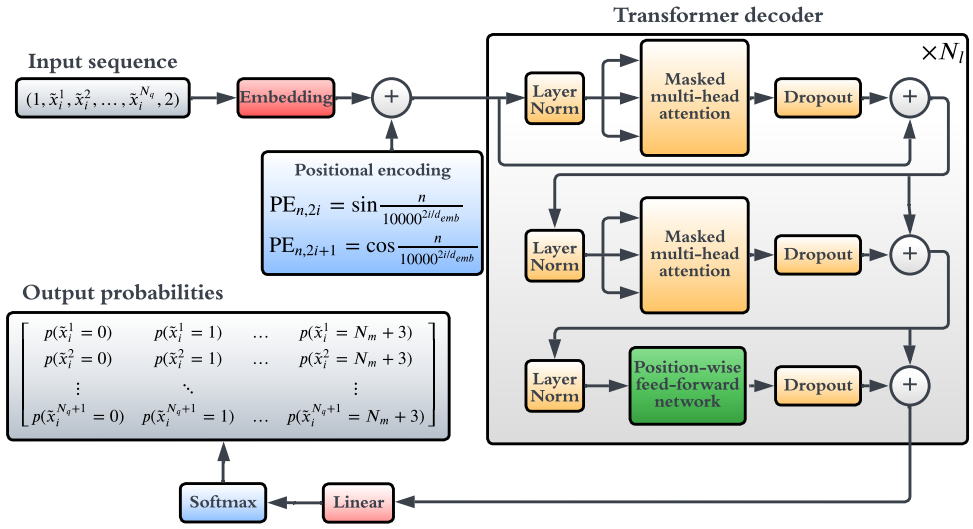

Figure 2: AQT transformer decoder consisting of N_l layers of two successive masked multi-headed self attention layers followed by a feed-forward layer with ReLu activation.

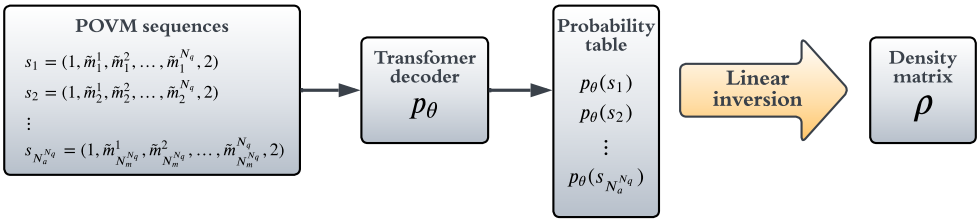

Figure 3: POVM inversion to obtain a reconstruction of the density matrix ρ. The probability of all possible outcome sequences under the model distribution is computed to build the probability table.

5 Experiments

In this project, one of the main goals in addition to understanding the AQT framework, was to benchmark it against standard linear inversion and maximum likelihood estimation. We extend the empirical results of the original work [21] by testing AQT on new quantum states with different forms and amounts of entanglement. We also provide a key comparison to standard techniques in terms of the quantum fidelity which was not included in the original work.

Quantum fidelity (which lies between 0 and 1) is a measure of the "closeness" of two quantum states given by

$$F_Q(\rho_1, \rho_2) = \left(\mathrm{Tr}\left[\sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}}\right]\right)^2 \tag{6}$$

In the case of pure states which we explore in this project, this is simplified to $F_Q(\rho_1, \rho_2) = |\langle\psi_1|\psi_2\rangle|^2$, where $\rho_1 = |\psi_1\rangle\langle\psi_1|$ and $\rho_2 = |\psi_2\rangle\langle\psi_2|$. We also note that although the original work had some qualitative results and evaluation metrics, the accuracy of the reconstruction of the density matrix was only measured in terms of the classical fidelity

$$F_C(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i=1}^{N_m^{N_q}} \sqrt{\mathbf{p}_1(\mathbf{s}_i)\mathbf{p}_2(\mathbf{s}_i)} \tag{7}$$

Classical fidelity only provides an upper bound on the quantum fidelity, and it has been noted that the discrepancy between the two can be substantial [21, 25].

We generated new POVM datasets by simulating the following quantum states using OpenQasm simulator [26] in an open-source framework for quantum computing, Qiskit [27] :

Greenberger-Horne-Zeilinger (GHZ) [28] For a system with N_q qubits, the GHZ state is defined as

$$|\text{GHZ}\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes N_q} + |1\rangle^{\otimes N_q}) \tag{8}$$

It is considered to be a maximally entangled state and is of high interest in quantum information due to its ability to demonstrate non-classical correlations and its relation to Bell's theorem [29]. It is widely used in several quantum communication and cryptography protocols.

W [17] The W states are another class of entangled states with a different multipartitite entanglement structure. For a system of N_q qubits, the W state is defined as

$$|\text{W}\rangle = \frac{1}{\sqrt{N_q}}(|100\ldots0\rangle + |010\ldots0\rangle + \cdots + |00\ldots01\rangle) \tag{9}$$

These also have wide applications in quantum information for tasks such as quantum teleportation [30] and superdense coding [31]. These states exhibit very different correlations in the measurement outcomes than GHZ states, thus making them ideal for testing QST methods.

Equal superposition These states do not have any entanglement and therefore, do not exhibit correlations among measurement outcomes of different qubits. We wanted to understand the impact of a lack of correlations in the input sequence on the reconstruction ability of AQT. Thus, we also prepared the system in an equal superposition state,

$$|+\rangle = \frac{1}{\sqrt{2^{N_q}}}(|0\rangle^{\otimes N_q} + |00\ldots01\rangle + |00\ldots10\rangle + \cdots + |1\rangle^{\otimes N_q}) = \frac{1}{\sqrt{2^{N_q}}}\sum_{i=0}^{2^{N_q}-1}|i\rangle \tag{10}$$

For each of the quantum states described above, we prepared POVM measurement datasets with 2700 and 72900 samples for 3 and 6 qubit systems respectively. The number of samples is determined by the POVM set and the number of unique measurements that can be performed on a system with N_q qubits. For the Pauli6 POVM, this is equal to $3^{N_Q} \times n_{shots}$ and we set $n_{shots} = 100$. We experimented with several different hyperparameter settings and found that the settings listed in table 2 resulted in high quantum fidelities of the reconstructed density matrices across many states. We also compared AQT with and without the post-processing procedure which constructs a new density matrix by minimizing the number of negative eigenvalues of the original reconstructed matrix.

We present the results of the experiments in table 1. We first note that AQT outperforms linear inversion for many states suggesting the probability table constructed using the transformer $\mathbf{p}_T(\mathbf{x})$ is a closer approximation to $\mathbf{p}_\rho(\mathbf{x})$ than a standard frequency-based estimate $\tilde{\mathbf{p}}(\mathbf{x})$, and that the model is able to exploit correlations in the measurement outcomes among different qubits. We do note that calculating the quantum fidelity between the ideal and the AQT reconstructed matrix often results in a value greater than 1. This occurs since the reconstructed density matrix does not satisfy the positive semi-definiteness and trace conditions. We do observe a slight improvement due to the post-processing minimization procedure which is run after the initial reconstruction. Finally, we note that for most of the different states studied, reconstructions by Gaussian MLE resulted in the highest quantum fidelity while satisfying the positivity and trace conditions.

From the empirical studies, we observe that AQT outperforms standard linear inversion in some cases, but when compared to MLE-based approaches, the advantages are unclear. When considering the computational complexity, we note that Gaussian-MLE has a time-complexity of $\mathcal{O}(d^4)$ in the worst-case and $\mathcal{O}(d^3)$ in the case of Pauli measurements, where $d = 2^{N_q}$. When comparing this to AQT, the linear inversion step requires $\mathcal{O}(n^3)$ operations, where $n = N_m^{N_q}$. Thus, both AQT and Gaussian-MLE scale exponentially with system size.

We also perform a qualitative analysis of the reconstructed density matrices using figure 4 where we plot the absolute values of each element. We observe that for the 6-qubit GHZ state, an inaccurate reconstruction directly results in a loss of quantum fidelity due to the missing peaks at matrix elements corresponding to $\{|0\rangle^{\otimes 6}\langle 1|^{\otimes 6}, |1\rangle^{\otimes 6}\langle 0|^{\otimes 6}\}$.

State (qubits)	Linear inversion	Linear MLE	Gaussian MLE [10]	AQT	AQT +min
GHZ (3)	0.934	0.992	**0.999**	0.994	0.995
W (3)	0.946	0.991	**0.999**	1.051	1.026
$\|+\rangle$ (3)	0.959	0.996	**0.999**	0.985	**0.999**
GHZ (6)	0.919	0.992	**0.999**	0.499	0.499
W (6)	0.925	0.992	**0.999**	1.018	1.015
$\|+\rangle$ (6)	0.959	0.996	**0.999**	1.006	1.002

Table 1: Quantum fidelities between the ideal and reconstructed density matrices. Gaussian MLE reconstructions result in highest quantum fidelity in most cases. We note that values greater than 1 for AQT and AQT+min are a result of the reconstructed density matrices not satisfying the positive semi-definiteness and trace conditions.

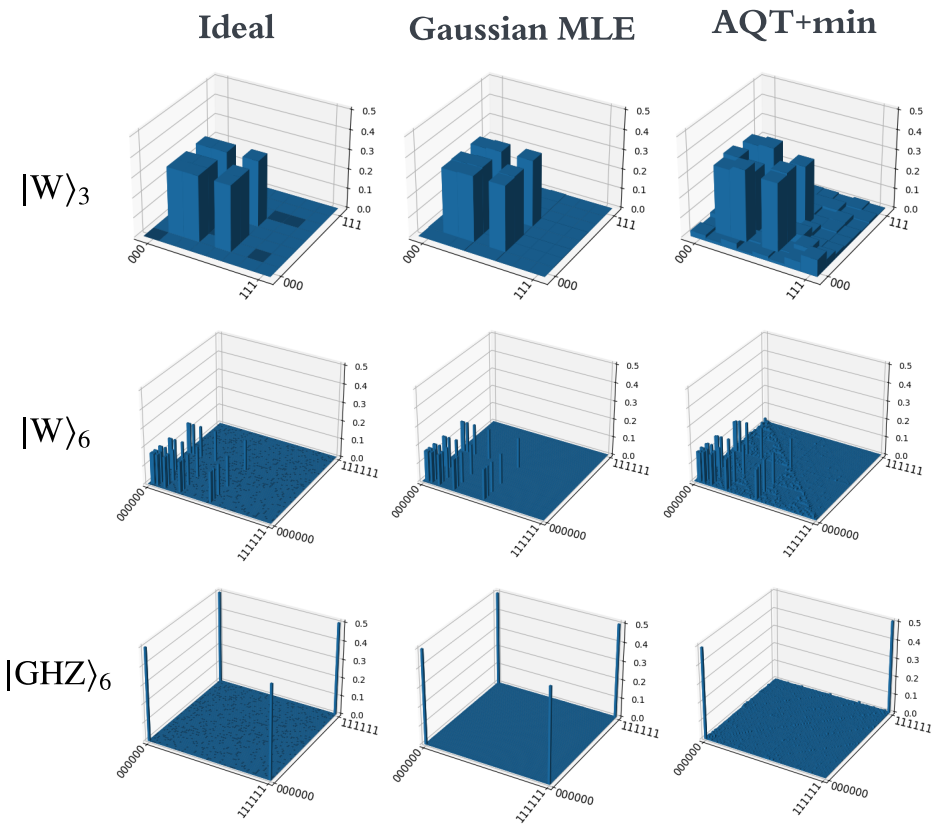

Figure 4: Absolute values of elements of the ideal and reconstructed density matrices for the 3 qubit W, and 6 qubit W and GHZ states. We observed a failure of AQT to reconstruct the GHZ state density matrix in the 6 qubit case.

6 Conclusion and Future Work

In this project, we explored the recently proposed framework of attention-based quantum state tomography. We performed key empirical analysis missing from the original work by calculating

the quantum fidelities of the reconstructed density matrices, and comparing AQT to standard linear inversion and MLE-based approaches on different quantum states. We observed that although AQT outperforms standard linear inversion is most cases, using MLE reconstruction still results in a more accurate and physically valid density matrix. We also note that the density matrices reconstructed by AQT often do not satisfy the positivity and trace conditions, and that this can be clearly observed when using quantum fidelity instead of classical fidelity as an evaluation metric for the reconstruction.

Although we do present some discussion on the complexity of AQT and MLE approaches, empirical evaluation on larger system with more qubits would lead to a better understanding of the accuracy and scaling of these methods in terms of both sample and computational complexity. During the experiments, we observed that a systematic hyperparameter optimization could lead to minor improvements in the quantum fidelity of the reconstructed density matrices, and may help solidify the key takeaways from the empirical analysis. Due to lack of access, we were unable to perform evaluations using actual quantum hardware, but it represents an interesting opportunity for future work due to the operation and measurement noise associated with actual quantum devices. Additionally, in future, incorporation of the physical constraints of the density matrices i.e. positive semi-definiteness and unit trace into the AQT framework could be the key to improve the reconstruction performance significantly.

Code

The work was conducted using a forked version of the open-source package linked with the original work github.com/KimGroup/AQT. We made several additions and changes to the original codebase to perform the empirical analysis presented above and the modified codebase is attached to this report. The (*\notebooks*) directory contains all newly added IPython notebooks to prepare the different quantum states, perform the tomography experiments, collect measurement outcome data, perform standard linear inversion and MLE (*ibm_get_data.ipynb*), compute the quantum fidelities and visualize the results (*results.ipynb*). The (*\circuits*) directory contains newly added state preparation circuit for the W state. The file (*aqt.py*) was modified to allow setting of hyperparameters, timing various components of the AQT pipeline and saving density matrices before and after post-processing. Missing documentation and docstrings were added throughout many files in the repository including (*aqt.py*, *fidelity.py*) and most importantly in (*ann.py*) to explain the computation graph of the transformer model. All experiments were conducted using newly generated simulation datasets, and all results were developed using newly added analysis notebooks.

References

[1] Richard P Feynman. Simulating physics with computers. In *Feynman and computation*, pages 133–153. CRC Press, 2018.

[2] Prashant S Emani, Jonathan Warrell, Alan Anticevic, Stefan Bekiranov, Michael Gandal, Michael J McConnell, Guillermo Sapiro, Alán Aspuru-Guzik, Justin T Baker, Matteo Bastiani, et al. Quantum computing at the frontiers of biological sciences. *Nature Methods*, 18(7):701–709, 2021.

[3] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. *Chemical reviews*, 119(19):10856–10915, 2019.

[4] Dylan Herman, Cody Googin, Xiaoyuan Liu, Alexey Galda, Ilya Safro, Yue Sun, Marco Pistoia, and Yuri Alexeev. A survey of quantum computing for finance. *arXiv preprint arXiv:2201.02773*, 2022.

[5] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[6] Lars S Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob FF Bulmer, Filippo M Miatto, Leonhard Neuhaus, Lukas G Helt, Matthew J Collins, et al. Quantum computational advantage with a programmable photonic processor. *Nature*, 606(7912):75–81, 2022.

[7] Marcus Cramer, Martin B Plenio, Steven T Flammia, Rolando Somma, David Gross, Stephen D Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. Efficient quantum state tomography. *Nature communications*, 1(1):1–7, 2010.

[8] J Řeháček, Z Hradil, and M Ježek. Iterative algorithm for reconstruction of entangled states. *Physical Review A*, 63(4):040303, 2001.

[9] Alexander I Lvovsky. Iterative maximum-likelihood reconstruction in quantum homodyne tomography. *Journal of Optics B: Quantum and Semiclassical Optics*, 6(6):S556, 2004.

[10] John A. Smolin, Jay M. Gambetta, and Graeme Smith. Efficient method for computing the maximum-likelihood quantum state from measurements with additive gaussian noise. *Phys. Rev. Lett.*, 108:070502, Feb 2012.

[11] Juan Carrasquilla. Machine learning for quantum matter. *Advances in Physics: X*, 5(1):1797528, 2020.

[12] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[13] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, 2018.

[14] Roger G Melko, Giuseppe Carleo, Juan Carrasquilla, and J Ignacio Cirac. Restricted boltzmann machines in quantum physics. *Nature Physics*, 15(9):887–892, 2019.

[15] Mohamed Hibat-Allah, Martin Ganahl, Lauren E Hayward, Roger G Melko, and Juan Carrasquilla. Recurrent neural network wave functions. *Physical Review Research*, 2(2):023358, 2020.

[16] Mohamed Hibat-Allah, Roger G Melko, and Juan Carrasquilla. Supplementing recurrent neural network wave functions with symmetry and annealing to improve accuracy. *arXiv preprint arXiv:2207.14314*, 2022.

[17] Wolfgang Dür, Guifre Vidal, and J Ignacio Cirac. Three qubits can be entangled in two inequivalent ways. *Physical Review A*, 62(6):062314, 2000.

[18] Tobias Schmale, Moritz Reh, and Martin Gärttner. Efficient quantum state tomography with convolutional neural networks. *npj Quantum Information*, 8(1):1–8, 2022.

[19] Andrea Rocchetto, Edward Grant, Sergii Strelchuk, Giuseppe Carleo, and Simone Severini. Learning hard quantum distributions with variational autoencoders. *npj Quantum Information*, 4(1):1–7, 2018.

[20] Shahnawaz Ahmed, Carlos Sánchez Muñoz, Franco Nori, and Anton Frisk Kockum. Quantum state tomography with conditional generative adversarial networks. *Physical Review Letters*, 127(14):140502, 2021.

[21] Peter Cha, Paul Ginsparg, Felix Wu, Juan Carrasquilla, Peter L McMahon, and Eun-Ah Kim. Attention-based quantum tomography. *Machine Learning: Science and Technology*, 3(1):01LT01, 2021.

[22] Juan Carrasquilla, Di Luo, Felipe Pérez, Ashley Milsted, Bryan K Clark, Maksims Volkovs, and Leandro Aolita. Probabilistic simulation of quantum circuits using a deep-learning architecture. *Physical Review A*, 104(3):032610, 2021.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[25] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[26] Andrew W Cross, Lev S Bishop, John A Smolin, and Jay M Gambetta. Open quantum assembly language. *arXiv preprint arXiv:1707.03429*, 2017.

[27] Qiskit: An open-source framework for quantum computing, 2021.

[28] Daniel M Greenberger, Michael A Horne, and Anton Zeilinger. Going beyond bell's theorem. In *Bell's theorem, quantum theory and conceptions of the universe*, pages 69–72. Springer, 1989.

[29] Daniel M Greenberger, Michael A Horne, Abner Shimony, and Anton Zeilinger. Bell's theorem without inequalities. *American Journal of Physics*, 58(12):1131–1143, 1990.

[30] Charles H Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K Wootters. Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. *Physical review letters*, 70(13):1895, 1993.

[31] Charles H Bennett and Stephen J Wiesner. Communication via one-and two-particle operators on einstein-podolsky-rosen states. *Physical review letters*, 69(20):2881, 1992.

7 Appendix

No. of qubits N_Q	3	6
Learning Rate	10^{-3}	10^{-3}
Epochs	100	100
Batch Size	100	100
N_l	2	4
d_{emb}	16	128
n_{heads}	4	4

Table 2: Hyperparameter settings observed to result in highest quantum fidelity of the reconstructed density matrices.