**ML4GST**

Author 1, Author 2, Author 3

**Abstract**This is the abstract of the paper. It summarizes the main points and findings of the research.

# 1. Introduction

Content for Introduction

# 2. Problem Setting and Definitions

**Background Theory of Gate Set Tomography.** Different from traditional quantum process tomography, which implicitly assumes near zero state preparation and measurement (SPAM) errors as shown in Figure 1. Gate set tomography relaxes this assumption by directly incorporating gates as both preparation and measurement operators, or formally as preparation and measurement fiducials. In quantum computer characterization setting, rather than probing each individual gate using traditional process tomography, gate set tomography aims to simultaneously reconstruct the full gate set using maximum likelihood method [1]. By measuring the outcomes prescribed by a list of gate sequences that acts to amplify errors of each gate as shown in Figure 2, one can run an optimization algorithm to find out all the parameterized process matrices within the gate set. It is precisely due to this ‘all in one’ tomographic method, that gate set tomography has the highest reconstruction accuracy versus traditional state tomography and process tomography that are largely plagued by the problem of SPAM errors. However, the trade-off of gate set tomography is immediately obvious, where way more computational resources have to be used to solve for this ‘simultaneous maximum likelihood’ across all gate sequences. This can be understood by the fact that maximizing likelihood function in GST is highly non-convex, in stark contrast to state and process tomography, where each observable probability is a linear function of the parameter [1]. Based on this observation, using deep learning techniques to capture complex non-linear relationship would be a natural choice.

**Super Operator Formalism.** Similar to the typical Dirac notation in Hibert space, where a row column vector is represented by a bra , and column vector by a ket . We denote superbra as and superket as . In quantum tomography settings, this conveniently maps a quantum state in the form of density matrix in the -dimensional Hibert space into complex -dimensional vector in Hilbert-Schmidt space, and the inner product .

**Pauli Transfer Matrix (PTM).** In this paper, we use PTM as our super operator, as it is a popular choice in quantum tomography. The PTM basis in Hilbert-Schmidt space has the following properties,

1. Hermiticity:
2. Orthonormality:
3. Traceless for : and

For single qubit, the normalized PTM basis would be . It is also because of this choice of basis, PTM vector and super operator are always real.

As an example, we write a single qubit density matrix as , represent by a real column vector, where each coefficient of can be found by taking the inner product .

To find the measurement probability of projecting onto the computational basis we can perform the standard dot product. First, we write the projectors as row vectors,

And then do a standard dot product (or trace) to obtain the measurement probabilities of getting 0 and 1,

Naturally, for any quantum state vector, we have the super operator that describes (noisy) quantum channel, which is not necessarily unitary and/or orthogonal. For any quantum operator , we have the PTM as,

Where applying a quantum operation/channel to a quantum state is represented as left multiplying a matrix to a vector,

For instance, the PTM of a single qubit rotational gate along x-axis by is given by,

Corresponding to a quantum operation , for is the unitary single qubit operator.

**Deep Neural Network Based Tomography.** The goal of tomography in any general setting is to learn a latent space that maximally matches with the observed data . The nature of this latent spacediffers from application to application. For instance, in medical imaging, by aggregating information from multiple 2D sinograms of CT scan, one can reconstruct a full 2D/3D CT image, whereas the dimension of (reconstructed 2D/3D CT image) is the same or higher than the observed data (2D sinograms). Notably, both and reside in the pixel space, that requires little to no transformation when passed to typical neural network models like convolutional neural network (CNN) [2], [3], [4] or diffusion model[5], [6], [7].

This is vastly different from quantum tomography, where and do not reside in the same space. In the case of quantum state(process) tomography, latent space will be the density(process) matrix that takes the form of a square matrix, where the observed data refers to the probabilities measured from the quantum device, conditioned on a certain measurement and/or preparation operators.

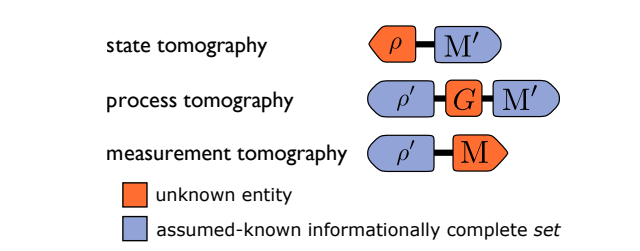
The same mismatch between latent space and observed space continues in gate set tomography, which prevents direct implementation of typical image deep generative models. Instead, an intermediate function has to be used in order to map the latent space to the observed space, an analytical function that maps the neural network output to the expected probabilities. In contrary to most existing publications that directly use deep neural network to reconstruct the full density or process matrix in quantum state or process tomography settings, we aim to predict physical error parameters instead in gate set tomography, and then use analytical function to reconstruct the full process matrices afterwards, as a way to ensure completely positive and trace preserving (CPTP) condition is met and remove the necessity of gauge fixing.

**Proposal.** We try to alleviate this issue by proposing transformer model based deep neural network, which excels in encoding and processing sequences [8], that has also shown promising results in quantum state tomography setting recently [9], [10].

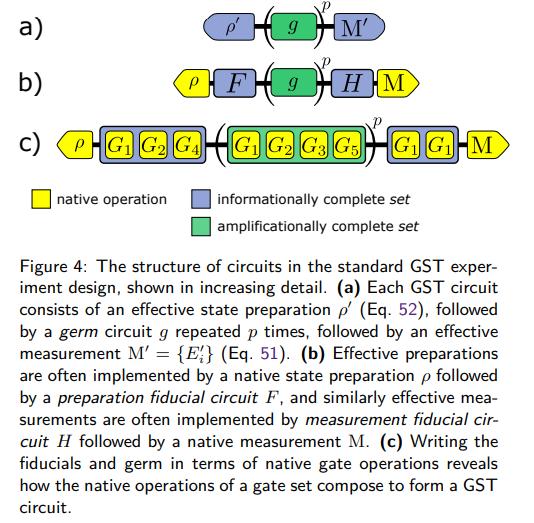
**Introduction to Transformer.** It has been almost 8 years since the publish of the famous paper that invented transformer architecture [8], and now the world is revolutionized by the incredible success and capabilities that GPT-4 [11] and other large language models (LLM) have offered. Here we briefly explain what a transformer is, specifically the encoder block that we used in our paper.

Figure 3 shows the full transformer architecture that is used for text generation in the original implementation, and it can be further divided into an encoder (left) and decoder block (left). Within each block, the main component that empowers transformer is the multi-head attention layer. As its name implies, attention layer is to ‘pay attention’ to the sequences, elements or structures of the input data, similar to what humans do. Mathematically, this is done by constructing matrix-like arrays to assign probability among tokens and performing dot product between the arrays to obtain a weighted sum. Attention mechanism is most commonly seen in LLM to attend sequences of input text, or alternatively, attending structure of images in computer vision [12].

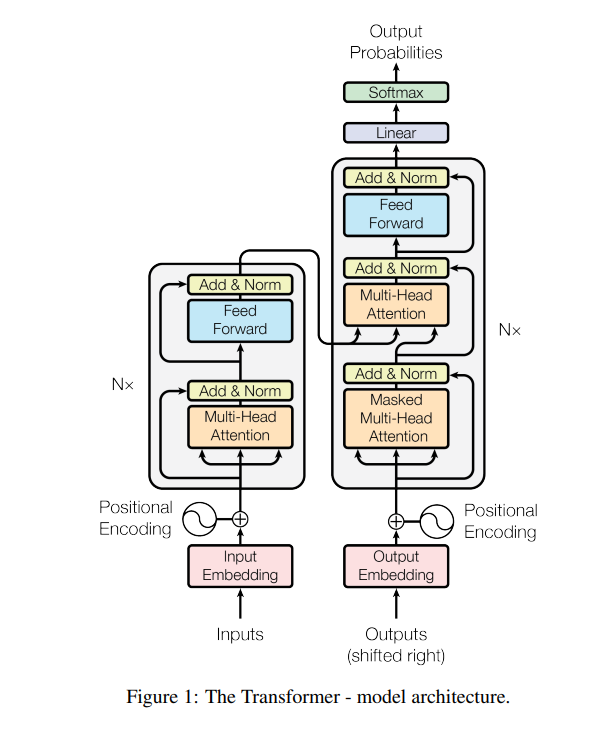
In our paper we only make use of the encoder block to encode input data and subsequently use simple feed-forward network for regression.



Figure



Figure



Figure

# 3. Related Work

**Neural Network Based Quantum Tomography.** The advancement of neural network algorithms and computer hardware in the last decade have brought forth numerous novel applications in the industry, such as autonomous robotics via reinforcement learning [ref], image generation [ref] via diffusion model, text generation via large language model [ref] and so much more. Riding on this trend, quantum physics community has borrowed these techniques from the industry for quantum tomography. Earlier work mainly focuses on using restricted Boltzmann machine for simple quantum state tomography tasks that can be represented by pure states. Later works employ deep neural networks for more difficult tasks such as general density matrix reconstruction [ref] and process matrix construction [ref]. The methods being used ranging from simple feedforward network to more advanced models such as conditional generative adversarial network and transformer model.

**Gate Set Tomography.** There are few works that focus on efficient scaling based on the original of gate set tomography paper. Notably, one paper [13] proposes compressive gate set tomography via Riemannian optimization. However, no known works to date study gate set tomography from a deep neural network perspective.

# 4. Method

Our ML4GST neural network consists of the following main components: 1) separate embedding layers for both integer-encoded gate sequences and measured probabilities, 2) separate positional encoding for both gate sequences and measured probabilities, 3) cross attention layer to aggregate information from two branches, 4) transformer block to encode the aggregated information, 5) fully connected layer to output physical error parameters. A schematic of the neural network architecture is shown in Fig

# 4.1 Separate branch for embedding gate sequences and measured probabilities

In gate set tomography, each gate sequence outputs a set of measured probabilities. By aggregating the information of multiple pairs of (gate sequence, measured probabilities), one can extract the information of the process matrix of each gate used. The gate sequences are first preprocessed by integer encoding and zero padding, where each gate is mapped to a unique integer and the gate sequences are zero padded to match the longest gate sequence in the dataset. The encoded gate sequences are then passed to an embedding layer that is used in typical transformer setting. On the other hand, the measured probabilities are passed to a fully connected layer and subsequent reshaping that emulates the effect of an embedding layer.

This way, both branches will have an extra learnable feature dimension that are ready to be processed by a transformer block later on.

# 4.2 Separate branch for positional encoding

We use the standard sine and cosine positional encoding for both embedded gate sequences and measured probabilities, that are flattened beforehand. As each individual pair of (gate sequence, measured probabilities) yields little tomographic information, we instead group multiple pairs together to increase the receptive field. The positional encoding then runs through element wise for this flattened embedded gate sequences and measured probabilities at each branch.

# 4.3 Cross attention layer

After embedding and positional encoding at each branch, a cross attention layer is used to attend to the relationship between the grouped gate sequences and measured probabilities. We choose cross attention instead of simple concatenation to avoid the vast data shape mismatch between gate sequences and measured probabilities that can possibly drown out the training signal.

# 4.4 Fully connected layers

Finally, multiple fully connected layers are used after the transformer block, to output predicted physical error parameters.

# 5. Experiments

We test out ML4GST using the open-source python package pyGSTi [ref], with the capability of: 1) customizing process matrix of each individual gate within a gate set, 2) selecting appropriate fiducials for the customized gate set, 3) generating appropriate gate sequences for GST experiment, 4) simulating measured counts for the gate sequences.

# 5.1 pyGSTi simulation settings

**1-Qubit XYI model.** pyGSTi python package uses Pauli Transfer Matrix (PTM) as default process matrix throughout GST implementation. Here we replaced the built-in single qubit XYI model pack with our custom PTM, specifically the X and Y rotational gates. These custom gates are parametrized with physical error parameters and in our case, the over-rotational angles and depolarizing errors. We then use the built-in function to find suitable fiducials based on the custom X and Y rotational gates. After that, we run the built-in single qubit XYI GST experiment function to generate gate sequences and simulated measured counts. Number of shots is set to 10k, maximum sequence length to , and sampling error to binomial.

# 5.2 Training details

**Grouping Data.** As mentioned briefly in the method, the gate sequences generated by pyGSTi are converted from strings to unique integers, and subsequently zero padded to match the maximum length of a sequence in the dataset, whereas simulated measured counts are normalized into probabilities. After that, both gate sequences and probabilities datasets are divided into groups, specified by a hyperparameter ‘group\_size’. Typically, the last group will not be divisible and will have a remainder. We simply repeat the elements inside the last group instead of zero padding to preserve overall data quality.

**Curriculum Learning.** We make use of curriculum learning to further divide the whole dataset into parts, again specified by a hyperparameter ‘part\_size’. The dataset is sorted in ascending order based on the non-zero length of gate sequence. This ensures the model learns global features from shorter gate sequences in the beginning and then progressively fine-tune predictions in later stages when it sees longer gate sequences. This learning methodology is similar to the algorithm implemented in the original GST paper [1], where the authors iteratively adding longer sequences accumulatively during optimization. We instead opt for non-accumulative approach in the curriculum learning.

**Analytical PTM Reconstruction.** Based on the predicted physical error parameters, namely the over-rotational angles and depolarizing errors, we analytically reconstruct PTMs corresponding to the gates within the gate set.

**Computing Loss.** As each grouped data outputs one set of physical error parameters prediction, it also has its own set of reconstructed PTMs. We compute probabilities analytically for all the gate sequences within a group, using the same set of reconstructed PTMs. This procedure is performed iteratively group by group within a particular stage set by curriculum learning. Finally, we compute mean squared error loss between the ground-truth probabilities and the reconstructed probabilities. Fig 2. shows the overall data pipeline from input to MSE loss.

# 6. Results

**Choice of Loss Function.**

The original GST paper [1] uses two loss functions for long-sequence GST optimization, the log likelihood function, and the estimator.

where denotes the index of a circuit, and let be the number of outcomes of , the total number of times circuit was repeated, the number of times outcome was observed, the probability predicted by the model of getting outcome from circuit , and is the corresponding observed frequency. The authors made use of estimator as a proxy of during optimization except the last phase, as it is more computationally efficient. Whereas is used in the final phase to steer the estimate to comply with the true statistical derivation. Here, we further simply the estimator to mean-squared error (MSE) loss, that has also been used in simpler linear GST setting. Alternatively, MSE loss can be seen from the perspective of reducing the likelihood function to a normal distribution by invoking central limit theorem [14].

where = is the sampling variance in the measurement.

**Convergence Analysis.**

For depolarizing errors, because we use the tanh activation function at the neural network output layer and subsequently taking absolute value in the custom training loop, the plots shown below will generally have the predicted values jumping between positive and negative. This is intended, as we want the predicted depolarizing error values to be close to and centered at zero, where tanh activation function is the prime candidate.

For both predicted depolarizing errors and over rotational angles, the predicted values exhibit oscillatory behavior at the beginning of each stage of curriculum learning, where an entirely new set of data was fed into the neural network for further training. This is indicated at epoch 90, 190, 263, corresponding to the start of stage 2, 3 and 4. Fig. shows the convergence behavior for depolarizing errors and over rotational angles.

Additionally, we showed that without curriculum learning, the model fails to within the normalized number of epoch, which is equal to the number of epoch for each stage in the curriculum learning. Fig. shows the convergence trajectories without curriculum learning.

**Benchmarking.**

To show that our predicted values are in good agreement with the ground-truth values from simulation, we choose KL divergence**,**  estimator and full Log L function as benchmark. We compare the benchmark results among three cases, ground-truth values, predicted values with curriculum learning and predicted values without curriculum learning (CL) as shown in Table 1.

We first look at MSE, KL divergence and estimator, which all are the functions that measure the distance between probability distributions. It is by no surprise that MSE has the lowest percentage error, as it was used as the loss function during training. We can also see a consistent trend among the three benchmarks, where the results without CL are much worse than the ones with CL, verifying the necessity of CL during training. Besides looking at just the percentage error, the error ratio between the known ‘bad and ‘good’ fit can roughly tell us how large the training signal would be, whereas bigger ratio usually refers to larger training signal. It can be seen that MSE and -Log L functions have large error ratio, meanwhile KL divergence and estimator yield small ratios.

Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | With CL | Without CL | Ground-Truth | % Error Ratio (W.O CL/ CL) |
| MSE (Training) | 1.9668e-05 (-0.08%) | 2.1339e-05 (-8.41%) | 1.9683e-05 (0%) | 105.1 |
| KL divergence | 5.2119e-05 (-2.35%) | 5.6215e-05 (-10.39%) | 5.0923e-05 (0%) | 4.4 |
| estimator | 0.003118 (-2.06%) | 0.003380 (-10.64%) | 0.003055 (0%) | 5.2 |
| -Log L | 16.11541 (-1.86e-5%) | 16.11554 (-9.93e-4%) | 16.11538 (0%) | 53.4 |

# 7. Conclusion

Content for Conclusion

# A. Appendices

Content for Appendices

[1] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, “Gate Set Tomography,” *Quantum*, vol. 5, p. 557, Oct. 2021, doi: 10.22331/q-2021-10-05-557.

[2] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, “CNN-Based Projected Gradient Descent for Consistent CT Image Reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1440–1453, Jun. 2018, doi: 10.1109/TMI.2018.2832656.

[3] D. P. Clark and C. T. Badea, “Convolutional regularization methods for 4D, x-ray CT reconstruction,” in *Medical Imaging 2019: Physics of Medical Imaging*, SPIE, Mar. 2019, pp. 574–585. doi: 10.1117/12.2512816.

[4] E. Kang, J. Min, and J. C. Ye, “A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction,” *Medical Physics*, vol. 44, no. 10, pp. e360–e375, 2017, doi: 10.1002/mp.12344.

[5] B. Huang, L. Zhang, S. Lu, B. Lin, W. Wu, and Q. Liu, “One Sample Diffusion Model in Projection Domain for Low-Dose CT Imaging.” arXiv, Dec. 07, 2022. doi: 10.48550/arXiv.2212.03630.

[6] W. Xia, C. Niu, W. Cong, and G. Wang, “Sub-volume-based Denoising Diffusion Probabilistic Model for Cone-beam CT Reconstruction from Incomplete Data.” arXiv, Mar. 24, 2023. doi: 10.48550/arXiv.2303.12861.

[7] W. Xia, Q. Lyu, and G. Wang, “Low-Dose CT Using Denoising Diffusion Probabilistic Model for 20$\times$ Speedup.” arXiv, Sep. 29, 2022. doi: 10.48550/arXiv.2209.15136.

[8] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jan. 11, 2024. [Online]. Available: https://proceedings.neurips.cc/paper\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[9] P. Cha, P. Ginsparg, F. Wu, J. Carrasquilla, P. L. McMahon, and E.-A. Kim, “Attention-based quantum tomography,” *Mach. Learn.: Sci. Technol.*, vol. 3, no. 1, p. 01LT01, Nov. 2021, doi: 10.1088/2632-2153/ac362b.

[10] H. Ma, Z. Sun, D. Dong, C. Chen, and H. Rabitz, “Tomography of Quantum States from Structured Measurements via quantum-aware transformer.” arXiv, Nov. 17, 2023. doi: 10.48550/arXiv.2305.05433.

[11] OpenAI *et al.*, “GPT-4 Technical Report.” arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2303.08774.

[12] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

[13] R. Brieger, I. Roth, and M. Kliesch, “Compressive Gate Set Tomography,” *PRX Quantum*, vol. 4, no. 1, p. 010325, Mar. 2023, doi: 10.1103/PRXQuantum.4.010325.

[14] D. Greenbaum, “Introduction to Quantum Gate Set Tomography.” arXiv, Sep. 09, 2015. doi: 10.48550/arXiv.1509.02921.