Q-TAPE: A TASK-AGNOSTIC PRE-TRAINED APPROACH FOR QUANTUM PROPERTIES ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Properties estimation for quantum systems is crucial for addressing quantum many-body problems in physics and chemistry. Recently, task-specific deep learning models have exhibited an enhanced capacity to estimate the properties, surpassing the performance of conventional statistical approaches. However, with rapid escalation of quantum computers, existing learning-based models fall short in learning from explosion of quantum data generated by the systems under different physical conditions. Inspired by the triumphs of Large Language Models in Natural Language Processing and Computer Vision, we introduce Q-TAPE, a task-agnostic pre-trained model that 1) facilitates learning of the rich information from diverse quantum systems with different physical conditions in a fully unsupervised fashion; 2) delivers high performance with limited training data, mitigating the cost for quantum data collection and reducing the time for convergence for different supervised tasks. Extensive experiments demonstrate the promising efficacy of Q-TAPE in various tasks including classifying quantum phases of matter on Rydberg atom model and predicting two-body correlation function on anisotropic Heisenberg model. Source code will be made publicly available.

1 INTRODUCTION

Precise estimation of quantum system properties is essential for verifying and evaluating quantum technologies (Huang et al., 2020; Gočanin et al., 2022). However, traditional tomographic techniques struggle for a generic quantum systems due to the exponential complexity inherent in describing quantum many-body systems (Gebhart et al., 2023). Fortunately, physical systems of interest such as those generated by the dynamics of a local Hamiltonian, are not generic, since their particular structure guarantees that the full complexity of Hilbert space is in principle not required for their accurate description (Carrasquilla et al., 2019). Numerous methods have emerged to characterize quantum systems from traditional Density functional theory (DFT) (Hohenberg & Kohn, 1964), Quantum Monte Carlo (QMC) (Ceperley & Alder, 1986), to advanced variational methods e.g. Tensor Networks (TNs) (Orús, 2019) and Neural Network Quantum States (NNQS) (Zhang & Di Ventra, 2023). All these strategies converge on a singular aim: *accurately characterize desired properties of the quantum state using as few identical copies and measurements as possible.*

Variational methods have increasingly become pivotal in addressing quantum many-body problems in recent years (Carleo et al., 2019; Miles et al., 2023). Some methods have already integrated quantum variational algorithms with classical machine learning models, demonstrating feasibility in text (Yang et al., 2022) and image classification (Qi et al., 2023). These methods endeavor to create classical parametric representations for many-body wave functions, with parameters being refined based on the expectation values of relevant observable estimators (Huang et al., 2022b). One notable approach within this domain is Tensor Networks (TNs). Specific frameworks like the Matrix Product State (MPS) (Perez-Garcia et al., 2006) and Projected Entangled Pair States (PEPS) (Corboz, 2016) break down the wave function into multiple tensor components. Another research trajectory leverages the neural networks to serve as universal function for approximating quantum system properties (Carleo et al., 2019; Carrasquilla et al., 2019; Zhang & Di Ventra, 2023). Instead of regarding the properties estimation as a optimization problem, NNQS-based methods frame it as a learning task. These methods optimize variational parameters by leveraging extensive training data pertinent to the studied quantum system. Compared with the TNs, this class of methods can more easily display non-local correlations, allowing in principle to capture quantum states with

higher entanglement (Huang et al., 2022b). Moreover, NNQS can concurrently model multiple quantum states and extract underlying features using akin model structures. However, it's crucial to recognize the present limitations in these variational techniques. TNs and its variants suffer the issue of generalizability, often indicating an inability to harness knowledge from a range of states to effectively reduce sample complexity without compromising estimation accuracy. This limitation is not exclusive to TNs and can be observed in methodologies like the classical shadow (Huang et al., 2020). On the other hand, while NNQS holds immense potential, it remains nascent. The application of advanced machine learning techniques for quantum physics is still an ongoing process. Existing NNQS-based models still fall short in learning from explosion of quantum data generated by the systems govern by different physical conditions. Fortunately, the recent progress in the field of machine learning, i.e. Large Language Models (LLMs), is expected to mitigate this problem.

The power of emerging LLMs (Radford et al., 2018; Brown et al., 2020) can be attributed to their capability to engage in unsupervised pre-training from extensive corpora. The pre-training equips LLMs with notably versatility, facilitating their application to various downstream tasks. In parallel, thanks to the increasing scale of the quantum devices, a vast amount of quantum data are produced by quantum measurement (Brydges et al., 2019). Such data holds intricate details about the quantum system. The emergent challenge is designing a versatile model, analogous to LLMs, that undergoes extensive pre-training to master these quantum intricacies.

Inspired by the concept of LLMs, we introduce a novel task-agnostic pre-trained model named Q-TAPE that can estimate the properties of the quantum system leveraging vast quantum data. We introduce a structured quantum dataset tailored for both digital and analog quantum computers. The quantum data, akin to corpora in linguistic models, serve as the foundation for our model's pre-training. The pre-training is fully unsupervised, empowering the model learn the underlying pattern of the examined quantum system across diverse quantum systems govern by different physical conditions. For the downstream tasks, we fine-tune Q-TAPE on two typical properties estimation tasks including classifying quantum phases of matter and predicting two-body correlation function. We also consider two types of quantum model including the Ryberg atom model and the anisotropic Heisenberg model. The results show its promising power for tackling properties estimation problems especially in scenarios constrained by limited data availability. The contributions are:

1) We delineate a comprehensive set of quantum data readily accessible for digital and analog quantum simulators, as well as for classical simulators of moderate size. These data are used for training a versatile pre-trained model through a fully unsupervised approach. For specialized tasks, the option to gather specific quantum system attributes as the supervised labels is also available.

2) Unlike the development of task-specific models reliant on restricted, task-specific labeled quantum data, Q-TAPE pursues an optimization objective focused on maximizing the expected log likelihood of measurement bit strings. This approach is entirely unsupervised and task-agnostic, making it possible for Q-TAPE to capture the useful pattern of a series of quantum systems. We empirically find that the incorporation of pre-training can more accurately classify quantum phases and predict correlation function on a resource-limited device when limited measurement records are available.

3) The amassed knowledge during the pre-training phase seamlessly transfers to the model when employed for solving specific properties estimation problems. To embed the batch-style discrete measurement records to a continuous space, a trainable LSTM embedding layer is attached to the transformer decoder. The LSTM-Transformer architecture provides an innate framework for handling diverse quantum data stemming from experiments conducted under varying physical conditions, enabling predictions of the quantum properties to which the system is subjected, including those that may exceed the capabilities of modern NISQ hardware for direct simulation.

2 Preliminaries of Quantum State and Quantum Measurement

We introduce basic definitions and annotations of quantum computing. We refer to the work by (Nielsen & Chuang, 2010) for details. We put the details on related work to Appendix A.

Quantum State and Density Operator. The quantum bit named as ***qubit*** is the basic unit of the quantum system. We call the ensemble of all qubits in a (sub)system the ***quantum state***. The qubit is in superposition and becomes deterministic once performing measurement on it. How a quantum state is described mathematically depends on the chosen basis state. For example, by using two

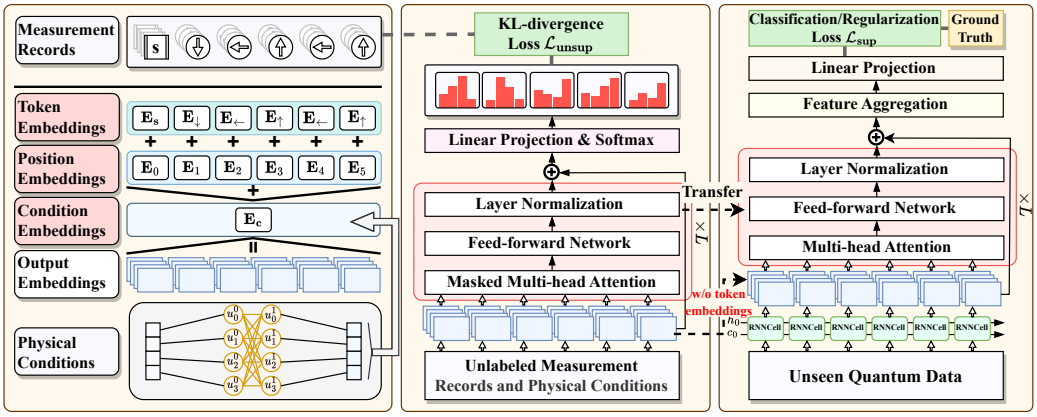

Figure 1: **Pre-training and fine-tuning for Q-TAPE. a)** The output embeddings are the summation of token embeddings, condition embeddings and position embeddings. Three embeddings correspond to encode discrete measurement records, continuous physical conditions and qubit positions, respectively. The token embeddings are replaced with the LSTM embeddings while fine-tuning. **b)** The main part of the model is a multi-layers transformer decoder. Pre-training Q-TAPE is entirely unsupervised. The output target is to approximate the *classical* distribution of the wave function. **c)** The model for fine-tuning and pre-training share the same structure. The pre-trained parameters are transferred to fine-tuning Q-TAPE. All the parameters are optimized by a task's supervised loss.

orthogonal ***computational basis states***[1] $|0\rangle = \begin{bmatrix}1\\0\end{bmatrix}$ and $|1\rangle = \begin{bmatrix}0\\1\end{bmatrix}$, one qubit can be described mathematically as a linear combination $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle = \begin{bmatrix}\alpha\\\beta\end{bmatrix}$ in the space $\mathbb{C}^2$, where $\alpha, \beta \in \mathbb{C}$ are the ***amplitudes*** satisfying $|\alpha|^2 + |\beta|^2 = 1$. An alternate formulation for describing the quantum state is possible using a tool known as the ***density operator*** or ***density matrix***. For example, the density matrix of $|0\rangle$ is $\rho_0 = |0\rangle\langle 0| = \left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right)$ where $\langle 0|$ denotes the conjugate transpose of $|0\rangle$. For a generic L-qubit quantum state with generic basis, it can be described by the so called ***wave function***:

$$|\mathbf{\Phi}\rangle = \sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} \mathbf{\Psi}(\sigma_1, \ldots, \sigma_L)|\sigma_1, \ldots, \sigma_L\rangle, \tag{1}$$

where $\mathbf{\Psi} : \mathbb{Z}^L \to \mathbb{C}$ maps a fixed configuration $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$ of L qubits to a complex number which is the amplitude satisfying $\sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} |\mathbf{\Psi}(\sigma_1, \ldots, \sigma_L)|^2 = 1$, and $\sigma_i \in \{1, \ldots, M\}$ is one of the M possible outcomes by performing quantum measurement on the i-th qubit. It is formulated in a complex Hilbert space where the vector representation of the quantum state $|\mathbf{\Phi}\rangle \in \mathbb{C}^{M^L}$ and its density matrix $|\mathbf{\Phi}\rangle\langle\mathbf{\Phi}| \in \mathbb{C}^{M^L \times M^L}$, which becomes astronomical for large L.

Quantum Measurement. Quantum measurement is a way to observe the quantum system to find out what is going on inside the system, and convert some of the quantum information into classical information that humans can understand. Quantum measurement is described by a set of ***measurement operators*** $\{\mathbf{O}_m\}_{m=1}^{M}$ satisfying $\sum_m \mathbf{O}_m = \mathbf{I}$, where M is the total number of measurement operators. Measuring a qubit leads to collapse of the wave function and produces potentially yield different outcomes. The possible outcomes correspond to the indices m of measurement operators. Concretely, upon measuring the qubit ρ, the probability of getting result m is given by $p(m) = \mathrm{tr}(\rho \mathbf{O}_m)$. For a quantum state with L qubits, performing quantum measurement independently on L qubits is easy to be implemented. The most common strategy is to measure each of the qubits of the quantum system in *parallel* (Leibfried et al., 1996; Jullien et al., 2014). According to the born rule of quantum mechanics, such measurement procedure outputs a measurement string $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$ where $\sigma_i \in \{1, \ldots, M\}$ with probability $|\mathbf{\Psi}(\sigma_1, \ldots, \sigma_L)|^2$ given in Eq. 1.

3 Q-TAPE

As shown in Fig. 1, our model involves two steps: pre-training and fine-tuning. For pre-training, the model is fed with unlabeled $\mathcal{D}_p$, and undergoes fully unsupervised training. Subsequently, the pre-

[1]Computational basis states are also referred to as the Z-basis states in some literature.

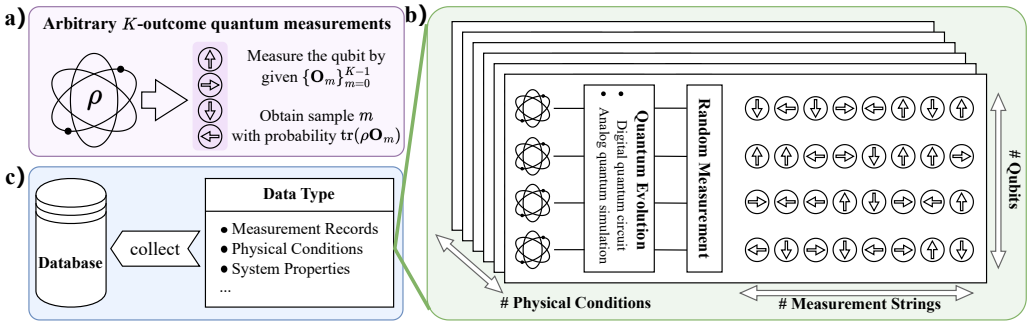

Figure 2: **Process of generating the quantum dataset. a)** For each qubit of the quantum system, we perform quantum measurement using operators $\{\mathbf{O}_m\}_{m=1}^{M}$ and obtain an integer outcome m with probability $p(m)$. **b)** Consider the quantum system govern by different physical conditions. Quantum measurements are performed on an ensemble of identical quantum states evolved under each of fixed physical conditions. Measurement can be done parallel for all the qubits of single copy of the quantum state and outputs a measurement string. This process is applicable and feasible to existing digital and analog quantum computers. **c)** The collected data are structured and packed into a series of tensors, which can be efficiently stored into classical devices and easy to process.

trained parameters are transferred to the supervised learning phase, where all the parameters are fine-tuned using labeled data $\mathcal{D}_t$ and for various downstream tasks. Finally, we evaluate the Q-TAPE's performance using dataset $\mathcal{D}_e$. It is important to note that each downstream fine-tuning model possesses separate parameters, even though they initially share the same pre-trained parameters. One of the most notable aspects of our model is the consistent structural similarity between its components in both pre-training and fine-tuning, with only a few small modifications when handling different downstream tasks. We first give a big picture of the proposed Q-TAPE.

3.1 Motivations and Overview

Q-TAPE is analogue to the Large Language Models (LLMs). Our strategy for building quantum datasets is conceptually equivalent to corpus used to train LLMs. Informally speaking, there is a great conceptual agreement between the type of input data in quantum dataset and that in NLP. Each measurement outcome σ_i of single qubit is analogue to the token and the number of the possible outcomes M is likely to the vocabulary size $|\mathcal{V}|$. A measurement string $\boldsymbol{\sigma}$, which resembles the sentence in texts, is a projection of the entire quantum system with correlative effects among them. The collection of measurement records $\mathbf{R}_i$ comprised of many measurement strings from various physical conditions are equivalent to the corpus gathered from various sources and genres.

These concepts have also been expressed implicitly in Sharir et al. (2020); Hibat-Allah et al. (2020); Cha et al. (2021); Zhang & Di Ventra (2023). However, they either adopted task-specific designs or relied on extensive quantum data pre-processing, potentially introducing biases. Our model, in contrast, derives inspiration from LLMs but tailors the approach specifically for quantum data from current Noisy Intermediate Scale Quantum (NISQ) devices, where the data type and data collection strategy are described in Sec 3.2 and details can be found in Appendix B. Given the genetated datasets, we first discuss how to unsupervisely pre-train Q-TAPE in Sec. 3.3. Afterwards the pre-trained parameters are shared to Q-TAPE with a supervised loss, which is presented in Sec. 3.4.

3.2 Description of the Quantum Dataset Generated From Simulation

In this section, we provide an exposition on the definition and specific details of quantum dataset. We first provide the definition of the quantum dataset in Def. 1 in which the procedures of quantum dataset generation are provided. An easy-to-understand flowchart is also provided in Fig. 2.

Definition 1 (**Quantum Dataset**). *The quantum dataset $\mathcal{D} = \{\mathbf{s}_i\}$ consists of measurement records of quantum states and essential characteristic variables of the quantum system. Each sample $\mathbf{s}_i = (\mathbf{R}_i, \mathbf{c}_i, \mathbf{p}_i)$ contains the measurement records $\mathbf{R}_i$, the physical condition variables $\mathbf{c}_i$ and the (optional) system property variables $\mathbf{p}_i$. Let L denote the number of qubits of quantum systems, K represent the number of copies of the quantum state and M denote the number of possible outcomes by performing measurement on the single qubit. We explain their meaning in detail below.*

1) $\mathbf{c}_i \in \mathbb{R}^C$ *represents the physical condition variables controlling the evolution of the quantum system. These variables can be directly obtained when initializing quantum experiments. The types of the variables are system size, coupling strength or the coefficients of the Pauli string, etc.*

2) The measurement records, denoted as $\mathbf{R}_i \in \mathbb{Z}^{K\times L}$*, are outcomes generated by the quantum measurement. We generate an ensemble of* K *identical quantum states evolved under a fixed physical condition determined by* $\mathbf{c}_i$*. Afterwards quantum measurement is performed independently on each qubit in parallel using a set of measurement operators* $\{\mathbf{O}_m\}_{m=1}^M$*. Performing measurement once on* L *qubits results in a measurement string, represented as* $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$ *where each* $\sigma_l \in \{1, \ldots, M\}$*. The measurement procedures above are performed on each of these* K *copies. Finally, we collect* $K \times L$ *measurement outcomes and store them within* $\mathbf{R}_i$*.*

3) (Optional) Certain system property $\mathbf{p}_i \in \mathbb{R}^P$ *that represents the statistics of the quantum system conditioned on* $\mathbf{c}_i$*, such as the entanglement entropy, purity, correlation function, etc. The exact values of* $\mathbf{p}_i$ *can be calculated by classical post-processing after a number of measurement records are obtained. We treat these properties as supervised labels such that they are only necessary for fine-tuning the machine leaning model.*

3.3 Unsupervised Cross-System Pre-training

Unlike (Czischek et al., 2022; Zhang & Di Ventra, 2023), we do not use task-specific supervised loss or a sampling-based loss to pre-train Q-TAPE. Instead, it is pre-trained in a fully unsupervised manner across quantum systems govern by different physical conditions, as illustrated in Fig. 1b.

Quantum Data for Fine-tuning. For both the Rydberg atom model and the anisotropic Heisenberg model, the quantum dataset $\mathcal{D}_p = \{\mathbf{R}_i, \mathbf{c}_i\}_{i=1}^{N_p}$ used for pre-training is constructed using the strategy discussed in Sec. 3.2. Here we provide how to reorganize the data to adapt for pre-training Q-TAPE. Let K_p be the number of measurement strings used for pre-training. We stack all the input measurement records $\{\mathbf{R}_i\}_{i=1}^{N_p}$ along the first dimension and output $\mathbf{E}_{\text{in}} \in \mathbb{Z}^{N_pK_p\times L}$, where each row $\boldsymbol{\sigma}_b \in \mathbb{Z}^L$ is a measurement sequence. We also construct the matrix $\mathbf{C}_{\text{in}} \in \mathbb{R}^{N_pK_p\times C}$ where each row is the physical condition $\mathbf{c}_b \in \mathbb{R}^C$ determining the system from which the $\boldsymbol{\sigma}_b$ is generated. We vary the value of N_p and K_p to evaluate the Q-TAPE's performance on different training size. Concretely, we consider $N_p \in \{25, 64, 100\}$ and $K_p \in \{64, 128, 256, 512, 1024\}$ for the Rydberg atom model, and $N_p \in \{20, 50, 90\}$ and $K_p \in \{64, 128\}$ for the anisotropic Heisenberg model. For each training iteration, we randomly sample B_p rows of $\mathbf{E}_{\text{in}}$ and $\mathbf{C}_{\text{in}}$. Such that the input of the model is $\{(\boldsymbol{\sigma}_b, \mathbf{c}_b) | \boldsymbol{\sigma}_b \in \mathbf{E}_{\text{in}}, \mathbf{c}_b \in \mathbf{C}_{\text{in}}\}$ with batch size B_p.

Input Embeddings. To handle various of downstream tasks, the input embedding should unambiguously capture the hidden patterns of the quantum system. As depicted in Fig. 1a, we consider three types of embeddings as the input of our model: token embeddings, condition embeddings and position embeddings. Since each element of the measurement string $\boldsymbol{\sigma}_b$ is a discrete integer $\sigma \in \{1, \ldots, M\}$ which resembles to the token in NLP, We use learned embeddings to convert the measurement string $\boldsymbol{\sigma}_b$ with additional start token s and output the token embeddings $\mathbf{E}_t \in \mathbb{R}^{B_p\times(L+1)\times d}$ where d is the feature dimension and B_p is the batch size. Although the token embeddings maintain most information of the quantum systems, we empirically find that encoding the physical condition into the model can improve the performance. A Feed-Forward Network (FFN) with one hidden layer is used to embed the physical condition $\mathbf{c}_b$ into the feature vector $\mathbf{E}_c \in \mathbb{R}^{B_p\times d}$. It is a sentence-level embedding which will be added to all of the L measurement tokens, so that we call it the global embedding. Subsequently, the input embeddings are the (broadcasting) summation given as $\mathbf{E}_{\text{out}} = \mathbf{E}_t + \mathbf{E}_c + \mathbf{E}_p$ where $\mathbf{E}_p$ is the embeddings of the positional encoding as the same as (Vaswani et al., 2017). The embedding $\mathbf{E}_{\text{out}}$ are then processed by deeper layers which will be discussed in detail below.

Model Architecture. As depicted in Fig. 1b, the main part of Q-TAPE is a multi-layer transformer decoder which originates from (Vaswani et al., 2017). The input is the embedding $\mathbf{E}_{\text{out}}$ and the output is $\mathbf{H} \in \mathbb{R}^{B_p\times(L+1)\times d}$, which are high-order representations all the corresponding measurement strings in a batch. Please refer to (Vaswani et al., 2017) for more details on transformer. We primarily report results on the model size which is 8 heads, 4 layers and 128 hidden dimensions. For pre-training, given a fixed qubit configuration $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$, Q-TAPE attempts to approximate the *classical* distribution $p(\sigma_1, \ldots, \sigma_L) = |\boldsymbol{\Psi}(\sigma_1, \ldots, \sigma_L)|^2$ in Eq. 1. The decoder is used to

approximate such joint distribution by factorizing it into a product of conditional probabilities:

$$p(\sigma_1, \ldots, \sigma_L|\mathbf{c}) = \prod_{i=1}^{N} p(\sigma_i|\sigma_{i-1}, \ldots, \sigma_1, \mathbf{c}), \tag{2}$$

which is achieved by masked attention mechanism. We pre-train the Q-TAPE using $B_p = 512$ and optimize the parameters by minimizing the average negative log-likelihood loss over dataset $\mathcal{D}_{\text{pre}}$:

$$\mathcal{L}_{\text{unsup}} = \frac{1}{B_p K_p} \sum_{(\boldsymbol{\sigma}, \mathbf{c}) \in \mathcal{D}_{\text{pre}}} -\log p(\sigma_1, \ldots, \sigma_L|\mathbf{c}), \tag{3}$$

which corresponds to the maximization of (conditional) likelihoods concerning the observed measurement outcomes. It is entirely unsupervised, enabling the model to be trained on extensive quantum datasets that encompass a wide range of physical conditions. To maintain the physical validity that restricts the output distribution to be normalized, a general strategy is employed to fix the last layer as the linear projection with *softmax* activation function, such that the output distribution satisfies $\sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} p(\sigma_1, \ldots, \sigma_L) = 1$ (see Appendix C for proof).

3.4 Fine-tuning for Properties Estimation

Fine-tuning is a straightforward process due to the inherent flexibility of the self-attention mechanism in the Transformer architecture. This flexibility enables Q-TAPE to effectively model a wide range of downstream tasks, whether it involves classifying quantum phases of matter or predicting the entanglement entropy of quantum states. This adaptability is achieved simply by replacing the relevant inputs and outputs as needed. Rather than the *two-step* model (Wang et al., 2022) that uses the pre-trained model to generate new measurement records conditioning on the physical variables and then predicts quantum properties based on classical shadow (Huang et al., 2020). Q-TAPE is an *end-to-end* task-agnostic pre-trained model to provide properties estimation for the quantum system.

Quantum Data for Fine-tuning and Input Embeddings. The dataset $\mathcal{D}_f = \{(\mathbf{R}_j, \mathbf{c}_j), \mathbf{p}_j\}_{i=j}^{N_f}$ are generated using the random seed different from the seed for generating $\mathcal{D}_p$. Then we split $\mathcal{D}_f$ to construct train/test dataset $\mathcal{D}_t$/$\mathcal{D}_e$. We make sure the sampled physical conditions for pre-training will not appear in fine-tuing, i.e. $\mathbf{c}_j \notin \{\mathbf{c}_i\}$ for $j \in \{1, \ldots, N_f\}$. Note that the physical conditions for fine-tuning are sampled from the same distribution as the pre-training. The details about the data collection can be found in Appendix B. Unlike the pre-training phase that the input measurement records is a sentence-level vector $\boldsymbol{\sigma}_b \in \mathbb{Z}^L$, the input of fine-tuning becomes a batch of measurement records $\mathbf{X}_i \in \mathbb{Z}^{L \times K_f}$ where K_f is the number of measurement strings. The reason for such change can be explained through both intuitive and rational perspectives. Intuitively, single measurement string is only a glimpse and cannot reflect the whole picture of the quantum system. Rationally, predicting the properties of the quantum system in classical computers generally requires exponential number of measurements with respect to the system size L (Huang et al., 2022a). Even though for some quantum system with low entanglement, the number stills grows quasi-polynomially with L (Huang et al., 2022b). Accordingly, the input of the model is replaced with $\{(\mathbf{X}_j, \mathbf{c}_j), \mathbf{p}_j\}_{j=1}^{B_t}$ where the tuple $(\mathbf{X}_j, \mathbf{c}_j)$ is the input, $\mathbf{p}_j$ is the corresponding label and B_t is the batch size used for supervised fine-tuning. However, the learned embeddings for embedding the measurement string $\boldsymbol{\sigma}_i$ is not feasible for the batch-style records $\mathbf{X}_j$. To deal with it, a Long Short-Term Memory (LSTM) layer is attached in front of the decoder, as depicted in Fig. 1c. The LSTM layer converts the discrete measurement records $\mathbf{X}_j$ and outputs the high-order embeddings $\mathbf{E}_{\text{rnn}} \in \mathbb{R}^{B_t \times L \times d}$. The additional embeddings including physical condition embeddings and positional embeddings are conserved and transferred from pre-training. Thus the output embedding is given as $\mathbf{E}_{\text{out}} = \mathbf{E}_{\text{rnn}} + \underbrace{\mathbf{E}_c + \mathbf{E}_p}_{\text{transferred}}$.

Feature Aggregation and Output Projection. The output of the L-layer transformer decoder is $\mathbf{H} \in \mathbb{R}^{B_t \times L \times d}$ where d is the hidden dimension of the LSTM and the transformer. For downstream tasks, the decoder is initialized with the same pre-trained parameters and is fine-tuned separately. To obtain the feature representation for each of the B_f systems govern by physical condition $\mathbf{c}_j$, a feature aggregation layer is attached after the last multi-head attention layer. This layer converts the hidden feature $\mathbf{H}$ along the second axis and output $\mathbf{H}' \in \mathbb{R}^{B_t \times d}$. Finally, additional linear projection layer is employed to project the feature into the space $\mathbb{R}^P$ used for prediction, along with a task-dependent activated function which is taken to be *tanh* for predicting the correlation function,

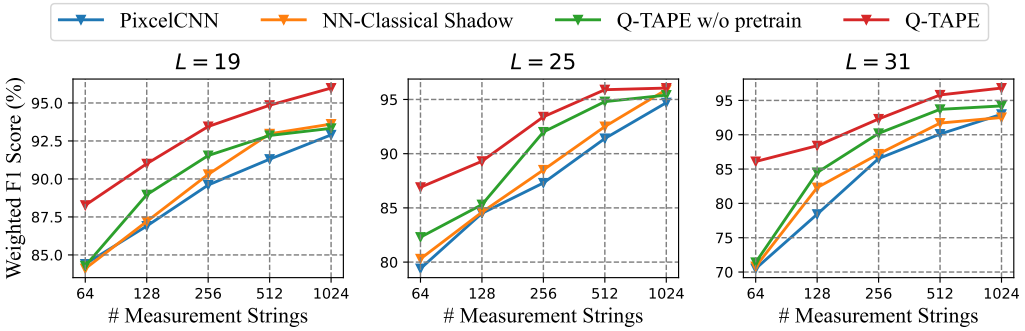

Figure 3: Comparison of weighted F1 score w.r.t. number of measurement strings on Rydberg atom model.

Table 1: Classification accuracy of quantum phases of matter on the Rydberg atom model with varied system size L and varied training size N_t, where K_f is fixed to be 1024. The best results are highlighted in **bold**.

Method	$L = 19$			$L = 25$			$L = 31$		
	$N_t = 25$	$N_t = 64$	$N_t = 100$	$N_t = 25$	$N_t = 64$	$N_t = 100$	$N_t = 25$	$N_t = 64$	$N_t = 100$
RBF Kernel	91.75	92.29	93.25	88.43	92.27	94.2	**88.32**	90.79	92.75
NTK	92.12	92.58	93.79	89.17	94.14	95.39	86.99	92.03	92.71
PixelCNN	92.18	92.79	92.98	88.91	91.59	94.73	85.29	92.21	92.98
Neural-Classical shadow	91.73	92.64	93.61	90.57	91.32	95.91	86.38	91.79	92.51
Q-TAPE	**94.14**	**93.38**	**95.95**	**93.95**	**96.51**	**96.05**	87.95	**94.95**	**96.67**
Q-TAPE w/o pretrain	93.80	92.89	93.35	90.85	95.35	95.27	87.45	92.77	94.32

since we have the prior that each element of the label $\mathbf{p}_j$ is in the range $[-1, 1]$ (See Appendix B for details). While the *log-softmax* is adopted for classifying quantum phases of matter.

Learning Objective. The properties estimation for the quantum system are treated as the supervised learning tasks. Tow types of tasks are considered in this paper, including classifying quantum phases of matter and predicting correlation function. The former belongs to the regression task, while the latter can be regarded as a classification task. For each supervised task, we maintain a consistent architecture within Q-TAPE. We seamlessly integrate task-specific inputs and ground-truth labels into Q-TAPE and proceed to fine-tune all model parameters in an end-to-end manner. Given that the training samples are $\{(\mathbf{X}_j, \mathbf{c}_j), \mathbf{p}_j\}_{j=1}^{B_t}$ where B_t is the batch size. For classifying quantum phases of matter, $\mathbf{p}_j$ is the one-hot label. We minimize the observed data negative log-likelihood which yields a supervised loss for classification (with C classes):

$$\mathcal{L}_{\text{sup}} = -\frac{1}{B_t} \sum_{j \in \{1,\ldots,N_t\}} \sum_{u=1}^{C} \mathbb{I}\left[\mathbf{p}_{j,u} = 1\right] \log\left(f_{\boldsymbol{\theta}}\left(\mathbf{X}_j, \mathbf{c}_j\right)\right), \tag{4}$$

where $\mathbb{I}[\cdot]$ is an indicator function, N_t is the size of training dataset and $f_{\boldsymbol{\theta}}(\cdot)$ denotes the prediction of the model with parameters $\boldsymbol{\theta}$ to be optimized. For predicting the correlation, $\mathbf{p}_j$ is the continuous valued label. We adopt the Root Mean Square Error (RMSE) loss:

$$\mathcal{L}_{\text{sup}} = \sqrt{\tilde{\mathcal{L}}_{\text{sup}}}, \quad \tilde{\mathcal{L}}_{\text{sup}} = \frac{1}{B_t} \sum_{j \in \{1,\ldots,N_t\}} \sum_{u=1}^{C} \left(f_{\boldsymbol{\theta}}\left(\mathbf{X}_j, \mathbf{c}_j\right)_u - \mathbf{p}_{j,u}\right)^2. \tag{5}$$

Detailed task-specific description arew given in Sec. 4 for the respective subsections.

4 Experiments

In this section, we present Q-TAPE fine-tuning results on two quantum properties estimation tasks including predicting correlation functions and classifying quantum phases of matter, where the former belongs to the regression task, while the latter can be regarded as a classification task. Two types of quantum datasets generated from two different quantum models are considered – the Rydberg atom model (Bernien et al., 2017) and the anisotropic Heisenberg model (Kranzl et al., 2023).

As baseline methods, we basically consider the classical shadow (Huang et al., 2020) – a learning-free protocol for constructing the representation of an unknown quantum state. Besides, we compare

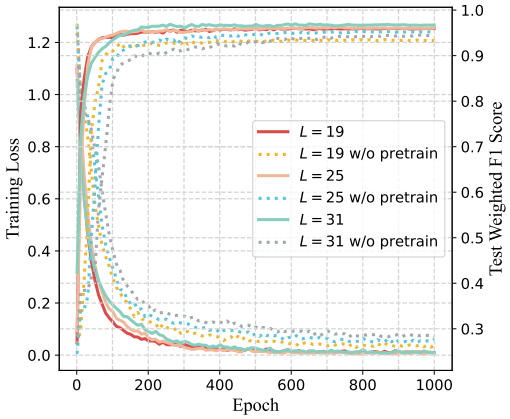

Figure 4: The evolution of training loss and test weighted F1 score with increasing training epochs where $N_t = 100$ and $K_f = 128$.

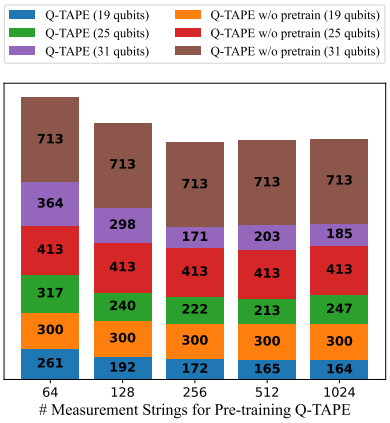

Figure 5: The required number of epochs for each respective model to attain 90% of its peak weighted F1 score where $N_t = 100$.

with some kernel methods including Radial Basis Function (RBF) Kernel (Huang et al., 2022b) and Neural Tangent Kernel (NTK) (Huang et al., 2022b). We further consider some advanced deep learning based methods, such as PixelCNN (Sharir et al., 2020) and a classical shadow based generative model (NN-classical shadow) (Wang et al., 2022) for comparison.

4.1 Classifying Quantum Phases of Matter On Rydberg Atom model

We first consider the Rydberg atom model with different system size L in $(19, 25, 31)$. We pre-train the Q-TAPE for different system sizes separately with a fixed number of sampled physical conditions $N_p = 100$. Each physical condition variable $\mathbf{c}_i$ is a 4-dimensional vector denoted as $[L_i, \Delta_i, \Omega_i, R_0/a_i]^\top$ where Δ is the detuning of a laser, Ω is the Rabi frequency and R_0/a is the interaction range. The values of these four variables can be obtained directly before initializing the (simulated) quantum experiments. For each physical condition we generate $N_p = 1024$ measurement strings based on Pauli-6 measurement operators, such that the total number of possible measurement outcomes is $M = 6$. Then Q-TAPE is pre-trained with dataset $\mathcal{D}_p$. The pre-trained parameters are transferred to fine tune the model, where the number of sampled physical conditions $N_t \in \{25, 64, 100\}$ and the number of measurement strings $K_f \in \{64, 128, 256, 512, 1024\}$ for $\mathcal{D}_t$ used for supervised training both vary. We fix the size of $\mathcal{D}_e$ for evaluation to be $N_e = 400$. Following (Bernien et al., 2017), we consider three categories of quantum phase, i.e., Disorder, Z_2, Z_3 to establish the label $\mathbf{p}_j$, which is a 3-dimensional one-hot vector calculated by `Bloqade.jl` (blo, 2023). More details about the data generation can be found in Appendix B.

We also take evaluation without pre-taining the Q-TAPE: all the parameters are initialized randomly using a uniform distribution $[-1, 1]$. We split the dataset $\mathcal{D}_f$ randomly with different training sizes and varied number of measurement strings and use *accuracy* and *weighted F1 score* as metrics for 3-class classification. The results are listed in Tab. 1 and Q-TAPE achieves the best mean accuracy except for one setting $L = 31$ with $N_t = 25$. Fig. 3 shows the performance on varied input measurement string K_f. Q-TAPE achieves the best F1 score across all systems and in particular, outperforms by a large margin when $K_f = 64$. The results indicate that pre-trained Q-TAPE can handle the input when a few number of measurement records are available, which is greatly instrumental due to the expensive and time-consuming (simulated) physical experiments. We further plot the training dynamics of pre-trained Q-TAPE throughout the training epochs in Fig. 4 for each $L \in \{19, 25, 31\}$, where we also plot the curve for random-initialized Q-TAPE for comparison. Meanwhile, the required number of epoch for the model to attain 90% of its peak weighted F1 score is provided in Fig. 5: within the same system size L, the pre-trained Q-TAPE converges faster than the non-pre-trained version, with a lower training error and a higher test weighted F1 score.

4.2 Predicting correlation Function on Anisotropic Heisenberg Model

Next we consider a regression task – predicting correlation on the anisotropic Heisenberg model. This quantum model inherits the long-range interactions between every two quantum sites, leading to a complex dynamics which is hard to be described by modern computing techniques (Orús, 2019).

Table 2: Root MSE of predicting the correlation on the anisotropic Heisenberg model with varied system size L and training size N_t. K_f is fixed to 64. The best results are in **bold**.

Method	$L = 8$			$L = 10$			$L = 12$		
	$N_t = 20$	$N_t = 50$	$N_t = 90$	$N_t = 20$	$N_t = 50$	$N_t = 90$	$N_t = 20$	$N_t = 50$	$N_t = 90$
Classical Shadow	0.2015±0.013	0.1954±0.016	0.1967±0.015	0.2015±0.011	0.1997±0.012	0.2015±0.021	0.1991±0.027	0.2064±0.022	0.2117±0.019
RBF Kernel	0.2085±0.025	0.2077±0.023	0.2081±0.019	0.2104±0.014	0.2131±0.019	0.2079±0.017	0.2039±0.034	0.1931±0.024	0.2157±0.026
NTK	0.2062±0.018	0.2064±0.026	0.2052±0.017	0.2095±0.013	0.2085±0.018	0.2097±0.018	0.2141±0.031	0.1922±0.024	0.2105±0.022
PixelCNN	0.2257±0.015	0.2357±0.019	0.2239±0.024	0.2393±0.011	0.2289±0.023	0.2108±0.024	0.2390±0.024	0.2297±0.035	0.2267±0.038
Neural-Classical Shadow	0.2069±0.022	0.2098±0.015	0.2057±0.012	0.2078±0.017	0.2054±0.017	0.1959±0.013	0.2037±0.029	0.2021±0.019	0.2102±0.026
Q-TAPE	**0.1761±0.032**	**0.1612±0.022**	**0.1697±0.025**	**0.1986±0.011**	**0.1949±0.012**	**0.1893±0.023**	**0.1989±0.023**	**0.1787±0.021**	**0.1769±0.015**
Q-TAPE w/o pretrain	0.2043±0.027	0.2057±0.036	0.1949±0.027	0.2179±0.015	0.1984±0.013	0.1981±0.025	0.2040±0.028	0.2097±0.031	0.2026±0.027

Table 3: Ablation study results on condition embedding and LSTM embedding. We consider $N_t = 64$ with $K_f = 128$ for the Rydberg model, and $N_t = 50$ with $K_f = 64$ for the Heisenberg model.

Rydberg	$L = 19$	$L = 25$	$L = 31$	Heisenberg	$L = 8$	$L = 10$	$L = 12$
original	**93.38**	**96.51**	**94.95**	original	**0.1612**	**0.1949**	**0.1787**
w/o cond. embed.	93.29	95.96	93.52	w/o cond. embed.	0.1906	0.2095	0.1981
w/o LSTM embed.	90.75	92.18	89.65	w/o LSTM embed.	0.1929	0.1997	0.1904

We restrict the system size L in $(8, 10, 12)$ due to memory limitations and get the ground states of quantum systems with different physical conditions by eigenvalue decomposition. A number of measurement records $K_p = 64$ along with their physical variables using Pauli-6 measurement operators, such that the total number of measurement outcomes is $M = 6$. Then we pre-train the Q-TAPE for varied number of system size independently with training size $N_p = 90$. The data for fine-tuning are generated in a roughly same manner as the fine-tuning. The difference is that we calculate true values of the two-body correlation functions and collect them as the supervised labels, which is a $L \times L$ continuous-valued matrix where each entry is in the range $[-1, 1]$. This matrix is vectorized and stored to be the supervised labels. Correspondingly, the loss per training example is defined as the mean squared error among the $L \times L$ entries between the prediction and the true label.

We vary the number of generated training samples $N_t \in \{20, 50, 90\}$ and fix the measurement strings $K_f = 64$. The RMSE on $N_e = 10$ is reported in Tab. 2. We can see that Q-TAPE achieves the best performance in all settings. Learning-based models often fail to surpass the predictive accuracy of learning-free classical shadow. Intriguingly, the pre-trained Q-TAPE stands out with a remarkable improvement over baseline models. This suggests that the pre-trained Q-TAPE effectively extracts necessarily information and useful patterns for predicting correlations between qubits.

4.3 Further Discussion

We study the effects of condition embedding and the LSTM embedding on both Rydberg atom model and anisotropic Heisenberg model. Note that we replace the LSTM with a fully connected layer with same input/output dimension. The results are in Tab. 3, where the results consistently show that both embedding techniques contribute to some positive effects and suggest that these two techniques can both help to leverage useful information from input quantum data.

5 Conclusion and Outlook

This paper proposes a task-agnostic pre-trained approach for estimation of the properties of the quantum systems via vast quantum data. A transformer encoder, enables to learn useful hidden information in a fully unsupervised pre-training procedure. The pre-trained parameters can be transferred to solving downstream tasks, leading to more effective classifying quantum phases and predicting correlation function on a resource-limited device given limited measurement records.

Limitations. In the present work, we focus on classifying quantum phases of matter and predicting correlation functions for experiments. Though Q-TAPE can be used as a flexible model for other quantum many body problems, such as reconstruct the density matrix, predicting entanglement entropy, etc. Furthermore, we only consider pre-training the model using fixed number of measurement strings. It is intriguing to see how the effects of varied number of measurement strings influences the model's performance. Additionally, the current model may not effectively characterize quantum systems controlled by time-dependent Hamiltonians. The proposed approach is currently limited to estimating properties for time-independent Hamiltonians. Exploring the adaptation of pre-training methods from traditional deep learning for time-series data could be a promising direction in investigating quantum time evolution in many-body systems.

REFERENCES

Bloqade.jl: Package for the quantum computation and quantum simulation based on the neutral-atom architecture., 2023. URL https://github.com/QuEraComputing/Bloqade.jl/.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B AkashNarayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

Alejandro Bermúdez, Luca Tagliacozzo, Germán Sierra, and P Richerme. Long-range heisenberg models in quasiperiodically driven crystals of trapped ions. *Physical Review B*, 95(2):024431, 2017.

Hannes Bernien, Sylvain Schwartz, Alexander Keesling, Harry Levine, Ahmed Omran, Hannes Pichler, Soonwon Choi, Alexander S Zibrov, Manuel Endres, Markus Greiner, et al. Probing many-body dynamics on a 51-atom quantum simulator. *Nature*, 551(7682):579–584, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tiff Brydges, Andreas Elben, Petar Jurcevic, Benoît Vermersch, Christine Maier, Ben P Lanyon, Peter Zoller, Rainer Blatt, and Christian F Roos. Probing rényi entanglement entropy via randomized measurements. *Science*, 364(6437):260–263, 2019.

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

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

Juan Carrasquilla, Giacomo Torlai, Roger G Melko, and Leandro Aolita. Reconstructing quantum states with generative models. *Nature Machine Intelligence*, 1(3):155–161, 2019.

David Ceperley and Berni Alder. Quantum monte carlo. *Science*, 231(4738):555–560, 1986.

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.

Philippe Corboz. Variational optimization with infinite projected entangled-pair states. *Physical Review B*, 94(3):035133, 2016.

Stefanie Czischek, M Schuyler Moss, Matthew Radzihovsky, Ejaaz Merali, and Roger G Melko. Data-enhanced variational monte carlo simulations for rydberg atom arrays. *Physical Review B*, 105(20):205108, 2022.

G Mauro D'Ariano, Matteo GA Paris, and Massimiliano F Sacchi. Quantum tomography. *Advances in imaging and electron physics*, 128:206–309, 2003.

Yuxuan Du, Yibo Yang, Tongliang Liu, Zhouchen Lin, Bernard Ghanem, and Dacheng Tao. Shadownet for data-centric quantum system learning. *arXiv preprint arXiv:2308.11290*, 2023.

Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. *Nature Reviews Physics*, 2(7):382–390, 2020.

Valentin Gebhart, Raffaele Santagati, Antonio Andrea Gentile, Erik M Gauger, David Craig, Natalia Ares, Leonardo Banchi, Florian Marquardt, Luca Pezzè, and Cristian Bonato. Learning quantum systems. *Nature Reviews Physics*, 5(3):141–156, 2023.

Aleksandra Gočanin, Ivan Šupić, and Borivoje Dakić. Sample-efficient device-independent quantum state verification and certification. *PRX Quantum*, 3(1):010317, 2022.

James Gubernatis, Naoki Kawashima, and Philipp Werner. *Quantum Monte Carlo Methods*. Cambridge University Press, 2016.

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.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

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.

Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022a.

Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V Albert, and John Preskill. Provably efficient machine learning for quantum many-body problems. *Science*, 377(6613):eabk3333, 2022b.

T Jullien, P Roulleau, B Roche, A Cavanna, Y Jin, and DC Glattli. Quantum tomography of an electron. *Nature*, 514(7524):603–607, 2014.

Florian Kranzl, Stefan Birnkammer, Manoj K Joshi, Alvise Bastianello, Rainer Blatt, Michael Knap, and Christian F Roos. Observation of magnon bound states in the long-range, anisotropic heisenberg model. *Physical Review X*, 13(3):031017, 2023.

Dietrich Leibfried, DM Meekhof, BE King, CH Monroe, Wayne M Itano, and David J Wineland. Experimental determination of the motional quantum state of a trapped atom. *Physical review letters*, 77(21):4281, 1996.

Laura Lewis, Hsin-Yuan Huang, Viet T Tran, Sebastian Lehner, Richard Kueng, and John Preskill. Improved machine learning algorithm for predicting ground state properties. *arXiv preprint arXiv:2301.13169*, 2023.

Cole Miles, Rhine Samajdar, Sepehr Ebadi, Tout T Wang, Hannes Pichler, Subir Sachdev, Mikhail D Lukin, Markus Greiner, Kilian Q Weinberger, and Eun-Ah Kim. Machine learning discovery of new phases in programmable quantum simulator snapshots. *Physical Review Research*, 5(1): 013026, 2023.

Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

Román Orús. Tensor networks for complex quantum systems. *Nature Reviews Physics*, 1(9):538–550, 2019.

David Perez-Garcia, Frank Verstraete, Michael M Wolf, and J Ignacio Cirac. Matrix product state representations. *arXiv preprint quant-ph/0608197*, 2006.

Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, and Min-Hsiu Hsieh. Pre-training tensor-train networks facilitates machine learning with variational quantum circuits. *arXiv preprint arXiv:2306.03741*, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of physics*, 326(1):96–192, 2011.

Or Sharir, Yoav Levine, Noam Wies, Giuseppe Carleo, and Amnon Shashua. Deep autoregressive models for the efficient variational simulation of many-body quantum systems. *Physical review letters*, 124(2):020503, 2020.

Y-Y Shi, L-M Duan, and Guifre Vidal. Classical simulation of quantum many-body systems with a tree tensor network. *Physical review a*, 74(2):022320, 2006.

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.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Haoxiang Wang, Maurice Weber, Josh Izaac, and Cedric Yen-Yu Lin. Predicting properties of quantum systems with conditional generative models. *arXiv preprint arXiv:2211.16943*, 2022.

Steven R White. Density matrix formulation for quantum renormalization groups. *Physical review letters*, 69(19):2863, 1992.

Dian Wu, Lei Wang, and Pan Zhang. Solving statistical mechanics using variational autoregressive networks. *Physical review letters*, 122(8):080602, 2019.

Ya-Dong Wu, Yan Zhu, Yuexuan Wang, and Giulio Chiribella. Learning and discovering quantum properties with multi-task neural networks. *arXiv preprint arXiv:2310.11807*, 2023.

Tailong Xiao, Jingzheng Huang, Hongjing Li, Jianping Fan, and Guihua Zeng. Intelligent certification for quantum simulators via machine learning. *npj Quantum Information*, 8(1):138, 2022.

Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen. When bert meets quantum temporal convolution learning for text classification in heterogeneous computing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8602–8606. IEEE, 2022.

Yuan-Hang Zhang and Massimiliano Di Ventra. Transformer quantum state: A multipurpose model for quantum many-body problems. *Physical Review B*, 107(7):075147, 2023.

Yan Zhu, Ya-Dong Wu, Ge Bai, Dong-Sheng Wang, Yuexuan Wang, and Giulio Chiribella. Flexible learning of quantum states with generative query neural networks. *Nature Communications*, 13 (1):6222, 2022.

A Related Work

A.1 Properties Estimation of Quantum Systems

Determining the properties of the quantum system is a long-standing problem in quantum physics (D'Ariano et al., 2003). The main challenge is that the complexity of describing the quantum system using classical computers scales exponentially with respect to the system size (Nielsen & Chuang, 2010). Even though, in fact, the quantum systems studied in physical experiments generally can be described by a limited number of physical variables. This restriction leads to the studied quantum systems occupy only a small part of the exponentially large Hilbert space (Carrasquilla et al., 2019), such that they can be characterized by some classical methods within an acceptable error.

Traditional algotithms including the QMC (Ceperley & Alder, 1986) and DFT (Hohenberg & Kohn, 1964) has made success for investigating the electronic structure (or nuclear structure), principally the ground state of many-body systems, such as atoms, molecules, and the condensed phases (Gubernatis et al., 2016). However, these methods have scalability issues and are difficult to be used to deal with large-scale quantum many body problems. An alternative is a class of TNs methods (Orús, 2019) based on variational method and shows unprecedented performance in analyzing the characteristics of ground state. This class of methods approximates the wave function by decomposition of the high-order tensor into multiple low-rank tensors. It is then possible to predict properties of the quantum state by the classical parametric representations of the many-body wave function.

With the continuous development of machine learning technologies, the tensor used in TNs is replaced with a neural network as a parametric function approximator of many-body wave functions. Different neural network ansatz corresponds to solve quantum many-body problems with different physical structures. For example, the 1-D array chain and the 2-D chain could be processed by the recurrent neural networks (RNNs) (Gubernatis et al., 2016) and the convolutional neural networks (CNNs) (Wu et al., 2019; Sharir et al., 2020), respectively.

A.2 Neural Representation of Quantum States

A comprehensive characterization of modern quantum devices entails the retrieval the useful information from measurements on identically prepared copies of quantum states (Huang et al., 2022b). Machine learning has been introduced to learn from experimental data and then applying that knowledge to predict physical properties and even reconstruction of the quantum state (Carleo et al., 2019; Gebhart et al., 2023). Compared with the learning-free tensor network-based methods (Shi et al., 2006; White, 1992; Schollwöck, 2011), machine learning can effectively address some quantum many-body problems that would be too hard to solve using classical processing alone (Huang et al., 2022b). Numerous studies have been carried out testing different types of architectures; examples include restricted Boltzmann machine (RBM) (Carleo & Troyer, 2017), recurrent neural networks (RNNs) (Carrasquilla et al., 2019), convolutional neural networks (CNNs) (Wu et al., 2019; Sharir et al., 2020), and transformers (Cha et al., 2021; Wang et al., 2022; Zhang & Di Ventra, 2023).

In terms of the optimization strategy for training the parameters, the methods can be separated in to two categories. The first is to use sampling-based optimization techniques (Cha et al., 2021; Zhang & Di Ventra, 2023) such as Variational Monte Carlos (VMC) (Ceperley & Alder, 1986). This class of methods aim to reconstruct the entire quantum state wave function with high fidelity. The second class is to use end-to-end auto-differentiation techniques to update the parameters (Eisert et al., 2020; Xiao et al., 2022; Wu et al., 2023; Du et al., 2023). This class of methods only concern certain type properties of the quantum system without reconstruct the entire quantum state wave function. Although some efforts claim that the properties of quantum systems can be predicted by pre-trained models, their optimization object is often supervised. Such that they are difficult to deal with large-scale quantum data sets (often without supervised labels). The recent work proposed by Zhu et al. (2022) implements a similar pre-training strategy for learning of quantum states, whereas our approach differs from it by avoiding assumptions about knowing the prior frequency about the measurement strings.

Our work is closely related to second class of methods. While ours employ a unsupervised pre-training phase to extract the hidden information of the quantum systems govern by different pa-

rameters. We find empirically that this scheme can make the model perform better under a limited number of copies of quantum states and measurements.

B Details of the Quantum Dataset Generation

In a word, a quantum dataset is a collection of data that describes quantum systems and their evolution. The collection of quantum data must take into account the following factors: 1) the method of data collection must be feasible on quantum devices and not contradict the disciplines of quantum mechanics; 2) the process of data collection is completely automated and does not require experienced experts to organize and label it and 3) the data must be structured and can be stored on resource-limited classical devices, thus can be easy to be processed by the machine learning techniques without further post-processing. The quantum dataset we established satisfies these three points. It is also worth mentioning that our model can be used as a centralized infrastructure to process all these data uniformly, thanks to the unsupervised learning design of the model.

In this paper, we conduct simulated experiments to generate the quantum dataset in classical computers. For the anisotropic Heisenberg model, quantum measurement is performed using the Pauli-6 measurement operators such that $M = 6$, whereas computational basis measurement operators are employed for the Rydberg atom model leading to $M = 2$. Assume that variables $\mathbf{c}_i$ describing the physical condition lives in a finite continuous space $\mathbb{F}$ within the physical restriction. Let $\mathcal{D}_p = \{\mathbf{R}_i, \mathbf{c}_i\}_{i=1}^{N_p}$ denote the quantum dataset used for pre-training and $\mathcal{D}_f = \{(\mathbf{R}_i, \mathbf{c}_i), \mathbf{p}_i\}_{i=1}^{N_f}$ for fine-tuning, where $|\mathcal{D}_p| = N_p$ and $|\mathcal{D}_f| = N_f$. For pre-training the model, we first uniformly sample a number of points $\{\mathbf{c}_i | \mathbf{c}_i \in \mathbb{F}\}_{i=1}^{N_p}$. Afterwards we conduct simulated experiments for each $\mathbf{c}_i$ and collect the corresponding measurement records. The system property $\mathbf{p}_i$ is not needed since the pre-training phase is fully supervised. While for fine-tuning, we replace another random seed and sample N_f physical conditions also within space $\mathbb{F}$, resulting in $\{\mathbf{c}_j | \mathbf{c}_j \in \mathbb{F}\}_{j=1}^{N_f}$. Note that We also collect the measurement records for each $\mathbf{c}_j$. The difference part is that we additionally calculate the system property $\mathbf{p}_j$ based on the collected measurement records and use it as supervised labels. We further split the $\mathcal{D}_f$ into $\mathcal{D}_t$ and $\mathcal{D}_e$ for training and evaluation respectively with varied separation ratio. Please refer to the corresponding subsections in Sec. 4 and Appendix B for details of the experimental configurations for dataset generation.

B.1 Rydberg Atom Model

Rydberg atom model is a programmable quantum simulators capable of preparing interacting qubit systems (Bernien et al., 2017). Such quantum model can be effectively described as a two-level quantum system consisting the ground state $|g\rangle$ ($|0\rangle$) and the Rydberg state $|r\rangle$($|1\rangle$). The quantum dynamics of this model is governed by the Hamiltonian

$$H_{\text{Rydberg}} = \sum_i \frac{\Omega}{2}\sigma_x^i - \sum_i \Delta n_i + \sum_{i<j} \frac{V_0}{|\vec{x}_i - \vec{x}_j|} n_i n_j \tag{6}$$

where σ_x is the PauliX matrix, Ω is the Rabi frequency, Δ is the detuning of a laser, V_0 is the Rydberg interaction constant, i, j is the Rydberg interaction constant and $\vec{x}_i$ is the position vector of the site i. $n_i = |r_i\rangle \langle r_i|$ is the occupation number operator at site i, and $\sigma_x^i = |g_i\rangle\langle r_i| + |r_i\rangle\langle g_i|$ describes the coupling between the ground state $|g_i\rangle$ and the Rydberg state $|r_i\rangle$ at position i.

We follow the recent work in (Wang et al., 2022) to generate the quantum dataset. We refer the readers to their paper for details. Here we briefly introduce the main procedures. We consider the Rydberg atom model with system size L in $\{19, 25, 31\}$. We fix the interaction constant $V_0 = 862690 \times 2\pi$ MHz μm^6 and vary the value of $\Omega \in [0, 5]$ and $\Delta \in [-10, 15]$ to get different physical conditions $\mathbf{c}$, where $\mathbf{c}$ is a 4-dimensional vector in the form $[L, \Delta, \Omega, R_0/a]$, where R_0/a denote the interaction range with $R_0 = (V_0/\Omega)^{1/6}$. Then the approximate ground state for diffident physical condition is prepared by the tool `Bloqade.jl` (blo, 2023). This tool can also output the measurement strings and the true phase of each physical condition. The measurement operators are chosen to be the computational basis $\{|0\rangle\langle 0|, |1\rangle\langle 1|\}$ for the quantum measurement, such that the total number of the possible outcomes is $M = 2$. In this paper, three different phases are considered including the Disordered phase, Z_2 Ordered phase and Z_3 Ordered phase. We sample $N_p = 100$ physical conditions with $K_p = 1024$ measurement strings for pre-training, and $N_t \in \{25, 64, 100\}$

physical conditions with $K_t \in \{64, 128, 256, 512, 1024\}$ for fine-tuning. The number of physical conditions for evaluation is fixed to be $N_e = 400$. The supervised labels for fine-tuning are one-hot encoded vectors of the true phases such that the dimension (number of classes) of $\mathbf{p}$ is 3. Note that it is ensured that the sampled physical conditions for pre-training will not appear in fine-tuing.

B.2 Anisotropic Heisenberg Model

Exploring the effects of these long-range interactions of the quantum system is essential for understanding the quantum mechanics (Bermúdez et al., 2017). In this paper, we consider the recent progress for the long-range interactions with the experimentally realized power-law exponent of the anisotropic Heisenberg model (Kranzl et al., 2023). The dynamics of the anisotropic Heisenberg model is determined by the Hamiltonian

$$H_{\text{Heisenberg}} = \frac{1}{3} \sum_{i<j} J_{ij} (\sigma_x^i \sigma_x^j + \sigma_y^i \sigma_y^j + h \sigma_z^i \sigma_z^j), \tag{7}$$

where $\sigma_{x,y,z}^i$ is the Pauli matrix operated on the i-th site, h determines the Ising interactions between the magnons, and J_{ij} is the long-range interaction strength satisfying $J_{ij} = J/|i-j|^\alpha$. We follow the configuration of (Kranzl et al., 2023) to geenrate the quantum dataset. The values of h and J are fixed with 1 and 369 rad/s, and we vary the value of $\alpha \in (1, 2]$ uniformly. It is extremely hard to characterize the quantum system with long-range interactions using the existing computing techniques. Thus we restrict the system size $L \in \{8, 10, 12\}$. For all the system we consider the number of measurement strings used for pre-training as $K_p = 64$ and vary the number of sampled physical condition N_p in $\{20, 50, 90\}$. The number of sampled physical conditions for evaluation is set to be $N_e = 10$. The physical condition $\mathbf{c}$ is defined as a vector whose dimension $C = L^2$, in which each element is the coupling strength J_{ij} for $i, j \in \{1, \ldots, L\}$. The problem of finding the ground state is viewed as the eigenvalue decomposition problem and we obtain the ground state for each sampled physical condition by the `scipy` (Virtanen et al., 2020) built-in functions. The measurement records and the true values of the two-body correlation function are obtained using the `pennylane` (Bergholm et al., 2018) toolbox. For the Anisotropic Heisenberg Model, we consider the Pauli-6 POVM measurement operators with $M = 6$ outcomes, which are given as $M_{\text{Pauli-6}} = \left\{\frac{1}{3} \times |0\rangle\langle 0|, \frac{1}{3} \times |1\rangle\langle 1|, \frac{1}{3} \times |+\rangle\langle +|, \frac{1}{3} \times |-\rangle\langle -|, \frac{1}{3} \times |r\rangle\langle r|, \frac{1}{3} \times |l\rangle\langle l|\right\}$, and $\{|0\rangle, |1\rangle\}$, $\{|+\rangle, |-\rangle\}$, $\{|r\rangle, |l\rangle\}$ stand for the eigenbases of the Pauli operators σ_z, σ_x, and σ_y, respectively. The output two-body correlation function is a $L \times L$ matrix and each element of the matrix is the expectation value of the observable

$$O_{ij} = \frac{1}{3} \left(\sigma_x^i \sigma_x^j + \sigma_y^i \sigma_y^j + \sigma_z^i \sigma_z^j\right). \tag{8}$$

Thus each element can be written as $\text{tr}(\rho O_{ij})$ in the range $[-1, 1]$, where ρ is the density matrix of the ground state for each sampled physical condition. We flatten the correlation function matrix to be the L^2-dimensional continuous-valued vector and treat it as the supervised label for fine-tuing.

C Poof of the Normalized Output Distribution

In the main text, we claim that the output (classical) distribution satisfies

$$\sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} p(\sigma_1, \ldots, \sigma_L) = 1, \tag{9}$$

as long as the last linear projection layer uses the *softmax* activated function. The proof is given below.

The softmax activated function is performed on the model's output, which is the product of conditional probabilities $p(\sigma_1, \ldots, \sigma_L|\mathbf{c}) = \prod_{i=1}^{N} p(\sigma_i|\sigma_{i-1}, \ldots, \sigma_1, \mathbf{c})$. It is easy to check the claim holds for $L = 1$. Given that the claim also holds for $L = k$. For $L = k + 1$, the following equation then be hold:

$$\sum_{i=1}^{N} p(\sigma_i|\sigma_{i-1}, \ldots, \sigma_1) = 1. \tag{10}$$

Such that

$$
\begin{aligned}
&\sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} |\mathbf{\Psi}(\sigma_1, \ldots, \sigma_{k+1})|^2 \\
&= \sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} \prod_{i=1}^{k+1} |\mathbf{\Psi}_i(\sigma_i|\sigma_{i-1}, \ldots, \sigma_1)|^2 \\
&= \sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} \left(\prod_{i=1}^{k} |\mathbf{\Psi}_i(\sigma_i|\sigma_{i-1}, \ldots, \sigma_1)|^2 \right) \sum_{j=k+1}^{N} |p(\sigma_j|\sigma_{j-1}, \ldots, \sigma_1)|^2 \\
&= \sum_{\sigma_1=1}^{M} \cdots \sum_{\sigma_L=1}^{M} |\mathbf{\Psi}(\sigma_1, \ldots, \sigma_k)|^2 \\
&= 1
\end{aligned}
\tag{11}
$$

The proof then complete.

D Additional Numerical Results

D.1 Results of Predicting the Entanglement Entropy

We take an additional downstream task: predicting the second-order Rényi entanglement entropy $-\log(\mathrm{tr}(\rho_A^2))$ for the anisotropic Heisenberg model, where A is the left-half subsystem with system size $L/2$ of the L-qubit quantum system. The number of training size is set to be $N_t = 90$ and the predicted RMSE results are given in Tab. 4. It can be observed that pre-training remains effective for predicting the entanglement entropy of the anisotropic Heisenberg model.

Table 4: The RMSE of predicting the second-order Rényi entanglement entropy for the anisotropic Heisenberg model. We sample $N_p = 100$ physical conditions with $K_p = 1024$ measurement strings for pre-training.

Method	$L = 8$					$L = 10$					$L = 12$				
	$K_f = 64$	$K_f = 128$	$K_f = 256$	$K_f = 512$	$K_f = 1024$	$K_f = 64$	$K_f = 128$	$K_f = 256$	$K_f = 512$	$K_f = 1024$	$K_f = 64$	$K_f = 128$	$K_f = 256$	$K_f = 512$	$K_f = 1024$
Classical Shadow	1.58282	1.56688	1.50989	1.40270	1.22974	1.72379	1.71451	1.73135	1.72740	1.68556	2.89481	2.90874	2.91391	2.90773	2.89722
RBF Kernel	0.07322	0.07160	0.07670	0.07692	0.07706	0.02539	0.02257	0.02242	0.02002	0.01983	0.08710	0.08242	0.08104	0.07081	0.07032
NTK	0.07117	0.06799	0.08834	0.08708	0.08690	0.02497	0.02221	0.02129	0.01996	0.01947	0.08432	0.08249	0.08071	0.07998	0.07381
PixelCNN	0.07198	0.07091	0.06849	0.06687	0.06784	0.01907	0.01892	0.01948	0.01952	0.02089	0.07406	0.07145	0.07107	0.06895	0.06677
Neural-Classical Shadow	0.06860	0.06415	0.06403	0.06315	0.06221	0.01844	0.01747	0.01664	0.01662	0.01657	0.07261	0.06858	0.06573	0.06156	0.05924
Q-TAPE	**0.06302**	**0.06141**	**0.06104**	**0.05998**	0.06072	**0.01698**	**0.01623**	**0.01534**	**0.01517**	**0.01520**	**0.05861**	**0.05812**	**0.05648**	**0.05623**	**0.05597**
Q-TAPE w/o Pretrain	0.06649	0.06295	0.06228	0.06071	**0.06034**	0.01711	0.01662	0.01696	0.01655	0.01532	0.06624	0.06542	0.06381	0.06042	0.05931

D.2 Model Sensitivity to the Number of Measurements

In Sec. 4, we study the relationship between the number of measurements and the classification accuracy of quantum phase of matters on Rydberg atom model. It is empirically evident in Fig. 3 that achieving linear growth in classification accuracy requires an exponential increase in the number of measurements per training example. Beyond the scaling related to number of measurements, we dive into further research on the scaling relationship between accuracy and the size of the training set (i.e., the number of sampled physical conditions which determine the dynamics of the quantum system). We constrain the number of measurement per example to 256 (since we find that a large value makes the accuracy reach saturation) and the results on the 31-qubit system are listed in the Tab. 5.

As evident from Tab. 5, accuracy approximately exhibits linear growth w.r.t. training size. This finding consistents with theoretical results presented in (Huang et al., 2022b; Lewis et al., 2023), which demonstrate that there exists a polynomial scaling relationship between model performance and the size of training dataset.

Table 5: Classification accuracy of quantum phases of matter on the Rydberg atom model with varied training size N_t, where $L = 31$ and $K_f = 256$. The results are averaged over 3 runs with different random seeds.

	$N_t = 20$	$N_t = 40$	$N_t = 60$	$N_t = 80$
Q-TAPE	82.05	87.24	89.16	90.63
Q-TAPE w/o pretrain	79.17	81.78	85.96	88.47

D.3 Fine Tuning the Model with Out-of-Distribution Dataset

In this section, we consider fine tuning the Q-TAPE with out-of-distribution (OOD) dataset, which means the dataset used for fine-tuning and the dataset used for pre-training come from different distributions.

Here, we consider two different configurations to make the fine-tuning dataset out-of-distribution from the pre-training one: the first is to re-generate the fine-tuning data by modifying the physical variables and the second is to fine tune the model based on the parameters transferred from the model pretrained on fewer qubits. In the following, we consider the Rydberg atom model.

Table 6: Classification accuracy of quantum phases of matter on the 31-qubit Rydberg atom model. The pre-trained parameters are transferred from the model trained on smaller system size. The training size is set to be $N_t = 100$, and the number of measurements $K_f = 1024$.

Q-TAPE (pre-trained on 19-qubit system)	95.74
Q-TAPE (pre-trained on 25-qubit system)	96.13
Q-TAPE (pre-trained on 31-qubit system)	96.67
Q-TAPE w/o pre-train	94.32

First, we take the evaluation that fine-tuning the model on 31-qubit system by using he parameters pre-trained on 19 and 25-qubit system. Note that the number of qubits is also a physical variable and we want to see if model parameters trained on small-scale systems could transfer and help model characterize larger-scale systems. The results are listed in Tab. 6. It is evident that pre-trained parameters transferred from small-scale systems is also useful for large-scale systems.

Table 7: Classification accuracy of quantum phases of matter on the 19-qubit Rydberg atom model. The training size is set to be $N_t = 100$, and the number of measurements $K_f = 1024$.

	no OOD	OOD
Q-TAPE	95.95	84.82
Q-TAPE w/o pre-train	93.35	94.23

Second, we modify the detuning of a laser from $[-10, 15]$ (which is exactly used in the paper) to $[-20, -10] \cup [15, 25]$ to generate OOD fine-tuning dataset, on Rydberg atom model with 19 qubits. The classification accracy are listed in Tab. 7. The pre-trained one fails to perform better than the Q-TAPE w/o pre-train. The main reason is that the modified detuning values habe driven the quantum evolution into a very different dynamics and the pre-trained model learns less knowledge about it. The question of whether pre-trained Q-TAPE remains beneficial for OOD quantum datasets in other settings remains open, and will be further explored in our future work.