



RydbergGPT

Downloaded from: <https://research.chalmers.se>, 2025-12-25 01:06 UTC

Citation for the original published paper (version of record):

Fitzek, D., Hong Teoh, Y., Fung, C. et al (2025). RydbergGPT. Machine Learning: Science and Technology, 6(4). <http://dx.doi.org/10.1088/2632-2153/ae1d0b>

N.B. When citing this work, cite the original published paper.



PAPER • OPEN ACCESS

RydbergGPT

To cite this article: David Fitzek *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 045057

View the [article online](#) for updates and enhancements.

You may also like

- [Towards instance-wise calibration: local amortized diagnostics and reshaping of conditional densities \(LADaR\)](#)
Biprateep Dey, David Zhao, Brett H Andrews et al.
- [An atomic cluster expansion potential for twisted multilayer graphene](#)
Yangshuai Wang, Drake Clark, Sambit Das et al.
- [Detecting model misspecification in cosmology with scale-dependent normalizing flows](#)
Aizhan Akhmetzhanova, Carolina Cuesta-Lazaro and Siddharth Mishra-Sharma



PAPER

RydbergGPT

OPEN ACCESS

RECEIVED
20 October 2024REVISED
31 October 2025ACCEPTED FOR PUBLICATION
7 November 2025PUBLISHED
1 December 2025

Original Content from
this work may be used
under the terms of the
Creative Commons
Attribution 4.0 licence.

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



David Fitzek^{1,2,*} , Yi Hong Teoh³, H P Cyrus Fung³, Gebremedhin A Dagnew³, Ejaaz Merali³,
M Schuyler Moss³ , Benjamin MacLellan³ and Roger G Melko^{3,4}

¹ Department of Microtechnology and Nanoscience, Chalmers University of Technology, 412 96 Gothenburg, Sweden

² Volvo Group Trucks Technology, 405 08 Gothenburg, Sweden

³ Department of Physics and Astronomy, University of Waterloo, 200 University Ave. West, Waterloo, Ontario N2L 3G1, Canada

⁴ Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada

* Author to whom any correspondence should be addressed.

E-mail: fitzek@protonmail.com

Keywords: machine learning, quantum computing, neutral atom arrays, Rydberg atoms, generative pretrained transformer, machine learning in physics

Abstract

We introduce a generative pretrained transformer (GPT) designed to learn the measurement outcomes of a neutral atom array quantum computer. Based on a vanilla transformer, our encoder–decoder architecture takes as input the interacting Hamiltonian, and outputs an autoregressive sequence of qubit measurement probabilities. Its performance is studied in the vicinity of a quantum phase transition in Rydberg atoms in a square lattice array. We explore the model’s generalization capabilities by demonstrating that it can accurately predict ground-state measurement outcomes for Hamiltonian parameter values that were not included in the training data. We evaluate three model variants, each trained for a fixed duration on a single NVIDIA A100 GPU, by examining their predictions of key physical observables. These results establish performance benchmarks for scaling to larger RydbergGPT models. These can act as benchmarks for the scaling of larger RydbergGPT models in the future. Finally, we release RydbergGPT as open-source software to facilitate the development of foundation models for diverse quantum computing platforms and datasets.

1. Introduction

Generative models have emerged as a key technology for predicting the probabilistic behavior of a quantum system. Their most basic task is to produce a target sequence representing qubit measurement outcomes; e.g. projective measurements distributed according to the Born rule of quantum mechanics. A wide variety of architectures have been explored for this purpose, from restricted Boltzmann machines (Carleo and Troyer 2017, Gao and Duan 2017, Nomura *et al* 2017, Czischek *et al* 2018, Torlai and Melko 2018, Carrasquilla *et al* 2019, Torlai *et al* 2019, Carrasquilla and Torlai 2021, Viteritti *et al* 2022), to recurrent neural networks (Carrasquilla and Melko 2017, Carrasquilla and Torlai 2021, Hibat-Allah *et al* 2021, 2023, 2024, Morawetz *et al* 2021, Czischek *et al* 2022, Moss *et al* 2024) and transformers (Carrasquilla *et al* 2021, Cha *et al* 2022, Wang *et al* 2022, Viteritti *et al* 2023, 2024, Zhang and Di Ventura 2023, Lange *et al* 2024, Sprague and Czischek 2024). While the above models have been trained to output target sequences drawn from a single parameterized distribution, the utility of generative models can be expanded by adopting modern encoder–decoder architectures. In this case, the encoder takes a source sequence and maps it to a context vector. This is then passed to the decoder, which combines it with other inputs and maps it to the target sequence.

The most powerful architecture to emerge recently is the attention-based transformer (Vaswani *et al* 2017), variations of which underlie most of the large language models (LLMs) being developed by academia and industry today (Brown *et al* 2020, Gozalo-Brizuela and Garrido-Merchan 2023, Touvron *et al* 2023, Wu *et al* 2023, Zhang *et al* 2023, Nakaji *et al* 2024). A key challenge in physics that can be tackled by the vanilla encoder–decoder transformer structure is the task of predicting the output of a quantum

computer given its experimental settings. In the simplest example, the source sequence is a description of an interacting Hamiltonian, and the target sequence is the conditional probability of each qubit measurement outcome, selected from a dictionary of size two. Viewing the task as a language problem, the self-attention mechanism encodes context, or correlations, between qubit measurement outcomes. Self-attention has been shown to be a suitable mechanism of encoding the possible long-range correlations—including entanglement—found in quantum systems without the need to pass hidden or latent vectors along the sequence (Wang *et al* 2022, Zhang and Di Ventra 2023, Sprague and Czischek 2024). The impressive demonstrations of scaling in LLMs gives hope that the transformer architecture will benefit quantum computers as they continue to grow in size and capability (Kaplan *et al* 2020).

In this work, we train an attention-based transformer to learn the distribution of qubit measurement outcomes corresponding to a Hamiltonian describing an array of interacting Rydberg atoms. Using the encoder–decoder architecture, we use simulated Rydberg occupation data drawn from a two-dimensional qubit array to train three different models. Once trained, each model can produce new data via autoregressive sampling of the decoder. In the case of a Rydberg atom array in its ground state, the autoregressive output can be interpreted as the complete quantum wavefunction, allowing for the calculation of many physical observables of interest. We report results for estimators of the energy and other physical observables. By varying the Hamiltonian that is input to the encoder, we find that the transformer is able to accurately predict qubit measurement outcomes in regions outside of the training regime. Our results show potential that attention-based transformer models are well-positioned to provide a powerful approach for predicting the properties of quantum many-body systems.

2. Rydberg atom array physics

Rydberg atom arrays are a powerful platform for quantum information processing, where interacting qubits are encoded with the electronic ground state $|g\rangle$ and excited (Rydberg) state $|r\rangle$ for each atom (Jaksch *et al* 2000, Lukin *et al* 2001, Endres *et al* 2016, Browaeys and Lahaye 2020, Wu *et al* 2021). We focus on this platform due to its experimental relevance as a leading quantum computing architecture, the rich physics of its phase transitions which provide an excellent test case for our approach, and its positive real-valued ground state wavefunctions that enable complete information extraction from projective measurements. The governing Hamiltonian defining the Rydberg atom array interactions has the following form:

$$\hat{H} = \sum_{i < j} C_6 V_{ij} \hat{n}_i \hat{n}_j - \delta \sum_{i=1}^N \hat{n}_i - \frac{\Omega}{2} \sum_{i=1}^N \hat{\sigma}_i^x, \quad (1)$$

$$C_6 = \Omega \left(\frac{R_b}{a} \right)^6, \quad V_{ij} = \frac{a^6}{\|\mathbf{r}_i - \mathbf{r}_j\|^6}, \quad (2)$$

where $\hat{\sigma}_i^x = |g\rangle_i \langle r|_i + |r\rangle_i \langle g|_i$, the occupation number operator $\hat{n}_i = \frac{1}{2} (\hat{\sigma}_i + \mathbb{1}) = |r\rangle_i \langle r|_i$ and $\hat{\sigma}_i = |r\rangle_i \langle r|_i - |g\rangle_i \langle g|_i$. The experimental settings of a Rydberg atom array are controlled by the detuning from resonance δ , Rabi frequency Ω , lattice length scale a and the positions of the atoms $\{\mathbf{r}_i\}_i^N$. From equation (2), we obtain a symmetric matrix \mathbf{V} , that encapsulates the relevant information about the lattice geometry, and derive the Rydberg blockade radius R_b , within which simultaneous excitations are penalized (Lukin *et al* 2001, Wu *et al* 2021). Finally, for the purposes of our study, the atom array is considered to be affected by thermal noise, in equilibrium at a temperature $T = 1/\beta$. The experimental settings are thus captured by the set of parameters $\mathbf{x} = (\Omega, \delta/\Omega, R_b/a, \mathbf{V}, \beta\Omega)$. By adjusting the free parameters in the Rydberg Hamiltonian, the system can be prepared in various phases of matter, separated by distinct phase transitions (Endres *et al* 2016, Ebadi *et al* 2021, Samajdar *et al* 2021, Xu *et al* 2021a, Kalinowski *et al* 2022, Miles *et al* 2023).

Rydberg atom array experiments are able to perform projective measurements of the qubit states in the occupation basis. At $T = 0$, projective measurements in the occupation basis provide complete information about the quantum state because the ground state wavefunction of the Rydberg Hamiltonian is positive and real-valued. This property allows for full reconstruction of the wavefunction from measurement statistics alone, without the phase ambiguity typically encountered in quantum state tomography (Bravyi *et al* 2007, Endres *et al* 2016, Xu *et al* 2021a, Merali *et al* 2023). In the next section, we consider these projective measurements as training vectors for an unsupervised learning strategy for our generative model. The goal of the generative model after training will be to predict, given a Hamiltonian, the distribution of projective measurements in the quantum system.

3. Transformer architecture

In order to learn the qubit measurement distribution of the Rydberg atom array, we employ the transformer encoder–decoder architecture (Vaswani *et al* 2017). These generative models are universal sequence-to-sequence function approximators and, granted sufficient tunable parameters θ , can encode arbitrarily complex functions (Yun *et al* 2020). Transformers belong to the class of generative models known as autoregressive models. Autoregressive models learn conditional probability distributions over the N individual variables, such as qubit states $\sigma = \{\sigma_i\}_i^N$, $p_\theta(\sigma_i | \sigma_{i-1}, \dots, \sigma_1)$. The joint distribution that arises from these conditionals can be used to represent the likelihood of measurement outcomes,

$$p_\theta(\sigma) = \prod_{i=1}^n p_\theta(\sigma_i | \sigma_{i-1}, \dots, \sigma_1). \quad (3)$$

This is the decoder output, which can be sampled to infer any number of qubit measurements. This output can be related to a pure quantum wavefunction for example through the Born rule $p_\theta(\sigma) = |\Psi_\theta(\sigma)|^2$. Additionally, under the assumption of a positive real-valued wavefunction, the transformer's output captures the full quantum state, i.e. $\Psi_\theta(\sigma) = \sqrt{p_\theta(\sigma)}$ (Wu *et al* 2021, Moss *et al* 2024), where θ denotes the model parameters.

An encoder that maps the experimental Hamiltonian settings to a context vector can be used to condition this decoder. The encoder–decoder architecture that we use is illustrated in figure 1 and follows the overall design introduced by Vaswani *et al* (2017).

Note that the source sequence for our encoder, the experimental Hamiltonian settings \mathbf{x} , have a natural graph representation. The interactions \mathbf{V} can be considered as the edge attributes and the remaining variables node attributes. The field of machine learning has proposed a multitude of models for processing such graphs (Kipf and Welling 2017). In our architecture, we choose to process the graph with a convolutional graph neural network (Kipf and Welling 2017, Wang *et al* 2019). This allows us to map the graph, which scales quadratically in the number of atoms N , to a sequence representation that scales linearly. To further capture correlations, we feed the sequence to the vanilla transformer encoder, consisting of multiple self multi-head attention and feedforward blocks, and obtain the context vector. The decoder autoregressively parameterizes the equilibrium state of the system with respect to the context via an attention mechanism between the context and the intermediate state within the decoder. The choice of hyperparameters for the architecture is outlined in table 1.

A powerful property of autoregressive models is the ability to perform independent and identically distributed sampling without the use of Markov chain Monte Carlo. To perform sampling, we input the experimental settings of interest to our encoder and obtain the context vector as output. Following that, the decoder, conditioned on this context, takes as input a starting token and samples the first element of the sequence. Then, we iteratively sample the i th token conditioned on all tokens preceding it. We traverse the lattice of atoms systematically with the snake path convention, shown in the bottom right of figure 1.

4. Training transformer models

We follow a data-driven training procedure, where we train transformers on a synthetic dataset \mathcal{D} composed of pairs of source and target sequences. The former corresponds to experimental settings \mathbf{x} , input to the encoder, and the latter corresponds to binary qubit measurement data, input to the decoder. Our target sequences are simulated Rydberg occupation measurements, produced via Stochastic Series Expansion quantum Monte Carlo (QMC) method described in Merali *et al* (2023).

We employ our QMC to produce uncorrelated samples, which we collect across a wide range of settings. Specifically, we choose values of the inverse temperature $\beta\Omega$ within the set $\{0.5, 1, 2, 4, 8, 16\}$, the detuning parameter δ/Ω over the range $[-0.364, 3.173]$, the Rydberg blockade R_b/a within $\{1.05, 1.15, 1.3\}$, and the linear system size L for the values $\{5, 6, 11, 12, 15, 16\}$. Without loss of generality, we take $\Omega = 1$ henceforth. For each configuration, we generate a dataset containing $|\mathcal{D}| = 10^5$ entries, each corresponding to a binary measurement sampled from the QMC simulation. The datasets are selected to explore δ/Ω values near a quantum phase transition in the square-lattice Rydberg atom arrays, which occurs at approximately $\delta/\Omega = 1.1$ (Merali *et al* 2023).

A transformer architecture with parameters θ is trained in an unsupervised manner with a dataset \mathcal{D} by minimizing the Kullback–Leibler divergence, resulting in the loss function:

$$\mathcal{L}(\theta) \approx -\frac{1}{|\mathcal{D}|} \sum_{\sigma \in \mathcal{D}} \ln p_\theta(\sigma). \quad (4)$$

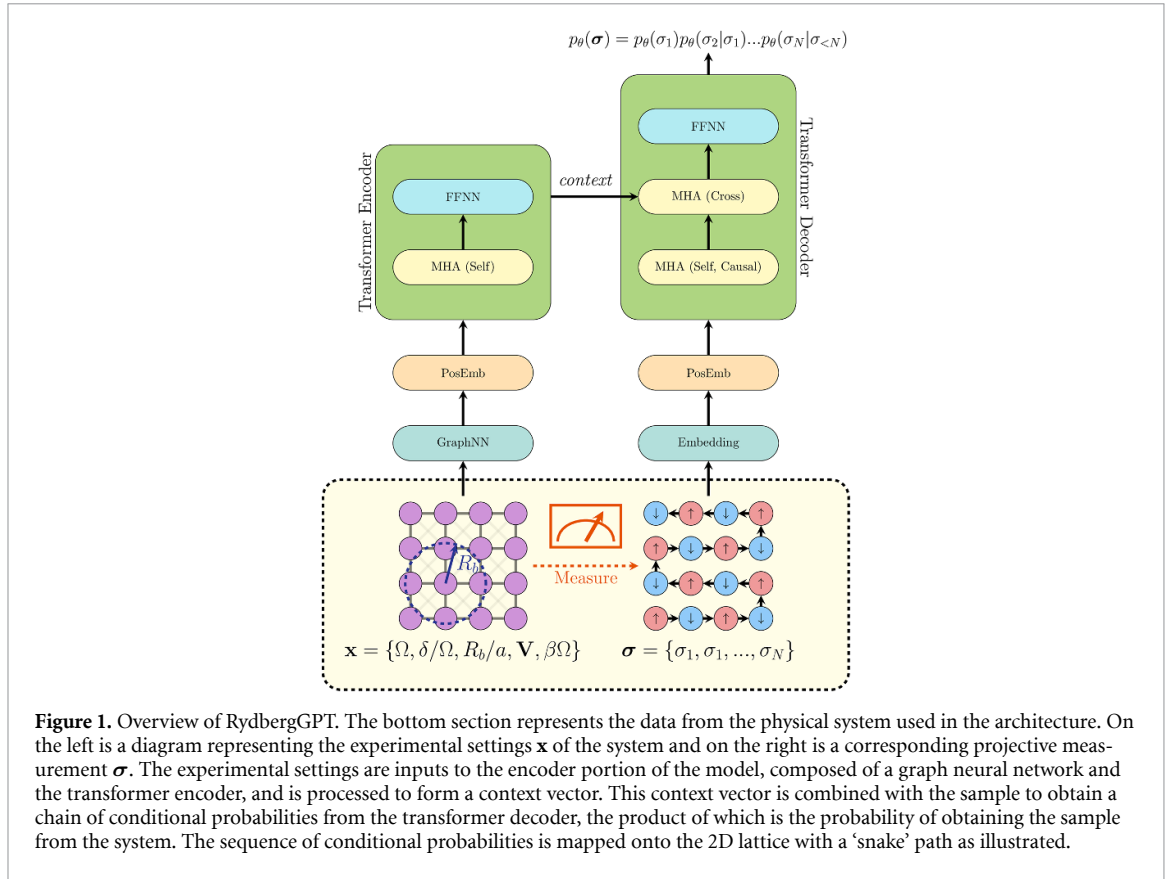


Table 1. The transformer encoder–decoder architecture and training parameters. Including the dimension of the feedforward network layer (d_{ff}), the model dimension (d_{model}), the graph dimension (d_{graph}) as well as the total number of trainable parameters.

Parameter	Value
Neural network architecture	
d_{ff}	128
d_{model}	32
d_{graph}	64
num heads	8
num blocks encoder	1
num blocks decoder	3
num graph layers	2
trainable params	66 562
Training hyperparameters	
batch size	1024
optimizer	AdamW
dropout	0.1
learning rate	0.001
learning rate schedule	Cosine annealing warm start
T_0	1
T_{mult}	2
η_{min}	0.00 001
dataset buffer	50

This loss is minimized when the decoder output and the target distributions are equal. We initialize each model using the method introduced by Glorot and Bengio (2010). We use the AdamW optimizer (Loshchilov and Hutter 2019), which is an extension to stochastic gradient descent with an adaptive learning rate and weight regularization, in combination with a cosine annealing warm restart learning rate scheduler (Loshchilov and Hutter 2017). Details for the training hyperparameters can be found in table 1.

Table 2. Number of tokens, in units of 10^8 , contained in our dataset for various sizes L . Each token corresponds to the measurement outcome of a single Rydberg qubit. For each configuration, i.e. size and experimental settings, we have a total of 10^5 samples, with each sample containing a total of $N = L \times L$ measurements.

Lattice size, L	5	6	11	12	15	16
Tokens (10^8)	1.5	2.16	7.26	8.64	13.5	15.4

Table 3. Details on the training of each model, including epochs trained, the amount of tokens in the training dataset, in units of 10^8 , and the total number of tokens processed during training, in units of 10^{10} . The models were each trained for a constant amount of time, 85 hours, on an Nvidia A100.

Model	M_1	M_2	M_3
Epochs	116	40	17
Tokens in dataset (10^8)	3.66	19.6	48.4
Tokens processed (10^{10})	4.25	7.82	8.23

In this manuscript, we train three models with different subsets of the simulated data. Specifically, the models are all trained on data at $R_b/a = 1.15$ and across the values for δ/Ω and $\beta\Omega$. However, each model is trained with different sets of linear size L . Model M_1 is trained with data for systems of size $L = 5, 6$, model M_2 with $L = 5, 6, 11, 12$ and model M_3 with $L = 5, 6, 11, 12, 15, 16$, refer to table 2 for the total number of tokens, i.e. measurement outcome of a single qubit, for each lattice size. As we generate 10^5 samples for each lattice size, the larger lattice sizes have more tokens compared to the smaller ones.

We conducted training on a single A100 GPU for 85 hours. The size of the training datasets for all three models is on the order of 10^8 tokens. During training the models processed approximately 10^{10} tokens. We summarize the relevant training information in table 3.

5. Inference with trained models

In this section, we perform inference using the trained model and study the disordered-to-checkerboard phase transition of the Rydberg atom array, governed by the dimensionless detuning δ/Ω . First, samples are drawn from the model in accordance to the method described at the end of section 3. The samples are then used to obtain predictions of physical observables. We compare our model's predictions with estimates obtained from QMC simulations (Merali *et al* 2023). For Rydberg atom arrays, QMC is able to obtain an estimate of the observables that can be considered exact, within statistical errors. In showing that our model's predictions of these observables are in good agreement with the QMC values, we illustrate the efficacy of data-driven training as a technique for predicting Rydberg atom array measurement outcomes for ground states. Further, we explore our trained models' ability to generalize away from the ground state and also to system sizes not represented in the training data.

In order to construct our estimators below, we draw samples from our model decoder. In our implementation, the complexity of sampling with respect to the system size $N = L \times L$ is of order $\mathcal{O}(N^3)$. However, this complexity can theoretically be reduced to $\mathcal{O}(N^2)$ by caching the appropriate intermediate values in the attention calculation.

5.1. Physical observables

Physical quantities that are diagonal in the occupation basis can be directly computed from the samples drawn from the model. One such quantity is the staggered magnetization for the square-lattice Rydberg atom array. The staggered magnetization is significant as it is the order parameter for the disorder-to-checkerboard quantum phase transition (Samajdar *et al* 2020), and it can be calculated with:

$$\langle \hat{\sigma}^{\text{stag}} \rangle \approx \frac{1}{N_s} \sum_{\sigma \sim p_\theta(\sigma)} \frac{1}{N} \left| \sum_{i=1}^N (-1)^i (n_i(\sigma) - 1/2) \right|, \quad (5)$$

where i runs over all $N = L \times L$ atoms using the snake path convention of figure 1. Here, $n_i(\sigma) = \langle \sigma | r_i | \sigma \rangle$ is the occupation number operator acting on atom i in a given configuration σ .

In contrast, we consider an off-diagonal observable, where we must make use of the ground state wave function amplitudes of the inferred samples $\Psi_\theta(\sigma) = \sqrt{p_\theta(\sigma)}$. As an example, we examine the spatially averaged expectation value of $\hat{\sigma}^x$, which is defined as

$$\langle \hat{\sigma}^x \rangle \approx \frac{1}{N_s} \sum_{\sigma \sim p_\theta(\sigma)} \frac{1}{N} \sum_{\sigma' \in \text{SSF}(\sigma)} \frac{\Psi_\theta(\sigma')}{\Psi_\theta(\sigma)}, \quad (6)$$

where single spin flip (SSF)(σ) is the set of configurations that are connected to σ by a SSF.

Finally, we consider an estimate of the ground state energy $\langle E \rangle$, which is defined as

$$\langle E \rangle \approx \frac{1}{N_s} \sum_{\sigma \sim p_\theta(\sigma)} \frac{\langle \sigma | \hat{H} | \Psi_\theta \rangle}{\langle \sigma | \Psi_\theta \rangle}. \quad (7)$$

This observable can be efficiently evaluated for local non-diagonal operators (Carleo and Troyer 2017, Torlai *et al* 2018, Melko *et al* 2019, Hibat-Allah *et al* 2020). More detailed derivations of equations (6) and (7) can be found in [appendix](#).

It is important to emphasize that these estimates of off-diagonal observables are appropriate only when the output of the decoder can be interpreted as the normalized wavefunction amplitude, $\Psi_\theta(\sigma) = \sqrt{p_\theta(\sigma)}$. In particular, this correspondence is valid when the system is a pure state and the wavefunction can be assumed to be real and positive in the occupation basis, as is the case for the Rydberg Hamiltonian in its $T=0$ ground state. Importantly, the use of the decoder output as a wave function amplitude is no longer valid for finite temperature states. While the above estimates can still be computed, $\Psi_\theta(\sigma)$ is not able to capture the physics of the thermal state, as is demonstrated in our results which follow.

5.2. Results

We now demonstrate the results of the trained models across parameter regimes and observables. In figure 2 we illustrate the performance of model M_1 on the three observables introduced in section 5.1 as a function of δ/Ω , for $R_b/a = 1.15, \beta\Omega = 16.0$. The estimators are calculated over $N_s = 10^5$ samples. The predictions are benchmarked against the QMC observables calculated from 10^5 decorrelated samples, which we take as the ground truth. At $\beta\Omega = 16$, the temperature is sufficiently low for the Rydberg atom array to be in its ground state to a good approximation. The model shows good agreement with the QMC estimates for the energy and staggered magnetization. As for the x -magnetization, we observe a slight deviation from the QMC estimates in the tails. We conclude that the trained model is able to generalize well to dimensionless detunings, δ/Ω , that lie outside of the training data set, when the system is restricted to the ground state.

Next, we investigate how well the model generalizes away from the ground state. In figure 3 we illustrate predictions made by model M_1 with the same strategy, except temperature $T = 1/\beta$ is varied instead of the detuning. We see a strong disagreement between the exact observables and the model-obtained estimates at high temperatures. In our architecture, the decoder represents a probability distribution over configurations in the Rydberg occupation basis, which can be mapped one-to-one onto a positive wavefunction of a pure state. As mentioned, the $T=0$ ground state of the Rydberg Hamiltonian equation (1) is known to take such a form, and thus the model performs well for the ground state, e.g. low temperatures. However, as the temperature is increased, the system becomes a well-mixed Gibbs state, which can no longer be accurately represented with a positive wavefunction. As such, the output of the decoder is incapable of accurately representing this state. The model is merely approximating a probability distribution that results in identical occupation-basis measurement likelihoods. This results in the incorrect estimate of off-diagonal observables, which require the wave function interpretation of the decoder output $\Psi_\theta(\sigma) = \sqrt{p_\theta(\sigma)}$, as in equation (6). To appropriately capture such high-temperature states, one might employ more sophisticated methods such as a complex density matrix (Kothe and Kirton 2023) or minimally entangled typical thermal states (METTS) (White 2009, Hendry *et al* 2022).

This is particularly evident in figure 3 at the limit of large temperature. Here, the system tends to the fully-mixed state where all states are equally probable. The positive wavefunction that shares the same computational basis probability as this state is the equal superposition of all states. This state corresponds to the maximum eigenstate of the x -magnetization, leading to the divergence in figure 3 between the exact QMC and the trained model estimators at high temperatures, where $\langle \hat{\sigma}^x \rangle$ becomes significantly over-estimated. This is further supported by the fact that the high-temperature estimates of the staggered magnetization are somewhat accurate. Again, this behavior is due to the fact that this diagonal estimator relies only on the computational basis probabilities with which certain configurations are sampled.

Finally, we assess the ability of our transformer to generalize to larger system sizes beyond those included in the training set. For this we use all three trained models, M_1 , M_2 and M_3 . Figure 4 presents the energy evaluations from each of our three trained models, averaged over 10^4 samples. These models were trained on consecutively larger datasets as previously described in table 3. We first observe that the model M_1 trained on the smallest dataset, containing only data from lattice sizes $L = 5, 6$, captures the $L = 5$ energy accurately but struggles to generalize to larger system sizes. Interestingly, the models trained

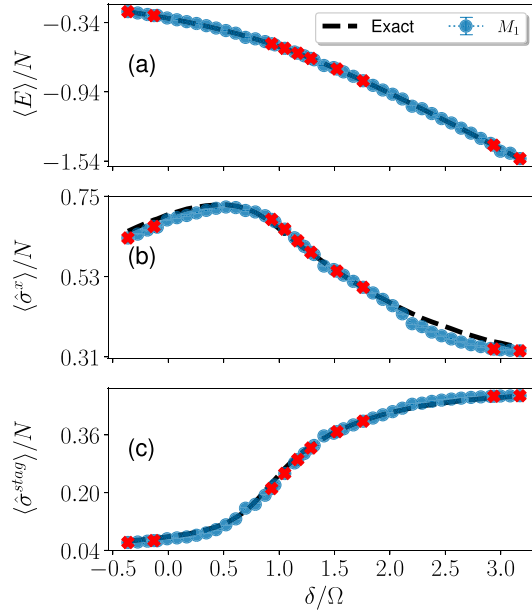


Figure 2. Model predictions of multiple observables across the disordered-to-checkerboard phase transition, which is governed by the dimensionless detuning, δ/Ω . The observables include, (a) energy, (b) site-averaged x -magnetization and (c) the staggered magnetization. Here, blue points correspond to observables estimated from projective measurements sampled from the trained model (i.e. out-of-distribution), red points indicate parameters contained within the training dataset (i.e. in-distribution), and dotted black lines correspond to the exact observables.

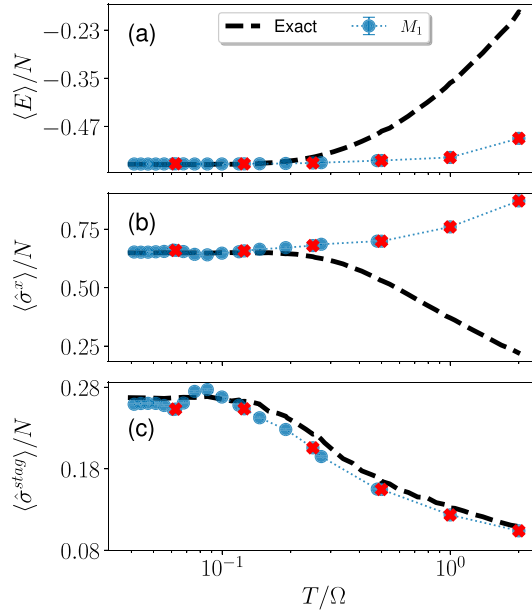


Figure 3. Model predictions of multiple observables as the dimensionless temperature T/Ω is increased. The observables include, (a) energy, (b) site-averaged x -magnetization, and (c) the staggered magnetization. Again, blue points correspond to observables estimated from projective measurements sampled from the trained model (i.e. out-of-distribution), red points indicate parameters contained within the training dataset (i.e. in-distribution), and dotted black lines correspond to the exact observables.

on more extensive datasets do not accurately capture the properties of their respective training sets, with the energy estimates deviating significantly from expected values.

These results illustrate how physical estimators are affected by bias in the training datasets. As shown in table 2, we always accumulate 10^5 samples for each lattice size. Therefore, the number of training tokens or measurements varies for each lattice size; more precisely, the dataset is biased towards larger lattice sizes. This explains why models M_2 and M_3 perform worse than M_1 for the lattice sizes $L = 5, 6$, since the fraction of measurements for smaller sizes is less represented in their training dataset. However,

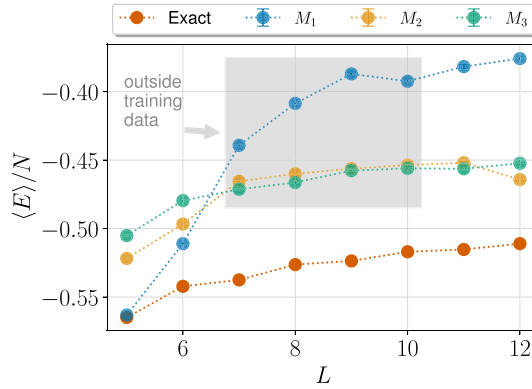


Figure 4. Extrapolation over lattice size L . The ground-state energy, as estimated by quantum Monte Carlo simulations and the models trained on three different datasets, demonstrates how training data affect the ability to extrapolate to system sizes not included in the training.

they perform better for larger lattice sizes, where a larger relative number of training tokens are available.

6. Conclusion

In this paper, we have explored the training of attention-based transformer models for the problem of mapping a quantum Hamiltonian to the qubit measurement outcomes of the corresponding quantum state. We construct an encoder–decoder architecture suitable for Rydberg atom arrays, where the encoder input is a representation of the interacting Hamiltonian and the decoder output represents measurements of the Rydberg occupation.

Since transformers have been developed as autoregressive models for natural language processing, they naturally represent probability distributions, such as Born-rule-distributed measurement outcomes. In the case where a quantum state can be assumed to be pure, real, and positive, this probability distribution has a one-to-one mapping to the wavefunction, giving the output of the decoder a powerful interpretation. In the ground state of Rydberg atom arrays governed by Hamiltonian equation (1), this assumption is valid, and we use it to demonstrate that a trained transformer model can accurately reproduce physical estimators in this case.

We have further explored the behavior of the transformer model when the quantum state is no longer real and positive, i.e. in a thermally mixed state at higher temperature. We clearly see the breakdown in our ability to interpret the output of the decoder as the corresponding quantum state, as the physical estimators significantly deviate from their exact values for large T . However, we propose that our decoder architecture could be modified to represent a density matrix (Kothe and Kirton 2023) or METTS (White 2009, Hendry *et al* 2022), which would be able to represent noisy, thermal, and dynamical states, extending the capabilities of the model.

Finally, we attempt to demonstrate the scaling potential of the encoder–decoder transformer. Namely, whether the transformer model, when trained on a dataset composed of a smaller many-body system, is able to make predictions for larger system sizes. Our results show signatures of generalization, however, it also suggest that we have not attained the necessary scale in our transformer to exploit its full potential. Hopefully, by training models with more data on a diverse set of lattice sizes, systematic improvements in model performance will be observed. In fact, such systematic scaling is the hallmark of large transformers used in industry, giving reason for optimism (Kaplan *et al* 2020).

More generally, just like in modern LLMs, our transformer architecture should allow for vastly improved performance with increases in the number of trainable parameters θ , larger datasets, and more computational resources. In addition, since these models are designed to represent physical qubit systems, a number of architectural improvements should be explored. For example, there may be more efficient ways to use tokens than to represent a dictionary of only two individual qubit measurements. In particular, transformers designed for natural language processing typically manage dictionaries comprising tens to hundreds of thousands of tokens (Brown *et al* 2020). An alternative approach involves using patches of qubits as tokens, which can reduce some computational demands (Sprague and Czischek 2024). However, this method introduces constraints on the system geometries, potentially limiting the

applicability to arbitrary quantum array configurations. This balance between computational efficiency and system flexibility remains a critical area for future research.

Notably, all three models discussed in this study were trained using only a single NVIDIA A100 GPU over a duration of 85 hours. Looking forward, we can use this as a benchmark as the size and capability of such transformer models is increased towards the goal of predicting the behavior of larger quantum systems. Inspired by the scaling observed in LLMs within industry today, we hypothesize that significantly larger transformers trained on vastly more data may be able to make highly useful predictions, e.g. about the behavior of a quantum computer in parameter regimes for which no data is currently available.

Finally, while our models in this study learned from synthetic data, our transformer architecture could instead be trained on qubit measurement data obtained from real-experimental devices. This approach would mark the beginning of a new type of *foundation model*, trained on real quantum computer output to generate predictions on a wide variety of unseen input settings. Like today's LLMs, the capabilities of such large quantum models could increase dramatically with increased number of parameters, data set size, and compute, potentially ushering in a new era of scaling in AI and quantum computer co-design.

The RydbergGPT code and trained model weights are available at Fitzek and Teoh (2024) and the dataset is available at Fitzek et al (2024).

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

We acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Perimeter Institute for Theoretical Physics. This research was supported in part by Grant NSF PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP). Computational support was provided by the facilities of the Shared Hierarchical Academic Research Computing Network SHARCNET: (www.sharcnet.ca) and Compute/Calcul Canada. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Economic Development, Job Creation and Trade. D F acknowledges the Knut and Alice Wallenberg (KAW) Foundation for funding through the Wallenberg Centre for Quantum Technology (WACQT). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through Grant Agreement No. 2022-06725.

Appendix. Estimating expectation values of off-diagonal operators

In this work, we estimate the ground state expectation values of two off-diagonal operators, $\langle \hat{\sigma}^x \rangle$ and $\langle E \rangle$, using samples inferred from our trained GPT model. Here, we provide more detail for how to arrive at equations (6) and (7).

Considering a general, off-diagonal operator \hat{O} , we can write down its expectation value with respect to the probability distribution encoded in the trained GPT model,

$$\langle \hat{O} \rangle = \frac{\langle \Psi_\theta | \hat{O} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle},$$

where $\Psi_\theta(\sigma) = \sqrt{p_\theta(\sigma)}$ is the wave function interpretation of the joint probability distribution. Inserting the identity in the numerator and the denominator, and multiplying by 1, the above expectation value can be equivalently written as

$$\langle \hat{O} \rangle = \frac{\sum_{\sigma} |\langle \Psi_\theta | \sigma \rangle|^2 \times \frac{\langle \sigma | \hat{O} | \Psi_\theta \rangle}{\langle \sigma | \Psi_\theta \rangle}}{\sum_{\sigma'} |\langle \Psi_\theta | \sigma' \rangle|^2}.$$

One can recognize $\frac{|\langle \Psi_\theta | \sigma \rangle|^2}{\sum_{\sigma'} |\langle \Psi_\theta | \sigma' \rangle|^2}$ as some normalized probability $P(\sigma)$ which can be approximated using importance sampling. Here we also define the local estimator of our operator \hat{O} ,

$$\hat{O}_{\text{loc}}(\sigma) = \frac{\langle \sigma | \hat{O} | \Psi_\theta \rangle}{\langle \sigma | \Psi_\theta \rangle},$$

such that the expectation value of \hat{O} can be estimated as

$$\langle \hat{O} \rangle \approx \frac{1}{N_s} \sum_{\sigma \sim p_\theta(\sigma)} \hat{O}_{\text{loc}}(\sigma).$$

If we take $\hat{O} = \hat{H}$, then we have exactly recovered equation equation (7).

In order to arrive at equation (6), we will consider what happens when the off-diagonal operator \hat{O} acts on the bra $\langle \sigma |$ in the definition of $\hat{O}_{\text{loc}}(\sigma)$.

If our off-diagonal operator \hat{O} is purely off-diagonal, it will map a given configuration configuration σ to a linear combination of some set of other configurations $\{\sigma'\}$, that is $\hat{O}|\sigma\rangle = \sum_{\{\sigma'\}} C_O(\sigma')|\sigma'\rangle$, and we can rewrite our expression for $\hat{O}_{\text{loc}}(\sigma)$ as,

$$\hat{O}_{\text{loc}}(\sigma) = \sum_{\sigma' \in \text{SSF}(\sigma)} C_O(\sigma') \frac{\Psi_\theta(\sigma')}{\Psi_\theta(\sigma)},$$

where we have used the fact that $\Psi_\theta(\sigma) = \langle \sigma | \Psi_\theta \rangle$.

We can note that if $\hat{O} = \hat{\sigma}^x$, then $\text{SSF}(\sigma)$ is the set of configurations connected to σ by a SSF and $C_O(\sigma') = 1$. Then it is clear that the inner sum in equation equation (6) takes the form of the equation above.

ORCID iDs

David Fitzek  0000-0003-4268-5485

M Schuyler Moss  0000-0001-8885-7030

Roger G Melko  0000-0002-5505-8176

References

- Bravyi S, DiVincenzo D P, Oliveira R I and Terhal B M 2007 The complexity of stoquastic local hamiltonian problems
- Browaeys A and Lahaye T 2020 Many-body physics with individually controlled Rydberg atoms *Nat. Phys.* **16** 132–42
- Brown T et al 2020 Language Models are few-Shot Learners *Advances in Neural Information Processing Systems* vol 33, ed H Larochelle, M Ranzato, R Hadsell, M F Balcan and H Lin (Curran Associates, Inc.) pp 1877–901
- Carleo G and Troyer M 2017 Solving the quantum many-body problem with artificial neural networks *Science* **355** 602–6
- Carrasquilla J, Luo Di, Pérez F, Milsted A, Clark B K, Volkovs M and Aolita L 2021 Probabilistic simulation of quantum circuits using a deep-learning architecture *Phys. Rev. A* **104** 032610
- Carrasquilla J and Roger G M 2017 Machine learning phases of matter *Nat. Phys.* **13** 431–4
- Carrasquilla J and Torlai G 2021 How to use neural networks to investigate quantum many-body physics *PRX Quantum* **2** 040201
- Carrasquilla J, Torlai G, Melko R G and Aolita L 2019 Reconstructing quantum states with generative models *Nat. Mach. Intell.* **1** 155–61
- Cha P, Ginsparg P, Felix W, Carrasquilla J, McMahon P L and Kim E-A 2022 Attention-based quantum tomography *Mach. Learn.: Sci. Technol.* **3** 01LT01
- Czischek S et al 2022 Data-enhanced variational monte carlo simulations for Rydberg atom arrays *Phys. Rev. B* **105** 205108
- Czischek S, Gärttner M and Gasenzer T 2018 Quenches near ising quantum criticality as a challenge for artificial neural networks *Phys. Rev. B* **98** 024311
- Ebadi S et al 2021 Quantum phases of matter on a 256-atom programmable quantum simulator *Nature* **595** 227–32
- Endres M, Bernien H, Keesling A, Levine H, Anschuetz E R, Krajenbrink A, Senko C, Vuletic V, Greiner M and Lukin M D 2016 Atom-by-atom assembly of defect-free one-dimensional cold atom arrays *Science* **354** 1024–7
- Fitzek D and Rydberggpt Y H T (2024 available at: <https://github.com/PIQuLL/RydbergGPT/tree/main>)
- Fitzek D, Teoh Y H, Fung H P, Dagnew G A, Ejaaz Merali M S M, MacLellan B and Melko R G 2024 Pennylane datasets for rydberggpt (available at: <https://pennylane.ai/datasets/other/rydberggpt>)
- Gao X and Duan L-M 2017 Efficient representation of quantum many-body states with deep neural networks *Nat. Commun.* **8** 662
- Glorot X and Bengio Y 2010 Understanding the difficulty of training deep feedforward neural networks *Proc. 13th Int. Conf. on Artificial Intelligence and Statistics (Proc. of Machine Learning Research, Chia Laguna Resort vol 9)*, (Sardinia, Italy, 13–15 May 2010) ed Y W Teh and M Titterton (PMLR) (available at: <https://proceedings.mlr.press/v9/glorot10a.html>), pp 249–56
- Gozalo-Brizuela R and Garrido-Merchan E C 2023 Chatgpt is not all you need. a state of the art review of large generative ai models
- Hendry D, Chen H and Feiguin A 2022 Neural network representation for minimally entangled typical thermal states *Phys. Rev. B* **106** 165111
- Hibat-Allah M, Ganahl M, Hayward L E, Melko R G and Carrasquilla J 2020 Recurrent neural network wave functions *Phys. Rev. Res.* **2** 023358

- Hibat-Allah M, Inack E M, Wiersema R, Melko R G and Carrasquilla J 2021 Variational neural annealing *Nat. Mach. Intell.* **3** 952–61
- Hibat-Allah M, Melko R G and Carrasquilla J 2023 Investigating topological order using recurrent neural networks *Phys. Rev. B* **108** 075152
- Hibat-Allah M, Melko R G and Carrasquilla J 2024 Supplementing recurrent neural network wave functions with symmetry and annealing to improve accuracy
- Jaksch D et al 2000 Fast quantum gates for neutral atoms *Phys. Rev. Lett.* **85** 2208–11
- Kalinowski M, Samajdar R, Melko R G, Lukin M D, Sachdev S and Choi S 2022 Bulk and boundary quantum phase transitions in a square rydberg atom array *Phys. Rev. B* **105** 174417
- Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Jeffrey W and Amodei D 2020 Scaling laws for neural language models
- Kipf T N and Welling M 2017 Semi-supervised classification with graph convolutional networks *5th Int. Conf. on Learning Representations, ICLR 2017 — Conf. Track Proc.*
- Kothe S and Kirton P 2023 Liouville space neural network representation of density matrices
- Lange H, Bornet G, Emperauger G, Chen C, Lahaye T, Kienle S, Browaeys A and Bohrdt A 2024 Transformer neural networks and quantum simulators: a hybrid approach for simulating strongly correlated systems
- Loshchilov I and Hutter F 2017 SGDR: stochastic gradient descent with warm restarts *Int. Conf. on Learning Representations* (available at: <https://openreview.net/forum?id=Skq89Scxx>)
- Loshchilov I and Hutter F 2019 Decoupled weight decay regularization *Int. Conf. on Learning Representations* (available at: <https://openreview.net/forum?id=Bkg6RiCqY7>)
- Lukin M D, Fleischhauer M, Cote R, Duan L M, Jaksch D, Cirac J I and Zoller P 2001 Dipol mbles *Phys. Rev. Lett.* **87** 037901
- Melko R G, Carleo G, Carrasquilla J and Cirac J I 2019 Restricted Boltzmann machines in quantum physics *Nat. Phys.* **15** 887–92
- Merali E, De Vlucht I J S and Roger G Melko 2023 Stochastic series expansion quantum monte carlo for rydberg arrays
- Miles C, Samajdar R, Ebadi S, Wang T T, Pichler H, Sachdev S, Lukin M D, Greiner M, Weinberger K Q and Kim E-A 2023 Machine learning discovery of new phases in programmable quantum simulator snapshots *Phys. Rev. Res.* **5** 013026
- Morawetz S, De Vlucht I J S, Carrasquilla J and Melko R G 2021 U(1)-symmetric recurrent neural networks for quantum state reconstruction *Phys. Rev. A* **104** 012401
- Moss M S, Ebadi S, Wang T T, Semeghini G, Bohrdt A, Lukin M D and Melko R G 2024 Enhancing variational monte carlo simulations using a programmable quantum simulator *Phys. Rev. A* **109** 032410
- Nakaji K et al 2024 The generative quantum eigensolver (GQE) and its application for ground state search
- Nomura Y, Darmawan A S, Yamaji Y and Imada M 2017 Restricted boltzmann machine learning for solving strongly correlated quantum systems *Phys. Rev. B* **96** 205152
- Samajdar R, Ho W W, Pichler H, Lukin M D and Sachdev S 2020 Complex density wave orders and quantum phase transitions in a model of square-lattice rydberg atom arrays *Phys. Rev. Lett.* **124** 103601
- Samajdar R, Ho W W, Pichler H, Lukin M D and Sachdev S 2021 Quantum phases of rydberg atoms on a kagome lattice *Proc. Natl Acad. Sci.* **118** e2015785118
- Shijie W, Irsoy O, Steven L, Dabrovolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D and Mann G 2023 BloombergGPT: a large language model for finance
- Sprague K and Czischek S 2024 Variational Monte Carlo with large patched transformers *Commun. Phys.* **7** 90
- Steven W R 2009 Minimally entangled typical quantum states at finite temperature *Phys. Rev. Lett.* **102** 190601
- Torlai G et al 2019 Integrating neural networks with a quantum simulator for state reconstruction *Phys. Rev. Lett.* **123** 230504
- Torlai G and Melko R G 2018 Latent space purification via neural density operators *Phys. Rev. Lett.* **120** 240503
- Torlai G, Mazzola G, Carrasquilla J, Troyer M, Melko R and Carleo G 2018 Neural-network quantum state tomography *Nat. Phys.* **14** 447–50
- Touvron H et al 2023 *Edouard Grave and Guillaume Lample* (Open and Efficient Foundation Language Models)
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Łukasz and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* pp 5999–6009
- Viteritti L L, Ferrari F and Becca F 2022 Accuracy of restricted Boltzmann machines for the one-dimensional $J_1 - J_2$ Heisenberg model *SciPost Phys.* **12** 166
- Viteritti L L, Rende R and Becca F 2023 Transformer variational wave functions for frustrated quantum spin systems *Phys. Rev. Lett.* **130** 236401
- Viteritti L L, Rende R, Parola A, Goldt S and Becca F 2024 Transformer wave function for the shastry-sutherland model: emergence of a spin-liquid phase
- Wang H, Weber M, Izaac J and Lin C Y-Y 2022 Predicting properties of quantum systems with conditional generative models
- Wang Y, Sun Y, Liu Z, Sarma S E, Bronstein M M and Solomon J M 2019 Dynamic graph cnn for learning on point clouds *ACM Trans. Graph.* **38** 1–12
- Wu X, Liang X, Tian Y, Yang F, Chen C, Liu Y-C, Tey M K and You Li 2021 A concise review of Rydberg atom based quantum computation and quantum simulation *Chin. Phys. B* **30** 020305
- Xu W, Venkatramani A V, Cantú S H, Šumarac T, Klüsener V, Lukin M D and Vuletić V 2021 Fast preparation and detection of a rydberg qubit using atomic ensembles *Phys. Rev. Lett.* **127** 050501
- Yun C, Bhojanapalli S, Rawat A S, Reddi S J and Kumar S 2020 Are transformers universal approximators of sequence-to-sequence functions?
- Zhang C et al 2023 A complete survey on generative ai (AIGC): is chatgpt from gpt-4 to gpt-5 all you need?
- Zhang Y-H and Di Ventra M 2023 Transformer quantum state: a multipurpose model for quantum many-body problems *Phys. Rev. B* **107** 075147