

Interpretable Artificial Intelligence Analysis of Strongly Correlated Electrons

Changkai Zhang (张昌凯)^{*} and Jan von Delft*Arnold Sommerfeld Center for Theoretical Physics, Center for NanoScience, and Munich Center for Quantum Science and Technology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany*

(Received 16 December 2025; accepted 15 September 2026; published 29 October 2026)

Artificial intelligence (AI) has become an exceptionally powerful tool for analyzing scientific data. In particular, attention-based architectures have demonstrated a remarkable capability to capture complex correlations and to furnish interpretable insights into latent, otherwise inconspicuous patterns. This progress motivates the application of AI techniques to the analysis of strongly correlated electrons, which remain notoriously challenging to study using conventional theoretical approaches. Here, we propose AI workflows for analyzing snapshot datasets from tensor-network simulations of the two-dimensional (2D) Hubbard model over a broad range of temperature and doping. The 2D Hubbard model is an archetypal strongly correlated system, hosting diverse intriguing phenomena including Mott insulators, anomalous metals, and high- T_c superconductivity. Our AI techniques yield fresh perspectives on the intricate quantum correlations underpinning these phenomena and facilitate universal omnimetry for ultracold-atom simulations of the corresponding strongly correlated systems.

DOI: [10.1103/jsgs-9tx8](https://doi.org/10.1103/jsgs-9tx8)

I. INTRODUCTION

The invention of artificial intelligence (AI) has revolutionized the way we interrogate and interpret scientific data. The attention scheme [1–5] has been proven highly effective in transduction tasks in conjunction with recurrent or convolutional networks. Afterward, the transformer [6]—an architecture built solely upon the attention scheme—was demonstrated to be compelling in capturing global dependencies in sequential data. Over the past decade, transformerlike architectures have dramatically enhanced the capability of the AI models across various domains, including natural language processing (NLP) [7–12], computer vision [13–16], bioinformatics [17–19], and numerous other areas [20–23]. Compared with alternative designs, the attention mechanism in the transformer excels particularly at encoding the correlation structure of the input data. This feature motivates the application of the transformer models in studying strongly correlated electrons.

Strongly correlated systems [24–27] are governed by considerably strong interactions, inducing collective behaviors that defy descriptions hinged on individual (quasi)particles. Canonical examples include Mott insulators [28–30], high- T_c superconductors [31–34], heavy-fermion materials [35–38], fractional quantum Hall systems [39–41], spin liquids [42–45], and quark-gluon plasmas [46–49]. The consequent high levels of quantum entanglement and correlations render these systems notoriously challenging for conventional theo-

retical approaches. With advances in computational hardware, a handful of numerical algorithms—among them quantum Monte Carlo [50–53], dynamical mean-field theory [54–57], density matrix renormalization group [58–62], and various ground-state [63–70] or finite-temperature [71–79] tensor network (TN) methods—have been devised to tackle strongly correlated systems. Moreover, quantum simulation apparatuses based on ultracold atoms [80–85] have achieved substantial progress in emulating strongly interacting lattice systems. Together, these techniques offer valuable many-body data from which the AI models can learn and distill meaningful insights.

Among the plethora of strongly correlated electron systems, the two-dimensional (2D) Hubbard model [86,87] stands out as a paradigmatic arena for a variety of intriguing phenomena, such as Mott physics, anomalous metals, and high- T_c superconductivity. The Hubbard model encapsulates the essential physics of itinerant electrons on a lattice with strong on-site Coulomb repulsion. Over the past few decades, the 2D Hubbard model has been subject to intensive investigations both numerically [52,53,70,78,79,88–98] and experimentally [80–85,99–108]. Robust antiferromagnetic (AFM) orders have been confirmed near half-filling [70,93,97,109–112], while in the doped regime—especially with carrier hopping beyond neighboring sites—diverse charge and spin orders, often coexisting with or competing against pairing tendencies, have been identified [70,78,93–95,113]. These properties broadly echo the observations in the cuprate superconductors.

Despite this decent progress, the majority of the existing researches focused on local and low-order spin and/or charge correlations, especially two-point correlators. However, mounting evidence indicates that high-order [108,114,115], nonlocal [116,117], polaronic [82,100,118,119] or otherwise stringlike [99,120,121] correlations play a pivotal role in

^{*}Contact author: cxzh@hku.hk

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

deciphering the complicated phase diagram of the 2D Hubbard model. This recognition highlights the promise of the AI techniques for the global vision of the underlying quantum correlations.

In this article, we study the 2D Hubbard model using specifically designed AI models. We start with assembling a dedicated dataset of snapshots across categories of temperatures and doping levels by sampling the thermal density matrix via TN simulations. Then, we propose two AI architectures classifying snapshots into the respective categories: the *pro architecture*, an analog of the encoder-only transformer, and the *core architecture*, a streamlined variant in which the feedforward transformations are deferred until after the attention stack. This semilinear design constitutes a key methodological innovation of our work and achieves performance comparable to that of the *pro architecture*, while offering better support for parallelism and improved interpretability.

Next, we perform multiple analyses on the trained core model. We use a confusion analysis to measure the quality of the classification tasks and obtain insights into the aggregate strength of quantum correlations in each category. Exploiting the *semilinear* structure of the attention stack in the core architecture, we propose an interpretation in terms of an effective Markovian dynamics, demonstrating the alignment of the attention design with intrinsic features of the physical system. Further examinations on the orthogonality relationships of the embedding and the attention maps are provided in the Supplemental Material [122].

Finally, we demonstrate an application of our core AI models as a universal omnimeter for ultracold-atom quantum simulations. The AI classifiers produce probabilistic scores (logits) for each category, which serve as posterior likelihoods conditioned on a snapshot acquired in the experiment. Averaging these outputs over a snapshot ensemble from repeated observations thus provides an empirical probability distribution over the categories. Once the categories are calibrated with predetermined physical quantities, the expectation values weighted by the probability distribution yield an accurate estimate of the corresponding quantities for the ensemble.

II. LATTICE MODEL AND DATASET

Lattice models serve as common platforms for the physics of crystalline materials, wherein charge carriers reside on and hop between discrete lattice sites. In many materials of interest, itinerant electrons predominantly occupy the outermost s orbital for transport. Consequently, the local Hilbert space at each lattice site is spanned by four basis: empty $|\emptyset\rangle$, spin-up $|\uparrow\rangle$, spin-down $|\downarrow\rangle$, and doubly occupied $|\uparrow\downarrow\rangle$ state.

In our study, we focus on the quintessential 2D Hubbard model on an 8×8 square lattice with open boundary conditions, defined via the following Hamiltonian:

$$\mathcal{H} = - \sum_{i,j,\sigma} t_{ij} [c_{i\sigma}^\dagger c_{j\sigma} + \text{H.c.}] + U \sum_i n_{i\uparrow} n_{i\downarrow}. \quad (1)$$

Here, $c_{i\sigma}^\dagger$ ($c_{i\sigma}$) creates (annihilates) an electron with spin σ on site i and $n_{i\sigma} = c_{i\sigma}^\dagger c_{i\sigma}$ denotes the corresponding number operator. The first term in Eq. (1) describes the kinetic energy

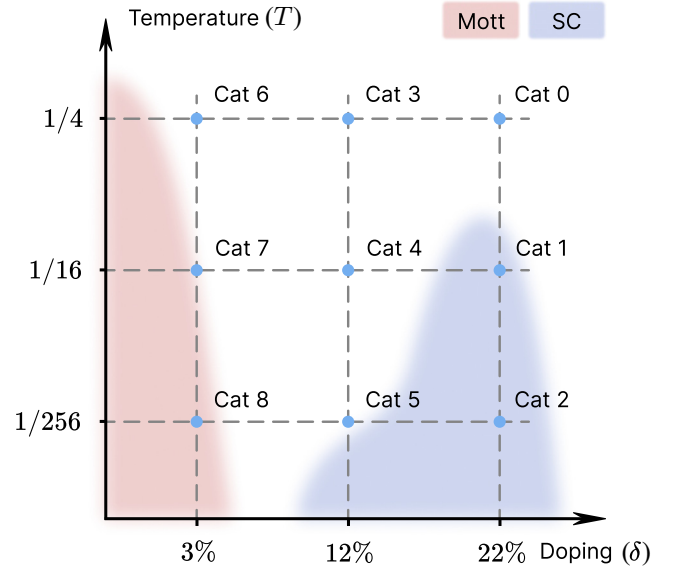


FIG. 1. A schematic depiction of the locations in phase space for the nine categories (Cat), created by combining three choices of temperatures (high, medium, and low) with three doping regimes (overdoped, medium doped, and underdoped). The red and blue freehand-shaded areas mark the AFM Mott insulating phase and the high- T_c superconducting phase, respectively, as expected for the Hubbard model. The charge doping varies with temperature (see also Fig. 5); precise values are provided in the Supplemental Material [122].

associated with electron hopping between sites i and j with amplitude t_{ij} , while the second term accounts for the on-site Coulomb repulsion with strength U . Throughout this work, we consider the minimal Hubbard model, where $t_{ij} = 1$ for nearest-neighbor pairs and $t_{ij} = 0$ otherwise. Also, we set $U = 10$ as established to be realistic for cuprate materials [123,124]. In many elemental metals, electron interactions are effectively weak ($U \approx 0$) due to electric-field screening by the surrounding lattice ions, yielding conventional metallic behavior at half-filling (one electron per site). In contrast, in materials like high- T_c cuprates, the on-site Coulomb repulsion becomes abnormally strong for electrons, opening a large energy gap that penalizes double occupancy. Hence, electron motion is substantially hindered by a large potential barrier and the system exhibits Mott insulating behavior at the macroscopic level.

Extra charge carriers can be introduced by adding (electron doping) or removing (hole doping) electrons relative to half-filling. Upon sufficient doping, charge transport in the material sets in and superconductivity may emerge, signified by enhanced pairing correlations at low temperatures. This evolution underlies the schematic phase diagram as shown in Fig. 1, where red and blue shaded areas mark the AFM Mott insulating phase and the high- T_c superconducting dome, respectively.

The (unnormalized) thermal density matrix $\rho = e^{-\beta\mathcal{H}}$ characterizes the statistical state of the lattice system, with inverse temperature $\beta = 1/T$. Note that ρ admits a Taylor expansion at high temperature (small β) for a given

Hamiltonian, and that

$$\rho(2\beta) = \rho(\beta)\rho(\beta). \quad (2)$$

Accordingly, one may cool the system down by repeatedly squaring the thermal density matrix starting from a high-temperature construct. This idea underlies the exponential Tensor Renormalization Group (XTRG) method [76,77,79,101,125], which offers a comprehensive thermal description of the lattice system over a broad temperature range. We thus employ XTRG to produce thermal density matrices at high, medium, and low temperatures at overdoped, medium-doped, and underdoped regions (hole doped), yielding nine categories as indicated in Fig. 1.

We then perform standard sitewise sampling on each thermal density matrix [79] to obtain snapshots (Fock bases of the many-body Hilbert space) of the lattice system. The snapshots, each consisting of 8×8 cells containing $|\emptyset\rangle$, $|\uparrow\rangle$, $|\downarrow\rangle$, or $|\uparrow\downarrow\rangle$, are next flattened according to row-major order into a sequence of 64 elements. For each location in the phase space, we generate 1000 snapshots, yielding a dataset of 9 categories and 9000 snapshots in total. This dataset is randomly partitioned into training (90%) and test (10%) subsets for the subsequent AI workflows.

III. ARCHITECTURES

Our AI architectures originate from the transformer paradigm while being tailored for categorical classification. Given an input snapshot, the objective is to infer a probability distribution over categories. Rather than appending a dedicated CLS token [8,16] as a label, we adopt a streamlined design that endows the involved attention mechanism with a distinctive and physically meaningful interpretation.

Figure 2 depicts schematic layouts of the models deployed in this study. Both architectures share the same foundational components—input codecs, a stack of multihead attention, feedforward networks (FFNs), and a terminal linear classification head. The pro architecture (right) is inspired by an encoder-only transformer, whereas the core architecture (left) leaves out the feedforward networks between attention blocks, thereby giving enhanced parallelism and improved interpretability.

Our exposition below assumes familiarity with standard components and techniques in the practices of the transformer architecture [6], and hence will focus on our distinctive designs. Full technical details can be found in the Supplemental Material [122].

A. Tokenization and input codecs

The tokenization is straightforward in our setting, as the *vocabulary* (local Hilbert space) comprises only four distinct *words* (local states). We therefore assign 0, 1, 2, and 3 to the empty, spin-up, spin-down, and doubly occupied states, respectively. Under this encoding, each snapshot in the dataset now becomes a sequence of integers (tokens). Formally, let $\mathcal{S} = \{0, 1, 2, 3\}$ denote the tokenized local state space. The input sequence becomes $\vec{\sigma} \in \mathcal{S}^L$, where L is the flattened sequence length.

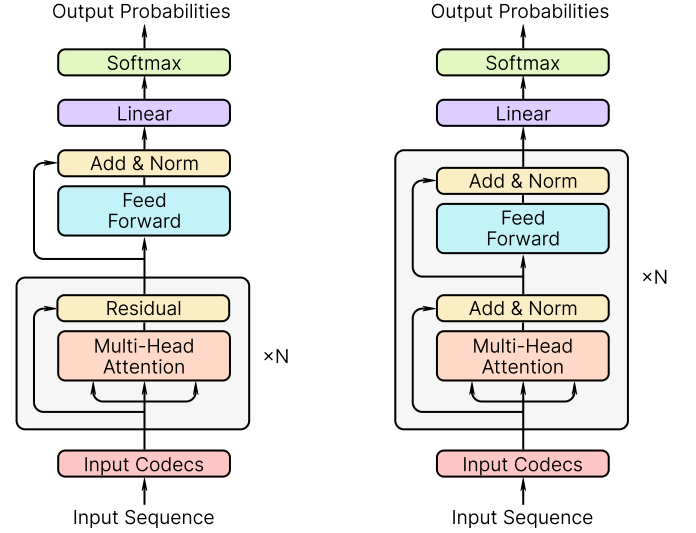


FIG. 2. Schematic illustrations of the core (left) and the pro (right) architecture for classification of sequential inputs. Both architectures comprise input codecs, multihead attention blocks, feedforward networks, and a final linear classification head. The pro architecture is an analog of the encoder-only transformer, while the core architecture leaves out the feedforward networks between attention blocks, which enhances parallelism and improves interpretability.

The input codecs accept and map each tokenized sequence into the model's latent parameter space in two stages. Each token σ is first transformed into a d_{model} -dimensional embedding vector $e(\sigma) \in \mathbb{R}^{d_{\text{model}}}$ via a learnable embedding module; this embedding depends solely on the token (local state) and is agnostic to its location in the sequence. To preserve positional information, we then add a positional vector $q_i \in \mathbb{R}^{d_{\text{model}}}$ for site i to each embedding vector. For this purpose, we employ the sinusoidal positional encoding [6], which has proved effective across a wide range of applications.

The input codecs thus assemble a feature matrix Σ with elements $\Sigma_i^\mu = e^\mu(\sigma_i) + q_i^\mu$ for each snapshot, where i indexes lattice sites and μ indexes latent dimensions. Contingent on the dataset under consideration, d_{model} should be adjusted for a balance of expressivity against overfitting. Note that in Ref. [6] the input embeddings are multiplied by a factor of $\sqrt{d_{\text{model}}}$ to scale up the weights; in contrast, this operation is empirically detrimental in our application domain, plausibly due to the exceedingly small vocabulary size relative to the sequence length.

B. Locality-biased attention

The (multihead) attention mechanism plays a central role in harnessing global correlation awareness for both architectural designs. We adopt the prevalent scaled dot-product attention scheme [1,6] to acquire raw attention scores, and subsequently impose a locality bias for an improved training profile. Even though positional information has been encoded amid the input codecs, we find that the prototypical attention setup, which is primarily designed for one-dimensional sequences, struggles in perceiving 2D spatial relationships. Hence, an explicit locality bias assists in this regard.

We start with linear projections of the input embeddings into query, key, and value vectors

$$\mathcal{Q}_i = \Sigma_i W_Q, \quad \mathcal{K}_i = \Sigma_i W_K, \quad \mathcal{V}_i = \Sigma_i W_V, \quad (3)$$

with W_Q , W_K , and W_V being learnable weight matrices. Here, we suppress the latent-space index μ and take matrix multiplications implicit. The attention between sites i and j thus reads

$$A_{ij} = \text{softmax}_j(\mathcal{Q}_i \mathcal{K}_j / \mathfrak{T}), \quad (4)$$

where $\mathfrak{T}$ denotes the *model temperature* (conceptually distinct from the physical temperature) that controls the sharpness of the attention distribution. In our exercises, $\mathfrak{T} = \sqrt{d_k}$ (see below for the definition of d_k) works reasonably well.

The multihead attention is realized by partitioning the latent space into h subspaces, each with dimension $d_k = d_{\text{model}}/h$. Attention is computed independently within each head, after which the head outputs are concatenated and linearly transformed to yield the final raw attention. In our actual practice, the multihead configurations fail to outperform their single-head counterpart, likely attributable to the limited size of the training set.

Many realistic physical systems exhibit locality: Objects only significantly influence their immediate neighbors, leading to a decay of the interactions and correlations with spatial separation. Accordingly, we apply a locality bias to the raw attention scores

$$\mathcal{A}_{ij} = \text{softmax}_j(A_{ij} \circ G_{ij}), \quad (5)$$

where G_{ij} is a handcrafted bias function that decays with the physical distance d_{ij} between sites i and j , and the circle $\circ$ denotes elementwise (Hadamard) multiplication. This locality bias encourages the mechanism to focus on nearby sites, effectively accelerating the convergence during the training process. In our implementation, we choose a Gaussian kernel

$$G_{ij} = \exp\{-d_{ij}^2/2\zeta^2\}, \quad (6)$$

with standard deviation $\zeta = \lambda/2$ and λ a characteristic length scale of the system (e.g., $\lambda = 8$ for the 8×8 square lattice considered here). The eventual performance is not highly affected by the specific choice of the bias function G_{ij} . For instance, a power-law decay kernel works almost equally well.

The locality bias serves as a tunable inductive prior. Its functional form and characteristic range can be adapted to the anticipated correlation structure of the physical system. For systems in which long-range correlations are expected to be dominant, the bias may be weakened or removed. As an example, in the synthetic derangement benchmark presented in the Supplemental Material [122], the architectures are trained without a locality bias to capture correlations between spatially separated columns.

Finally, as a standard technique to stabilize the gradient propagation, we apply a residual connection [126] by adding a Σ_i to the output of the attention block as

$$\text{attn}(\Sigma_i) = \Sigma_i + \sum_j \mathcal{A}_{ij} \mathcal{V}_j. \quad (7)$$

Afterward, layer normalization [127] is employed in the pro architecture, whereas the core architecture omits this step for reasons that will become clear in the ensuing interpretation.

C. Feedforward and classification

The feedforward networks constitute one of the principal sources of nonlinearity in the model. Each FFN is a sitewise fully connected three-layer perceptron comprising an input layer, a hidden layer, and an output layer. The input and output layers have width d_{model} , while the hidden layer has width d_{hidden} . A ReLU activation is applied between the two affine maps to introduce nonlinearity. Concretely, the FFN reads

$$\text{FFN}(\Sigma_i) = \text{ReLU}(\Sigma_i W_1 + \mathbf{b}_1) W_2 + \mathbf{b}_2, \quad (8)$$

where W_1 , W_2 , $\mathbf{b}_1$, and $\mathbf{b}_2$ are learnable weights and biases. The same linear maps are shared across all sites i , whereas different FFN blocks carry independent parameters. A residual connection and layer normalization follow each FFN.

The arrangement of FFNs constitutes the pivotal difference between the pro and core architectures. The pro variant inserts an FFN of dimension $d_{\text{hidden}} = d_{\text{ff}}$ after each attention block, while the core variant defers nonlinearity to a single FFN of dimension $d_{\text{hidden}} = N \times d_{\text{ff}}$ after the entire stack of N attention blocks. We call the latter design *semilinear* attention stack (for the reason that will become clear in the impending interpretation).

Under this parametrization, the core and pro architectures contain an almost equivalent amount of learnable model parameters. However, the postponed feedforward network markedly reduces the depth of nonlinearity within the core model, which benefits the upcoming interpretation since physical objects commonly propagate linearly. Also, centralizing the FFN boosts parallelism during both training and inference.

The classification head ingests the abstract embedding Σ_i^{ff} processed through the preceding attention and feedforward networks, and returns a categorical distribution over the target labels. Concretely, the logit $y_{i,c}$ and the corresponding probability $p_{i,c}$ are computed as

$$y_{i,c} = \Sigma_i^{\text{ff}} W_c + \mathbf{b}_c, \quad p_{i,c} = \text{softmax}_c(y_{i,c}), \quad (9)$$

where W_c and $\mathbf{b}_c$ denote the learnable weight and bias associated with category c . Each site i produces its own distribution; this is warranted by the attention mechanism, which injects into each site contextual information aggregated from all other sites. The overall prediction is obtained by the argmax of the averaged per-site distributions $p_c = \text{avg}_i p_{i,c}$ over all lattice sites.

D. Hyperparameters

For the XTRG-generated snapshot dataset of the Hubbard model, we adopt the following hyperparameters: embedding dimension $d_{\text{model}} = 128$; single-head attention $h = 1$; feedforward dimension $d_{\text{ff}} = 1024$; and $N = 2$ attention blocks. Increasing either the number of heads or the number of attention blocks empirically induces severe overfitting and should be considered only after expanding the dataset.

Furthermore, we implement an ablation toggle that trivializes all attention blocks by hard-setting $A_{ij} = 1$ when enabled.

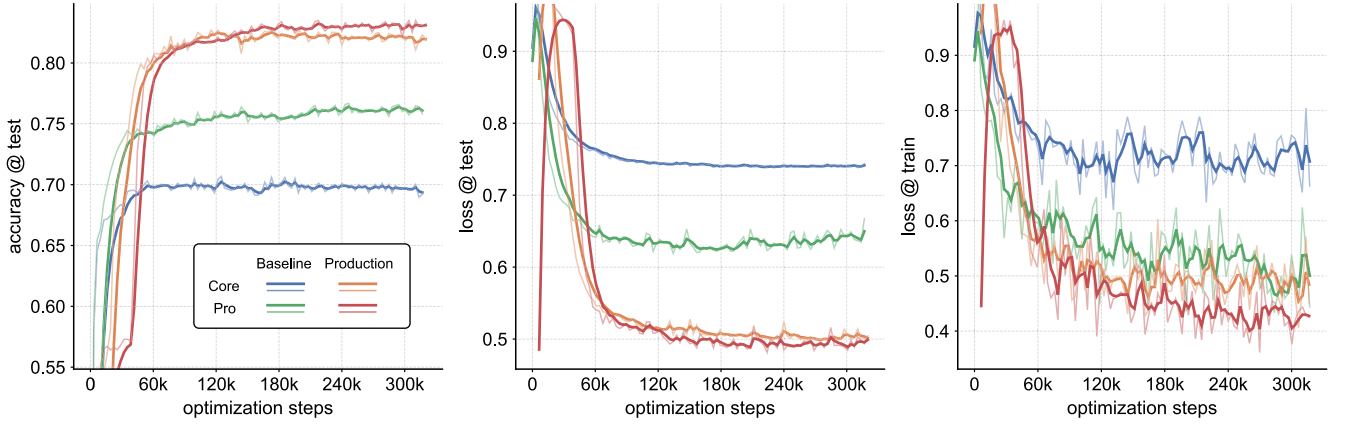


FIG. 3. Training profiles and benchmarks of the production and baseline models following the core and pro architectures. Metrics are displayed as raw data (thin lines with muted color) and with an exponential smoothing factor $\alpha = 0.4$ (thick lines with deep color). For baseline models (with trivialized attention), the pro architecture consistently outperforms the core variant across all metrics, congruous with the anticipated benefits of elevated nonlinearity in the pro model. In contrast, for production models (with full-functional attention), the core architecture achieves merely negligible gaps in performance, indicating an alignment of the semilinear attention with the intrinsic properties of the dataset.

Under this switch, the effective attention reduces to the fixed locality kernel, thereby delegating the classification task entirely to the FFNs. This toggle is useful both as a baseline and during the warm-up phase of training.

IV. TRAINING AND BENCHMARKS

Hardware. All the production and baseline models were trained on an NVIDIA 3090 GPU. The project utilizes PyTorch version 2.2.2 and CUDA 11.8.

Initialization. All learnable parameters are initialized with Xavier (Glorot) initialization [128]. For production runs, the first 200 epochs serve as a warm-up phase during which the ablation toggle is enabled, effectively pretraining the FFNs with trivialized attention.

Batch and epochs. We use a batch size of 256 and train for a total of 10 000 epochs. Checkpoints are saved every 100 epochs, and the one with the lowest validation loss is selected as the final deliverable.

Objective. We optimize the ubiquitous cross-entropy loss as our optimization objective. The loss is computed as the negative log-likelihood averaged over all sites:

$$\text{loss} = \text{avg}_i \left[- \sum_c \xi_c \ln p_{i,c} \right], \quad (10)$$

where ξ_c denotes the ground-truth one-hot label for category c (broadcasted across all sites), or specifically

$$\xi_c = \begin{cases} 1, & \text{for correct category } c, \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Optimizer. We use the standard Adam optimizer [129] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-9}$. The learning rate is fixed to 5×10^{-6} . We find that both architectural designs are sensitive to this setting: Materially larger or smaller values tend to induce premature plateaus at elevated loss.

Regularization. We apply dropouts [130] with a rate of 0.01 to input codecs, attention blocks, FFNs, and all residual

connections. Contrary to the common practice in the NLP applications, we disable the label smoothing [131] as over-confidence is not a primary concern for a physically generated dataset.

Training yield. Owing to stochastic initialization, training outcomes exhibit variability. Empirically, roughly one in seven attempts attains a top-performing model.

Benchmarks. Figure 3 summarizes the training profiles of four models we trained: the production and baseline models under the core and pro architectures. The baselines are trained with the aforementioned ablation toggle switched on, such that their attention mechanisms are effectively disabled.

Benchmarks are reported as test-set accuracy and loss, along with training loss. Test accuracy is defined as the fraction of correctly classified snapshots in the held-out test set. All four models achieve stable convergence given sufficient training, and generalization is satisfactory as indicated by the proximity of training and test losses.

Moreover, the production models consistently surpass their baseline counterparts, attesting to the efficacy of the attention mechanism. For the baselines, the pro variant outperforms the core across all metrics, consistent with the anticipated benefits of enhanced nonlinearity. In contrast, in production runs, the core model closes the gap to within negligible differences, indicating that FFNs interleaved between attention blocks are largely redundant and that the semilinear attention stack aligns well with the intrinsic structure of the dataset.

V. INTERPRETATION

For decades, achieving a principled interpretation of the internal mechanism of AI models has been one of the highest endeavors in the field [132]. Unlike opaque black-box approaches [133], the attention mechanism offers a natural lens on a model's focus and decision-making. The calculated attention scores $\mathcal{A}_{ij}$ are typically construed as a measure of *importance* of token j to token i (or equivalently, the *attention* paid by token i to token j) [2,3]. This heuristic has been widely

utilized in a variety of application domains [4,134–142] for analysis, diagnosis, and debugging.

However, a comprehensive study [143] showed that attention weights often fail to provide consistent or exclusive explanations of model predictions; in particular, alternative attention patterns can yield essentially identical performance. These observations have ignited a prolonged debate [143–151] on whether—or to what extent—the attention meaningfully reveals a model’s reasoning process.

Despite the ongoing controversy, consensus remains that attention maps furnish at least *an* (if not *the*) explanation for the inner workings of the model [146]. Thereafter, further aggregation methods, such as attention rollout [152–154] and attention flow [152,155,156], have been proposed to propagate attention scores across multiple layers. The attention rollout, in particular, essentially performs a layerwise matrix multiplication of the (residual-augmented) attention matrices. Considering the fact that these matrices are all row stochastic, it becomes natural to interpret them as Markovian transition kernels [153]. This probabilistic viewpoint forms the basis of our interpretation.

Our interpretation focuses exclusively on the core architecture, as it depends critically on the *semilinear* nature of the attention stack (the precise meaning of which will be clarified in the upcoming subsections).

A. Classical and quantum Markov process

Before heading to the interpretation, we first formalize the relevant constructs for both classical and quantum Markov processes on the lattice system. For the classical scenario, consider a discrete-time Markov dynamics in which, at each update, the state s_i at site i may overwrite the state s_j at site j ; the corresponding transition probability is $\mathcal{A}_{ij}$. All lattice sites update synchronously in one time step.

Suppose that a collection of observables is associated with each local state, and write Σ_i^μ for the μ th observable evaluated on the state at site i . Under the Markov evolution described above, the observables update after one step as

$$\Sigma_i^\mu \leftarrow \sum_j \mathcal{A}_{ij} \Sigma_j^\mu. \quad (12)$$

For the quantum scenario, the local state can be associated with a (pure) local density matrix $\rho_i = |\sigma_i\rangle\langle\sigma_i| \equiv |i\rangle\langle i|$. The quantum Markov process is specified by a completely positive trace-preserving map $\mathcal{E}$ (also known as a quantum channel) comprising a set of Kraus operators

$$K_j = \sum_i \sqrt{\mathcal{A}_{ij}} |j\rangle\langle i|, \quad (13)$$

whose action on the local state is

$$\mathcal{E}(\rho_i) = \sum_j K_j \rho_i K_j^\dagger = \sum_j \mathcal{A}_{ij} \rho_j. \quad (14)$$

Assign to each ρ_i the same family of observables Σ_i^μ . Under the channel $\mathcal{E}$, these observables also evolve according to Eq. (12). Therefore, in both classical and quantum constructs, Eq. (12) captures the one-step evolution of observables under the Markov process with the transition kernel $\mathcal{A}_{ij}$.

B. Interpretation of the attention stack

We now make explicit the link between the attention stack in the core architecture and the Markovian description above. Viewing the embedded features as observables, the right-hand side of Eq. (12) coincides with the application of the attention matrix $\mathcal{A}_{ij}$ to the input embedding Σ_i^μ . In conjunction with the embedding projector W_V and the residual connection, the output of a stack of $N = 2$ attention blocks can be expressed as

$$\begin{aligned} \Sigma_i^{\text{attn}} &= \Sigma_i + \sum_k \mathcal{A}_{ik}^{(1)} \Sigma_k W_V^{(1)} \\ &\quad + \sum_j \mathcal{A}_{ij}^{(2)} \left[\Sigma_j + \sum_k \mathcal{A}_{jk}^{(1)} \Sigma_k W_V^{(1)} \right] W_V^{(2)} \\ &= \Sigma_i + \sum_j \mathcal{A}_{ij}^{(1)} \Sigma_j W_V^{(1)} + \sum_j \mathcal{A}_{ij}^{(2)} \Sigma_j W_V^{(2)} \\ &\quad + \sum_{j,k} \mathcal{A}_{ij}^{(2)} \mathcal{A}_{jk}^{(1)} \Sigma_k W_V^{(1)} W_V^{(2)}, \end{aligned} \quad (15)$$

where $\mathcal{A}_{ij}^{(\ell)}$ and $W_V^{(\ell)}$ denote, respectively, the attention matrix and the embedding projector from the ℓ th attention block. On the right-hand side of the final equality, four terms appear: The first term carries the original embedding; the second and third terms propagate features according to the first and second attention kernels; and the final term captures the cascaded propagation through both blocks. These terms can be interpreted as four distinct *Markovian propagation modes* acting on the input Σ_i . The same expansion generalizes analogously to deeper stacks.

For trivial embedding projectors $W_V^{(\ell)} = \mathbb{1}$ (where $\mathbb{1}$ is the identity matrix), Eq. (15) reduces to

$$\Sigma_i^{\text{attn}} = \sum_j \mathcal{R}_{ij} \Sigma_j = \sum_j \left[\prod_\ell (\mathbb{1} + \mathcal{A}^{(\ell)})_{ij} \right] \Sigma_j, \quad (16)$$

where $\mathcal{R}_{ij}$ is precisely the standard attention rollout [152] (without normalization). We refer to this limiting case as a *linear* attention stack. However, in practice, the embedding projectors are generally nontrivial, yielding the *semilinear* attention stack. This is the unique feature of the core architecture; in contrast, in the pro variant, the layer normalization and the FFN introduce nonlinearity between attention blocks.

This perspective furnishes an interesting interpretation of the core architecture. First, the input codecs learn a feature embedding whose components can be construed as physically relevant observables attached to each local state. The attention stack then effects a superposition of Markovian propagation modes that evolve these observables across the lattice, while residual connections preserve the original features. Finally, the classification head operates on the resultant evolved features (observables). In essence, the model learns an *effective* Markov dynamics—encoded in the attention kernels and embedding projections—that best aligns the propagated observables with the downstream classification objective. This structural interpretation provides a concrete diagnostic handle on the learned attention dynamics, though further connections to specific microscopic observables and effective theories of the Hubbard model remain to be established.

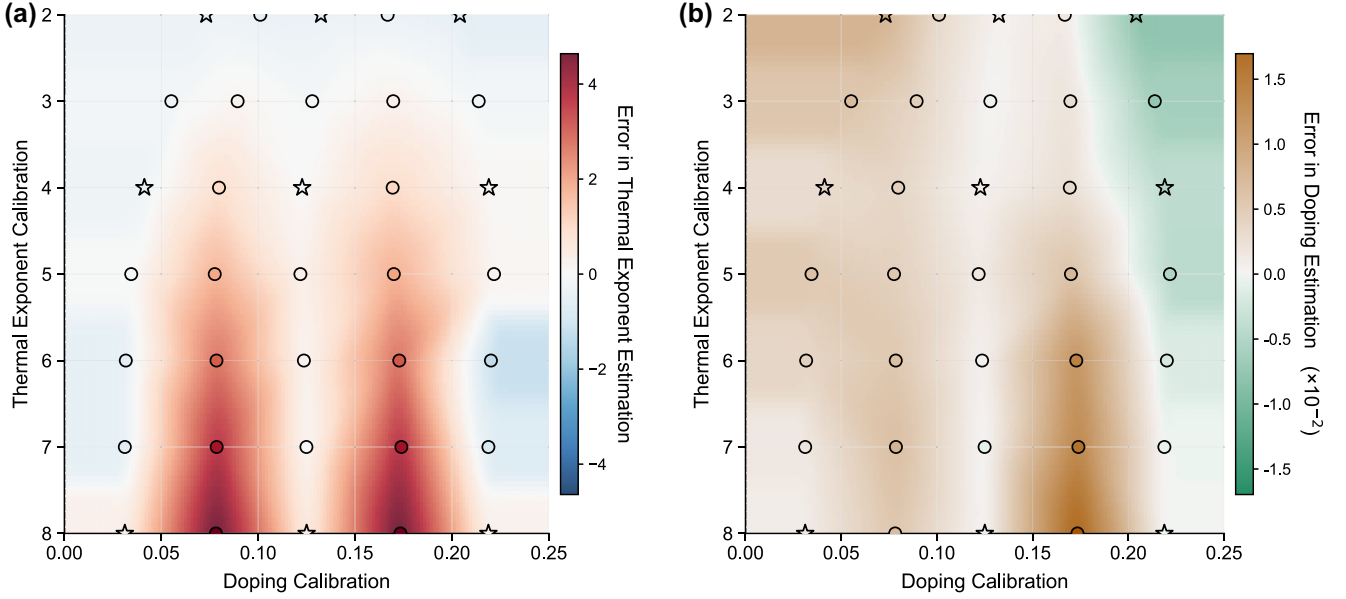


FIG. 5. Error in the omnimeter estimation of (a) the thermal exponent and (b) charge doping. Open circles/pentagrams mark the locations in phase space of the snapshot ensembles under evaluation, with pentagrams (circles) indicating data included (not included) in the training set. Color scales are obtained via interpolation. Overall performance is good, except for the bands at the unseen doping levels [around 7% and 17% in panel (a) and around 17% in panel (b)].

VII. UNIVERSAL OMNIMETRY

One of the straightforward applications of our AI classifier is to perform measurements on an arbitrary ensemble of snapshots. Each category in the dataset is affiliated with a set of known observables, and the classifier outputs a distribution over categories, thereby inducing an estimate of the corresponding observables. We refer to this procedure as *universal omnimetry*, since a single classifier acting on the same raw snapshot ensemble can simultaneously infer all observables affiliated with its categories.

For a demonstration of this new technique, we augment the dataset with additional snapshots drawn from regions of phase space that were not included in the training set (i.e., *unseen* by the model). A sufficiently generalizable model should then produce a distribution over categories that reflects the resemblance of the correlation patterns in the input ensemble against those learned during training. Thus, the performance of the omnimeter serves as a probe of the underlying correlation structure across distinct partitions of phase space.

The workflow starts with a *calibration* of observables for all categories in the training set. Let $\omega_c^{(\alpha)}$ denote the α th observable affiliated with category c . In our study of the finite-temperature Hubbard model, the affiliated observables are temperature and charge doping. As XTRG algorithm produces thermal density matrices at temperatures $T = 1/2^{n_T}$, we utilize the *thermal exponent* n_T as a representative observable in place of temperature. The calibrated categories thereby provide the reference domain in which the omnimeter associates ensemble-level snapshot patterns with temperature and doping. Additional observables can, of course, be accommodated, provided the dataset supports and a corresponding calibration is available.

Next, given an ensemble of snapshots $\{x\}$ generated under fixed conditions, the core classifier returns, for each snapshot, a categorical distribution $p_c(x)$ over c . Averaging these distributions across the ensemble yields a collective distribution measuring the probability that this ensemble corresponds to category c . The target observable for the ensemble is then estimated by the distribution-weighted average of the calibration:

$$\langle \omega^{(\alpha)} \rangle = \sum_c \omega_c^{(\alpha)} \text{avg}_x p_c(x). \quad (17)$$

Figure 5(a) shows the error of the omnimeter's estimates of the thermal exponent n_T for ensembles drawn from various locations in phase space. Overall performance is satisfactory—particularly at doping levels partially covered in the training set. Notably, two red bands appear at unseen doping levels (around 7% and 17%), with inferred temperatures systematically higher than the ground truth. This bias is plausibly attributed to correlation patterns in those regions that were absent during training; the model interprets these as signatures of elevated thermal fluctuations and hence predicts higher temperature. This behavior, in turn, corroborates that the classifier has genuinely learned to associate correlation patterns with thermodynamic conditions. In particular, the nontrivial variation of the error landscape across phase space indicates that the model responds to variations in the underlying correlation structure, rather than merely performing a nearest-grid interpolation of the calibrated labels. Within the phase-space coverage of its calibration data, the omnimeter should therefore be understood as a correlation-aware diagnostic rather than a simple interpolation scheme.

Figure 5(b) reports the error in estimating charge doping. Performance is again strong, aside from a distinct band at the

unseen doping of around 17%. A similar mechanism applies: The model finds that the unseen patterns resemble those at around 22% doping, consistent with both doping levels lying within the superconducting regime.

These artifacts can be mitigated by enriching the calibration set in the affected portions of phase space. The missing doping levels in the nine-category calibration illustrate that reliable generalization requires representative sampling of the correlation patterns in the target regime. In practice, the calibration grid can be tailored to the resolution required for a given thermometry or doping extraction task. As demonstrated in the Supplemental Material [122], a 25-category omnimeter that includes these doping levels substantially suppresses the artifacts and reduces the relative error in both temperature and doping to below 10%. The reference simulations need only be performed once, after which the trained omnimeter can be applied to many experimental snapshot ensembles. A more comprehensive evaluation will be presented in an upcoming dedicated technical report.

Thermometry remains a central challenge in ultracold-atom experiments [85,106], and our AI omnimeter offers a competitive upgrade. Modern quantum gas microscopy [82,83] produces ensembles of site-resolved snapshots of the analog cold-atom simulator, which can be compared directly with our numerical snapshot dataset. Whereas current thermometers often rely on matching hand-selected metrics and correlators [85,106,107,157], our approach automatically discovers discriminative patterns and aggregates them into robust temperature estimates. We therefore anticipate that this approach can materially enhance the reliability of thermometry in ultracold-atom platforms.

VIII. SUMMARY AND OUTLOOK

In this article, we establish an end-to-end technological stack for AI-assisted analysis of strongly correlated electron systems on a lattice. The workflow starts with tensor-network simulations that generate thermal density matrices and, in turn, an extensive snapshot dataset. This dataset is then processed by our tailored AI architectures featuring locality-biased, semilinear attention with principled interpretability grounded in effective Markovian dynamics and strong capacity to capture latent correlation patterns. The trained model is subsequently subjected to a comprehensive confusion analysis, revealing the prominence and uniqueness of

correlation structures across thermodynamic conditions. Finally, the model is deployed as an omnimeter to infer multiple observables from arbitrary ensembles of snapshots. In its present supervised implementation, the omnimeter operates within the physical domain covered by reliable and representative calibration data; within this domain, however, the costly reference simulations need only be performed once, after which the trained model can be efficiently applied to many experimental snapshot ensembles.

Our research demonstrates the viability of bespoke AI technologies for interrogating challenging strongly correlated quantum systems. The approach is versatile and readily extends to lattice models for diverse physical scenarios. Moreover, the tailored architectures are especially useful for sequential datasets with small *vocabulary* but long *sequence length*, a characteristic feature of many scientific data such as DNA/RNA chains, protein structures, or other high-resolution imaging. The observation that the core model attains performance comparable to the pro variant suggests further opportunities to optimize the transformer architecture. Besides, the universal omnimetry furnishes a generic measurement methodology for quantum many-body experiments equipped with site-resolved quantum microscopy [82,83]. Beyond classifiers, alternative AI paradigms—e.g., generative models—merit applications in, e.g., autoregressive neural quantum states [158–162] or neural transformer backflow [163–165]. We therefore anticipate that this work opens a promising new avenue for the study of strongly correlated systems and will motivate further researches along these lines.

ACKNOWLEDGMENTS

We thank Annabelle Bohrdt, Fabian Döschl, Hannah Lange, Roger Melko, and Henning Schlömer for helpful discussions and feedback. This research was funded in part by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy EXC-2111 (Project No. 390814868), and is part of the Munich Quantum Valley, supported by the Bavarian state government through the Hightech Agenda Bayern Plus.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

-
- [1] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, in *International Conference on Learning Representations* (ICLR, 2015), [arXiv:1409.0473](#).
 - [2] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, and P. Blunsom, Reasoning about entailment with neural attention, in *International Conference on Learning Representations* (ICLR, 2016), [arXiv:1509.06664](#).
 - [3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *Proceedings of the*

- 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei (PMLR, 2015), Vol. 37, pp. 2048–2057.
- [4] Y. Wang, M. Huang, X. Zhu, and L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by J. Su, K. Duh, and X. Carreras (Association for Computational Linguistics, Austin, TX, 2016), pp. 606–615.
- [5] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, Structured attention networks, in *International Conference on Learning Representations* (ICLR, 2017).

- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017), Vol. 30.
- [7] L. Dong, S. Xu, and B. Xu, Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition, in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Piscataway, NJ, 2018), pp. 5884–5888.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), edited by J. Burstein, C. Doran, and T. Solorio (Association for Computational Linguistics, Minneapolis, MN, 2019), pp. 4171–4186.
- [9] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, FastSpeech: Fast, robust and controllable text to speech, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Red Hook, NY, 2019), Vol. 32.
- [10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, Language models are few-shot learners, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Red Hook, NY, 2020), Vol. 33, pp. 1877–1901.
- [11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, Conformer: Convolution-augmented transformer for speech recognition, in *Proceedings of Interspeech 2020* (International Speech Communication Association, 2020), pp. 5036–5040.
- [12] S. Latif, S. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, M. Usama, and J. Qadir, Transformers in speech processing: Overcoming challenges and paving the future, *Comput. Sci. Rev.* **58**, 100768 (2025).
- [13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, in *Computer Vision—ECCV 2020*, edited by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Lecture Notes in Computer Science (Springer International Publishing, Cham, 2020), Vol. 12346, pp. 213–229.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations (ICLR, 2021)*, arXiv:2010.11929.
- [15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, 2021), pp. 9992–10002.
- [16] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, Transformers in vision: A survey, *ACM Comput. Surv.* **54**, 1 (2022).
- [17] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, Perceiver: General perception with iterative attention, in *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang (PMLR, 2021), Vol. 139, pp. 4651–4664.
- [18] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, *et al.*, Highly accurate protein structure prediction with AlphaFold, *Nature (London)* **596**, 583 (2021).
- [19] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, *Proc. Natl. Acad. Sci. USA* **118**, e2016239118 (2021).
- [20] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, Decision transformer: Reinforcement learning via sequence modeling, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates, Red Hook, NY, 2021), Vol. 34, pp. 15084–15097.
- [21] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, Transformers in time series: A survey, in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, edited by E. Elkind (International Joint Conferences on Artificial Intelligence Organization, 2023), pp. 6778–6786.
- [22] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, *Proc. AAAI Conf. Artif. Intellig.* **35**, 11106 (2021).
- [23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks, in *International Conference on Learning Representations (ICLR, 2018)*.
- [24] E. Dagotto, Complexity in strongly correlated electronic systems, *Science* **309**, 257 (2005).
- [25] *Strongly Correlated Systems: Theoretical Methods*, edited by A. Avella and F. Mancini, Springer Series in Solid-State Sciences Vol. 171 (Springer, Berlin, 2012).
- [26] *Strongly Correlated Systems: Numerical Methods*, edited by A. Avella and F. Mancini, Springer Series in Solid-State Sciences Vol. 176 (Springer, Berlin, 2013).
- [27] *Strongly Correlated Systems: Experimental Techniques*, edited by A. Avella and F. Mancini, Springer Series in Solid-State Sciences Vol. 180 (Springer, Berlin, 2015).
- [28] N. F. Mott, The basis of the electron theory of metals, with special reference to the transition metals, *Proc. Phys. Soc. London, Sect. A* **62**, 416 (1949).
- [29] D. B. McWhan, A. Menth, J. P. Remeika, W. F. Brinkman, and T. M. Rice, Metal-insulator transitions in pure and doped V_2O_3 , *Phys. Rev. B* **7**, 1920 (1973).
- [30] M. Imada, A. Fujimori, and Y. Tokura, Metal-insulator transitions, *Rev. Mod. Phys.* **70**, 1039 (1998).
- [31] J. G. Bednorz and K. A. Müller, Possible high T_c superconductivity in the Ba-La-Cu-O system, *Z. Phys. B: Condens. Matter* **64**, 189 (1986).

- [32] M. K. Wu, J. R. Ashburn, C. J. Torng, P. H. Hor, R. L. Meng, L. Gao, Z. J. Huang, Y. Q. Wang, and C. W. Chu, Superconductivity at 93 K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure, *Phys. Rev. Lett.* **58**, 908 (1987).
- [33] Y. Tokura, H. Takagi, and S. Uchida, A superconducting copper oxide compound with electrons as the charge carriers, *Nature (London)* **337**, 345 (1989).
- [34] A. Schilling, M. Cantoni, J. D. Guo, and H. R. Ott, Superconductivity above 130 K in the Hg-Ba-Ca-Cu-O system, *Nature (London)* **363**, 56 (1993).
- [35] J. Kondo, Resistance minimum in dilute magnetic alloys, *Prog. Theor. Phys.* **32**, 37 (1964).
- [36] K. Andres, J. E. Graebner, and H. R. Ott, $4f$ -virtual-bound-state formation in CeAl_3 at low temperatures, *Phys. Rev. Lett.* **35**, 1779 (1975).
- [37] F. Steglich, J. Aarts, C. D. Bredl, W. Lieke, D. Meschede, W. Franz, and H. Schäfer, Superconductivity in the presence of strong Pauli paramagnetism: CeCu_2Si_2 , *Phys. Rev. Lett.* **43**, 1892 (1979).
- [38] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono, Iron-based layered superconductor $\text{La}[\text{O}_{1-x}\text{F}_x]\text{FeAs}$ ($x = 0.05\text{--}0.12$) with $T_c = 26$ K, *J. Am. Chem. Soc.* **130**, 3296 (2008).
- [39] D. C. Tsui, H. L. Stormer, and A. C. Gossard, Two-dimensional magnetotransport in the extreme quantum limit, *Phys. Rev. Lett.* **48**, 1559 (1982).
- [40] R. B. Laughlin, Anomalous quantum Hall effect: An incompressible quantum fluid with fractionally charged excitations, *Phys. Rev. Lett.* **50**, 1395 (1983).
- [41] H. L. Stormer, D. C. Tsui, and A. C. Gossard, The fractional quantum Hall effect, *Rev. Mod. Phys.* **71**, S298 (1999).
- [42] P. W. Anderson, Resonating valence bonds: A new kind of insulator?, *Mater. Res. Bull.* **8**, 153 (1973).
- [43] F. D. M. Haldane, Nonlinear field theory of large-spin Heisenberg antiferromagnets: Semiclassically quantized solitons of the one-dimensional easy-axis Néel State, *Phys. Rev. Lett.* **50**, 1153 (1983).
- [44] Y. Shimizu, K. Miyagawa, K. Kanoda, M. Maesato, and G. Saito, Spin liquid state in an organic Mott insulator with a triangular lattice, *Phys. Rev. Lett.* **91**, 107001 (2003).
- [45] P. Szirmai, C. Mézière, G. Bastien, P. Wzietek, P. Batail, E. Martino, K. Mantulnikovs, A. Pisoni, K. Riedl, S. Cottrell, C. Baines, L. Forró, and B. Náfrádi, Quantum spin-liquid states in an organic magnetic layer and molecular rotor hybrid, *Proc. Natl. Acad. Sci. USA* **117**, 29555 (2020).
- [46] D. J. Gross and F. Wilczek, Ultraviolet behavior of non-Abelian gauge theories, *Phys. Rev. Lett.* **30**, 1343 (1973).
- [47] H. D. Politzer, Asymptotic freedom: An approach to strong interactions, *Phys. Rep.* **14**, 129 (1974).
- [48] P. Koch, B. Müller, and J. Rafelski, Strangeness in relativistic heavy ion collisions, *Phys. Rep.* **142**, 167 (1986).
- [49] K. G. Wilson, The origins of lattice gauge theory, *Nucl. Phys. B* **140**, 3 (2005).
- [50] R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, Monte Carlo calculations of coupled boson-fermion systems. I, *Phys. Rev. D* **24**, 2278 (1981).
- [51] G. Sugiyama and S. Koonin, Auxiliary field Monte-Carlo for quantum many-body ground states, *Ann. Phys.* **168**, 1 (1986).
- [52] M. Qin, C.-M. Chung, H. Shi, E. Vitali, C. Hubig, U. Schollwöck, S. R. White, and S. Zhang (Simons Collaboration on the Many-Electron Problem), Absence of superconductivity in the pure two-dimensional Hubbard model, *Phys. Rev. X* **10**, 031016 (2020).
- [53] H. Xu, C.-M. Chung, M. Qin, U. Schollwöck, S. R. White, and S. Zhang, Coexistence of superconductivity with partially filled stripes in the Hubbard model, *Science* **384**, eadh7691 (2024).
- [54] M. Capone and G. Kotliar, Competition between d -wave superconductivity and antiferromagnetism in the two-dimensional Hubbard model, *Phys. Rev. B* **74**, 054513 (2006).
- [55] G. Knizia and G. K.-L. Chan, Density matrix embedding: A simple alternative to dynamical mean-field theory, *Phys. Rev. Lett.* **109**, 186404 (2012).
- [56] B.-X. Zheng and G. K.-L. Chan, Ground-state phase diagram of the square lattice Hubbard model from density matrix embedding theory, *Phys. Rev. B* **93**, 035126 (2016).
- [57] T. I. Vanhala and P. Törmä, Dynamical mean-field theory study of stripe order and d -wave superconductivity in the two-dimensional Hubbard model, *Phys. Rev. B* **97**, 075112 (2018).
- [58] S. R. White, Density matrix formulation for quantum renormalization groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
- [59] S. R. White and D. J. Scalapino, Density matrix renormalization group study of the striped phase in the 2D $t\text{--}J$ model, *Phys. Rev. Lett.* **80**, 1272 (1998).
- [60] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, *Ann. Phys.* **326**, 96 (2011).
- [61] E. Stoudenmire and S. R. White, Studying two-dimensional systems with the density matrix renormalization group, *Annu. Rev. Condens. Matter Phys.* **3**, 111 (2012).
- [62] A. Gleis, J.-W. Li, and J. von Delft, Controlled bond expansion for density matrix renormalization group ground state search at single-site costs, *Phys. Rev. Lett.* **130**, 246402 (2023).
- [63] F. Verstraete and J. I. Cirac, Renormalization algorithms for quantum-many body systems in two and higher dimensions, [arXiv:cond-mat/0407066](https://arxiv.org/abs/cond-mat/0407066).
- [64] J. Jordan, R. Orús, G. Vidal, F. Verstraete, and J. I. Cirac, Classical simulation of infinite-size quantum lattice systems in two spatial dimensions, *Phys. Rev. Lett.* **101**, 250602 (2008).
- [65] H. C. Jiang, Z. Y. Weng, and T. Xiang, Accurate determination of tensor network state of quantum lattice models in two dimensions, *Phys. Rev. Lett.* **101**, 090603 (2008).
- [66] T. Barthel, C. Pineda, and J. Eisert, Contraction of fermionic operator circuits and the simulation of strongly correlated fermions, *Phys. Rev. A* **80**, 042333 (2009).
- [67] C. V. Kraus, N. Schuch, F. Verstraete, and J. I. Cirac, Fermionic projected entangled pair states, *Phys. Rev. A* **81**, 052338 (2010).
- [68] P. Corboz, R. Orús, B. Bauer, and G. Vidal, Simulation of strongly correlated fermions in two spatial dimensions with fermionic projected entangled-pair states, *Phys. Rev. B* **81**, 165104 (2010).
- [69] P. Corboz, J. Jordan, and G. Vidal, Simulation of fermionic lattice models in two dimensions with projected entangled-pair states: Next-nearest neighbor Hamiltonians, *Phys. Rev. B* **82**, 245119 (2010).
- [70] C. Zhang, J.-W. Li, D. Nikolaidou, and J. von Delft, Frustration-induced superconductivity in the $t\text{--}t'$ Hubbard model, *Phys. Rev. Lett.* **134**, 116502 (2025).
- [71] S. R. White, Minimally entangled typical quantum states at finite temperature, *Phys. Rev. Lett.* **102**, 190601 (2009).

- [72] E. M. Stoudenmire and S. R. White, Minimally entangled typical thermal state algorithms, *New J. Phys.* **12**, 055026 (2010).
- [73] W. Li, S.-J. Ran, S.-S. Gong, Y. Zhao, B. Xi, F. Ye, and G. Su, Linearized tensor renormalization group algorithm for the calculation of thermodynamic properties of quantum lattice models, *Phys. Rev. Lett.* **106**, 127202 (2011).
- [74] P. Czarnik, L. Cincio, and J. Dziarmaga, Projected entangled pair states at finite temperature: Imaginary time evolution with ancillas, *Phys. Rev. B* **86**, 245101 (2012).
- [75] P. Czarnik and J. Dziarmaga, Variational approach to projected entangled pair states at finite temperature, *Phys. Rev. B* **92**, 035152 (2015).
- [76] B.-B. Chen, L. Chen, Z. Chen, W. Li, and A. Weichselbaum, Exponential thermal tensor network approach for quantum lattice models, *Phys. Rev. X* **8**, 031082 (2018).
- [77] H. Li, B.-B. Chen, Z. Chen, J. von Delft, A. Weichselbaum, and W. Li, Thermal tensor renormalization group simulations of square-lattice quantum spin models, *Phys. Rev. B* **100**, 045110 (2019).
- [78] Q. Li, Y. Gao, Y.-Y. He, Y. Qi, B.-B. Chen, and W. Li, Tangent space approach for thermal tensor network simulations of the 2D Hubbard model, *Phys. Rev. Lett.* **130**, 226502 (2023).
- [79] C. Zhang and J. von Delft, Finite-temperature study of the Hubbard model via enhanced exponential tensor renormalization group, *Phys. Rev. B* **114**, 125114 (2026).
- [80] A. Mazurenko, C. S. Chiu, G. Ji, M. F. Parsons, M. Kanász-Nagy, R. Schmidt, F. Grusdt, E. Demler, D. Greif, and M. Greiner, A cold-atom Fermi-Hubbard antiferromagnet, *Nature (London)* **545**, 462 (2017).
- [81] C. S. Chiu, G. Ji, A. Mazurenko, D. Greif, and M. Greiner, Quantum state engineering of a Hubbard system with ultracold fermions, *Phys. Rev. Lett.* **120**, 243201 (2018).
- [82] J. Koepsell, J. Vijayan, P. Sompet, F. Grusdt, T. A. Hilker, E. Demler, G. Salomon, I. Bloch, and C. Gross, Imaging magnetic polarons in the doped Fermi-Hubbard model, *Nature (London)* **572**, 358 (2019).
- [83] J. Koepsell, S. Hirthe, D. Bourgund, P. Sompet, J. Vijayan, G. Salomon, C. Gross, and I. Bloch, Robust bilayer charge pumping for spin- and density-resolved quantum gas microscopy, *Phys. Rev. Lett.* **125**, 010403 (2020).
- [84] T. Chalopin, P. Bojović, D. Bourgund, S. Wang, T. Franz, I. Bloch, and T. Hilker, Optical superlattice for engineering Hubbard couplings in quantum simulation, *Phys. Rev. Lett.* **134**, 053402 (2025).
- [85] M. Xu, L. H. Kendrick, A. Kale, Y. Gang, C. Feng, S. Zhang, A. W. Young, M. Lebrat, and M. Greiner, A neutral-atom Hubbard quantum simulator in the cryogenic regime, *Nature (London)* **642**, 909 (2025).
- [86] J. Hubbard, Electron correlations in narrow energy bands V. A perturbation expansion about the atomic limit, *Proc. R. Soc. London, Ser. A* **296**, 82 (1967).
- [87] P. W. Anderson, The resonating valence bond state in La_2CuO_4 and superconductivity, *Science* **235**, 1196 (1987).
- [88] J. P. F. LeBlanc, A. E. Antipov, F. Becca, I. W. Bulik, G. K.-L. Chan, C.-M. Chung, Y. Deng, M. Ferrero, T. M. Henderson, C. A. Jiménez-Hoyos, *et al.*, Solutions of the two-dimensional Hubbard model: Benchmarks and results from a wide range of numerical algorithms, *Phys. Rev. X* **5**, 041041 (2015).
- [89] B.-X. Zheng, C.-M. Chung, P. Corboz, G. Ehlers, M.-P. Qin, R. M. Noack, H. Shi, S. R. White, S. Zhang, and G. K.-L. Chan, Stripe order in the underdoped region of the two-dimensional Hubbard model, *Science* **358**, 1155 (2017).
- [90] N. J. Robinson, P. D. Johnson, T. M. Rice, and A. M. Tsvelik, Anomalies in the pseudogap phase of the cuprates: Competing ground states and the role of umklapp scattering, *Rep. Prog. Phys.* **82**, 126501 (2019).
- [91] M. Qin, T. Schäfer, S. Andergassen, P. Corboz, and E. Gull, The Hubbard model: A computational perspective, *Annu. Rev. Condens. Matter Phys.* **13**, 275 (2022).
- [92] H.-C. Jiang and T. P. Devereaux, Superconductivity in the doped Hubbard model and its interplay with next-nearest hopping t' , *Science* **365**, 1424 (2019).
- [93] Y.-F. Jiang, J. Zaanen, T. P. Devereaux, and H.-C. Jiang, Ground state phase diagram of the doped Hubbard model on the four-leg cylinder, *Phys. Rev. Res.* **2**, 033073 (2020).
- [94] C.-M. Chung, M. Qin, S. Zhang, U. Schollwöck, and S. R. White, Plaquette versus ordinary d -wave pairing in the t' -Hubbard model on a width-4 cylinder, *Phys. Rev. B* **102**, 041106 (2020).
- [95] Y.-F. Jiang, T. P. Devereaux, and H.-C. Jiang, Ground-state phase diagram and superconductivity of the doped Hubbard model on six-leg square cylinders, *Phys. Rev. B* **109**, 085121 (2024).
- [96] B. Ponsioen, S. S. Chung, and P. Corboz, Period 4 stripe in the extended two-dimensional Hubbard model, *Phys. Rev. B* **100**, 195141 (2019).
- [97] A. Wietek, Y.-Y. He, S. R. White, A. Georges, and E. M. Stoudenmire, Stripes, antiferromagnetism, and the pseudogap in the doped Hubbard model at finite temperature, *Phys. Rev. X* **11**, 031007 (2021).
- [98] A. Bohrdt, C. S. Chiu, G. Ji, M. Xu, D. Greif, M. Greiner, E. Demler, F. Grusdt, and M. Knap, Classifying snapshots of the doped Hubbard model with machine learning, *Nat. Phys.* **15**, 921 (2019).
- [99] C. S. Chiu, G. Ji, A. Bohrdt, M. Xu, M. Knap, E. Demler, F. Grusdt, M. Greiner, and D. Greif, String patterns in the doped Hubbard model, *Science* **365**, 251 (2019).
- [100] G. Salomon, J. Koepsell, J. Vijayan, T. A. Hilker, J. Nespolo, L. Pollet, I. Bloch, and C. Gross, Direct observation of incommensurate magnetism in Hubbard chains, *Nature (London)* **565**, 56 (2019).
- [101] B.-B. Chen, C. Chen, Z. Chen, J. Cui, Y. Zhai, A. Weichselbaum, J. von Delft, Z. Y. Meng, and W. Li, Quantum many-body simulations of the two-dimensional Fermi-Hubbard model in ultracold optical lattices, *Phys. Rev. B* **103**, L041107 (2021).
- [102] J. Koepsell, D. Bourgund, P. Sompet, S. Hirthe, A. Bohrdt, Y. Wang, F. Grusdt, E. Demler, G. Salomon, C. Gross, and I. Bloch, Microscopic evolution of doped Mott insulators from polaronic metal to Fermi liquid, *Science* **374**, 82 (2021).
- [103] P. Sompet, S. Hirthe, D. Bourgund, T. Chalopin, J. Bibo, J. Koepsell, P. Bojović, R. Verresen, F. Pollmann, G. Salomon, C. Gross, T. A. Hilker, and I. Bloch, Realizing the symmetry-protected Haldane phase in Fermi-Hubbard ladders, *Nature (London)* **606**, 484 (2022).
- [104] S. Hirthe, T. Chalopin, D. Bourgund, P. Bojović, A. Bohrdt, E. Demler, F. Grusdt, I. Bloch, and T. A. Hilker, Magnetically

- mediated hole pairing in fermionic ladders of ultracold atoms, *Nature (London)* **613**, 463 (2023).
- [105] M. Xu, L. H. Kendrick, A. Kale, Y. Gang, G. Ji, R. T. Scalettar, M. Lebrat, and M. Greiner, Frustration- and doping-induced magnetism in a Fermi-Hubbard simulator, *Nature (London)* **620**, 971 (2023).
- [106] T. Chalopin, P. Bojović, S. Wang, T. Franz, A. Sinha, Z. Wang, D. Bourgund, J. Obermeyer, F. Grusdt, A. Bohrdt, L. Pollet, A. Wietek, A. Georges, T. Hilker, and I. Bloch, Observation of emergent scaling of spin-charge correlations at the onset of the pseudogap, *Proc. Natl. Acad. Sci. USA* **123**, e2525539123 (2026).
- [107] G. Pasqualetti, O. Bettermann, N. Darkwah Oppong, E. Ibarra-García-Padilla, S. Dasgupta, R. T. Scalettar, K. R. A. Hazzard, I. Bloch, and S. Fölling, Equation of state and thermometry of the 2D $SU(N)$ Fermi-Hubbard model, *Phys. Rev. Lett.* **132**, 083401 (2024).
- [108] D. Bourgund, T. Chalopin, P. Bojović, H. Schlömer, S. Wang, T. Franz, S. Hirthe, A. Bohrdt, F. Grusdt, I. Bloch, and T. A. Hilker, Formation of individual stripes in a mixed-dimensional cold-atom Fermi-Hubbard system, *Nature (London)* **637**, 57 (2025).
- [109] X. Dong, E. Gull, and A. J. Millis, Quantifying the role of antiferromagnetic fluctuations in the superconductivity of the doped Hubbard model, *Nat. Phys.* **18**, 1293 (2022).
- [110] H. Xu, H. Shi, E. Vitali, M. Qin, and S. Zhang, Stripes and spin-density waves in the doped two-dimensional Hubbard model: Ground state phase diagram, *Phys. Rev. Res.* **4**, 013239 (2022).
- [111] B. Xiao, Y.-Y. He, A. Georges, and S. Zhang, Temperature dependence of spin and charge orders in the doped two-dimensional Hubbard model, *Phys. Rev. X* **13**, 011007 (2023).
- [112] S. d. A. Sousa-Júnior, N. C. Costa, and R. R. dos Santos, Half-filled extended Hubbard model on a square lattice: Phase boundaries from determinant quantum Monte Carlo simulations, *Phys. Rev. B* **109**, 165102 (2024).
- [113] S. Jiang, D. J. Scalapino, and S. R. White, Density matrix renormalization group based downfolding of the three-band Hubbard model: Importance of density-assisted hopping, *Phys. Rev. B* **108**, L161111 (2023).
- [114] A. Bohrdt, Y. Wang, J. Koepsell, M. Kánasz-Nagy, E. Demler, and F. Grusdt, Dominant fifth-order correlations in doped quantum antiferromagnets, *Phys. Rev. Lett.* **126**, 026401 (2021).
- [115] C. Miles, A. Bohrdt, R. Wu, C. Chiu, M. Xu, G. Ji, M. Greiner, K. Q. Weinberger, E. Demler, and E.-A. Kim, Correlator convolutional neural networks as an interpretable architecture for image-like quantum matter data, *Nat. Commun.* **12**, 3905 (2021).
- [116] L. W. Cheuk, M. A. Nichols, K. R. Lawrence, M. Okan, H. Zhang, E. Khatami, N. Trivedi, T. Paiva, M. Rigol, and M. W. Zwierlein, Observation of spatial charge and spin correlations in the 2D Fermi-Hubbard model, *Science* **353**, 1260 (2016).
- [117] T. A. Hilker, G. Salomon, F. Grusdt, A. Omran, M. Boll, E. Demler, I. Bloch, and C. Gross, Revealing hidden antiferromagnetic correlations in doped Hubbard chains via string correlators, *Science* **357**, 484 (2017).
- [118] F. Grusdt, M. Kánasz-Nagy, A. Bohrdt, C. S. Chiu, G. Ji, M. Greiner, D. Greif, and E. Demler, Parton theory of magnetic polarons: Mesonic resonances and signatures in dynamics, *Phys. Rev. X* **8**, 011046 (2018).
- [119] F. Grusdt and L. Pollet, $\mathbb{Z}_2$ parton phases in the mixed-dimensional $t-J_z$ model, *Phys. Rev. Lett.* **125**, 256401 (2020).
- [120] Z. Y. Weng, D. N. Sheng, Y.-C. Chen, and C. S. Ting, Phase string effect in the $t-J$ model: General theory, *Phys. Rev. B* **55**, 3894 (1997).
- [121] T.-L. Ho, Imaging the Holon string of the Hubbard model, *Proc. Natl. Acad. Sci. USA* **117**, 26141 (2020).
- [122] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/jsgs-9tx8> for an introduction to the transformer architecture for physicists, additional technical details, benchmarks and analysis, and a performance preview of a 25-category omnimeter.
- [123] M. Hirayama, Y. Yamaji, T. Misawa, and M. Imada, *Ab initio* effective Hamiltonians for cuprate superconductors, *Phys. Rev. B* **98**, 134501 (2018).
- [124] M. Hirayama, T. Misawa, T. Ohgoe, Y. Yamaji, and M. Imada, Effective Hamiltonian for cuprate superconductors derived from multiscale *ab initio* scheme with level renormalization, *Phys. Rev. B* **99**, 245155 (2019).
- [125] X. Lin, B.-B. Chen, W. Li, Z. Y. Meng, and T. Shi, Exciton proliferation and fate of the topological Mott insulator in a twisted bilayer graphene lattice model, *Phys. Rev. Lett.* **128**, 157201 (2022).
- [126] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2016), pp. 770–778.
- [127] J. L. Ba, J. R. Kiros, and G. E. Hinton, Layer normalization, [arXiv:1607.06450](https://arxiv.org/abs/1607.06450).
- [128] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Y. W. Teh and M. Titterton, Proceedings of Machine Learning Research, Vol. 9 (JMLR, Chia Laguna Resort, Sardinia, Italy, 2010), pp. 249–256.
- [129] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations (ICLR)*, 2015, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [130] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**, 1929 (2014).
- [131] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2016), pp. 2818–2826.
- [132] A. Bibal and B. Frénay, Interpretability of machine learning models and representations: An introduction, in *Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016)*, edited by M. Verleysen (CIACO, Bruges, Belgium, 2016), pp. 77–82.
- [133] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* **51**, 1 (2018).
- [134] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism, in

- Advances in Neural Information Processing Systems 29*, edited by D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (Curran Associates, Red Hook, NY, 2016), pp. 3504–3512.
- [135] A. Martins and R. Astudillo, From Softmax to Sparsemax: A sparse model of attention and multi-label classification, in *Proceedings of the 33rd International Conference on Machine Learning*, edited by M. F. Balcan and K. Q. Weinberger (PMLR, New York, NY, 2016), Vol. 48, pp. 1614–1623.
- [136] J. Lee, J.-H. Shin, and J.-S. Kim, Interactive visualization and manipulation of attention-based neural machine translation, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by L. Specia, M. Post, and M. Paul (Association for Computational Linguistics, Copenhagen, Denmark, 2017), pp. 121–126.
- [137] Q. Xie, X. Ma, Z. Dai, and E. Hovy, An interpretable knowledge transfer model for knowledge base completion, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers), edited by R. Barzilay and M.-Y. Kan (Association for Computational Linguistics, Vancouver, Canada, 2017), pp. 950–962.
- [138] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, Universal transformers, in *International Conference on Learning Representations (ICLR, 2019)*, [arXiv:1807.03819](https://arxiv.org/abs/1807.03819).
- [139] G. Brunner, Y. Liu, D. Pascual, O. Richter, M. Ciaramita, and R. Wattenhofer, On identifiability in transformers, in *International Conference on Learning Representations (ICLR, 2020)*, [arXiv:1908.04211](https://arxiv.org/abs/1908.04211).
- [140] H. Chen and Y. Ji, Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation, was presented at the NeurIPS 2019 Workshop on Robust AI in Financial Services (Vancouver, Canada, 2019), <https://sites.google.com/view/robust-ai-in-fs-2019>.
- [141] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, What does Bert look at? An analysis of Bert’s attention, in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, edited by T. Linzen, G. Chrupała, Y. Belinkov, and D. Hupkes (Association for Computational Linguistics, Florence, Italy, 2019), pp. 276–286.
- [142] A. Coenen, E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, and M. Wattenberg, Visualizing and measuring the geometry of BERT, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Red Hook, NY, 2019), Vol. 32.
- [143] S. Jain and B. C. Wallace, Attention is not explanation, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), edited by J. Burstein, C. Doran, and T. Solorio (Association for Computational Linguistics, Minneapolis, MN, 2019), pp. 3543–3556.
- [144] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, Attention interpretability across NLP tasks, [arXiv:1909.11218](https://arxiv.org/abs/1909.11218).
- [145] J. Vig, Visualizing attention in transformer-based language representation models, [arXiv:1904.02679](https://arxiv.org/abs/1904.02679).
- [146] S. Wiegreffe and Y. Pinter, Attention is not not explanation, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by K. Inui, J. Jiang, V. Ng, and X. Wan (Association for Computational Linguistics, Hong Kong, China, 2019), pp. 11–20.
- [147] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, Self-attention generative adversarial networks, in *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, 2019), Vol. 97, pp. 7354–7363.
- [148] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, Learning to deceive with attention-based explanations, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Association for Computational Linguistics, 2020), pp. 4782–4793.
- [149] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, and F. Wang, Why attentions may not be interpretable? in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD ’21)* (ACM Press, New York, 2021), pp. 25–34.
- [150] A. Galassi, M. Lippi, and P. Torroni, Attention in natural language processing, *IEEE Trans. Neural Networks Learn. Syst.* **32**, 4291 (2021).
- [151] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin, Is attention explanation? An introduction to the debate, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers), edited by S. Muresan, P. Nakov, and A. Villavicencio (Association for Computational Linguistics, Dublin, Ireland, 2022), pp. 3889–3900.
- [152] S. Abnar and W. Zuidema, Quantifying attention flow in transformers, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Association for Computational Linguistics, 2020), pp. 4190–4197.
- [153] T. Yuan, X. Li, H. Xiong, H. Cao, and D. Dou, Explaining information flow inside vision transformers using Markov chain, in *NeurIPS 2021 Workshop on eXplainable AI Approaches for Debugging and Diagnosis* (OpenReview, 2021), paper TTcf6QSDaQ.
- [154] L. Xu, X. Yan, W. Ding, and Z. Liu, Attribution rollout: A new way to interpret visual transformer, *J. Ambient Intell. Hum. Comput.* **14**, 163 (2023).
- [155] N. Metzger, C. Hahn, J. Siber, F. Schmitt, and B. Finkbeiner, Attention flows for general transformers, in *International Conference on Learning Representations (ICLR, 2023)*.
- [156] B. Azarkhalili and M. W. Libbrecht, Generalized attention flow: Feature attribution for transformer models via maximum flow, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers), edited by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar (Association for Computational Linguistics, Vienna, Austria, 2025), pp. 19954–19974.
- [157] T. Hartke, B. Oreg, N. Jia, and M. Zwierlein, Doublon-hole correlations and fluctuation thermometry in a Fermi-Hubbard gas, *Phys. Rev. Lett.* **125**, 113601 (2020).
- [158] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, Recurrent neural network wave functions, *Phys. Rev. Res.* **2**, 023358 (2020).

- [159] D. Luo, Z. Chen, K. Hu, Z. Zhao, V. M. Hur, and B. K. Clark, Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models, [Phys. Rev. Res. **5**, 013216 \(2023\)](#).
- [160] Y.-H. Zhang, Transformer quantum state: A multipurpose model for quantum many-body problems, [Phys. Rev. B **107**, 075147 \(2023\)](#).
- [161] E. Ibarra-García-Padilla, H. Lange, R. G. Melko, R. T. Scalettar, J. Carrasquilla, A. Bohrdt, and E. Khatami, Autoregressive neural quantum states of Fermi Hubbard models, [Phys. Rev. Res. **7**, 013122 \(2025\)](#).
- [162] H. Lange, G. Bornet, G. Emperauger, C. Chen, T. Lahaye, S. Kienle, A. Browaeys, and A. Bohrdt, Transformer neural networks and quantum simulators: A hybrid approach for simulating strongly correlated systems, [Quantum **9**, 1675 \(2025\)](#).
- [163] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, [Phys. Rev. Lett. **122**, 226401 \(2019\)](#).
- [164] A.-J. Liu and B. K. Clark, Neural network backflow for *ab initio* quantum chemistry, [Phys. Rev. B **110**, 115137 \(2024\)](#).
- [165] L. Zhang and D. Luo, Neural transformer backflow for solving momentum-resolved ground states of strongly correlated materials, [arXiv:2509.09275](#).