

Unveiling the physical meaning of transformer attention in neural network quantum states: A conditional mutual information perspective

Tianyu Ruan(阮天雨), Bowen Kan(阚博文), Yixuan Sun(孙艺轩), Honghui Shang(商红慧), Shihua Zhang(张世华), and Jinlong Yang(杨金龙)

Citation: Chin. Phys. B, 2026, 35 (**1**): 010301. DOI: 10.1088/1674-1056/ae1118
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Global receptive field transformer decoder method on quantum surface code data and syndrome error correction

Ao-Qing Li(李熬庆), Ce-Wen Tian(田策文), Xiao-Xuan Xu(徐晓璇), Hong-Yang Ma(马鸿洋), and Jun-Qing Liang(梁俊卿)
Chin. Phys. B, 2025, 34 (**3**): 030306. DOI: 10.1088/1674-1056/adab63

Quantum decoder design for subsystem surface code based on multi-head graph attention and edge weighting

Nai-Hua Ji(纪乃华), Hui-Qian Sun(孙汇倩), Bo Xiao(肖博), Ping-Li Song(宋平俐), and Hong-Yang Ma(马鸿洋)
Chin. Phys. B, 2025, 34 (**2**): 020309. DOI: 10.1088/1674-1056/ada1c7

Decoding topological XYZ^2 codes with reinforcement learning based on attention mechanisms

Qing-Hui Chen(陈庆辉), Yu-Xin Ji(姬宇欣), Ke-Han Wang(王柯涵), Hong-Yang Ma(马鸿洋), and Nai-Hua Ji(纪乃华)
Chin. Phys. B, 2024, 33 (**6**): 060314. DOI: 10.1088/1674-1056/ad342b

Variational quantum eigensolvers by variance minimization

Dan-Bo Zhang(张旦波), Bin-Lin Chen(陈彬琳), Zhan-Hao Yuan(原展豪), and Tao Yin(殷涛)
Chin. Phys. B, 2022, 31 (**12**): 120301. DOI: 10.1088/1674-1056/ac8a8d

Quantum computation and simulation with superconducting qubits

Kaiyong He(何楷泳), Xiao Geng(耿霄), Rutian Huang(黄汝田), Jianshe Liu(刘建设), and Wei Chen(陈炜)
Chin. Phys. B, 2021, 30 (**8**): 080304. DOI: 10.1088/1674-1056/ac16cf

---

SPECIAL TOPIC — AI + Physical Science

Unveiling the physical meaning of transformer attention in neural network quantum states: A conditional mutual information perspective

Tianyu Ruan(阮天雨)[1,2,†], Bowen Kan(阚博文)[3,4,†], Yixuan Sun(孙艺轩)[5,†], Honghui Shang(商红慧)[5,‡], Shihua Zhang(张世华)[1,2,§], and Jinlong Yang(杨金龙)[5,¶]

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190, *China*
[2] *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing* 100049, *China*
[3] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*
[4] *University of Chinese Academy of Sciences, Beijing* 101408, *China*
[5] *State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Anhui* 230026, *China*

(Received 30 July 2025; revised manuscript received 5 October 2025; accepted manuscript online 9 October 2025)

Transformer-based neural-network quantum states (NNQS) have shown great promise in representing quantum many-body ground states, offering high flexibility and accuracy. However, the interpretability of such models remains limited, especially in terms of connecting network components to physically meaningful quantities. We propose that the attention mechanism — a central module in transformer architectures — explicitly models the conditional information flow between orbitals. Intuitively, as the transformer learns to predict orbital configurations by optimizing an energy functional, it approximates the conditional probability distribution $p(x_n|x_1,\ldots,x_{n-1})$, implicitly encoding conditional mutual information (CMI) among orbitals. This suggests a natural correspondence between attention maps and CMI structures in quantum systems. To probe this idea, we compare weighted attention scores from trained transformer wavefunction ansatze with CMI matrices across several representative small molecules. In most cases, we observe a positive rank-level correlation (Kendall's tau) between attention and CMI, suggesting that the learned attention can reflect physically relevant orbital dependencies. This study provides a quantitative link between transformer attention and conditional mutual information in the NNQS setting. Our results provide a step toward explainable deep learning in quantum chemistry, pointing to opportunities in interpreting attention as a proxy for physical correlations.

Keywords: attention mechanism, quantum chemistry, many-body Schrödinger equation, entanglement entropy

PACS: 03.65.Ud, 03.67.–a, 07.05.Mh

DOI: 10.1088/1674-1056/ae1118 **CSTR:** 32038.14.CPB.ae1118

1. Introduction

Neural-network quantum states (NNQS) have become a powerful paradigm for approximating solutions of the electronic Schrödinger equation. Beginning with restricted Boltzmann machines for spin systems,[1] the field has rapidly advanced to include deep neural trial states such as FermiNet,[2] which introduced antisymmetry-preserving deep wavefunctions for electrons, and PauliNet,[3] which incorporated explicit correlation factors to improve accuracy. Convolutional architectures have also been developed, for example in the two-dimensional Hubbard model,[4] where convolutional filters were used to efficiently capture local correlations. Further developments include fermionic RBM ansatze optimized via variational Monte Carlo,[5] message-passing neural quantum states that leverage graph architectures to capture nonlocal correlations,[6] and autoregressive neural wavefunctions for electrons.[7] Collectively, these works demonstrated that expressive neural parameterizations can achieve near-chemical accuracy while scaling to Hilbert spaces far beyond the reach of traditional variational methods such as Hartree–Fock, full configuration interaction (FCI), or coupled-cluster theory.[8–10] More broadly, attention-based neural architectures have also been deployed in physics and materials domains, including crystal property prediction,[11] protein essentiality analysis,[12] and reinforcement-learning decoders for topological codes,[13] underscoring the versatility of attention across scientific applications.

Building on this foundation, transformer architectures[14] have recently emerged as effective NNQS backbones for correlated-electron systems. Several groups have reported

[†]These authors contributed equally to this work.
[‡]Corresponding author. E-mail: shh@ustc.edu.cn
[§]Corresponding author. E-mail: zsh@amss.ac.cn
[¶]Corresponding author. E-mail: jlyang@ustc.edu.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

competitive ground-state energies together with qualitatively interpretable attention patterns. For instance, PsiFormer[15] introduced a self-attention ansatz inspired by the vision transformer architecture to represent electronic wavefunctions in *ab initio* quantum chemistry. Wu *et al.*[16] proposed NNQS-transformer with autoregressive masking, Geier *et al.*[17] applied a transformer-based ansatz to moiré-electron systems, and Gu *et al.*[18] explored the doped Hubbard model, where attention heads appeared to capture physically relevant motifs. While these studies indicate that self-attention can encode aspects of electronic structure, their analyses remain largely qualitative, for example, correlating attention heads with orbital distances or geometric motifs without establishing rigorous connections to measurable quantities. Thus, despite the practical success of transformer NNQS, the role of attention in encoding quantum correlations remains phenomenological. This motivates our work, which establishes a quantitative link between attention maps and conditional mutual information computed from exact wavefunctions.

To address the interpretability challenge, we focus on the second-quantized representation of electronic structure, where orbital occupations provide a natural language for attention-based generative modeling. Within this setting, we propose a physically motivated hypothesis linking transformer attention maps to quantum information theory. Specifically, we conjecture that the attention scores learned by transformer-based NNQS correspond closely to the CMI between molecular orbitals — a measure that quantifies correlation under the second-quantized framework. This intuition is grounded in the model's generative structure: the transformer autoregressively models the joint distribution of orbital occupations by approximating conditional probabilities of the form $p(x_n|x_1,\ldots,x_{n-1})$. Through this learning objective, the model implicitly encodes the statistical dependence between the current orbital and its predecessors. The attention mechanism, as a central component in this process, offers an explicit computational pathway for capturing these conditional dependencies. Thus, we hypothesize that the attention maps generated during sampling reflect the underlying CMI structure of the system. If validated, this correspondence would provide a bridge between deep learning architectures and physically interpretable quantities, offering new insights into how quantum correlations are modeled within neural networks.

To evaluate this hypothesis, we adopt the QiankunNet framework,[16] a class of variational quantum states parameterized by transformer architectures. QiankunNet is, to our knowledge, one of the earliest and few NNQS models formulated directly in the second-quantized orbital-occupation representation with an autoregressive transformer backbone, making it a natural testbed for our study. This approach has been demonstrated to efficiently model complex many-body wavefunctions, offering a scalable and expressive alternative to traditional electronic-structure methods. Within this framework, we compute both the weighted attention scores from the transformer and the conditional mutual information matrices derived from FCI wavefunctions for several representative small molecules. Our empirical analysis reveals that for several small molecules, the attention-induced relevance rankings show positive alignment with those obtained via CMI. These results provide initial evidence that attention mechanisms can, under suitable conditions, reflect aspects of quantum correlations in a physically meaningful manner.

This work presents a quantitative exploration linking transformer attention maps with CMI, which is closely related to conditional entanglement entropy. Rather than claiming a universal law, our results suggest a promising perspective on how deep learning models — specifically those based on attention — can encode interpretable structures in quantum many-body systems. By grounding abstract neural mechanisms in measurable quantum information quantities, our approach contributes toward the broader goal of explainable AI in scientific modeling.

The remainder of the paper is organized as follows. Section 2 provides the necessary background on QiankunNet, von Neumann and mutual information entropy, and transformer attention mechanisms. Section 3 introduces our methodology for computing conditional mutual information and weighted attention scores. Section 4 presents a simplified theoretical analysis using a surrogate linear–autoregressive model to elucidate the relationship between attention and mutual information. Section 5 presents numerical experiments on several molecular systems, highlighting both the observed alignment and cases where it weakens. We conclude in Section 6 with a discussion of implications, limitations, and directions for further investigation.

2. Background

In this section, we present the foundational concepts that underpin our study: the transformer-based neural-network QiankunNet, the notion of quantum entanglement entropy and mutual information, and the attention mechanism used in transformer architectures.

2.1. QiankunNet

The QiankunNet is introduced to overcome the scalability limitations inherent in traditional *ab initio* electronic-structure methods, particularly within the formalism of second quantization. The many-body wavefunction is represented in polar form as $\Psi(x) = |\Psi|(x)\,\mathrm{e}^{\mathrm{i}\phi(x)}$, where both the amplitude and the phase are modeled using neural networks. The phase $\phi(x)$ is parameterized by a multi-layer perceptron (MLP)

$$\phi(x) = \sigma \circ L_d \circ \cdots \circ \sigma \circ L_1(x), \tag{1}$$

where d denotes the depth of the network, σ is a nonlinear activation function, and L_i represents the linear transformation at the i-th layer. The input vector $x = [x_1, \ldots, x_n]$ encodes the occupation states of n orbits. The amplitude $|\Psi|(x)$ is modeled in an autoregressive manner using a transformer, such that the squared modulus of the wavefunction satisfies

$$|\Psi|^2(x_1, \ldots, x_n) \approx |\Psi_\theta|^2(x) = \prod_{l=1}^{n} p_\theta(x_l | x_{<l}), \tag{2}$$

where p_θ is the conditional distribution learned by a transformer with parameters θ. The transformer architecture employed in QiankunNet consists of an input embedding layer, a stack of self-attention blocks, and a final linear output projection. Each transformer block includes the following components: (1) a multi-head self-attention layer with residual (skip) connection; (2) layer normalization; (3) a position-wise feedforward network (typically implemented as an MLP) with residual connection; (4) a second layer normalization step.

Autoregressive sampling proceeds as follows: given a partially generated sequence of orbital states $[x_1, \ldots, x_k]$, a special start token S is prepended to form the input sequence $[S, x_1, \ldots, x_k]$. The transformer processes this sequence and outputs a set of vectors $[z_0, z_1, \ldots, z_k]$, where z_i encodes a probability distribution over the vocabulary for predicting the token at position $i+1$. In particular, z_k predicts the orbital state x_{k+1}. Thus, the i-th token position consistently corresponds to the i-th orbital throughout the network, while the start token S serves only to bootstrap prediction of the first orbital.

Solving the many-body Schrödinger equation reduces to an eigenvalue problem in which the ground-state energy is obtained by minimizing the energy expectation value:

$$\langle \mathcal{H} \rangle_{\Psi_\theta} = \frac{\langle \Psi_\theta | \mathcal{H} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle}, \tag{3}$$

where $\mathcal{H}$ is the Hamiltonian and Ψ_θ is the trial wavefunction parameterized by the neural network. This objective is optimized using large-scale variational Monte Carlo (VMC), which samples electron configurations according to the probability density $|\Psi_\theta(x)|^2$. The QiankunNet framework offers both expressive representational power and scalability to a large Hilbert space. However, its architectural complexity featuring multiple nonlinear modules such as self-attention and feedforward networks makes it challenging to interpret how individual components contribute to the modeling of physically meaningful structures.

2.2. Entropy

Quantum information theory has long been applied to analyze electron correlations in quantum chemical systems based on traditional wavefunction calculations. In particular, orbital entanglement measures such as single-orbital entropy and orbital-pair mutual information have been used to characterize static and dynamic correlation, guide active-space selection, and reveal chemical bonding patterns.[19–22] Building on this tradition, we briefly review the basic definitions of entanglement entropy and mutual information, which will form the foundation for our conditional mutual information analysis in transformer-based NNQS.

In quantum information theory, the von Neumann entropy serves as a fundamental measure of uncertainty or mixedness in a quantum system. It plays a role analogous to that of Shannon entropy in classical information theory.[23,24] For a system described by a density matrix ρ, the von Neumann entropy is defined as

$$S(\rho) = -\mathrm{Tr}(\rho \ln \rho), \tag{4}$$

where Tr denotes the trace over the Hilbert space of the system. When a system is partitioned into two subsystems, A and B, the interaction between A and B can be described by the entanglement entropy, which is given by the von Neumann entropy of the reduced density matrix of each part

$$S_{A|B} = -\mathrm{Tr}(\rho_A \ln \rho_A) = -\mathrm{Tr}(\rho_B \ln \rho_B). \tag{5}$$

In the case of a pure state, the reduced density matrix is defined as

$$\rho_A = \mathrm{Tr}_B |\Psi\rangle\langle\Psi|, \quad \rho_B = \mathrm{Tr}_A |\Psi\rangle\langle\Psi|. \tag{6}$$

Here, ρ_A (ρ_B) is obtained by tracing out the states associated with subsystem B (A).

In quantum chemistry, understanding orbital interactions is essential for analyzing electronic structure, chemical bonding, molecular physical properties and reactivity. While conventional frameworks such as molecular orbital theory and ligand field theory often provide qualitative insights, quantum information theory offers a quantitative alternative through the notion of entanglement entropy. Specifically, consider a molecular system described by a Hilbert space spanned by k orbitals. One can partition the system into a single orbital and its complementary subsystem. The von Neumann entropy of the reduced density matrix associated with the single orbital quantifies how strongly it is entangled with the rest of the system. This quantity is referred to as the single-orbital entropy, and it serves as a localized measure of orbital correlation.

Let $\{|i\rangle\}$ denote the local basis of the selected orbital, and $\{|\boldsymbol{n}\rangle\}$the basis for the remaining orbitals. The reduced density matrix for the single orbital is then[25]

$$\rho_{i,i'} = \sum_{\boldsymbol{n}} \langle \boldsymbol{n}| \langle i|\Psi\rangle \langle \Psi|i'\rangle |\boldsymbol{n}\rangle, \tag{7}$$

and according to Eq. (5) above, its von Neumann entropy is given by

$$s_i = -\sum_{\alpha=1}^{4} \omega_{\alpha,i} \ln \omega_{\alpha,i}, \tag{8}$$

where $\omega_{\alpha,i}$ are the eigenvalues of the reduced density matrix. Each orbital admits four possible occupation states: empty $|0\rangle$, spin-up $|\uparrow\rangle$, spin-down $|\downarrow\rangle$ and doubly occupied $|\uparrow\downarrow\rangle$, resulting in a 4-dimensional reduced density matrix.

Similarly, one can extend the analysis to pairs of orbitals. By partitioning the system into a two-orbital subsystem and the remaining orbitals, the two-orbital entropy quantifies the degree of entanglement between the selected pair and the rest of the system. The von Neumann entropy of the reduced two-orbital density matrix is given by

$$s_{ij} = -\sum_{\alpha=1}^{16} \omega_{\alpha,ij} \ln \omega_{\alpha,ij}, \tag{9}$$

where $\omega_{\alpha,ij}$ are the eigenvalues of the 16-dimensional reduced density matrix for the orbital pair (i,j). The dimensionality arises from the tensor product of the four-level Hilbert spaces associated with each orbital. Following the formulation by Rissler *et al.*,[20] the mutual information between orbitals i and j is defined as

$$I_{ij} = \frac{1}{2}(s_i + s_j - s_{ij})(1-\delta_{ij}), \tag{10}$$

where δ_{ij} is the Kronecker delta function ensuring that $I_{ii} = 0$. This expression captures the total (quantum) correlation between the two orbitals, and serves as the quantum analogue of mutual information in classical information theory, up to a normalization factor.

2.3. Attention mechanism and attention map

In transformer and other attention-based architectures, the attention mechanism serves to model dependencies between different positions or tokens within a sequence. Given the query, key, and value projections Q,K,V and input token features $z^{\text{in}} = (z_1,\ldots,z_k)^{\text{T}}$, attention computes a weighted aggregation of information as

$$z^{\text{out}} = Az^{\text{in}}V, \quad A_{ij} = \frac{\exp\left(\frac{(Qz_i)\cdot(Kz_j)}{\tau}\right)}{\sum_k \exp\left(\frac{(Qz_i)\cdot(Kz_k)}{\tau}\right)}, \tag{11}$$

where A_{ij} denotes the normalized attention weight from token i to token j, and τ is a temperature parameter controlling the sharpness of the distribution. The attention matrix A captures the relative importance of each token j in constructing the output representation for token i, thereby encoding how contextual information flows across the sequence.

In practice, transformers employ multi-head attention to allow the model to jointly attend to information from different representation subspaces. Each attention head is associated with its own set of projection matrices (Q^l,K^l,V^l), where $l = 1,\ldots,H$ and produces an attention matrix A^l. The outputs of all heads are linearly transformed and aggregated as

$$z^{\text{out}} = \sum_l A^l z^{\text{in}} V^l O^l, \quad A^l_{ij} = \frac{\exp\left(\frac{(Q^l z_i)\cdot(K^l z_j)}{\tau}\right)}{\sum_k \exp\left(\frac{(Q^l z_i)\cdot(K^l z_k)}{\tau}\right)}, \tag{12}$$

where O^l denotes the output projection matrix for the l-th head, and A^l_{ij} encodes the attention weight from token i to j within that head.

To ensure causal structure in autoregressive generation, a binary mask matrix $M \in \{0,1\}^{k\times k}$ is applied to the attention matrix. This mask is lower triangular, with entries $M_{ij} = 1$ if $j \le i$, and zero otherwise. The final masked attention becomes

$$A = M \odot A, \tag{13}$$

where $\odot$ denotes the Hadamard (element-wise) product. This causal masking prevents tokens from attending to future positions, preserving the sequential nature of the generation process and enforcing directional information flow.

The attention map refers to a visualization of the attention matrix, offering an intuitive representation of how much attention each token allocates to others in the sequence. In the context of autoregressive models, attention maps reveal how the prediction at a given position depends on earlier tokens, thereby reflecting the model's internal notion of contextual relevance. Such visualizations are widely used to interpret transformer behavior across a range of domains, including natural language processing and quantum systems.[15,26–28] In particular, they provide a qualitative tool for assessing whether the model captures meaningful structural dependencies such as locality, symmetry, or long-range interactions embedded in the data.

3. Method

3.1. Conditional mutual information

In quantum chemistry, molecular orbitals are commonly constructed using the Hartree–Fock (HF) method, which relies on a single-electron approximation. While computationally efficient, this approximation neglects dynamic electron correlation effects and can result in significant deviations from the exact ground-state behavior. To overcome this limitation, the configuration interaction (CI) method provides a more accurate representation by expressing the many-electron wavefunction as a linear combination of Slater determinants. By performing a full variational optimization over this configurational space, the full configuration interaction (FCI) approach yields the exact ground-state wavefunction within a given basis set. This highly accurate wavefunction enables the calculation of reduced density matrices and entanglement-based quantities, such as mutual information between orbitals.

Given the ground-state wavefunction from a full configuration interaction (FCI) calculation, the CI coefficients can be used to construct reduced density matrices for one- to four-particle subsystems. These density matrices form the basis for computing von Neumann entropies and mutual information between orbitals. When the total spin projection s_z is a conserved quantum number, the one-orbital reduced density matrix becomes diagonal, while the two-orbital density matrix acquires a block-diagonal structure. Only those matrix elements that preserve both particle number and spin projection contribute nonzero values. Within the second-quantized framework, the one-orbital reduced density matrix for orbital i can be written as[21]

$$\rho_i = \begin{pmatrix} \langle 1-n_{i,\uparrow}-n_{i,\downarrow}+n_{i,\uparrow}n_{i,\downarrow}\rangle & 0 & 0 & 0 \\ 0 & \langle n_{i,\downarrow}-n_{i,\uparrow}n_{i,\downarrow}\rangle & 0 & 0 \\ 0 & 0 & \langle n_{i,\uparrow}-n_{i,\uparrow}n_{i,\downarrow}\rangle & 0 \\ 0 & 0 & 0 & \langle n_{i,\uparrow}n_{i,\downarrow}\rangle \end{pmatrix}. \tag{14}$$

Here, $n_{i,\sigma} = c^\dagger_{i,\sigma}c_{i,\sigma}$ is the spin-dependent number operator of spin σ, $c^\dagger_{i,\sigma}$ and $c_{i,\sigma}$ are the creation and annihilation operators of the i-th orbital. The form of the two-orbital density matrix is given in Table 1, in which the elements are as follows:

$$\begin{aligned} &O_1 = 1-n_\uparrow-n_\downarrow+n_\uparrow n_\downarrow, \quad O_2 = c_\downarrow - n_\uparrow c_\downarrow, \\ &O_3 = c_\uparrow - n_\downarrow c_\uparrow, \quad O_4 = c_\downarrow c_\uparrow, \\ &O_5 = c^\dagger_\downarrow - n_\uparrow c^\dagger_\downarrow, \quad O_6 = n_\downarrow - n_\uparrow n_\downarrow, \quad O_7 = c^\dagger_\downarrow c_\uparrow, \\ &O_8 = -n_\downarrow c_\uparrow, \quad O_9 = c^\dagger_\uparrow - n_\downarrow c^\dagger_\uparrow, \quad O_{10} = c_\downarrow c^\dagger_\uparrow, \\ &O_{11} = n_\uparrow - n_\uparrow n_\downarrow, \quad O_{12} = n_\uparrow c_\downarrow, \quad O_{13} = c^\dagger_\downarrow c^\dagger_\uparrow, \\ &O_{14} = -n_\downarrow c^\dagger_\uparrow, \quad O_{15} = n_\uparrow c^\dagger_\downarrow, \quad O_{16} = n_\uparrow n_\downarrow. \end{aligned} \tag{15}$$

Then, we can calculate the mutual information of orbital pairs using Eqs. (8)–(10).

In order to calculate the conditional mutual information between orbitals i and j (with $i > j$), we analyze the CI coefficients obtained from a full configuration interaction (FCI) calculation. For each possible occupation pattern of orbitals $1,2,\ldots,j-1,j+1,\ldots,i-1$, we identify and extract all configurations consistent with that occupation pattern. These selected configurations are then normalized and used to construct the reduced density matrix, from which the mutual information between orbitals i and j is evaluated. The weighted average is then performed on all such occupation patterns, where the weight corresponds to the total squared CI coefficients associated with each pattern in the ground state wavefunction:

$$I^{\mathrm{Cond}}_{ij} = \sum_{C(1,2,\cdots,j-1,j+1,\cdots,i-1)} P(C) I_{C,ij}. \tag{16}$$

Here, $I_{C,ij}$ is the mutual information calculated from the configuration under the occupation pattern C. The summation over C runs over all possible occupation states, and the weight is given by the sum of the squared CI coefficients, $P(C)$, for each configuration corresponding to the occupation pattern. Conceptually, this formulation captures the residual correlation between orbitals i and j after accounting for all prior orbital information, analogous to how self-attention mechanisms isolate directional dependencies in transformer models.

Table 1. The elements of two-orbital reduced density matrix of the i-th and j-th orbital expressed in terms of single-orbital operators. In the table, we define $\alpha/\beta = \langle O_{i,\alpha}O_{j,\beta}\rangle$.

$\rho_{\alpha\beta}$	$n=0,s_z=0$	$n=1,s_z=-\frac{1}{2}$	$n=1,s_z=\frac{1}{2}$	$n=2,s_z=-1$	$n=2,s_z=0$	$n=2,s_z=1$	$n=3,s_z=-\frac{1}{2}$	$n=3,s_z=\frac{1}{2}$	$n=4,s_z=0$
	$-,-$	$-,\downarrow$ $\downarrow,-$	$-,\uparrow$ $\uparrow,-$	$\downarrow,\downarrow$	$-,\updownarrow$ $\downarrow,\uparrow$ $\uparrow,\downarrow$ $\updownarrow,-$	$\uparrow,\uparrow$	$\downarrow,\updownarrow$ $\updownarrow,\downarrow$	$\uparrow,\updownarrow$ $\updownarrow,\uparrow$	$\updownarrow,\updownarrow$
$-,-$	1/1								
$-,\downarrow$ $\downarrow,-$		1/6 2/5 5/2 6/1							
$-,\uparrow$ $\uparrow,-$			1/11 3/9 9/3 11/1						
$\downarrow,\downarrow$				6/6					
$-,\updownarrow$ $\downarrow,\uparrow$ $\uparrow,\downarrow$ $\updownarrow,-$					1/16 2/15 3/14 4/13 5/12 6/11 7/10 8/9 9/8 10/7 11/6 12/5 13/4 14/3 15/2 16/1				
$\uparrow,\uparrow$						11/11			
$\downarrow,\updownarrow$ $\updownarrow,\downarrow$							6/16 8/14 14/8 16/6		
$\uparrow,\updownarrow$ $\updownarrow,\uparrow$								11/16 12/15 15/12 16/11	
$\updownarrow,\updownarrow$									16/16

3.2. Weighted attention score

In transformer architectures, each attention head computes token-to-token affinities independently. To accurately evaluate the flow of information between tokens, it is crucial to account not only for the attention weights themselves but also for how these weights interact with the projected feature representations. Specifically, in each head, the attention matrix $A^{m,l}$ is used to weight the projected value vectors $z^{(m)}V^{m,l}O^{m,l}$ (Eq. (12)). This observation suggests that the actual influence from token j to token i is modulated not only by the attention weight $A^{m,l}_{ij}$ but also by the magnitude of the corresponding projected vector. Motivated by this insight, we define a weighted attention score that integrates the attention weights with the norm of the associated value projections. Specifically, given an input sequence $[x_1,\ldots,x_k]$ with the task of predicting x_{k+1}, we compute for each attention head l within block m the attention score of orbital $k+1$ as

$$\bar{A}_{(k+1)j}=\sum_{m=1}^{M}\sum_{l=1}^{H}A^{m,l}_{(k+1)j}\,\|z_j^{(m)}V^{m,l}O^{m,l}\|_2, \tag{17}$$

where M is the number of transformer blocks, H is the number of heads per block, $z_j^{(m)}$ denotes the input representation of token j to block m, and $V^{m,l},O^{m,l}$ are the value and output projection matrices for head l in block m. This formulation captures not just the positional importance assigned by attention, but also the magnitude of information transmitted through each path. The final weighted attention score is obtained by taking the expectation of $\bar{A}_{(k+1)j}$ over the input distribution $[x_1,\ldots,x_k]$ sampled from the autoregressive transformer. This averaged matrix provides a more physically grounded measure of effective attention, and serves as the basis for comparison with conditional mutual information in subsequent analysis.

4. A surrogate model linking attention and mutual information

Motivation While the attention mechanism itself has a clear mathematical form, its combination with normalization, MLP layers, residual connections, and embeddings produces an over-parameterized system that is difficult to analyze directly. To gain intuition, we construct a linearized toy model that isolates the information-transmission role of attention and illustrates, in a simplified setting, how attention coefficients relate to conditional mutual information.

Setting Fix a target position i and predecessors $<i=\{1,\ldots,i-1\}$. Each site t has a binary occupancy variable $n_t\in\{0,1\}$, which can be viewed as the indicator extracted from the more general orbital state x_t used in the NNQS formulation (where x_t may take four possible values). We model

$$z_i=\sum_{t<i}A_{it}v_t,\quad \mathbb{P}(n_i=1\mid n_{<i})=\sigma(z_i), \tag{18}$$

where A_{it} plays the role of an attention coefficient, σ is the logistic sigmoid function $\sigma(u)=1/(1+\mathrm{e}^{-u})$, and v_t is a scalar value. Importantly, v_t is defined as a proportional embedding of the occupancy n_t:

$$v_t=\alpha_t n_t,\quad \alpha_t>0, \tag{19}$$

so that $v_t=0$ whenever $n_t=0$, and v_t can be nonzero only if $n_t=1$. Thus the effective contribution from site t is intrinsically conditioned on its occupation state. For $j<i$, define $\pi_j=\mathbb{P}(n_j=1\mid n_{<i,-j})$ and assume the boundedness condition

$$0<c\le\pi_j\le C<1\quad \text{a.s.} \tag{20}$$

We work in the small-coupling regime $|\Delta|\le\delta\ll 1$ (a.s.), where Δ denotes the difference between the logits z_i conditioned on $n_j=1$ and $n_j=0$, respectively. We denote $H(p)=-p\log p-(1-p)\log(1-p)$, and define $p_i^{(0)}=P(n_i=1\mid n_j=0)$ and $p_i^{(1)}=P(n_i=1\mid n_j=1)$.

4.1. Main statement and two-sided bounds

The conditional mutual information (CMI) can be written as the "Jensen gap"

$$\begin{aligned}I(n_i;n_j\mid n_{<i,-j}) \;=\; & H\big(\underbrace{\pi_j p_i^{(1)}+(1-\pi_j)p_i^{(0)}}_{=:\bar{p}_i}\big)\\ & -\Big[\pi_j H\big(p_i^{(1)}\big)+(1-\pi_j)H\big(p_i^{(0)}\big)\Big].\end{aligned} \tag{21}$$

Theorem 1 (CMI scaling in the linearized model) Under the setting above,

$$\begin{aligned}&I(n_i;n_j\mid n_{<i,-j})\\ &=\frac{1}{2}\,\mathbb{E}\Big[p_i^{(0)}\big(1-p_i^{(0)}\big)\,\pi_j(1-\pi_j)\,\Delta^2\Big]+\mathcal{O}(\delta^3),\end{aligned} \tag{22}$$

where the expectation is over $n_{<i,-j}$. Consequently,

$$\begin{aligned}&\frac{1}{2}\,\eta(1-\eta)\,\underline{\phi}(c,C)\,\mathbb{E}[\Delta^2]+\mathcal{O}(\delta^3)\\ &\le I(n_i;n_j\mid n_{<i,-j})\le\frac{1}{32}\,\mathbb{E}[\Delta^2]+\mathcal{O}(\delta^3)\end{aligned} \tag{23}$$

provided the margin condition $p_i^{(0)}\in[\eta,1-\eta]$ a.s. for some $\eta\in(0,\frac{1}{2})$; here

$$\underline{\phi}(c,C):=\min\{c(1-c),C(1-C)\}. \tag{24}$$

Proof Set $p:=p_i^{(0)}$, $\pi:=\pi_j$ and expand $p_i^{(1)}=\sigma(z_i^{(0)}+\Delta)$ around $\Delta=0$:

$$\begin{aligned}p_i^{(1)}&=p+s_1\Delta+\frac{1}{2}s_2\Delta^2+\mathcal{O}(\Delta^3),\\ s_1&=\sigma'(z_i^{(0)})=p(1-p),\\ s_2&=\sigma''(z_i^{(0)})=s_1(1-2p).\end{aligned}$$

Then

$$\bar{p}_i\;=\;\pi p_i^{(1)}+(1-\pi)p$$

$$= p + \pi s_1 \Delta + \frac{1}{2}\pi s_2 \Delta^2 + \mathcal{O}(\Delta^3).$$

Expand H at p:

$$\begin{aligned} H(\bar{p}_i) &= H(p) + H'(p)(\bar{p}_i - p) \\ &\quad + \frac{1}{2}H''(p)(\bar{p}_i - p)^2 + \mathcal{O}(\Delta^3), \\ H(p_i^{(1)}) &= H(p) + H'(p)(p_i^{(1)} - p) \\ &\quad + \frac{1}{2}H''(p)(p_i^{(1)} - p)^2 + \mathcal{O}(\Delta^3), \end{aligned}$$

with $H'(p) = \log\frac{1-p}{p}$ and $H''(p) = -\frac{1}{p(1-p)}$. Compute

$$I = H(\bar{p}_i) - \left[\pi H(p_i^{(1)}) + (1-\pi)H(p)\right].$$

The linear terms cancel:

$$H'(p)\left[\pi s_1 \Delta\right] - \pi H'(p)\left[s_1 \Delta\right] = 0,$$

and likewise the linear part of Δ^2 vanishes. The leading term is thus quadratic:

$$\begin{aligned} I &= \frac{1}{2}H''(p)\left[(\pi s_1 \Delta)^2 - \pi(s_1\Delta)^2\right] + \mathcal{O}(\Delta^3) \\ &= \frac{1}{2}\left(-\frac{1}{p(1-p)}\right) s_1^2 \Delta^2 \pi(\pi - 1) + \mathcal{O}(\Delta^3). \end{aligned}$$

Using $s_1 = p(1-p)$ gives

$$I = \frac{1}{2}p(1-p)\pi(1-\pi)\Delta^2 + \mathcal{O}(\Delta^3).$$

Taking expectation over $n_{<i,-j}$ proves the expansion. For the upper bound, $p(1-p) \le \frac{1}{4}$ and $\pi(1-\pi) \le \frac{1}{4}$ yield

$$I \le \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{1}{4}\,\mathbb{E}[\Delta^2] + \mathcal{O}(\delta^3) = \frac{1}{32}\,\mathbb{E}[\Delta^2] + \mathcal{O}(\delta^3).$$

For the lower bound, if $p \in [\eta, 1-\eta]$ a.s., then $p(1-p) \ge \eta(1-\eta)$ and $\pi(1-\pi) \ge \underline{\phi}(c,C)$, giving $I \ge \frac{1}{2}\eta(1-\eta)\,\underline{\phi}(c,C)\,\mathbb{E}[\Delta^2] + \mathcal{O}(\delta^3)$. □

Implication Intuitively, when both π_j and $|v_j|$ vary moderately across j, the ranking of CMI is expected to align with the attention ranking. This simplified analysis provides an information-theoretic rationale for the moderate rank correlations observed between attention maps and CMI in our experiments. It also highlights that exact equivalence is generally unattainable in full transformer architectures, where nonlinear activations and multi-head mixing tend to weaken such correspondence.

We next empirically examine this correspondence on several small molecular systems using the QiankunNet framework.

5. Experimental results

To assess the physical interpretability of the learned attention mechanism, we compare it against a theoretically grounded metric: conditional mutual information (CMI) between orbitals. At each autoregressive prediction step, i.e., when predicting orbital x_{k+1}, we extract the attention vector corresponding to x_{k+1} and interpret it as the model's relevance scores over the preceding orbitals. We then compare this attention-induced ranking with the CMI-derived ranking, which reflects the mutual information between x_{k+1} and each prior orbital x_j (with $j < k+1$), conditioned on all other preceding orbitals.

Empirically, we find that attention scores tend to assign higher weights to orbitals with stronger CMI values, suggesting that the model implicitly captures underlying physical correlations. While the alignment is not exact, the relative orderings of orbitals are frequently consistent. To quantify this observation, we compute the Kendall's τ rank correlation coefficient between the attention-based and CMI-based rankings. Details regarding the metric and implementation are provided in Appendix A2.

Data We evaluate our QiankunNet model on four representative small molecular systems: N_2, HN, HF, and LiH, with N_2 containing ten molecular orbitals and HN, HF, and LiH containing six molecular orbitals each. For N_2, the bond length is fixed at 2.0 Å, while for HN, HF, and LiH, the bond length is set to 1.2 Å. These systems are selected for their simplicity and diversity in orbital interactions, providing a suitable testbed for analyzing learned attention patterns.

Metrics We adopt Kendall's τ[29] as the primary measure of rank similarity between attention-derived and CMI-derived orderings (Appendix A1). This non-parametric statistic evaluates the consistency of pairwise relative rankings, with larger τ values indicating stronger agreement.

Visualization and metric-based evaluation Figure 1 provides side-by-side comparisons of weighted attention maps and CMI matrices for N_2, HN, HF, and LiH. We observe that the attention maps often identify the same most and least informative orbitals as those indicated by CMI. That is, the peaks and troughs in attention distributions tend to align well with the extrema in mutual information. The alignment of patterns indicates that the model captures physically meaningful dependencies in orbital relevance. However, when considering the full spectrum of values, there appears to be no simple linear or nonlinear functional relationship between attention scores and CMI values. This suggests that the transformer does not directly replicate the mutual information structure numerically.

As shown in Table 2, all four molecules yield τ values that are significantly higher than a random permutation baseline, indicating that attention preserves the relative relevance structure between orbitals, even without reproducing the exact CMI values. In this sense, attention does not encode the precise magnitude of correlations but tends to capture their hierarchy, highlighting which orbitals matter most. This makes attention a useful qualitative tool for interpreting what the model considers important in decision-making.

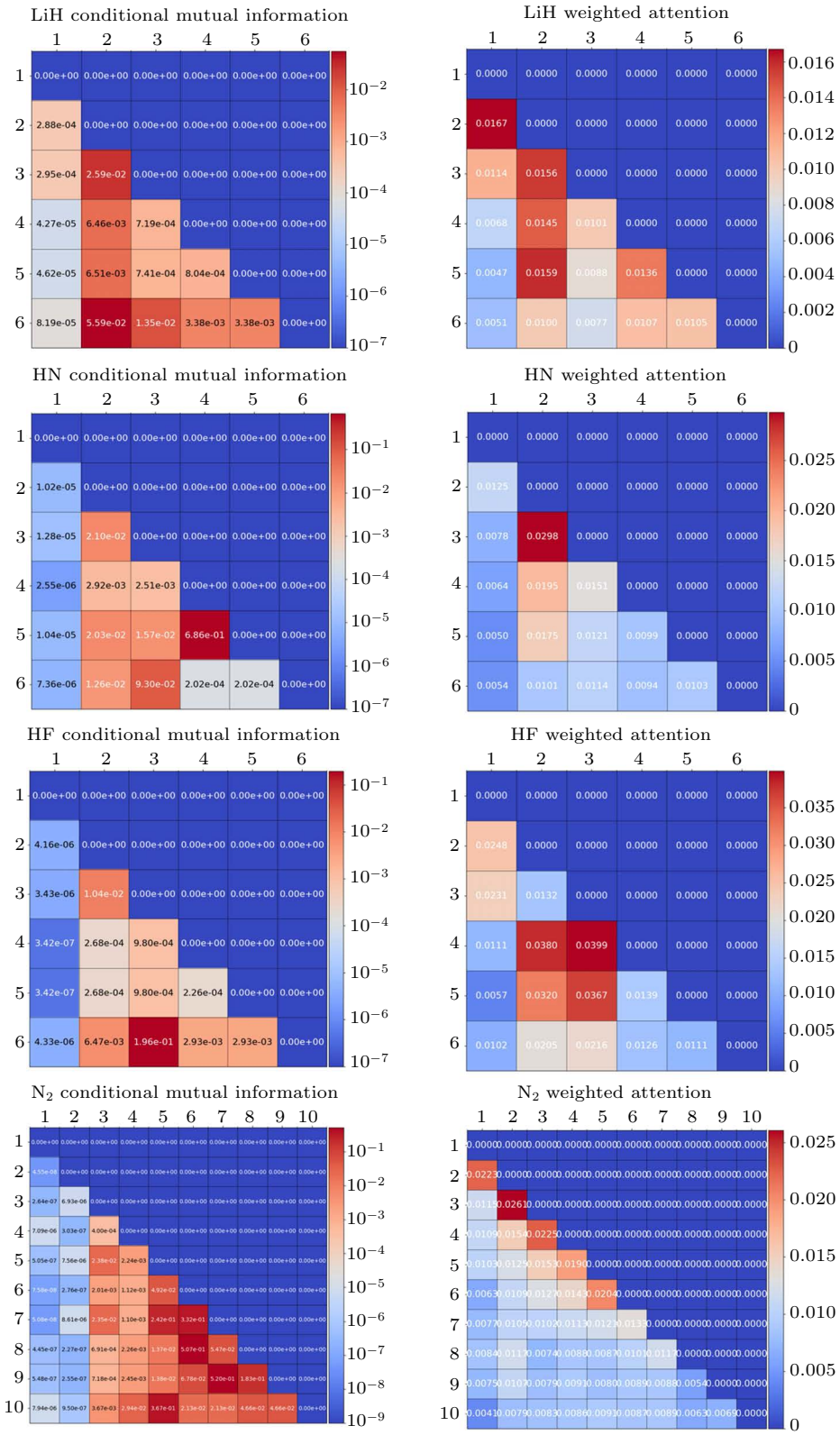


Fig. 1. Comparison between conditional mutual information and weighted attention maps for four molecular systems: LiH, HN, HF, and N_2 (top to bottom). Left: conditional mutual information (CMI) matrices, where each entry at row i, column j reflects the mutual information between orbitals i and j, conditioned on all prior orbitals except j. Right: corresponding attention maps from QiankunNet, weighted as described in Subsection 3.2; each entry shows the attention score assigned to token j when predicting orbital i.

Our experiments reveal that the degree of alignment is system dependent, and it can fluctuate across different molecules. A potential explanation is that attention is not the only mechanism contributing to conditional mutual information: nonlinear modules also play a key role, and their influence may grow with system complexity. In addition, as illustrated by the toy model in Section 4, both π and p affect orbital-level alignment, and real molecules are likely to exhibit

even richer variability.

Taken together, these findings show that attention maps, when interpreted through weighted contributions, provide quantitative evidence of a meaningful connection to conditional mutual information. While the degree of alignment is system dependent, the consistent positive trends across several molecules highlight attention's potential as a proxy for physical correlations and as an explanatory aid for NNQS in quantum chemistry.

Table 2. Kendall's τ between the CMI-based rankings and those derived from (i) weighted attention and (ii) random permutations. For each molecule, τ is averaged over all generation steps. Larger values indicate closer alignment with CMI, while the random baseline, obtained by comparing shuffled sequences with the CMI-derived target sequence, represents the level of agreement expected under no structural signal.

Molecule	Weighted attention	Random baseline
LiH	0.78	0.005
HN	0.73	–0.007
HF	0.49	–0.009
N_2	0.4625	0.006

6. Conclusion

Neural-network quantum states (NNQS) provide a powerful framework for modeling complex quantum systems, but their overparameterized nature and architectural complexity often obscure how physically meaningful structures are internally represented. In this work, we investigated the attention mechanism within QiankunNet and examined whether it bore a meaningful relation to information-theoretic quantities.

By introducing a weighted attention score and comparing it against conditional mutual information (CMI) computed from full configuration interaction (FCI) wavefunctions, we observed positive rank-level correlations between attention-based and CMI-based relevance orderings in several small molecular systems. This suggests that, under suitable conditions, attention can reflect aspects of quantum correlation. Beyond the empirical evidence, we outlined a conceptual link between attention contributions and CMI, providing an intuitive explanation for why such correlations may arise.

It is worth noting that our derivation relies on the autoregressive probability-distribution parameterization of QiankunNet, which is one of the very few NNQS architectures that naturally incorporates an explicit attention mechanism. This unique design makes such a rigorous formulation possible. Nevertheless, QiankunNet adopts a standard transformer backbone and sampling process without ad hoc modifications, ensuring that our observations are not artifacts of special design choices. While this analysis is architecture-specific, the underlying intuition that attention encodes conditional dependencies between orbitals may inspire extensions to other NNQS paradigms.

At the same time, our analysis has clear limitations:

Correlation versus sufficiency Our results rely on Kendall's τ rank correlation, which captures agreement in ordering but does not establish that the attention mechanism fully encodes physically interpretable quantities. The observed alignment may arise indirectly, and further theoretical or experimental probes are needed to clarify the underlying mechanisms.

Restricted evaluation Our experiments are limited to four small molecules with simple geometries and relatively few orbitals. Extending the study to larger and more complex systems is necessary to assess the generality of our observations.

This study presents a quantitative analysis linking transformer attention with conditional mutual information in the NNQS setting. The results provide a step toward explainable deep learning in quantum chemistry and highlight new opportunities for interpreting attention as a proxy for physical correlations.

Appendix A: Experimental details

A1. Kendall's τ

Given two ranked lists σ and π of length n, Kendall's $\tau(\sigma,\pi)$ is defined as

$$\tau = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where $n_0 = \frac{n(n-1)}{2}$ is the total number of possible pairs, n_c is the number of concordant pairs, i.e., pairs (i,j) such that the relative order in σ and π agree, n_d is the number of discordant pairs, where the orders disagree, n_1 and n_2 account for ties in σ and π, respectively.

Kendall's τ takes values in the range $[-1,1]$, where $\tau = 1$ indicates perfect agreement between the two rankings (all pairs are concordant), $\tau = -1$ indicates complete disagreement (all pairs are discordant, i.e., reverse order), and $\tau = 0$ suggests that the rankings are uncorrelated in their pairwise ordering.

For the lower triangular CMI matrix and the attention matrix, we compute their rank similarity for each row in which the number of non-zero elements is greater than or equal to two. The reported result is the average of the rank similarities across all such rows. The CMI-derived ranking serves as the target sequence. The random baseline is computed by comparing 1000 randomly permuted sequences with this target and averaging the resulting Kendall's τ values; it reflects the agreement expected in the absence of structural signal.

A2. Model setting

Basic setting

Device: Experiments were conducted on NVIDIA A100-SXM4-40GB GPUs with 40 GB of memory, using CUDA 12.3

and driver version 545.23.08. The networks are implemented using PyTorch.

Floating-point precision: `fp32`.

Random seed: 333.

System setting

Molecular system: LiH, HF, HN with a bond length of 1.2 Å. N_2 with a bond length of 2.0 Å.

Electron conservation: Restricted.

Model and optimizer configuration

Table A1. Model and optimizer configuration used during training.

Component	Configuration
Wavefunction type	complex
Model architecture	transformer
Embedding dimension of transformer (d_{model})	32
Number of layers	4
Attention heads	4
Dropout	0.0
Phase MLP hidden size	[512, 512, 512, 512]
Optimizer	AdamW
Learning rate	0.0003
Betas	(0.9, 0.99)
Epsilon	1×10^{-9}
Weight decay	0.00
Learning rate scheduler	disabled
Gradient clipping	disabled
Gradient accumulation	enabled (batch size 8192)

Training procedure and sampling

Training epochs: 30000.

Samples per epoch: 12288000.

Wavefunction optimization steps: 3000000.

Sampling algorithm: BFS.

Data availability statement

The paper and the supplementary materials contain the key data supporting the main findings of this study. Additional raw data generated during the current research are available from the corresponding author upon reasonable request. The code used for analysis is not publicly available but can be shared upon request for academic purposes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. T2222026), the CAS Project for Young Scientists in Basic Research (Grant No. YSBR-034) and the Robotic AIScientist Platform of the Chinese Academy of Sciences.

References

[1] Carleo G and Troyer M 2017 *Science* **355** 602
[2] Pfau D, Spencer J S, Matthews A G and Foulkes W M C 2020 *Phys. Rev. Res.* **2** 033429
[3] Hermann J, Schätzle Z and Noé F 2020 *Nat. Chem.* **12** 891
[4] Choo K, Neupert T and Carleo G 2019 *Phys. Rev. B* **100** 125124
[5] Choo K, Mezzacapo A and Carleo G 2020 *Nat. Commun.* **11** 2368
[6] Pescia G, Nys J, Kim J, Lovato A and Carleo G 2024 *Phys. Rev. B* **110** 035108
[7] Barrett T D, Malyshev A and Lvovsky A 2022 *Nature Machine Intelligence* **4** 351
[8] Szabo A and Ostlund N S 1996 *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (Courier Corporation, Mineola, NY)
[9] Roos B O, Taylor P R and Sigbahn P E 1980 *Chemical Physics* **48** 157
[10] Bartlett R J and Musiał M 2007 *Rev. Mod. Phys.* **79** 291
[11] Wang T, Chen J, Teng J, Shi J, Zeng X and Snoussi H 2023 *Chin. Phys. B* **32** 090703
[12] Yang P, Lu P and Zhang T 2023 *Chin. Phys. B* **32** 058902
[13] Chen Q H, Ji Y X, Wang K H, Ma H Y and Ji N H 2024 *Chin. Phys. B* **33** 060314
[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 *Advances in Neural Information Processing Systems* **30**
[15] von Glehn I, Spencer J S and Pfau D 2022 arXiv preprint arXiv:2211.13672
[16] Wu Y, Guo C, Fan Y, Zhou P and Shang H 2023 *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (SC '23)
[17] Geier M, Nazaryan K, Zaklama T and Fu L 2025 arXiv preprint arXiv:2502.05383
[18] Gu Y, Li W, Lin H, Zhan B, Li R, Huang Y, He D, Wu Y, Xiang T, Qin M, *et al.* 2025 arXiv preprint arXiv:2507.02644
[19] Legeza Ö and Sólyom J 2003 *Phys. Rev. B* **68** 195116
[20] Rissler J r m, Noack R M and White S R 2006 *Chem. Phys.* **323** 519
[21] Boguslawski K, Tecmer P, Barcza G, Örs Legeza and Reiher M 2013 *Journal of Chemical Theory and Computation* **7** 2959
[22] Szalay S, Pfeffer M, Murg V, Barcza G, Verstraete F, Schneider R and Legeza Ö 2015 *International Journal of Quantum Chemistry* **115** 1342
[23] von Neumann J 1955 *Mathematical Foundations of Quantum Mechanics* (Princeton, NJ: Princeton University Press)
[24] Shannon C E 1948 *The Bell system Technical Journal* **27** 379
[25] Boguslawski K and Tecmer P 2015 *International Journal of Quantum Chemistry* **19** 1289
[26] Jetley S, Lord N A, Lee N and Torr P H S 2018 arXiv preprint arXiv:1804.02391
[27] Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929
[28] Chefer H, Gur S and Wolf L 2021 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791
[29] Kendall M G 1938 *Biometrika* **30** 81