Lawrence Berkeley National Laboratory
LBL Publications

Title
Transformer-based operator learning framework for self-energy in strongly correlated systems

Permalink
https://escholarship.org/uc/item/6gw4q4mg

Journal
Physical Review B, 113(24)

ISSN
2469-9950

Authors
Zhu, Yuanran
Rosenberg, Peter
Huang, Zhen
et al.

Publication Date
2026-06-01

DOI
10.1103/k5s2-x6hq

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Transformer-based operator learning framework for self-energy in strongly correlated systems

Yuanran Zhu,[1,*] Peter Rosenberg,[2] Zhen Huang,[3] Hardeep Bassi,[4] Chao Yang,[5] and Shiwei Zhang[2]

[1] *Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[2] *Center for Computational Quantum Physics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*
[3] *Department of Mathematics, University of California, Berkeley, California 94720, USA*
[4] *Department of Applied Mathematics, University of California, Merced, California 95343, USA*
[5] *Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, USA, 94720.*

arXiv:2504.14483v3 [cond-mat.str-el] 4 Jun 2026

We introduce Σ-Attention, a Transformer-based operator-learning framework for approximating the self-energy operator of strongly correlated electronic systems. By creating a batched dataset that combines results from three complementary approaches: many-body perturbation theory, strong-coupling expansion, and exact diagonalization, each effective in specific parameter regimes, Σ-Attention is applied to learn an accurate approximation for the self-energy operator that is valid across a wide range of parameter regimes. This hybrid strategy leverages the strengths of existing methods while relying on the transformer's ability to generalize beyond individual limitations. More importantly, the scalability of the Transformer architecture allows the learned self-energy to be extended to systems with larger sizes, leading to much improved computational scaling. Using the 1D Hubbard model, we demonstrate that Σ-Attention can accurately predict the Matsubara Green's function of large systems with a wide range of coupling strength. Our framework offers a promising and scalable pathway for studying strongly correlated systems with many possible generalizations.

I. INTRODUCTION

Strongly correlated materials have long fascinated the condensed matter community due to their rich quantum phenomena, including high-temperature superconductivity, magnetism, and Mott transitions. These emergent properties arise from intricate electron–electron interactions that often render conventional single-particle approaches such as density functional theory (DFT) inadequate. As a result, accurately modeling strongly correlated systems remains one of the most challenging problems in theoretical and computational physics and chemistry.

Over the years, a variety of methods have been developed to study strongly correlated systems. These include dynamical mean-field theory (DMFT) [1–3], strong-coupling expansion (SCE) and cluster perturbation theory (CPT) [4–10], quantum cluster methods [11–13], various quantum Monte Carlo (QMC) techniques [14–20], and tensor-based approaches such as the density matrix renormalization group (DMRG) and its extensions [21–24]. All of these approaches provided valuable insights into different systems and have shown to be effective in different parameter regimes. However, each method has inherent limitations. For instance, SCE is typically restricted to very strong interactions, DMRG does not scale as well in extended systems, and QMC loses exactness in the presence of the sign problem. These challenges motivate the search for a unified framework capable of capturing the complex behavior across all regimes.

In recent years, the rapid development of machine learning (ML)—particularly deep learning—has opened new avenues for addressing complex many-body problems. Notable advances include neural-network-based representations of many-body/Kohn–Sham states or DFT functional [25–29], efficient calculation of two-electron integrals [30, 31], and Green's function-based approaches for studying many-body interactions [32–36]. In these studies, various ML models—such as convolutional neural networks (CNNs), variational autoencoders, and recurrent neural networks (RNNs)—have been employed to model quantum states and calculate observables. Among these, we are particularly interested in ML models that use the Transformer architectures. Originally developed for natural language processing [37–39], Transformer has demonstrated remarkable success in learning operators [40–43] and capturing long-range dependencies. Their inherent scalability and ability to handle variable input dimensions make them an attractive tool for studying strongly correlated systems, especially for addressing finite-size effects and improving computational scaling towards the thermodynamic limit with simulations.

Driven by these motivations, in this work, we introduce Σ-Attention, a transformer-based operator-learning framework designed to approximate the self-energy operator of strongly correlated electronic systems. Leveraging an Encoder-Only transformer as our ansatz, our approach learns a mapping that predicts the self-energy $\Sigma(k, \mathrm{i}\omega_n)$ from the non-interacting Green's function $G_0(k, \mathrm{i}\omega_n)$ and the two-body interaction v. A distinctive aspect of our method is the use of a batched dataset that integrates complementary data from many-body perturbation theory (MBPT), strong-coupling expansion

* yzhu4@lbl.gov

(SCE), and exact results from smaller sizes, e.g., by diagonalization (ED), each of which is effective in specific parameter regimes. This enables the transformer to learn an accurate approximation to the self-energy operator that remains valid across a wide range of parameter regimes. Consequently, our hybrid strategy combines the strengths of these established methods and exploits the transformer's generalization capability to extend predictions to systems with varying sizes and interaction strengths. We test the utility of this framework using a schematic 1D Hubbard model, where accurate numerical results in large system sizes are obtained by auxiliary-field quantum Monte Carlo [14, 15] calculations. We demonstrate that Σ-Attention yields high-fidelity predictions of the Green's function for systems with different sizes and interaction strengths, accurately capturing the metal-to-insulator transition over a wide range of U values.

The remainder of this paper is organized as follows. In Section II, we present the preliminary and background necessary for this study. Section III details the Σ-Attention framework, including its architecture and training strategies. In Section IV, we apply our method to the 1D Hubbard model and compare the predictions with benchmark results. Finally, Section V summarizes the main findings and outlines future research directions. Section A provides all technical details for the numerical simulations.

II. PRELIMINARY

Consider a generic fermionic quantum many-body system in a lattice:

$$\mathcal{H} = \sum_{ij,\sigma} t_{ij} c^\dagger_{i\sigma} c_{j\sigma} + \frac{1}{2} \sum_{ijkl} \sum_{\sigma\sigma'} v^{\sigma\sigma'}_{ijkl} c^\dagger_{i\sigma} c^\dagger_{j\sigma'} c_{k\sigma'} c_{l\sigma}, \quad (1)$$

where t_{ij} represents the single-particle Hamiltonian (including kinetic energy and any external potential) and $v^{\sigma\sigma'}_{ijkl}$ are the two-electron integrals encoding the interaction between electrons. The equilibrium properties of this interacting system are captured by the Matsubara Green's functions. Focusing on single-particle excitations, the single-particle Matsubara Green's function is defined as

$$G_{ij,\sigma}(\tau) = -\langle \mathcal{T} c_{i\sigma}(\tau) c^\dagger_{j\sigma}(0) \rangle,$$

where $\mathcal{T}$ is the time-ordering operator in imaginary time and $\langle\cdot\rangle := \mathrm{Tr}[e^{-\beta\mathcal{H}}]/Z$ denotes the ensemble average. By performing a Fourier transformation with respect to the imaginary time τ, one obtains the Green's function in frequency space $G_{ij,\sigma}(\mathrm{i}\omega_n)$ which is defined on the discrete Matsubara frequencies $\mathrm{i}\omega_n = (2n+1)\pi/\beta$. In this study, we focus on the system with equal spin-up and spin-down electrons and hereafter ignore the spin indices in the Green's function. The following discussion is valid for both the lattice Green's function $G_{ij}(\tau), G_{ij}(\mathrm{i}\omega_n)$ and the k-space Green's function $G(k,\tau), G(k,\mathrm{i}\omega_n)$ commonly used for periodic systems. Hence we will use simple notations G, G_0, Σ instead and write down their explicit expressions whenever needed.

Dyson's equation connects the non-interacting Green's function G_0, to the full (interacting) Green's function G, through the self-energy Σ, which encapsulates all the effects of interactions. In frequency space, Dyson's equation is typically written as

$$G(\mathrm{i}\omega_n) = G_0(\mathrm{i}\omega_n) + G_0(\mathrm{i}\omega_n)\,\Sigma(\mathrm{i}\omega_n)\,G(\mathrm{i}\omega_n), \quad (2)$$

or, equivalently, in its inverse form,

$$G^{-1}(\mathrm{i}\omega_n) = G_0^{-1}(\mathrm{i}\omega_n) - \Sigma(\mathrm{i}\omega_n). \quad (3)$$

By determining the self-energy $\Sigma(\mathrm{i}\omega_n)$, one can solve Dyson's equation to obtain the interacting Green's function, from which various physical observables can be derived. For example, the density of states (DOS) $\rho(\omega)$, which is key to identifying phenomena such as the metal-to-insulator transition, is computed by first performing an analytic continuation $\mathrm{i}\omega_n \to \omega + i0^+$ [44, 45] to obtain the retarded Green's function $G^R(\omega)$, and then using the relation

$$\rho(\omega) = -\frac{1}{\pi} \mathrm{Im}\, G^R(\omega). \quad (4)$$

For weakly correlated systems in which $|v_{ijkl}| \ll |t_{ij}|$, many-body perturbation theory (MBPT) is normally employed to approximate the self-energy. In the bare expansion approach, the self-energy is viewed as an operator that maps the non-interacting Green's function G_0 and the two-body interaction v (with v used as shorthand for v_{ijkl}) to the self-energy Σ. This mapping is expressed via the expansion series

$$\Sigma(\mathrm{i}\omega_n) = \Sigma[v, G_0](\mathrm{i}\omega_n) = \underbrace{\Sigma^{(1)}(\mathrm{i}\omega_n)}_{O(|v|)} + \underbrace{\Sigma^{(2)}(\mathrm{i}\omega_n)}_{O(|v|^2)} + \cdots,$$

where the nth-order contribution $\Sigma^{(n)}$ is represented by the corresponding Feynman diagrams. Resummation techniques can also be applied to obtain the renormalized self-energy operator $\Sigma = \Sigma[v, G](\mathrm{i}\omega_n)$. Consequently, Dyson's equation (2) must be solved self-consistently until convergence is reached.

Perturbative calculations typically fail for systems with intermediate or strong correlations. In such cases, bare approximations often yield non-physical results, while renormalized self-energy approaches—such as the second-order Born (2ndB) or GW methods—may simply fail to converge. Consequently, phenomena like the Mott transition cannot be captured by simple perturbation theory. To address these shortcomings, more advanced self-energy approximations have been developed, notably those based on Dynamical Mean-Field Theory (DMFT) [1] and its extensions, including Cellular DMFT (CDMFT) [3] and the Dynamical Cluster Approximation (DCA)[12, 13], which effectively incorporate short-range spatial fluctuations.

In general, DMFT-based approaches introduce a non-perturbative ansatz for the self-energy operator by matching the self-energy of the embedded impurity or dynamical clusters to the global lattice self-energy. In this work, we show that a transformer-based neural network can provide a novel ansatz, with the matching process ensuring that the ansatz is generally valid across different parameter regimes of the model.

III. METHOD

A. Problem setup

While the self-energy operator is traditionally derived using MBPT, the functional-integral approach [46] indicates that, even in the non-perturbative regime, the self-energy can be expressed self-consistently as an functional of the Green's function:

$$\Sigma = \Sigma[G] = \frac{\delta \Phi}{\delta G},$$

where Φ is the Luttinger-Ward functional. This suggests the existence of a *universal* self-energy operator that maps the interactive Green's function G and the two-body interaction v onto Σ. We further *assume* that this universality is also valid for the bare Green's function G_0 [47] . The aim of this work is to find NN approximations to the self-energy operator:

$$G_0(\mathrm{i}\omega_n), v \xrightarrow{\text{NN}} \Sigma(\mathrm{i}\omega_n): \quad \text{(Bare)}$$
$$G(\mathrm{i}\omega_n), v \xrightarrow{\text{NN}} \Sigma(\mathrm{i}\omega_n): \quad \text{(Renormalized)}$$

where the τ-space formulation can be similarly defined. In this work, we only test the bare ansatz since the subsequent computation of G from the NN-predicted self-energy is straightforward, requiring no self-consistent iterations. The postulated universality would require the approximated self-energy operator to be valid across systems of various sizes, for a particular form of two-body interaction (e.g., Coulomb) v, and for any $G_0(\mathrm{i}\omega_n)$. However, achieving this complete universality using NN is nearly impossible and, in many cases, unnecessary, as the physical properties of electronic systems can vary significantly depending on factors such as geometry, dispersion relations, and interaction strength. Therefore, we will focus our discussions on the 1D Hubbard model. Our goal with machine learning is to design an appropriate neural network and select a sufficiently large and representative dataset, so that, after proper training, the NN-approximated self-energy operator can generalize as broadly as possible to Hubbard models with different interaction strengths U and varying system sizes. In particular, achieving generalizability across system sizes is highly desirable, as it addresses the challenging finite-size effect problems. With that being said, the approach we developed should be generally applicable to a generic electronic system (1), given the right choice of training dataset.

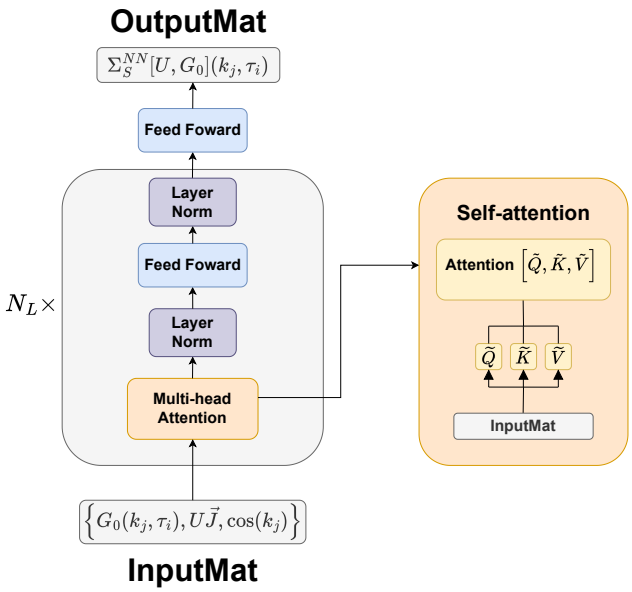

FIG. 1. Encoder-only transformer architecture employed as the neural network module for Σ-Attention. The model requires the bare Green's function G_0, interaction strength U, and positional encoding $\cos(k)$ as input, which are concatenated into a single matrix InputMat (see details in Section A 2); Shown on the right is a schematic of a single-head self-attention module. For full details on the multi-head extension, see Section A. The InputMat is multiplied by different learnable weight matrices W^Q, W^K, W^V to form $\tilde{Q}, \tilde{K}$, and $\tilde{V}$, respectively. This module generates its output via (5). A fully connected (feed forward) layer is then used to project the output to the target dimension of the scaled self-energy matrix prediction Σ_S^{NN}. The loss between the predicted self-energy and ground truth self-energy is then measured to update the total network parameters θ. Here, θ encompasses all parameters that parameterize the weight matrices in the self-attention block(s) and feed forward layer(s). Post-training, with the predicted Σ_S^{NN}, the Green's function $G(k, \tau)$ is solved for via (3).

In the following subsections, we discuss separately the design of the neural network and the selection of the training dataset, which are equally important in our construction and reflect the motivations highlighted above.

B. Encoder-Only Transformer for operator learning

To achieve system size generalizability, the neural network must be trainable on data $\{G_0, \Sigma\}$ from small systems and then be applicable to larger systems. Conventional neural network architectures, such as multilayer perceptrons (MLPs), are not well suited for this task because they require weight matrices and bias vectors whose dimensions depend on the input size. As the system size

changes, the dimensionality of the modeling weight matrices and bias vectors must be changed accordingly. This problem is well-addressed using Transformer models due to its dimensional-agnostic architecture. To see this, it is suffice to examine its core computational unit, namely the self-attention mechanism (see FIG. 1), which has the following computational formula:

$$\text{Attention}[\tilde{Q},\tilde{K},\tilde{V}] = \sigma\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{n_f}}\right)\tilde{V}, \tag{5}$$

where the input is an $n_p \times n_f$ matrix (InputMat). This matrix is projected through learnable weight matrices W^Q, W^K, W^V (dimension of which are $n_f \times n_d$, $n_f \times n_d$, and $n_f \times n_f$) to produce $\tilde{Q}, \tilde{K}, \tilde{V}$. For example, $\tilde{Q} = \text{InputMat} \cdot W^Q$. The crucial observation is that the dimensions of these weight matrices depend only on the *fixed* feature dimension n_f and hidden dimension n_d, but *not* on the physical dimension n_p. Consequently, the same set of weight matrices can process inputs of any physical dimension n_p, making the architecture independent of system size.

For our self-energy prediction task, let us give a simple example of using the above one-layer self-attention as the modeling ansatz for the operator. For this setting, the input matrix InputMat has dimensions $N_k \times (N_\tau + 2)$, composed of three components: (i) the bare Green's function values $G_0(k_i, \tau_j)$, (ii) the interaction strength U (broadcasted as a uniform column), and (iii) positional encodings $\cos(k_i)$. Here, k_i $(i = 1, 2, \ldots, N_k)$ and τ_j $(j = 1, 2, \ldots, N_\tau)$ represent uniformly sampled momentum and imaginary-time grid points on an equally spaced grid with $N_\tau = 101$ points (see Section A 2 for details of the positional encodings). In this example, the physical dimension $n_p = N_k$ varies with system size, while the feature dimension $n_f = N_\tau + 2$ and the hidden dimension n_d is a hyperparameter of the NN that remain fixed across all systems. Because the self-attention weight matrices depend only on n_f and n_d, a single trained model can accept Green's functions from systems of any size. The self-attention output only needs to pass through a linear projection (multiplying by a matrix of dimension $n_f \times (n_f - 2)$) to produce the final predictions of shape $N_k \times N_\tau$, which correspond directly to the self-energy $\Sigma(k_i, \tau_j)$ for all k and τ grids. In specific implementation, we actually use the Multi-Head Attention as the modeling ansatz for the self-energy. Nevertheless, the above analysis for the dimension-agnostic feature of the Transformer architecture still applies. The full architectural details, including the definition of Multi-Head Attention, training procedures, and data handling, are provided in Section A 2.

From an operator-learning perspective, the self-energy operator is approximated using the Transformer ansatz. The network is trained by minimizing the difference $\|\Sigma^{NN}[U, G_0] - \Sigma[U, G_0]\|$ across a variety of input pairs $\{G_0, U\}$. Since the exact self-energy for arbitrary U, G_0 is unknown, the training uses data where reliable approximations (from MBPT, SCE, or ED) are available. The trained network then recovers the self-energy operator in regimes where exact solutions are difficult to obtain.

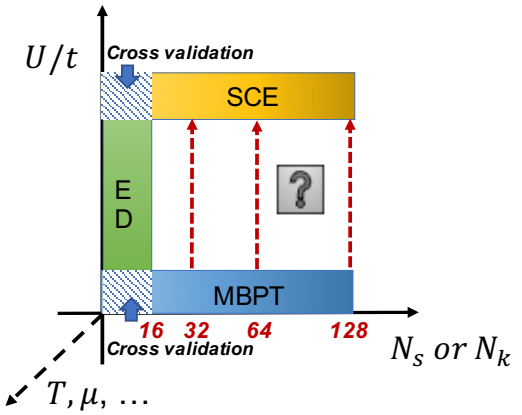

FIG. 2. Complementary datasets used for training Σ-Attention. In the shaded overlapping regime, two different approaches can be used to generate the same training data, allowing cross-validation and establishing the applicability range of perturbative methods such as MBPT and SCE. The area marked by a question mark (corresponding to large N_s and intermediate interaction strength) represents a regime that is typically challenging for current numerical methods, which we aim to explore using the generalizability of Σ-Attention. This paper only considers the half-filling case and fixed temperature. Nevertheless, system parameters such as temperature T and filling μ could be incorporated into the dataset to study T- or μ-induced phase transition or crossover. This awaits further investigations.

C. Complementary datasets for training NN

The Transformer architecture offers promising system-size scalability for representing the self-energy operator. However, achieving generalizability with respect to the interaction strength U—that is, ensuring that the self-energy learned at certain values of U can be extended to other values—is nontrivial. Relying solely on MBPT data is clearly insufficient for this purpose. To capture phenomena such as the metal-to-insulator transition, it is essential to provide the neural network with data spanning both weakly and strongly correlated regimes. This motivates the use of the following complementary datasets for training the neural network, each of which is generated by a different approach. The applicability ranges of these methods are summarized in FIG 2. All the calculation details are provided in Section A 1.

a. MBPT For weakly correlated systems where $U \ll t$, many-body perturbation theory (MBPT) using self-energy approximations such as the second-order Born (2ndB) and GW methods [48] provides quantitatively accurate estimates of both the Green's function and the self-energy. Owing to the perturbative nature of

MBPT, the computational cost of generating the training data $\{G_0, \Sigma^{\mathrm{2ndB/GW}}\}$ remains relatively low even for large systems. Consequently, as illustrated in FIG. 2, the MBPT data occupies the lower rectangular regime of the overall data space.

b. SCE In the regime of extremely strong correlations, where $U \gg t$, a perturbative approach known as the strong-coupling expansion (SCE) [4–7, 49] can be employed to obtain the Green's function by perturbing around the atomic limit. Unlike the weak-coupling MBPT, the SCE regards the Hubbard-U term as the non-perturbative core and expands in powers of t_{ij}. Consequently, this formulation does not directly approximate the conventional self-energy defined in (2), but rather its strong-correlation analogue $\Gamma(\mathrm{i}\omega_n)$. In practice, we use the SCE to compute the interacting Green's function G and then solve (3) to extract the corresponding self-energy Σ^{SCE} [4, 5, 49]. SCE is also a perturbative theory and hence occupies the upper rectangular regime in the overall data space as illustrated in FIG 2.

c. ED In the regime of intermediate interaction strength, perturbation theory often struggles, necessitating the use of alternative methods. Common approaches include quantum Monte Carlo (QMC)[14–16, 50], density matrix renormalization group (DMRG)[21, 51], and exact diagonalization (ED)[52, 53]. These methods can yield quantitatively accurate results within their respective domains, yet each has its limitations. Compared to perturbative techniques, these methods are computationally more demanding and are thus often limited to relatively small systems. As depicted in FIG. 2, they occupy the left rectangular regime of the data space. In this work, we utilize an ED solver [53] to generate this portion of the dataset. Exact results can be produced for larger system sizes in this particular case, for example using AFQMC. However, since our goal in this work is to develop and test the operator learning approach, we leave them for use in validation only.

All these complementary datasets are combined into a single training dataset to devise Σ-Attention to learn a global self-energy operator that is valid for any interaction strength U. As we can see, the assembly of the training data in our design is inherently *modular*; that is, any component of the combined dataset can be replaced if an alternative approach yields quantitatively accurate results within a given parameter range. For example, at half-filling, AFQMC can substitute for other methods since the fermionic sign problem is absent. Likewise, cluster perturbation theory [9, 10] is a viable alternative to the strong-coupling expansion. The guiding principle is that any alternative or new method must provide quantitatively accurate results—otherwise, incorporating inaccurate data would contaminate the data pool and lead to suboptimal training outcomes (see Appendix B 3 for further discussion). Additionally, the generation of these datasets should be computationally efficient, remaining within the comfort zone of the corresponding method, so that the benefits of training a neural network for generalization are fully realized.

Method	N_s	U/t
ED	4, 6, 8, ⋯, 16	0.5, 1.0, ⋯, 7.0
MBPT-I	32, 34, ⋯, 128	0.2, 0.3, ⋯, 1.5
MBPT-II	20, 24, 28	0.2, 0.3, 0.4, 0.5
SCE	20, 34, ⋯, 128	6.0, 6.5, 7.0

TABLE I. System parameters for each method used for generating the training data for Σ-Attention. Note that in the overlapping regime, MBPT and SCE results are benchmarked with ED to determine their validity ranges. Details can be found in Section A 2. When $20 \leq N_s < 32$, the finite-size effect will lead to inaccuracy of the 2ndB calculation for relatively large U. Hence we split the MBPT datasets into two parts as shown above.

On the other hand, N_k and U are not the only system parameters that determine the applicability range of different approaches. In principle, our framework can be generalized to predict and characterize phase transitions associated with any order parameters [54] , such as temperature T and filling μ. The corresponding bare ansatz would then be

$$G_0(\mathrm{i}\omega_n), U, T, \mu, \cdots \xrightarrow{\mathrm{NN}} \Sigma(\mathrm{i}\omega_n) \quad \text{(Bare)}.$$

Accordingly, one may consider incorporating additional datasets, such as those generated by high-temperature expansion [55, 56]. This, however, will not be addressed in this paper and is left for future work.

IV. APPLICATION TO HUBBARD MODEL

We apply Σ-Attention to 1D Hubbard model at half-filling to demonstrate the effectiveness of this operator learning approach. In particular, we focus on whether the model trained on the MBPT-SCE-ED combined datasets can be generalized to predict Green's function of large systems with intermediate interaction strength where the metal-to-insulator transition happens. The modeling Hamiltonian is given by

$$\mathcal{H} = -t \sum_{\langle ij \rangle \sigma} c^\dagger_{i\sigma} c_{j\sigma} + \sum_i U n_{i\uparrow} n_{i\downarrow} - \mu \sum_{i\sigma} n_{i\sigma} \quad (6)$$

where $\langle ij \rangle$ denotes the nearest-neighbor hopping, the chemical potential is set to be $\mu = U/2$, corresponding to the half-filling case. The periodic boundary condition is imposed for $N_s = 4, 8, 12, \cdots$, while the anti-periodic boundary condition is imposed for $N_s = 6, 10, 14, \cdots$ [57]. The inverse temperature is fixed to be $\beta = 20$. We should note that, while in this paper, we only test Σ-Attention in Hubbard model, the framework is generally applicable to systems with general two-body interaction v_{ijkl}.

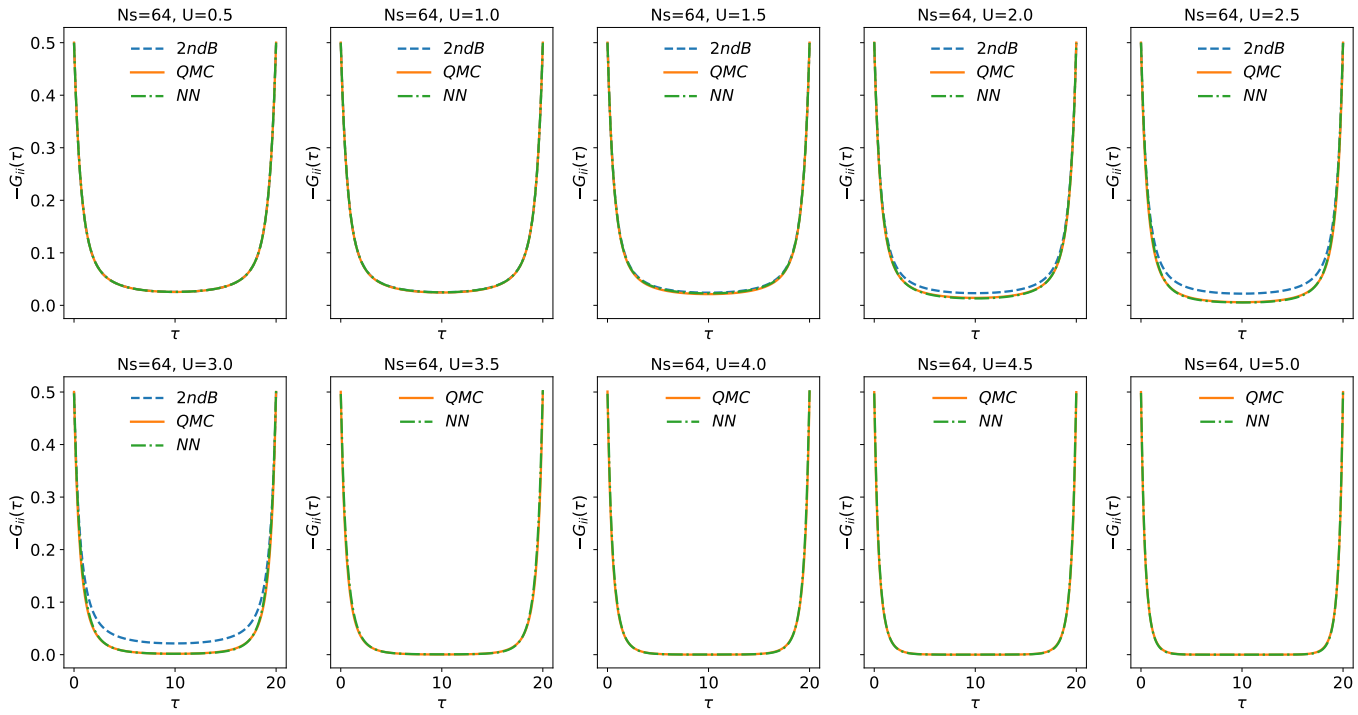

FIG. 3. Comparison of the onsite Green's function $G_{ii}(\tau)$ for Hubbard model with $N_s = 64$ computed using AFQMC, Σ-Attention, and second-order Born (2ndB) calculations. Note that due to the PBC, $G_{ii}(\tau)$ is the same for any site i. The self-consistent 2ndB calculation does not converge for $U > 3.0$, therefore we only displayed the results up to $U = 3.0$.

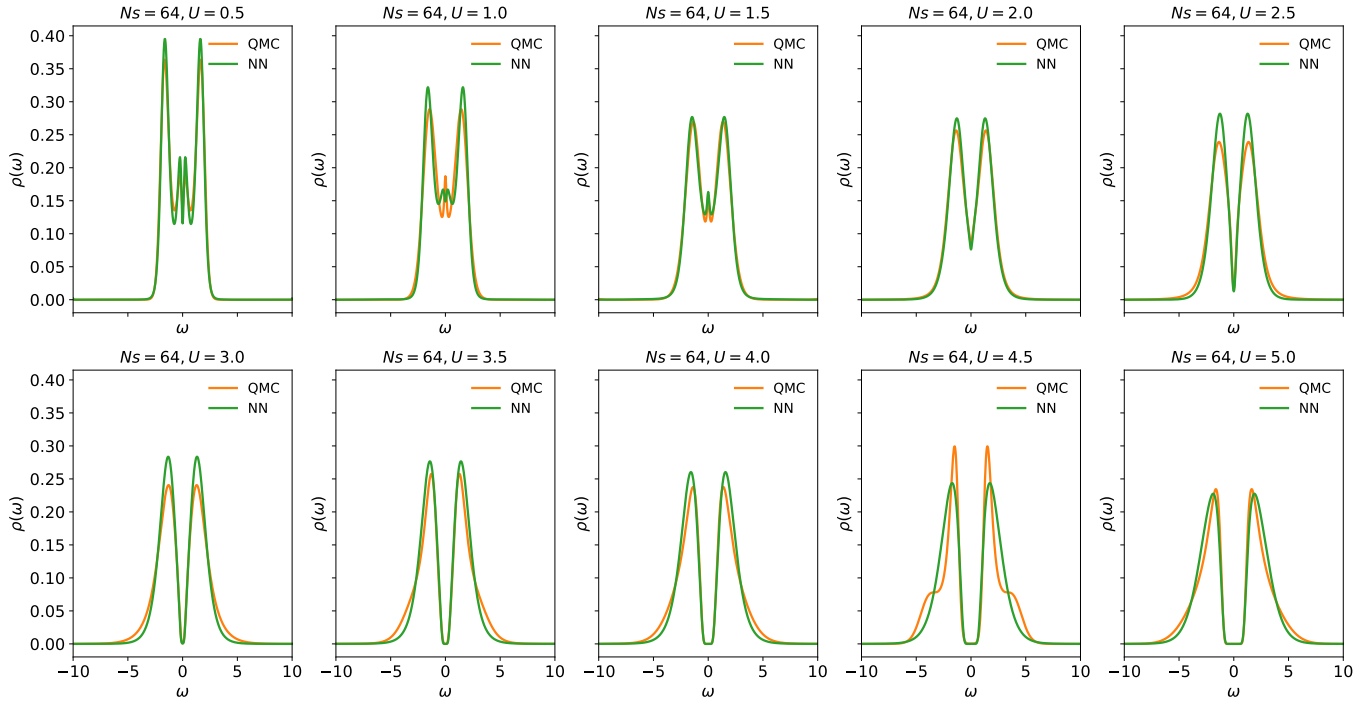

FIG. 4. Comparison of denisty of state (DOS) $\rho(\omega)$ for Hubbard model with $N_s = 64$ computed using the Green's function $G_{ii}(\tau)$ obtained by AFQMC and Σ-Attention. We see that the spectral gap opening is well-captured by Σ-Attention.

A. Numerical results

Table I summarizes the training dataset used in our study. After proper training of Σ-Attention (see Section A 2 for details), we test whether the network can accurately predict the Green's function dynamics for large systems with $N_s = 32, 64, 128$ in the intermediate coupling regime. The training dataset and target prediction domains are illustrated in Figure 2. Since the network takes G_0 and U as input and outputs the self-energy, the full Green's function $G(k, \mathrm{i}\omega_n)$ is computed by solving

$$G^{-1}(k, \mathrm{i}\omega_n) = G_0^{-1}(k, \mathrm{i}\omega_n) - \Sigma^{NN}[G_0, U](k, \mathrm{i}\omega_n). \quad (7)$$

Throughout training and prediction, $G_0(k, \mathrm{i}\omega_n)$ is chosen to be the Hartree–Fock (HF) Green's function. Since Σ-Attention produces output almost instantly, computing $G(k, i\omega_n)$ is extremely efficient and requires virtually no additional computational cost other than matrix inversion in (7). After completing the above procedures, we also apply the adaptive pole-fitting techniques developed in [45] to $G(k, i\omega_n)$ to obtain its *causal* approximation. This final step filters out numerical noise and ensures that the resulting Green's function satisfies the required causality conditions.

FIG 3 presents the final calculation result of $G_{ii}(\tau)$ for $N_s = 64$ and $U = 0.5, 1.0, \ldots, 5.0$. For benchmarking, we compare our results with AFQMC and MBPT outcomes where the later is obtained using the 2ndB self-energy approximation. At half-filling, AFQMC is sign-problem free, making it an ideal benchmark. As shown in the figure, Σ-Attention yields accurate predictions for the Green's function across all U values, whereas the 2ndB approximation deteriorates as U increases, as expected. The comparison between AFQMC results and the NN prediction is shown more clearly in FIG. 5. The overall prediction error,

$$\|G_{ii}^{NN}(\tau) - G_{ii}^{QMC}(\tau)\|_\infty,$$

for all the U values is on the order of 10^{-4} to 10^{-2}; similar error bound holds for the off-diagonal elements $G_{ij}(\tau)$. The additional numerical results for Hubbard model with $N_s = 32, 128$ are provided in Appendix B 1 and the conclusion is the same. To clearly see the metal-to-insulator transition predicted by Σ-Attention, we also compare the density of state (DOS) (4) obtained by the analytic continuation (AC) of the Matsubara Green's function using the Maximum entropy method [58]. The AC procedure is known to be ill-posed and the approximation error is inevitably amplified in $\rho(\omega)$. However, due to the high accuracy of the AFQMC results, the corresponding AC error is better controlled which makes it a reliable benchmark when comparing the main physical features of the spectral function. From FIG 4, we clearly see that the NN-predicted DOS well captures the opening of the spectral gap and have a overall shape agrees with the AFQMC results.

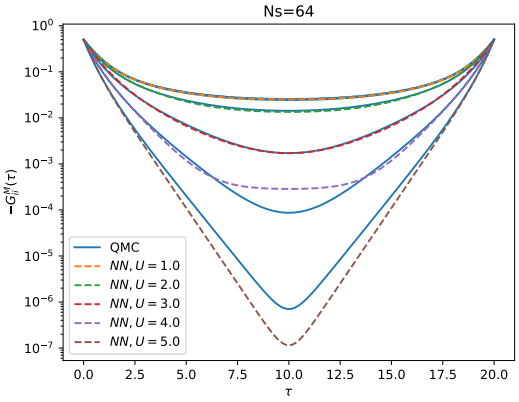

FIG. 5. Prediction of the onsite Green's function $G_{ii}(\tau)$ displayed on the log scale.

Based on the obtained numerical results, Σ-Attention demonstrates great scalability and delivers accurate predictions of Green's function dynamics for systems as large as $N_s = 128$ under intermediate interaction strengths. Comparatively, AFQMC simulations scale roughly as $\mathcal{O}(N_s^3)$ and $N_s = 128$ systems treated here therefore require modest computational resources, in the range of $\sim$ 300 CPU core-hours. In the current framework, the causality of the Matsubara Green's function is enforced through a post-processing step that applies numerical causal projection. A more principled approach would be to incorporate the relevant physical constraints directly into the neural-network ansatz itself. We leave the development of such physically informed architectures to future work.

V. CONCLUSION

In this work, we introduced Σ-Attention, a Transformer-based operator-learning framework designed to learn the self-energy of strongly correlated systems. By leveraging an encoder-only Transformer as an ansatz for approximating the self-energy operator, our approach combines complementary datasets from many-body perturbation theory, strong-coupling expansion, and exact diagonalization. This hybrid strategy allows Σ-Attention to learn an accurate approximation to the self-energy that remains valid across different parameter regimes. Our demonstration on the 1D Hubbard model at half-filling shows that Σ-Attention effectively captures the metal-to-insulator transition while addressing finite-size effects by extending predictions to larger system sizes. This framework not only harnesses the strengths of different established theoretical methods but also exploits the generalization capabilities of Transformer architectures to overcome their individual limitations.

Looking ahead, whether the proposed framework can generalize to non-half-filling cases remains an open question since the underlying physics becomes more com-

plex. We anticipate that it is likely to be insufficient if only using training data from the half-filled regime; however, incorporating representative and accurate training data from non-half-filled regimes could potentially enable accurate predictions. This is an important direction that we leave for future investigation. More broadly, the promising results of this study open avenues for other research directions in the future. Potential extensions include: (i) incorporating additional physical parameters (e.g., temperature, chemical potential) and the corresponding numerical simulation data into the framework; (ii) extending the method to predict two-particle Green's functions and higher-order correlation functions; (iii) generalizing to 2D and 3D systems; and (iv) applying the methodology to real material systems. We believe that Σ-Attention offers a scalable and flexible pathway for advancing the study of strongly correlated materials and may pave the way for new insights into a variety of complex quantum phenomena.

VI. ACKNOWLEDGEMENT

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Basic Energy Sciences, Scientific Discovery through Advanced Computing (SciDAC) program under Award Number DE-SC0022198. This work is also supported by the Center for Computational Study of Excited-State Phenomena in Energy Materials (C2SEPEM) at the Lawrence Berkeley National Laboratory, which is funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under Contract No. DE-AC02-05CH11231, as part of the Computational Materials Sciences Program. This work is partially supported by the Simons Targeted Grant in Mathematics and Physical Sciences on Moire Materials Magic (Z. H.). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES ERCAP0029462 (project m4022) and ASCR-ERCAP m1027. The Flatiron Institute is a division of the Simons Foundation. Y. Zhu gratefully acknowledges the invaluable discussions with colleagues and friends, including Lin Lin, Efekan Kokcu, Lei Zhang, Harish Bhat, Sergei Iskakov, Emanuel Gull, Vojtech Vlcek, Jason Kaye, Harrison LaBollita, Zhouquan Wan and Peizhi Mai.

VII. DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the authors upon reasonable request.

Appendix A: Numerical simulation details

1. Training Data

MBPT, SCE, and ED are employed to generate the combined training dataset for Σ-attention. The MBPT data are obtained from the equilibrium Matsubara Dyson's equation solver in the `NESSi` Package [59], using the second-order Born (2ndB) self-energy approximation for the Hubbard model. The SCE results are based on the analytical expression for the Green's function of the 1D Hubbard model derived by Pairault et al. [4]:

$$G^{-1}(k,\mathrm{i}\omega_n) = 2t\cos(k)+\left\{\frac{\mathrm{i}\omega_n}{(\mathrm{i}\omega_n)^2-\frac{U^2}{4}}+\frac{6t^2U^2\,\mathrm{i}\omega_n}{4\left[(\mathrm{i}\omega_n)^2-\frac{U^2}{4}\right]^3}+6t^3\cos(k)\left[\frac{(\beta U/4)\tanh(\beta U/4)}{\left[(\mathrm{i}\omega_n)^2-\frac{U^2}{4}\right]^2}+\frac{U^2\left(2(\mathrm{i}\omega_n)^2-\frac{U^2}{4}\right)}{4\left[(\mathrm{i}\omega_n)^2-\frac{U^2}{4}\right]^4}\right]\right\}^{-1}.$$

The corresponding self-energy is obtained by solving Dyson's equation (3) using the above $G(k,\mathrm{i}\omega_n)$ and the Hartree-Fock (HF) solution as $G_0(k,\mathrm{i}\omega_n)$. The `EDLib` [53] is the solver we used to generate the ED training data. Note that for the Hubbard model with $\beta = 20$, setting the eigenvalue number threshold to be $N = 20$ would include sufficient exited states for getting an accurate enough Green's function. The corresponding self-energy is also obtained by solving Dyson's equation (3). After finishing all these calculation, we frequently use `pydlr` [60, 61] to accurately interpolate discretized functions and perform transformations such as $G_{ij}(\tau) \leftrightarrow G_{ij}(\mathrm{i}\omega_n) \leftrightarrow G(k,\mathrm{i}\omega_n) \leftrightarrow G(k,\tau)$ for both the Green's function and the self-energy.

As illustrated in FIG. 2, in the overlapping regime, MBPT and SCE generated data are cross-checked with the ED result. Through this comparison, we determined that for the 1D Hubbard model (6) and $N_s \geq 32$, the MBPT result is valid for $U \leq 1.5$ (l_∞-error for $G(k,\tau)$ is $\sim 10^{-3}$). The SCE is valid for $U \geq 6.0$ (l_∞-error for $G(k,\tau)$ is $\sim 10^{-3}$).

2. Learning and Training details

Σ-attention takes as input the Hartree-Fock Green's function $G_0(k,\tau)$ (or equivalently $G_0(k,\mathrm{i}\omega_n)$) along with the Hubbard interaction U and outputs the predicted self-energy $\Sigma(k,\tau)$ (or $\Sigma(k,\mathrm{i}\omega_n)$). The entire neural network is implemented in PyTorch and trained on NERSC

GPUs. Because the main functionalities of PyTorch such as auto-differentiations only support real-valued data, we adopt the τ-space formulation, i.e.,

$$G_0(k,\tau), U \xrightarrow{\text{NN}} \Sigma(k,\tau).$$

For an inverse temperature $\beta = 20$, we discretize $G_0(k,\tau)$ using an equally spaced τ-grid $\{\tau_i\}_{i=1}^{N_\tau} \subset [0,\beta]$ with $N_\tau = 101$. The momentum k-points provide a natural positional encoding via $\cos(k_j)$, where

$$k_j = \frac{2\pi j}{N_s}, \quad j = -\frac{N_s}{2}, \ldots, \frac{N_s}{2},$$

and N_s is the total number of sites. The HF Green's function $G_0(k,\tau_i)$, the scalar U (broadcasted as an all-ones vector $\vec{J}$ of dimension $N_s \times 1$), and the positional encoding $\cos(k_j)$ are concatenated along the columns to form an input matrix:

$$\{G_0(k_j,\tau_i), U\vec{J}, \cos(k_j)\} \xrightarrow{\text{Concat}} \text{InputMat}_{N_s\times(N_\tau+2)}$$

where the feature dimension $n_f = N_\tau + 2 = 103$. Our encoder-only Transformer takes in InputMat and processes it through 2 identical *Multi-Head Attention* layers and one final projection layer to generate the final output: self-energy $\Sigma(k_i,\tau_j) \in \mathbb{R}^{N_s\times 101}$. Conceptually, the Multi-Head Attention is just the concatenation of the output of single-head Attention along the columns, which is projected back to the target output using a new matrix W^O. The corresponding computational rule therefore is [37]:

$$\text{MultiHead}(\text{InputMat}) = \text{Concat}(\text{Head}_1, \cdots, \text{Head}_{16})W^O$$
$$\text{Head}_i = \text{Attention}[\tilde{Q}_i, \tilde{K}_i, \tilde{V}_i]$$

where

$$\tilde{Q}_i = \text{InputMat}W_i^Q,$$
$$\tilde{K}_i = \text{InputMat}W_i^K,$$
$$\tilde{V}_i = \text{InputMat}W_i^V.$$

In our setting, each Multi-Head Attention has 16 heads and the hidden dimension of each head is $n_d = 16$. Consequently, the coefficient matrices W_i^Q, W_i^K, and W_i^V in each head are of dimension $\mathbb{R}^{103\times 16}$ (since the input feature dimension $n_f = 103$). The Multi-head Attention output projection matrix W^O has dimension $\mathbb{R}^{256\times 103}$, mapping the concatenated 256 head dimensions back to 103. Finally, a linear output layer with weight matrix $O \in \mathbb{R}^{103\times 101}$ transforms the hidden representation of dimension $\mathbb{R}^{N_s\times 103}$ into the final output $\Sigma(k_i,\tau_j) \in \mathbb{R}^{N_s\times 101}$.

Note that the weighted inner product in self-Attention has some similarities with the 2ndB (bare) self-energy:

$$\Sigma^{2ndB}(k,\tau) = \frac{U^2}{N_k^2}\sum_{pq} G_0(k-q,\tau)G_0(p+q,\tau)G_0(p,-\tau) \tag{A1}$$

Σ-Attention is trained in a supervised way by matching the NN-predicted self-energy with the ground-truth values obtained via MBPT, SCE, and ED within their respective ranges of applicability. Because the network is optimized on a diverse dataset $\{G_0, U\}$, the corresponding self-energy $\Sigma[U, G_0]$ exhibits varying magnitudes across different U values. This variability can cause optimization difficulties, as the error may fluctuate and eventually favor data points with larger U, leading to underfitting in the small-U regime. To address this and to ensure that the NN-predicted self-energy remains positive (and hence physically meaningful), we rescale the target self-energy according to

$$\Sigma_S[U,G_0](k,\tau) = \frac{\sqrt{|\Sigma(k,\tau)|}}{0.13\,U^2}.$$

where the constant 0.13 is obtained by numerical fitting to ensure the scaled $\Sigma_S[U,G_0](k,\tau)$ for different U-values will have the same magnitude as of $G_0(k,\tau)$, which will make the optimization easier. Thus, the direct output of Σ-attention is $\Sigma_S[U,G_0](k,\tau)$. The loss function used during optimization is a combined $l_1 + l_2$ loss, given by

$$\mathcal{L} = \sum_{ij}\left|\Sigma_S^{NN}[U,G_0](k_j,\tau_i) - \Sigma_S^{\text{target}}[U,G_0](k_j,\tau_i)\right| + \sum_{ij}\left|\Sigma_S^{NN}[U,G_0](k_j,\tau_i) - \Sigma_S^{\text{target}}[U,G_0](k_j,\tau_i)\right|^2.$$

where $\Sigma_S^{NN}(k_j,\tau_j)$ is the output of Σ-Attention that corresponds to the input G_0, U. During training, given the large combined dataset, we employ the randomized batch sweeping technique [62]. Namely, in each epoch, we randomly sample a single (or a small batch) input/output pair: $\{\text{InputMat}, \Sigma_S[U,G_0](k_j,\tau_i)\}$, i.e., a data point from the entire dataset, and perform optimization. For a sufficient amount of training epochs, the randomized training almost surely ensures that every data point from the dataset is visited. This approach shares the same spirit of the stochastic gradient descent and is particularly useful for operator-learning tasks [32, 33, 62]. The training epochs for examples in Section IV are set to be $N = 100000$, Adam optimizer with adaptive learning rate is used to optimize the NN. In PyTorch, this can be done simply using the function `torch.optim.lr_scheduler.CosineAnnealingLR`. FIG 6 illustrates the decay of the training loss for a representative run.

Appendix B: Additional numerical results

1. Hubbard model of different sizes

In this section, we present additional numerical results obtained by Σ-Attention for Hubbard model with system sizes $N_s = 32$ and $N_s = 128$. Figure 8 displays the approximated onsite Green's function $G_{ii}(\tau)$. Figure 9

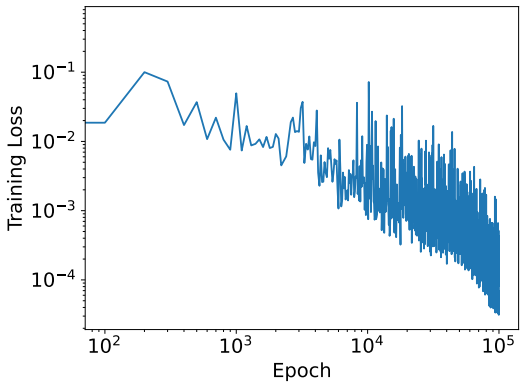

FIG. 6. Training loss decays for a representative run of Σ-Attention. The large error fluctuation is due to the randomized batch sweeping.

illustrates the first few off-site Green's functions $G_{0j}(\tau)$. Figures 10–11 present analogous results for the system size $N_s = 128$. The information we obtained from these figures is consistent with the main conclusions presented in the paper.

2. Effect of causal projection

In the main text, we apply the adaptive pole-fitting techniques developed in [45] to both the AFQMC and the NN-predicted Green's function to ensure its causality. Figure 7 compares the NN-predicted $G_{ii}(\tau)$ before and after the causal projection is applied. Quantitatively, we can see that the causal projection does not change the function significantly but effectively filters out numerical noise and ensures that the resulting Green's function satisfies the required causality conditions. This is done by doing discrete Lehmann representation fitting for the Matsubara Green's function data [45].

3. Importance of combined training datasets

To demonstrate the importance of each component in the combined MBPT+ED+SCE dataset and show each of them influences the final outcome of Σ-Attention. Figure 12 presents the NN prediction for $N_s = 64$ when one component is excluded from the training dataset. Compared with Figures 5, one notices that in all cases the prediction error increases. In particular, excluding the SCE data results in completely wrong predictions (the metal-to-insulator transition is not captured), clearly underscoring its importance in the learning process. This is as expected since the SCE result informs Σ-Attention of the insulting state as $U \to +\infty$.

[1] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions, Reviews of modern physics **68**, 13 (1996).

[2] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. Marianetti, Electronic structure calculations with dynamical mean-field theory, Reviews of Modern Physics **78**, 865 (2006).

[3] G. Kotliar, S. Y. Savrasov, G. Pálsson, and G. Biroli, Cellular dynamical mean field approach to strongly correlated systems, Physical review letters **87**, 186401 (2001).

[4] S. Pairault, D. Sénéchal, and A.-M. Tremblay, Strong-coupling expansion for the hubbard model, Physical review letters **80**, 5389 (1998).

[5] S. Pairault, D. Senechal, and A.-M. Tremblay, Strong-coupling perturbation theory of the hubbard model, The European Physical Journal B-Condensed Matter and Complex Systems **16**, 85 (2000).

[6] W. Metzner, Linked-cluster expansion around the atomic limit of the hubbard model, Physical Review B **43**, 8549 (1991).

[7] T. D. Stanescu and G. Kotliar, Strong coupling theory for interacting lattice models, Physical Review B—Condensed Matter and Materials Physics **70**, 205112 (2004).

[8] C. Gros and R. Valenti, Cluster expansion for the self-energy: A simple many-body method for interpreting the photoemission spectra of correlated fermi systems, Physical Review B **48**, 418 (1993).

[9] D. Sénéchal, D. Perez, and M. Pioro-Ladriere, Spectral weight of the hubbard model through cluster perturbation theory, Physical review letters **84**, 522 (2000).

[10] D. Sénéchal, D. Perez, and D. Plouffe, Cluster perturbation theory for hubbard models, Physical Review B **66**, 075129 (2002).

[11] T. Maier, M. Jarrell, T. Pruschke, and M. H. Hettler, Quantum cluster theories, Reviews of Modern Physics **77**, 1027 (2005).

[12] M. H. Hettler, M. Mukherjee, M. Jarrell, and H. R. Krishnamurthy, Dynamical cluster approximation: Nonlocal dynamics of correlated electron systems, Physical Review B **61**, 12739 (2000).

[13] A. Macridin, M. Jarrell, and T. Maier, Phase separation in the hubbard model using the dynamical cluster approximation, Physical Review B—Condensed Matter and Materials Physics **74**, 085104 (2006).

[14] S. Zhang, J. Carlson, and J. E. Gubernatis, Constrained path monte carlo method for fermion ground states, Physical Review B **55**, 7464 (1997).

[15] S. Zhang and H. Krakauer, Quantum monte carlo method using phase-free random walks with slater determinants, Physical review letters **90**, 136401 (2003).

[16] E. Gull, A. J. Millis, A. I. Lichtenstein, A. N. Rubtsov, M. Troyer, and P. Werner, Continuous-time monte carlo

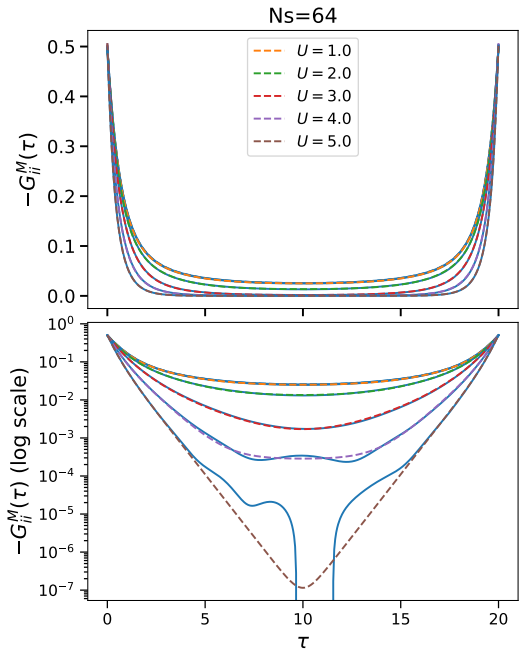

FIG. 7. Comparison of NN-predicted Green's function $G_{ii}(\tau)$ before and after the causal projection, where the solid line represents the original result and the dashed line represents the projected result. The data is displayed both on the regular (top) and log (bottom) scale.

methods for quantum impurity models, Reviews of Modern Physics **83**, 349 (2011).

[17] K. Van Houcke, E. Kozik, N. Prokof'ev, and B. Svistunov, Diagrammatic monte carlo, Physics Procedia **6**, 95 (2010).

[18] R. Rossi, Determinant diagrammatic monte carlo algorithm in the thermodynamic limit, Physical review letters **119**, 045701 (2017).

[19] A. E. Antipov, Q. Dong, J. Kleinhenz, G. Cohen, and E. Gull, Currents and green's functions of impurities out of equilibrium: Results from inchworm quantum monte carlo, Physical Review B **95**, 085144 (2017).

[20] R. Blankenbecler, D. Scalapino, and R. Sugar, Monte carlo calculations of coupled boson-fermion systems. i, Physical Review D **24**, 2278 (1981).

[21] U. Schollwöck, The density-matrix renormalization group, Reviews of modern physics **77**, 259 (2005).

[22] W.-Y. Liu, H. Zhai, R. Peng, Z.-C. Gu, and G. K. Chan, Accurate simulation of the hubbard model with finite fermionic projected entangled pair states, arXiv preprint arXiv:2502.13454 (2025).

[23] G. K.-L. Chan and S. Sharma, The density matrix renormalization group in quantum chemistry, Annual review of physical chemistry **62**, 465 (2011).

[24] G. K. Chan, A. Keselman, N. Nakatani, Z. Li, and S. R. White, Matrix product operators, matrix product states, and ab initio density matrix renormalization group algorithms, The Journal of chemical physics **145** (2016).

[25] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nature physics **14**, 447 (2018).

[26] Z.-A. Jia, B. Yi, R. Zhai, Y.-C. Wu, G.-C. Guo, and G.-P. Guo, Quantum neural network states: A brief review of methods and applications, Advanced Quantum Technologies **2**, 1800077 (2019).

[27] D. Pfau, J. S. Spencer, A. G. Matthews, and W. M. C. Foulkes, Ab initio solution of the many-electron schrödinger equation with deep neural networks, Physical review research **2**, 033429 (2020).

[28] B. Hou, J. Wu, and D. Y. Qiu, Unsupervised representation learning of kohn–sham states and consequences for downstream predictions of many-body effects, Nature Communications **15**, 9481 (2024).

[29] J. Nelson, R. Tiwari, and S. Sanvito, Machine learning density functional theory for the hubbard model, Physical Review B **99**, 075132 (2019).

[30] S. Liang, K. Kowalski, C. Yang, and N. P. Bauman, Effective many-body interactions in reduced-dimensionality spaces through neural network models, Physical Review Research **6**, 043287 (2024).

[31] S. Liang, K. Kowalski, C. Yang, and N. P. Bauman, Exploring the nexus of many-body theories through neural network techniques: the tangent model, arXiv preprint arXiv:2501.15792 (2025).

[32] H. Bassi, Y. Zhu, S. Liang, J. Yin, C. C. Reeves, V. Vlček, and C. Yang, Learning nonlinear integral operators via recurrent neural networks and its application in solving integro-differential equations, Machine Learning with Applications **15**, 100524 (2024).

[33] Y. Zhu, J. Yin, C. C. Reeves, C. Yang, and V. Vlček, Predicting nonequilibrium green's function dynamics and photoemission spectra via nonlinear integral operator learning, Machine Learning: Science and Technology **6**, 015027 (2025).

[34] F. Kakizawa, S. Terasaki, and H. Shinaoka, Physics-informed neural network model for quantum impurity problems based on lehmann representation, arXiv preprint arXiv:2411.18835 (2024).

[35] E. Agapov, O. Bertomeu, A. Carballo, C. B. Mendl, and A. Sander, Predicting interacting green's functions with neural networks, arXiv preprint arXiv:2411.13644 (2024).

[36] C. Venturella, J. Li, C. Hillenbrand, X. L. Peralta, J. Liu, and T. Zhu, Unified deep learning framework for many-body quantum chemistry via green's functions, arXiv preprint arXiv:2407.20384 (2024).

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, Advances in neural information processing systems **30** (2017).

[38] J. Ainslie, S. Ontanon, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, and L. Yang, Etc: Encoding long and structured inputs in transformers, arXiv preprint arXiv:2004.08483 (2020).

[39] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, *et al.*, Deepseek-vl: towards real-world vision-language understanding, arXiv preprint arXiv:2403.05525 (2024).

[40] G. Kissas, J. H. Seidman, L. F. Guilhoto, V. M. Preciado, G. J. Pappas, and P. Perdikaris, Learning operators with coupled attention, Journal of Machine Learning Research **23**, 1 (2022).

[41] E. Calvello, N. B. Kovachki, M. E. Levine, and A. M. Stuart, Continuum attention for neural operators, arXiv preprint arXiv:2406.06486 (2024).

[42] Z. Hao, Z. Wang, H. Su, C. Ying, Y. Dong, S. Liu, Z. Cheng, J. Song, and J. Zhu, Gnot: A general neural operator transformer for operator learning, in *International Conference on Machine Learning* (PMLR, 2023) pp. 12556–12569.

[43] N. Boullé and A. Townsend, A mathematical guide to operator learning, arXiv preprint arXiv:2312.14688 (2023).

[44] J. Fei, C.-N. Yeh, D. Zgid, and E. Gull, Analytical continuation of matrix-valued functions: Carathéodory formalism, Physical Review B **104**, 165111 (2021).

[45] Z. Huang, E. Gull, and L. Lin, Robust analytic continuation of green's functions via projection, pole estimation, and semidefinite relaxation, Physical Review B **107**, 075151 (2023).

[46] M. Potthoff, Non-perturbative construction of the Luttinger-Ward functional, Condens. Mat. Phys. **9** (2006).

[47] Technically speaking, this amounts to assume the existence of a functional $\tilde{\Phi}[G_0]$ such that $\Sigma = \delta\tilde{\Phi}/\delta G_0$. We emphasize that whether this is generally valid is an open question. In this work, we only use this ansatz to motivate the NN-representation of a universal self-energy operator valid both at weak and strong couplings.

[48] G. Stefanucci and R. van Leeuwen, *Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction* (Cambridge University Press, 2013).

[49] N. Dupuis and S. Pairault, A strong-coupling expansion for the hubbard model, International Journal of Modern Physics B **14**, 2529 (2000).

[50] M. Qin, H. Shi, and S. Zhang, Benchmark study of the two-dimensional hubbard model with auxiliary-field quantum monte carlo method, Physical Review B **94**, 085103 (2016).

[51] S. R. White, Strongly correlated electron systems and the density matrix renormalization group, Physics Reports **301**, 187 (1998).

[52] P. Fulde, *Electron correlations in molecules and solids*, Vol. 100 (Springer Science & Business Media, 1995).

[53] S. Iskakov and M. Danilov, Exact diagonalization library for quantum electron models, Computer Physics Communications **225**, 128 (2018).

[54] Presumably, the phase transition or crossover under the investigation of Σ-Attention should be at least second order. Otherwise, the network will likely generate a smooth approximation to the change of order parameter cross the critical point.

[55] M. Bartkowiak, J. Henderson, J. Oitmaa, and P. De Brito, High-temperature series expansion for the extended hubbard model, Physical Review B **51**, 14077 (1995).

[56] E. Perepelitsky, A. Galatas, J. Mravlje, R. Žitko, E. Khatami, B. S. Shastry, and A. Georges, Transport and optical conductivity in the hubbard model: A high-temperature expansion perspective, Physical Review B **94**, 235115 (2016).

[57] The anti-PBC is chosen to avoid the finite-size effect for system sizes $N_s = 6, 10, 14, \cdots$ since otherwise the metal-to-insulator transition cannot be seen due to this sampling of k-points in the first Brillouin zone.

[58] J. Kaufmann and K. Held, ana_cont: Python package for analytic continuation, Computer Physics Communications **282**, 108519 (2023).

[59] M. Schüler, D. Golež, Y. Murakami, N. Bittner, A. Herrmann, H. U. Strand, P. Werner, and M. Eckstein, Nessi: The non-equilibrium systems simulation package, Computer Physics Communications **257**, 107484 (2020).

[60] J. Kaye, K. Chen, and H. U. Strand, libdlr: Efficient imaginary time calculations using the discrete lehmann representation, Computer Physics Communications **280**, 108458 (2022).

[61] J. Kaye, K. Chen, and O. Parcollet, Discrete lehmann representation of imaginary time green's functions, Physical Review B **105**, 235115 (2022).

[62] Y. Zhu, Y.-H. Tang, and C. Kim, Learning stochastic dynamics with statistics-informed neural network, Journal of Computational Physics **474**, 111819 (2023).

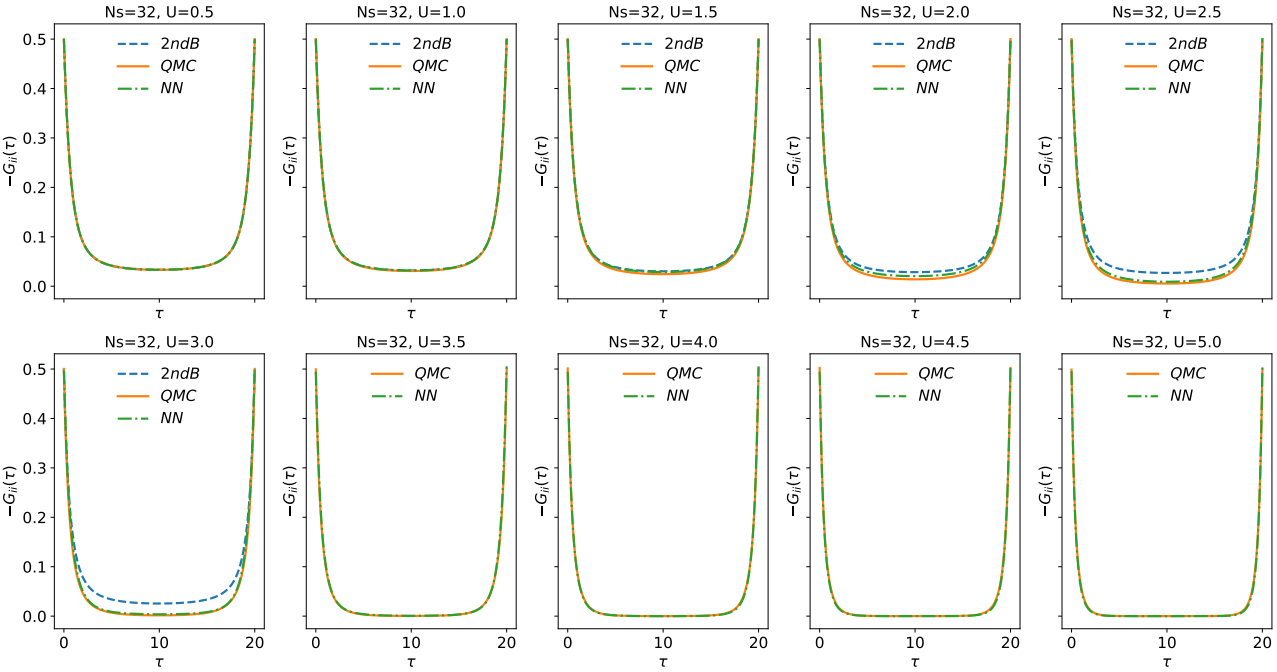

FIG. 8. Prediction of onsite Green's function $G_{ii}(\tau)$ for Hubbard model with $N_s = 32$.

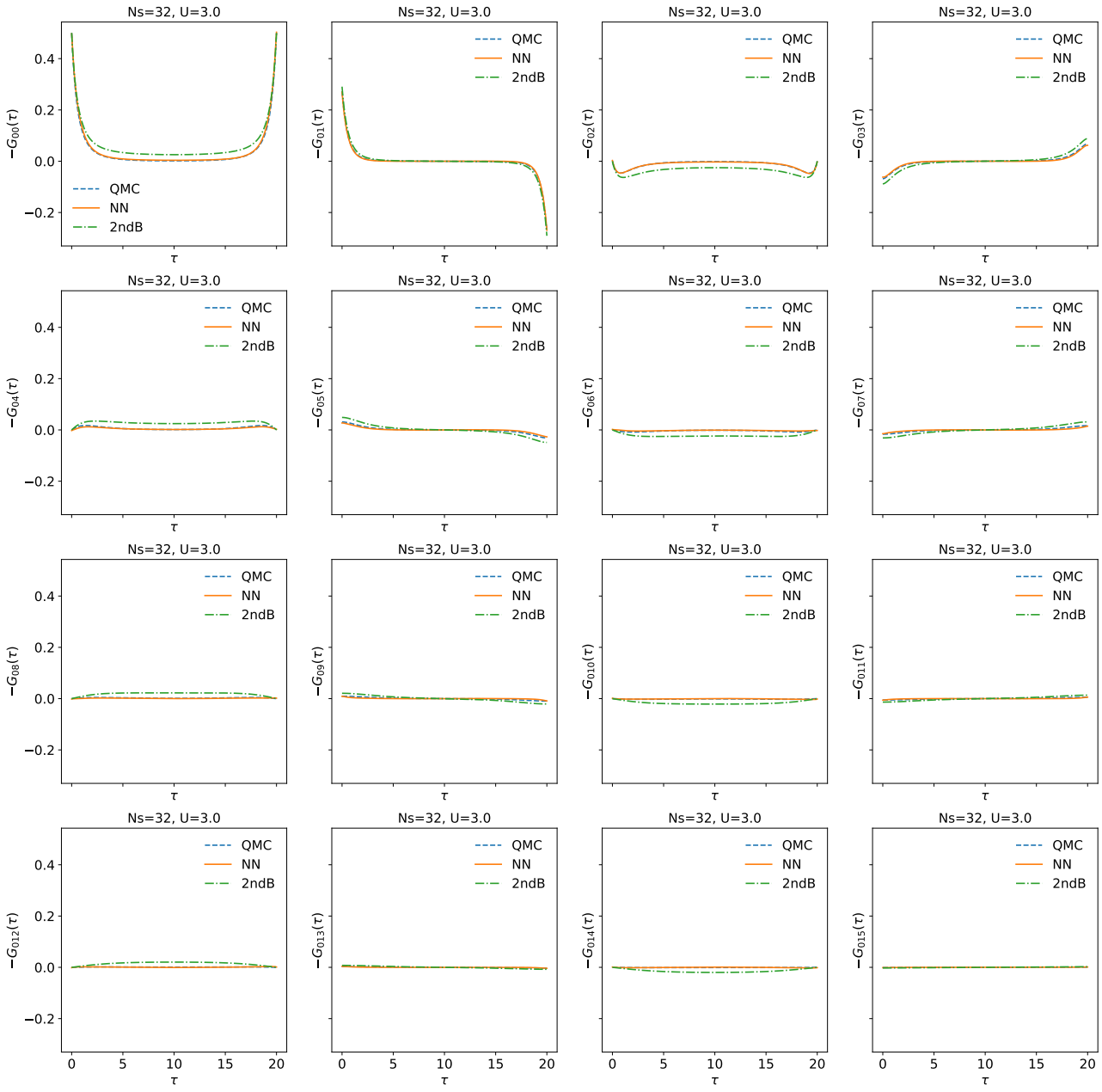

FIG. 9. Prediction of the off-site Green's function $G_{0j}(\tau)$ for Hubbard model with $N_s = 32$. The data shown corresponds to an interaction strength of $U = 3.0$. Similar trends are observed for other values of U.

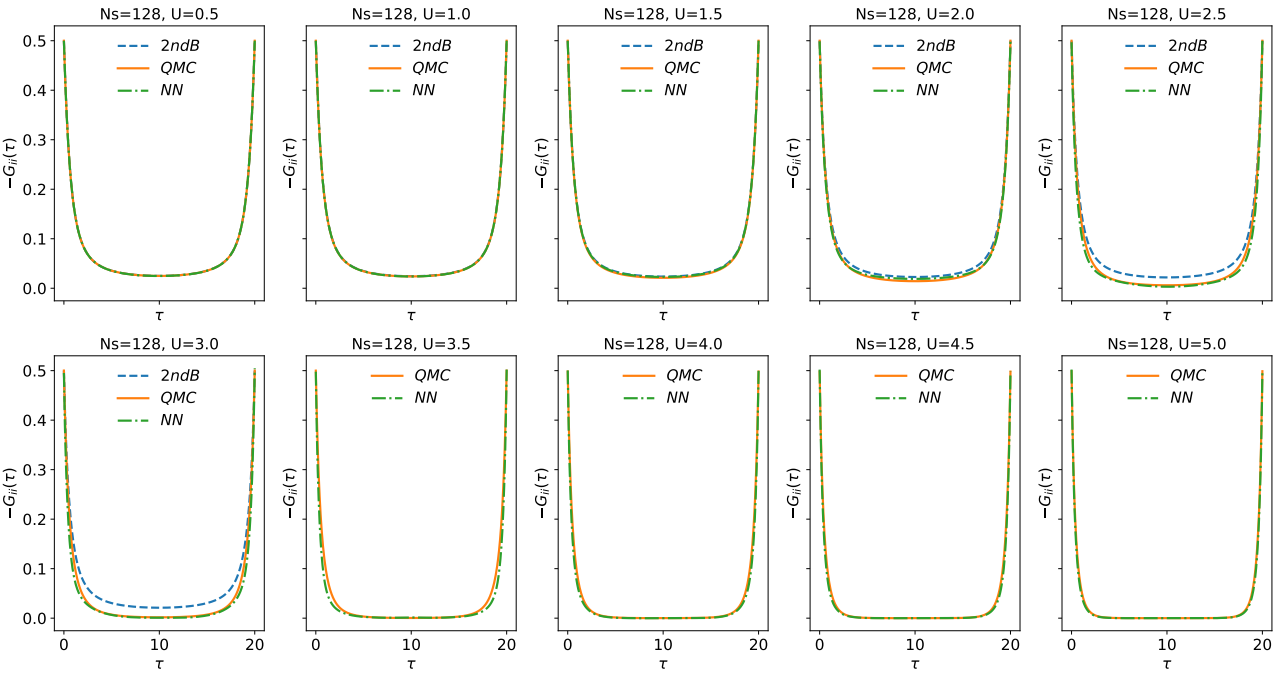

FIG. 10. Prediction of onsite Green's function $G_{ii}(\tau)$ for Hubbard model with $N_s = 128$.

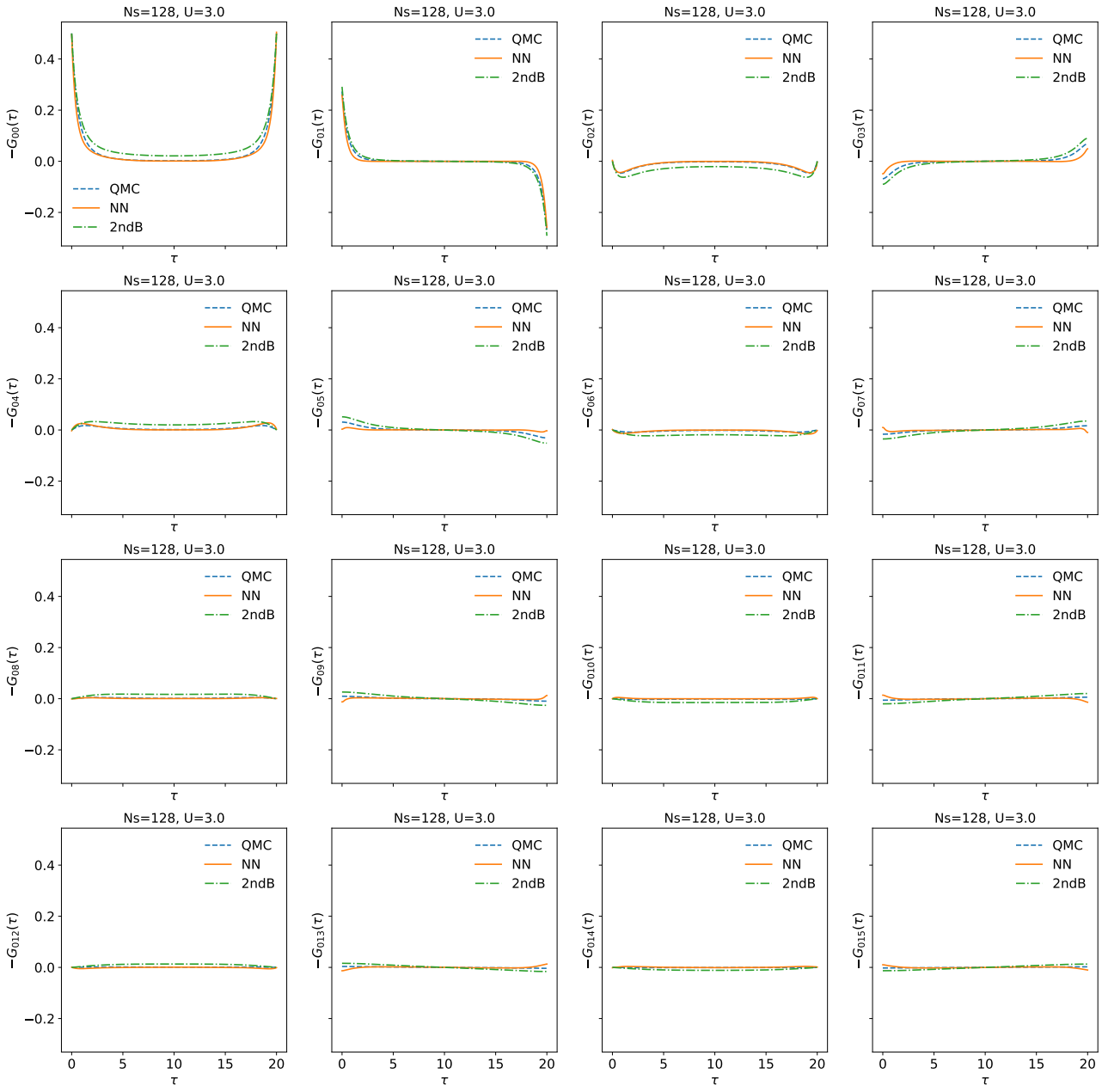

FIG. 11. Prediction of the off-site Green's function $G_{0j}(\tau)$ for Hubbard model with $N_s = 128$. The data shown corresponds to an interaction strength of $U = 3.0$. Similar trends are observed for other values of U.

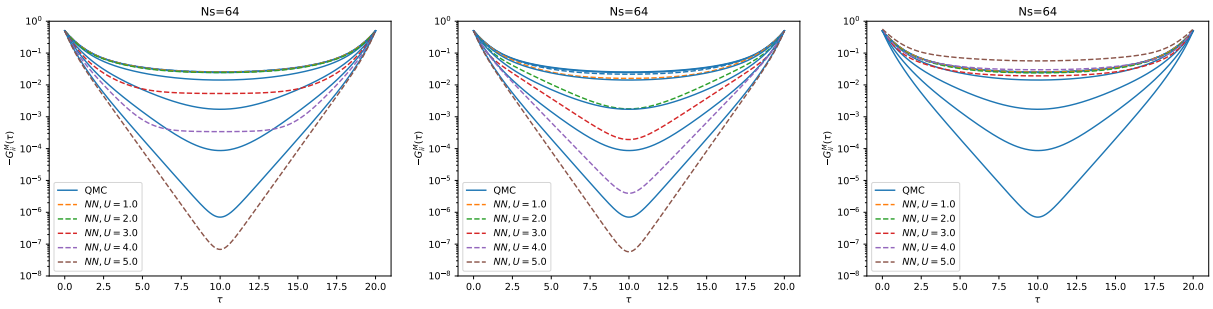

FIG. 12. Comparison of the onsite Green's function $G_{ii}(\tau)$ computed using different combinations of the training datasets. Specifically, they are: (Left) MBPT+SCE (Middle) ED+SCE (Right) ED+MBPT