

Quantum-enhanced neural exchange-correlation functionals

Igor O. Sokolov^{1,*}, Gert-Jan Both¹, Art D. Bochevarov², Pavel A. Dub³, Daniel S. Levine², Christopher T. Brown⁴,
Shaheen Acheche¹, Panagiotis Kl. Barkoutsos¹, and Vincent E. Elfving¹

¹*Pasqal, 24 av. Emile Baudot, 91120 Palaiseau, France*

²*Schrödinger Inc., 1540 Broadway, 24th Floor, New York, New York 10036, USA*

³*Schrödinger Inc., 9868 Scranton Rd., Suite 3200, San Diego, California 92121, USA*

⁴*Schrödinger Inc., 1 Main St., 11th Floor, Cambridge, Massachusetts 02142, USA*



(Received 12 September 2024; revised 22 August 2025; accepted 8 December 2025; published 20 January 2026)

Kohn-Sham density functional theory (KS-DFT) provides the exact ground-state energy and electron density of a molecule, contingent on the as-yet-unknown universal exchange-correlation (XC) functional. Recent research has demonstrated that neural networks can efficiently learn to represent approximations to that functional, offering accurate generalizations to molecules not present during the training process. With the latest advancements in quantum-enhanced machine learning (ML), evidence is growing that quantum neural network (QNN) models may offer advantages in ML applications. In this work, we explore the use of QNNs for representing XC functionals, enhancing and comparing them to classical ML techniques. We present QNNs based on differentiable quantum circuits as quantum (hybrid) models for XC in KS-DFT, implemented across various architectures. We assess their performance on one-dimensional (1D) and 3D systems. To that end, we expand existing differentiable KS-DFT frameworks and propose strategies for efficient training of such functionals, highlighting the importance of fractional orbital occupation for accurate results. We prove that, under specific assumptions, noise-induced errors do not accumulate across SCF iterations, and we confirm this result through numerical experiments. Our best QNN-based XC functional yields energy profiles of the H_2 and planar H_4 molecules that deviate by no more than 1 mHa from the reference DMRG and FCI/6-31G results, respectively. Moreover, they reach chemical precision on a system, H_2H_2 , not present in the training dataset, using only a few variational parameters. This work lays the foundation for the integration of quantum models in KS-DFT, thereby opening new avenues for expressing XC functionals in a differentiable way and facilitating computations of various properties.

DOI: [10.1103/m511-fys2](https://doi.org/10.1103/m511-fys2)

I. INTRODUCTION

Density functional theory (DFT) has emerged as a major direction of research in modern quantum chemistry, enabling an efficient prediction of the electronic structure and properties of molecules and materials [1–3]. It has become the dominant method for solving the Schrödinger equation in physics-based atomistic simulations [4,5]. DFT's effectiveness largely stems from the ability to express the ground-state density of interacting electrons via a system of noninteracting electrons, the so-called Kohn-Sham system. However, DFT can yield exact ground-state energy and electron density only if the exact exchange-correlation (XC) functional is provided. Despite the widespread applicability of the technique, the development of accurate and versatile DFT XC functionals remains a challenge [6–8]. Consequently, DFT faces problems

such as underestimation of reaction barriers, band gaps [9], and poor handling of degenerate states potentially due to delocalization and static correlation errors [4,10]. The universal functional capable of solving the aforementioned problems and capturing the diverse electronic properties of molecules, solids, and surfaces remains unknown. In its pursuit, hundreds of approximate XC functionals have been devised [11,12]. Many of these XC functionals are tailored to certain types of chemical systems and properties thereof and often require prior knowledge of the targeted data to ensure the success of the method [13–15]. An additional challenge when constructing such functionals lies in achieving a good balance between accuracy and computational efficiency [16,17].

A number of studies [18–27] propose the use of compelling capabilities of machine learning (ML) techniques as a solution for the efficient design of XC functionals that can be used for generic DFT calculations. ML algorithms, ranging from deep neural networks to sophisticated regression models, have demonstrated their potential to unlock hidden patterns in vast datasets of chemical systems, allowing for the estimation of a universal XC functional. This combination of quantum chemistry and artificial intelligence heralds a promising era for XC functional design with many demonstrations for different systems [28,29] indicating the increase of the reach of the DFT

*Contact author: sokolov.igor.ch@gmail.com

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

methods. The majority of these ML-based techniques employ differentiable programming as a widely used paradigm in deep learning (DL), which involves optimizing parameters using gradient-based methods during the training process [30]. In particular, existing chemistry codes are being refactored to incorporate automatic differentiation [31,32]. Training neural XC functionals within a differentiable DFT procedure was shown to provide superior generalization capabilities in one dimension (1D) [25,33] compared to the previous approach [34]. Extensions to 3D with variation of model inputs (e.g., density and its gradients), locality [28,32], and novel genetic approaches [35] followed, to name just a few. Nevertheless, a universal neural XC functional has not been trained yet [30]. In an effort to meet this challenge, we propose a novel method for learning representations of XC functionals.

Recently, a novel DL framework within the broader scope of quantum machine learning (QML) has been introduced [36]. This framework utilizes differentiable quantum circuits (DQCs), capitalizing on the unique methodologies inherent in quantum computation. This category of circuits facilitates the construction of quantum neural networks (QNNs), enabling efficient function representation [37] and derivative computation through analytic differentiation rules [38–40]. Various QNN models, which utilize different data encoding methods and distinct quantum circuit designs, have emerged. [41] A relevant quantum advantage in QML has been shown for situations where purely quantum data (quantum state) are used as the model input [42] and when the quantum model is learning from quantum experiments [43]. Given the predominance of classical data in ML, the appeal of QNNs lies in their high effective dimension and rich expressivity [37], which are determined by the unitary mappings of data to quantum states [44]. However, caution is warranted, as commonly used metrics for analyzing QNNs—such as expressivity and entangling capability—may not directly correlate with a model’s ability to perform well on a specific classical dataset. Despite sharing many issues with classical neural networks, such as problems with barren plateaus [45], as well as unique issues such as the absence of efficient backpropagation [46], advances have been made in addressing these through specific techniques [47,48] and efficient output extraction methods [49,50]. Moreover, systematic design principles for QNN architectures that are simultaneously trainable and provably nonclassically simulable have been proposed [51].

Many studies have explored the combination of DFT with variational quantum algorithms through auxiliary Hamiltonians [52], embedding schemes [53], or reduced density matrix methods [54], among others. However, to the best of our knowledge, the direct application of QML for the functional representation of XC has not been explored. Recent developments have further reinforced the connection between DFT and QML. Gaitan and Nori [55] established a mapping between qubit systems and fermionic lattice models, providing a foundation for employing ground-state and time-dependent DFT to study properties of large-scale quantum systems, including the determination of spectral gaps. On the other hand, Baker and Poulin [56] caution against overly optimistic expectations for QML, demonstrating that no QML algorithm can discover the universal functional in polynomial time unless QMA-complete problems reduce to the complexity class

BQP. Within the probably approximately correct framework of machine learning, QML can at best provide a polynomial reduction in oracle queries (training points) when learning the XC functional. We take an optimistic perspective, analogous to the situation in classical ML where theoretical arguments once predicted that training deep-learning models (e.g., due to numerous local minima) would not succeed. Nonetheless, the remarkable success of large language models has demonstrated that practical performance can exceed pessimistic theoretical limitations. Therefore, in this work, we investigate the synergy between QML and DFT for designing XC functionals [57]. As suggested in Ref. [58], our methodology incorporates quantum models as part of larger classical ML pipelines. The goal is to use techniques arising from the QML domain to process empirical data from DFT (or general quantum chemistry) calculations within the supervised ML approach to create more robust and versatile XC functionals that can be applied to a wider range of problems in the quantum chemistry domain.

Our main contributions are as follows:

- (1) Expanding Kohn-Sham density functional theory (KS-DFT) frameworks for quantum XC functional training and implementing in an open-source package QEX [59] written in JAX [60].
- (2) Demonstrating the importance of fractional occupation in 3D KS-DFT for improving accuracy.
- (3) Introducing QNNs models for XC representation, showcasing efficiency in 1D and 3D systems.
- (4) Showcasing that quantum models are at least on par with classical models in terms of accuracy and parameter efficiency on the tested molecular systems.
- (5) Providing error bounds for the KS-DFT self-consistent loop under reasonable assumptions in noisy settings, e.g., when using noisy QNN-based XC functionals.

This work is organized as follows: In Sec. II, we give the theoretical background of QML for KS-DFT and define our models for XC functionals. In Sec. III, we provide details of our framework implementation. In Sec. IV, we present our empirical study of the proposed architectures, analyze the effect of quantum and classical layers within the same backbone architecture, and discuss the required quantum resources and error bounds when considering typical hardware noise. Finally, in Sec. V, we discuss our conclusions and give an outlook on the development of useful quantum-enhanced XC functionals.

II. THEORY

In this section, we introduce the KS-DFT approach, which serves as the foundation for training and evaluating our quantum machine-learned XC functionals. We then delve into the utilization of quantum ML for training XC functionals. Subsequently, we discuss how neural XC functionals are embedded in KS-DFT and explore their architectural details.

A. Kohn-Sham density functional theory

Contemporary KS-DFT implementations [2] achieve the electronic ground state through a solution of the KS equations,

$$\hat{H}[n](\mathbf{r})\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}), \quad (1)$$

to self-consistency with respect to the electronic density $n(\mathbf{r})$. The Hamiltonian operator is given by $\hat{H}[n](\mathbf{r}) = -\nabla^2/2 + v_{\text{KS}}[n](\mathbf{r})$ as a function of the electron density $n(\mathbf{r})$. The latter is defined as the sum of probability densities over all N_{occ} occupied orbitals, $n(\mathbf{r}) = \sum_{i=1}^{N_{\text{occ}}} |\phi_i(\mathbf{r})|^2$, where ϕ_i is a one-electron KS orbital with corresponding energy ϵ_i . The KS potential is given by $v_{\text{KS}}[n](\mathbf{r}) = v(\mathbf{r}) + v_H[n](\mathbf{r}) + v_{\text{XC}}[n](\mathbf{r})$, which contains the external ionic Coulomb potentials $v(\mathbf{r})$, the Hartree potential $v_H[n](\mathbf{r})$, and the XC potential $v_{\text{XC}}[n](\mathbf{r})$. The XC potential is the functional derivative of the XC energy, E_{XC} , with respect to the electron density, i.e., $v_{\text{XC}}[n](\mathbf{r}) = \delta E_{\text{XC}}[n]/\delta n(\mathbf{r})$. The XC energy E_{XC} is defined as

$$E_{\text{XC}}[n] = \int d\mathbf{r} \epsilon_{\text{XC}}[n](\mathbf{r})n(\mathbf{r}), \quad (2)$$

where $\epsilon_{\text{XC}}[n](\mathbf{r})$ is the XC energy per electron. Finally, the total energy is obtained as the sum, $E[n] = T_S[n] + V[n] + E_H[n] + E_{\text{XC}}[n]$, of the noninteracting kinetic energy $T_S[n]$, the external potential energy $V[n]$, the Hartree energy $E_H[n]$, and the XC energy $E_{\text{XC}}[n]$.

In practice, given a finite basis set $\{\xi_i\}_{i=1}^{N_b}$, a linear combination of N_b basis functions, $\phi_i(\mathbf{r}) = \sum_{j=1}^{N_b} p_i^j \xi_j(\mathbf{r})$, is optimized to produce suitable orbitals ϕ . This involves the solution of Roothan's equation defined as

$$\mathcal{F}[n]\mathbf{p}_i = \epsilon_i S\mathbf{p}_i, \quad (3)$$

where $\mathbf{p}_i = (p_i^1, \dots, p_i^{N_b})^T$, $S_{ij} = \int d\mathbf{r} \xi_i(\mathbf{r})\xi_j^*(\mathbf{r})$ is the overlap matrix and $\mathcal{F}[n]$ is the Fock matrix as a functional of the electron density profile n , with elements $\mathcal{F}_{ij} = \int d\mathbf{r} \xi_i(\mathbf{r})\hat{H}[n](\mathbf{r})\xi_j^*(\mathbf{r})$. Equation (3) is then solved, e.g., numerically by an eigendecomposition of the Fock matrix to obtain the orbital coefficients $\{\mathbf{p}_i\}$ that correspond to some density profile n . Self-consistency is achieved when the electron density no longer changes within a specified tolerance over consecutive N_{KS} KS iterations [i.e., solutions of Eq. (3)]. The resulting energy E can then be compared to validated experimental data, with the quality of the result dependent on the choice of the (neural) functional used to represent $E_{\text{XC}}[n]$. If the exact XC functional was known, then Hohenberg-Kohn theorems [61] guarantee that the self-consistent solution yields the exact ground-state energy and density.

B. Embedding of neural XC functionals within KS-DFT

In Fig. 1, we summarize the approaches to incorporate our neural XC models of different localities in KS-DFT. In particular, within the established Kohn-Sham framework (see Fig. 2), neural XC models are integrated using Eq. (2). These models are represented by differentiable functions $f_\theta : (n, \mathbf{r}) \rightarrow \epsilon_{\text{XC}}(\mathbf{r})$ with parameters denoted by θ . Consequently, the machine-learned XC energy can be expressed in general as $E_{\text{XC}}[n] \approx E_{\text{XC}}^\theta[n] = \int d\mathbf{r} f_\theta(n, \mathbf{r})n(\mathbf{r})$, where $E_{\text{XC}}^\theta[n]$ approximates the exact $E_{\text{XC}}[n]$. In practice, an integration grid is used to compute the integral on a finite educated choice of N_{grid} grid points. The XC energy is integrated as

$$E_{\text{XC}}[n] \approx E_{\text{XC}}^{\theta,1}[n] = \sum_{i=1}^{N_{\text{grid}}} w_i f_\theta(n, \mathbf{r}_i) n(\mathbf{r}_i), \quad (4)$$

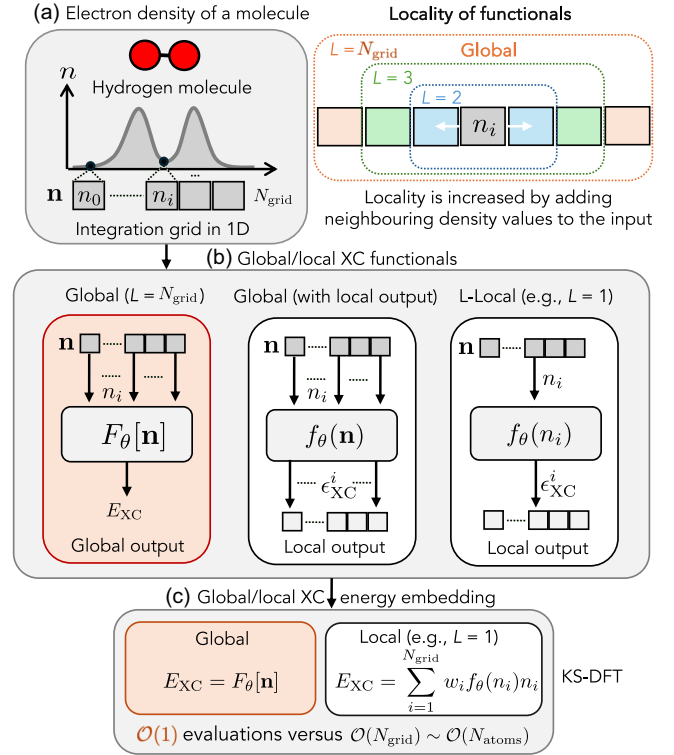


FIG. 1. (a) Electron density of a molecule is expressed on an integration grid. (b) The neural XC functional takes L neighboring values of density as input. (c) The output of such functional can be interpreted as local or total XC energy, with respective embeddings in KS-DFT [Eqs. (4) and (5)]. With the global approach, only $\mathcal{O}(1)$ calls to the functional is necessary to obtain XC energy instead of $\mathcal{O}(N_{\text{grid}})$.

where the grid weights $\{w_i\}_{i=1}^{N_{\text{grid}}}$ are computed for each grid type and chemical system. The values $\{\mathbf{r}_i\}_{i=1}^{N_{\text{grid}}}$ denote the Cartesian coordinates associated with grid points. We call this scheme the *local energy embedding* since local XC energy density $\epsilon_{\text{XC}}(\mathbf{r})$ is required [see Fig. 3(a)].

Another approach is to represent the total XC directly with a neural functional $F_\theta[n]$, bypassing the integration as

$$E_{\text{XC}}[n] \approx E_{\text{XC}}^{\theta,2}[n] = F_\theta[n], \quad (5)$$

where $F_\theta[n]$ learns implicitly the relation between the XC energy and the grid during training through $\{n_i\}_{i=1}^{N_{\text{grid}}}$ input density values. Both equations, Eqs. (4) and (5), provide distinct ways of embedding neural networks as XC functionals in KS-DFT frameworks. We call this approach the *global energy embedding* since the model requires the whole density profile to infer the total XC energy [see Fig. 3(a)]. Note that despite avoiding grid integration in the global energy embedding, both techniques require the computation of the local XC potential $v_{\text{XC}}[n](\mathbf{r})$ at every grid point.

Neural XC functionals are also characterized by various degrees of *locality* that represents the number of density points provided as inputs [34]. Locality was shown to greatly impact the ability to learn accurate XC functionals [25].

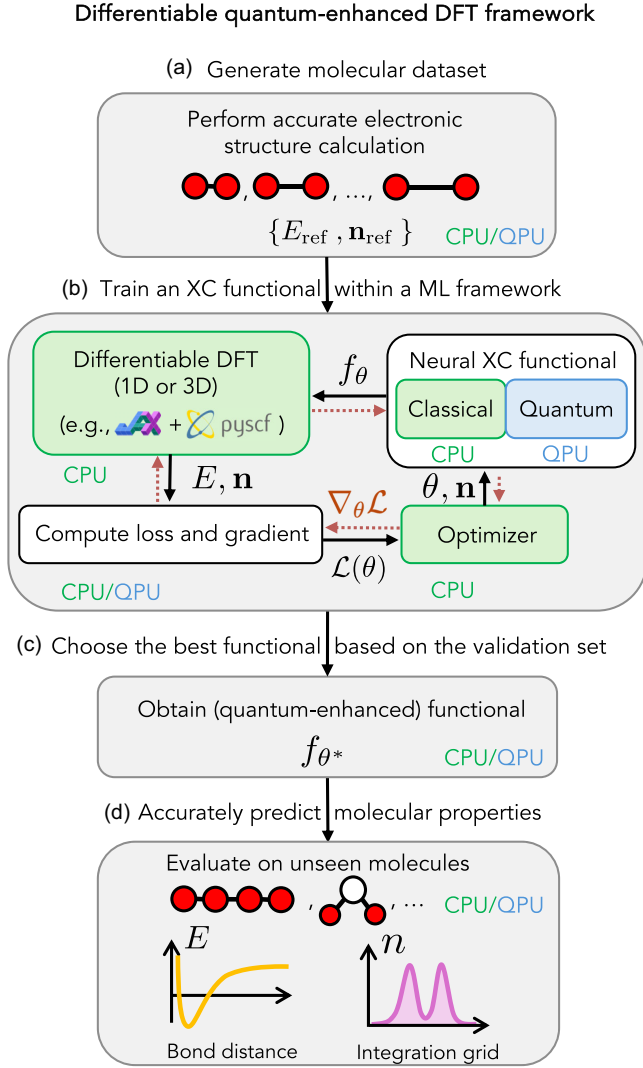


FIG. 2. Differentiable chemistry framework for training quantum-enhanced XC functionals. (a) The molecular dataset consists of reference energies E_{ref} and densities $\mathbf{n}_{\text{ref}}$ generated by an accurate *ab initio* method. (b) A differentiable (Q)NN f_{θ} outputs the XC energy. The corresponding potential is obtained via (automatic) differentiation. The Fock equations in KS-DFT are iteratively solved to obtain new orbitals until self-consistency. Each KS iteration contributes to the loss $\mathcal{L}$, resulting in regularization [25]. Trainable parameters θ are updated by a classical optimizer. (c) The converged parameters θ^* define the trained functional f_{θ^*} , which is subsequently used to evaluate the properties of any molecule in stage (d). Brown arrows in (b) indicate the gradient flow through the procedure. CPU and QPU denote classical (i.e., central and graphical) and quantum processing units.

In particular, L -local ($L \in [[1, N_{\text{grid}}]]$) models are represented by the functions

$$f_{\theta}^{\text{local}} : \mathbb{R}^L \rightarrow \mathbb{R}, \{n\}^L \rightarrow \epsilon_{\text{XC},i}^{\theta}, \quad (6)$$

where the model function has access only to some L neighboring density points (kernel size) at a time and is applied in succession on, e.g., $\lceil N_{\text{grid}}/L \rceil$ points [34]. In the limit where $L = 1$, the 1-local model is represented by the functions

$f_{\theta}^{\text{local}} : n_i \rightarrow \epsilon_{\text{XC},i}^{\theta}$. Local models are embedded in DFT via Eq. (4).

Global models are defined as a functional of the whole density,

$$F_{\theta}^{\text{global}} : \mathbb{R}^{N_{\text{grid}}} \rightarrow \mathbb{R}, \{n\}_{i=1}^{N_{\text{grid}}} \rightarrow E_{\text{XC}}^{\theta}, \quad (7)$$

where the functional represents the whole integrated XC energy and is embedded in DFT via Eq. (5). Alternatively, global models can also be embedded in DFT via Eq. (4) as

$$f_{\theta}^{\text{global}} : \mathbb{R}^{N_{\text{grid}}} \rightarrow \mathbb{R}^{N_{\text{grid}}}, \{n\}_{i=1}^{N_{\text{grid}}} \rightarrow \{\epsilon_{\text{XC},i}^{\theta}\}_{i=1}^{N_{\text{grid}}}, \quad (8)$$

where local XC energy density depends on the whole electron density. Next, we present the concrete realization of our XC models in classical and quantum settings.

C. Architectures of neural XC functionals

In this section, we propose a blueprint of basic model architectures that can couple classical and quantum layers. Neural XC functionals are typically represented as compositions of differentiable functions, implemented through stacks of neural network (NN) layers.

$$f_{\theta}(n, \mathbf{r}) = (f_{\theta_{N_l}}^{N_l} \circ f_{\theta_{N_l-1}}^{N_l-1} \circ \dots \circ f_{\theta_1}^1)(n, \mathbf{r}), \quad (9)$$

with parameters $\theta = (\theta_{N_l}, \dots, \theta_1)$ and N_l layers.

1. Classical architectures

The most general neural network is a multilayer perceptron (MLP) which can approximate any function given enough layers [62]. Since we are exploring the most general quantum models, MLPs are suitable candidates for comparative studies.

For our purpose, we implement MLPs with N_i inputs, N_l layers, and N_n neurons per layer. We introduce two MLP variants: Local MLP (LMLP) uses local density information at N_{grid} points to map:

$$f_{\theta}^{\text{LMLP}} : n_i \rightarrow \epsilon_{\text{XC},i}^{\theta}, \quad (10)$$

while global MLP is a functional of the whole density, mapping:

$$F_{\theta}^{\text{GMLP}} : \{n\}_{i=1}^{N_{\text{grid}}} \rightarrow E_{\text{XC}}^{\theta}. \quad (11)$$

Finally, another notable neural network is the global Kohn-Sham regularizer (GKSR) model, which combines local and global convolutional layers. GKSR incorporates inductive biases such as a self-interaction gate and a negativity transform, addressing common issues like self-interaction errors and imposing the negativity condition on the XC energy [25]. In our implementation, the GKSR model is represented by the function:

$$f_{\theta}^{\text{GKSR}} : \{n\}_{i=1}^{N_{\text{grid}}} \rightarrow \{\epsilon_{\text{XC},i}^{\theta}\}_{i=1}^{N_{\text{grid}}}. \quad (12)$$

In the following section, we discuss how XC functionals can be represented as quantum circuits.

2. Quantum architectures

A widely adopted strategy for QML models involves the utilization of a QNN, typically constructed using differentiable quantum circuits [36]. In this work, our quantum XC

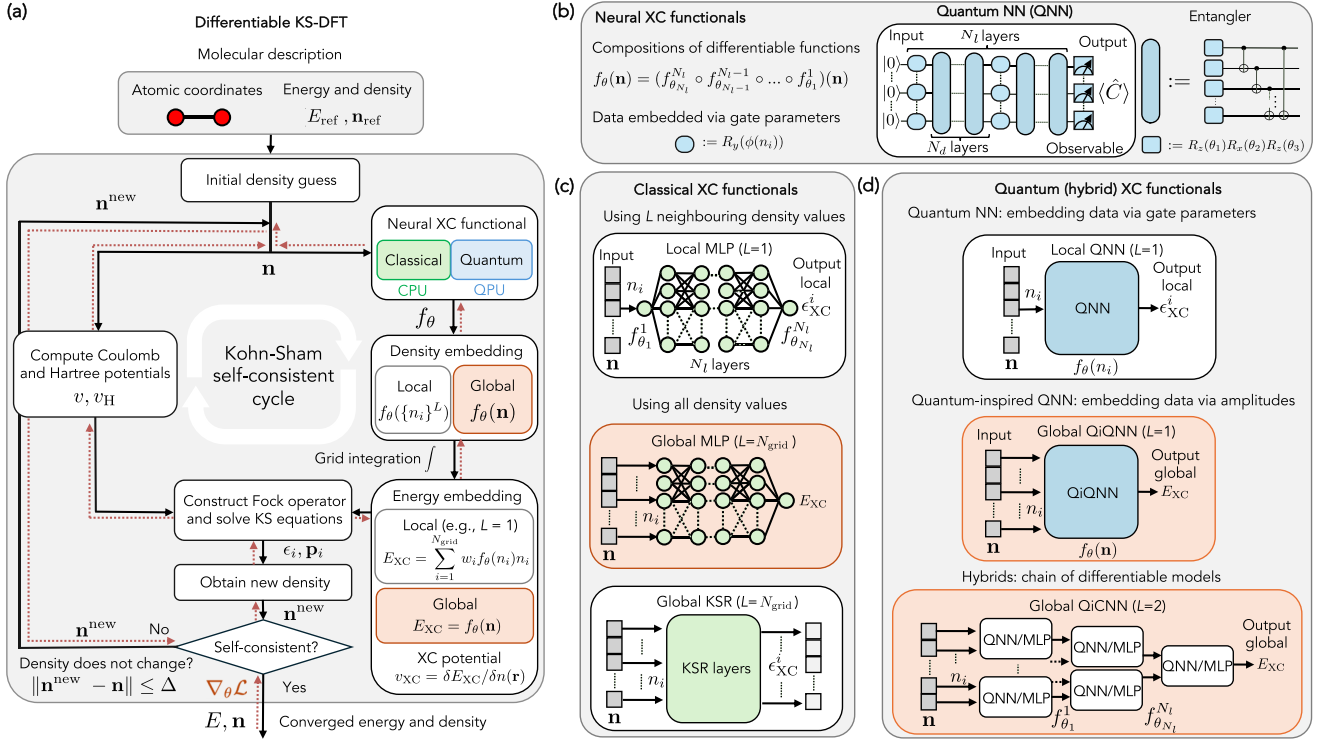


FIG. 3. (a) Method for training quantum-enhanced neural XC functionals within KS-DFT. Data describing a single molecule is passed to KS-DFT. An initial electron density $\mathbf{n}$ estimate is made and potentials v , v_H and v_{XC} are computed to construct the Fock operator. Neural XC functional f_θ is represented by a quantum and/or classical model, which takes L density values as inputs (i.e., $L = 1$ fully local and $L = N_{\text{grid}}$ global). The XC functional can be embedded in DFT locally (outputs the values on every grid point) or globally (outputs the total XC energy). Then, the KS equation and the orbital coefficients $\mathbf{p}$ are updated. The new density is computed and if the change is small enough, the procedure is stopped (self-consistency) and the trained XC functional is obtained. Otherwise, the new orbitals are kept for the next iteration. The loss function is computed and differentiated with respect to the model parameters. The red dashed arrows show the gradient flow. (b) Building blocks of neural XC functionals are defined as compositions of differentiable functions, which can be represented by QNNs. (c) Architectures of classical and (d) quantum (hybrid) XC functionals. Models differ by how many inputs can be taken at a time. The QiCNN architecture allows one to chain classical and quantum models. Orange indicates that the global energy embedding [see (a)] is used, and blue indicates quantum operations. Note that MLPs can be replaced by any classical architecture.

functionals are based on such QNNs, which consist of three blocks: a feature map U_f , an ansatz U_a , and a measurement of an operator $\hat{O}$. Typically, a N_q -qubit quantum computer is initialized at some initial state $|0\rangle$. Here we consider the zero computational basis state on all qubits. A feature map is used to embed the data (e.g., electron density n) in the Hilbert space, $U_f(n)|0\rangle$ in a differentiable manner for the computation of XC potential v_{XC} . The discretized density $\{n_i\}_{i=1}^{N_{\text{grid}}}$ values are embedded in the feature map as gate parameters. Then, an ansatz circuit $U_a(\theta_a)$ is used to act on the embedded data with local qubit operations and entangling gates that mediate interactions to represent the most general action on the data. Finally, the output of the model is extracted via projective measurements (shots) of an arbitrary cost operator $\hat{C} = \sum_{i=1}^{N_o} \theta_c^i \hat{O}_i$, where θ_c are fixed or trainable parameters associated with Pauli operators $\hat{O} \in \{I, \hat{X}, \hat{Y}, \hat{Z}\}^{\otimes N_q}$. We assume the parameters θ_c are nontrainable, unless specified otherwise. The cost operator is chosen to be the total magnetization, composed of local observables, $\hat{C} = \sum_{i=1}^{N_q} \hat{Z}_i$, which was shown to exhibit, at worst, polynomially vanishing gradients if the depth scales as $\mathcal{O}(\log(N_q))$, thereby mitigating cost-function-induced barren plateaus [63]. The representation

of such model is given by a function

$$f_\theta^q(n) = \langle 0|U^\dagger(n, \theta)\hat{C}U(n, \theta)|0\rangle, \quad (13)$$

where $U(n, \theta) = \prod_{l=1}^{N_l} (\prod_{d=1}^{N_d} U_a^{l,d}(\theta_a^{l,d})U_f^l(n))$ is the general representation of the model with N_d ansatz layers and N_l total repetition of ansatz and feature map, which implements the data reuploading strategy [64]. The repetitions of layers are akin to the number of layers of classical NNs and augment the expressivity of the model. Given enough layers, classical and quantum models can, in principle, approximate a given function [62,64,65]. In Appendix C, we give details of typical building blocks of QNNs.

To parallel the local MLP, we formulate a fully local QNN designed to take a single density point n_i , from density vector $\mathbf{n}$, as input and yield the expectation of an observable:

$$f_\theta^{\text{LQNN}} : n_i \rightarrow \epsilon_{XC,i}^\theta = \langle 0|U^\dagger(n_i, \theta)\hat{C}U(n_i, \theta)|0\rangle, \quad (14)$$

defined through Eq. (6) and integrated into DFT via Eq. (4). We explore local QNNs with both product [Eq. (16)] and Chebyshev [Eq. (17)] feature maps, named local PrQNN and local ChQNN, respectively. The ansatz adheres exactly to the structure specified in Eq. (C3). Note that for this model it is possible to consider the L -local case, f_θ^{LQNN} :

$\{n_i\}^L \rightarrow \epsilon_{\text{XC},i}^\theta = \langle 0|U^\dagger(n, \theta)\hat{C}U(n, \theta)|0\rangle$, where a collection of L neighboring densities is embedded in a single feature map. However, as shown in Ref. [25], only global functional (e.g., $L = N_{\text{grid}}$) is capable of best approximating exact XC functionals. Hence, in the next section, we develop model architectures that embed the whole density.

3. Feature maps for quantum models

The choice of feature map is crucial for the expressiveness of quantum XC functionals. We consider encoding data in circuit parameters and amplitudes using general angle embeddings as specific cases of feature maps:

$$U_f(\mathbf{n}) = \bigotimes_{k=1}^{N_q} e^{-\frac{i}{2}\varphi(n_k)\hat{G}_k}, \quad (15)$$

where $\hat{G}$ are generators (possibly multiqubit operators) and $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ is a data transformation function. We explore three established strategies for embedding data in quantum circuits. The choice of those strategies can be encouraged by viewing the feature maps selected for this work as Fourier series, which allows for a deeper understanding of their structure and behavior. According to Schuld *et al.* [44], such feature maps admit a Fourier series representation with a (potentially) exponential number of modes in the number of encoding gates or layers. This means that QNNs have an expressive power that may grow exponentially. It remains an open question whether the greater expressivity of QNNs provides a tangible advantage in representing XC functionals, even when such models cannot be reliably approximated classically for large N_q [66]. However, despite decades of research towards a universal XC functional, no single functional has proven accurate across all molecular systems. In this context, such new model types could offer an intriguing alternative.

Product features map. We explore the 1-local product features map that embeds the data value n_k in the rotation angles of R_y gates of all qubits:

$$U_{f,\text{pr}}^{(1)}(n_k) = \bigotimes_{l=1}^{N_q} e^{-\frac{i}{2}n_k\hat{Y}_l}. \quad (16)$$

Similarly to L -local models discussed earlier, an L -local product feature map can encode multiple data points within a single feature map. Nevertheless, this approach was not employed in the present work.

1-local Chebyshev feature map. A Chebyshev feature map utilizes a polynomial basis [44]:

$$U_{f,\text{ch}}^{(1)}(\mathbf{n}) = \bigotimes_{k=1}^{N_q} e^{-i \cdot k \cdot \arccos(n_k)\hat{Y}_k}. \quad (17)$$

However, normalization of the density to a unit vector, i.e., $\mathbf{n} \leftarrow \mathbf{n}/\|\mathbf{n}\|$, is necessary. It suppresses electron count information and thus renders this approach unsuitable for density-based models.

Amplitude feature map. Finally, the amplitude feature map maps the density vector to a quantum state as $|\mathbf{n}\rangle = U_{f,a}^{(N_{\text{grid}})}(\mathbf{n})|0\rangle$. However, this method presents two important drawbacks: The density is also normalized to a unit vector

which suppresses the information of the number of electrons and, in general, an efficient implementation algorithm for the unitary $U_{f,a}^{(N_{\text{grid}})}(\mathbf{n})$ does not exist for arbitrary vectors $\mathbf{n}$ [67].

4. Quantum-classical hybrid architectures

We propose to leverage advantages of classical NNs and combine them with quantum models. Notably, the number of elements N_{grid} in the electron density $\mathbf{n}$ scales linearly with the number of atoms [68]. Even for small diatomic molecules, there are hundreds of grid points for the smallest grid density available with the best integration grids [68]. This poses an issue with the feature maps based on angle encoding since the number of qubits N_q required to embed the entire density $\mathbf{n}$ equals its size N_{grid} , posing severe requirements for potential implementation on near-term quantum computers. Hence, we devise a model, global QNN (GQNN), that is composed of two layers:

$$f_\theta^{\text{GQNN}}(n) = (f_{\theta_c}^c \circ f_{\theta_q}^q)(n), \quad (18)$$

where f^c and f^q represent the classical and quantum layers with associated parameters θ_c and θ_q , respectively. In principle, any differentiable classical layer can be used to map the outputs of $f_{\theta_q}^q$ to XC energy. Here the quantum layer consists of L -local QNNs defined as

$$f_\theta^q: \{n_i\}^L \rightarrow \epsilon_{\text{XC},i} = \langle 0|U^\dagger(n, \theta)\hat{C}U(n, \theta)|0\rangle, \quad (19)$$

with a product feature map and an alternate ladder ansatz. We first partition the input $\mathbf{n}$ in $N_b = \lceil N_{\text{grid}}/L \rceil$ batches and then apply f_θ^q to each batch. This results in N_b outputs that become the inputs of a classical function f_c defined as

$$f_{\theta_c}^c: \{\epsilon_{\text{XC},i}\}_{i=1}^{N_b} \rightarrow E_{\text{XC}}^\theta, \quad (20)$$

where, for simplicity, the classical layer is the standard MLP with N_b input neurons and one output neuron (potentially with additional MLP layers). In general, any classical model architecture can be used to represent the classical layer.

5. Quantum-inspired architectures

Another possibility for designing XC functionals is to use simulated quantum models as quantum-inspired (Qi) solution. We introduce the global QiNN, which incorporates density through amplitude embedding [Eq. (15)] and utilizes the standard ansatz outlined in earlier sections. It is represented by the function

$$F_\theta^{\text{QiNN}}: (\{n\}_{i=1}^{N_{\text{grid}}}) \rightarrow E_{\text{XC}}^\theta, \quad (21)$$

where amplitude embedding feature map is used.

Finally, we propose global quantum-inspired convolutional NN architecture that implements the composition of quantum layers with angle embedding as

$$f_\theta^{\text{GQiCNN}}(n) = (f_{\theta_c}^c \circ f_{\theta_{N_1}}^{q,N_1} \circ \dots \circ f_{\theta_{q,1}}^{q,1})(n), \quad (22)$$

where the product feature map [Eq. (16)] is throughout all the layers. We do not employ Chebyshev embedding [Eq. (17)]

due to detrimental normalization of the data required, which suppresses the information about the number of electrons in our architectures. Similarly to the hybrid architecture (GQNN), we consider quantum layers that consist of L -local QNNs, defined in Eq. (19), with a product feature map and alternate ladder ansatz defined in Appendix C. The input is partitioned in $\mathbf{n}$ in $N_b = \lceil N_{\text{grid}}/L \rceil$ batches and then f_{θ}^q is applied to each batch. This is repeated N_l times until the output is a single scalar, which represents E_{XC} . If not (i.e., odd number of inputs), then the output of quantum layers is combined with an MLP as in Eq. (20) to output E_{XC} . Note that, although all models in this work are simulated, we classify this subset (QiNN and QiCNN) as quantum inspired because the amplitude encoding used in the global QiNN (i.e., $|\mathbf{n}\rangle = \sum_{j=1}^{N_q} \mathbf{n}_j |j\rangle$, where $\mathbf{n}$ is assumed to be normalized) typically requires an exponential number of gates [67,69], rendering it impractical on current NISQ devices. The QiCNN model is implemented as a sequence of QNNs in which the inputs to each layer are the expectation values from the previous layer, making the model more susceptible to error accumulation, unsuited for current noisy architectures.

III. METHODS

In this section, we discuss the implementation of our approach in differentiable chemistry frameworks and strategies for efficient training of neural XC functionals.

A. Implementation within differentiable chemistry frameworks

To evaluate the capabilities of quantum computers for machine learning XC functionals, we based our implementation on two differentiable frameworks.

Li *et al.* proposed an important method to facilitate the generalization of neural XC functionals: using the KS iterative procedure as a regularizer [25]. The novel idea is to compute the energy loss at every KS iteration and the electron density loss at the last KS iteration and then to differentiate through N_{KS} iterations [solutions of Eq. (1)] which would enable the model to converge faster to the reference, i.e., in fewer KS iterations, while regularizing its parameters. To allow this feature, a fully differentiable chemistry code in 1D in JAX [60] was developed [25] with the aim of demonstrating the regularization capacity of the framework [25]. Since it reduces the number of dimensions (to 1D) and hence simplifies the problem, this is an ideal setup for investigation of the capacity of QNN-based models to represent XC functionals. We adopt this exact framework with minimal changes to hyperparameters for all our preliminary results in 1D (i.e., only linear molecules).

However, realistic chemistry simulations require a full 3D treatment of molecular systems. To address this demand, Kasim *et al.* proposed a 3D differentiable chemistry framework implemented entirely in PyTorch [70], which enabled accurate training of functionals for various diatomic systems [28]. However, both specialized frameworks of Li *et al.* and Kasim *et al.* are not used by computational chemists on a regular basis. Therefore, in this work, one of our goals is to contribute to an existing effort that focuses on making a popular chemistry framework, PySCF [71], differentiable.

This differentiable framework is known as PySCF with automatic differentiation (PySCFAD) [31]. It is similar to Li's 1D framework, as it is also based on JAX, but offers the capability for an accurate treatment of chemical systems in 3D, making it suitable for a wide range of chemical simulations. Working within the scope of the PySCFAD framework, we replaced the calls to the standard XC functionals in differentiable SCF calculations used in KS-DFT (see Sec. II) by (quantum) NNs. This entails that we introduce necessary modifications to the already well-tested KS-DFT implementation.

To facilitate the training of neural XC functionals within KS-DFT in 3D, the spin components of the electron density can be constrained to be equal, i.e., $n_{\uparrow}(\mathbf{r}) = n_{\downarrow}(\mathbf{r})$ and $n(\mathbf{r}) = n_{\uparrow}(\mathbf{r}) + n_{\downarrow}(\mathbf{r})$, which defines the so-called spin-restricted case of KS-DFT (RKS-DFT).

In addition, standard RKS-DFT implementations fail to consistently represent the ground-state density, preventing the training of accurate XC models. Although the exact interacting ground-state wave function cannot, in general, be represented by a single-determinant wave function of a non-interacting KS system with binary orbital occupations (0 or 1), this is not a problem for the density itself. Still, RKS-DFT can fail numerically for open-shell systems (typically strongly correlated systems near dissociation), because the usual expression for the density as a sum over occupied orbitals becomes ill defined: KS valence orbitals may be (near-)degenerate, which can lead in practice to convergence issues in (R)KS-DFT. This, in turn, necessitates fractional occupation of the KS system, which can no longer be described by a single Slater determinant but instead must be treated as an ensemble. To address this limitation, especially evident in the challenging strongly correlated regime, we adopt the algorithm of Chai [72]. Other approaches can be employed with an increased computational cost [73]. Our choice was dictated by the simplicity of its integration in existing codes, with detailed proofs given in Ref. [72] and our implementation in Appendix A.

In summary, we extended two chemistry codes to incorporate QNN-based functionals within KS-DFT:

(1) For 1D cases, we utilized the KS-DFT implementation by Li *et al.* [25].

(2) For 3D cases, we employed the differentiable KS-DFT implementation from PySCFAD [31], to which we added the fractional occupation (FO) method developed by Chai [72].

The 1D framework simplifies the training of neural functionals by limiting spatial degrees of freedom, although it restricts the study to linear systems. This setup provides an ideal environment to explore quantum models in a simplified context. Conversely, the 3D framework permits the evaluation of similar systems within a comprehensive representation, allowing for a realistic analysis. To provide a quick overview, Figs. 2 and 3(a) summarize our framework in the following steps:

(1) Generate dataset. Electronic structure calculations are performed via the best available level of theory (e.g., full configuration interaction (FCI) [74] and DMRG) within a basis set (e.g., 6-31G* or larger) to create a dataset consisting of a set of energies and densities $\{(E_{\text{ref}}^i, \mathbf{n}_{\text{ref}}^i)\}_{i=1}^{N_{\text{mol}}}$, with $\mathbf{n}_{\text{ref}}^i$ evaluated on a selected grid of size N_{grid} , for N_{mol} molecular geometries.

(2) Train models. At every epoch, N_{KS} iterations are performed, each calling a neural functional f_{θ} that outputs the XC energy E_{XC} via local [Eq. (4)] or global embeddings [Eq. (5)]. The corresponding v_{KS} potential is obtained via (automatic) differentiation. Then, the Fock equations are solved to obtain new orbitals. This whole procedure continues until self-consistency. Each KS iteration contributes to the loss $\mathcal{L}$, resulting in regularization [25]. At each epoch, trainable parameters θ are updated by a classical optimizer.

(3) Select the best model. The converged parameters θ^* define the trained functional f_{θ^*} and are selected at one out of N_{epochs} epochs that provide the lowest loss on the validation set (i.e., selected geometries that are not present in the train and test sets).

(4) Predict properties. The selected XC functional f_{θ^*} can be used to evaluate molecular properties. The neural functional can be completely agnostic to the input size (N_{grid}) by design and hence used for any molecular system.

Our implementation is available on GitHub as part of the QEX software package [59]. In the next section, we describe each step of the framework in detail.

B. Training neural XC functionals

In this work, we adopt the standard supervised ML approach. We generate the training (reference) dataset $D = \{(E_{\text{ref}}^i, n_{\text{ref}}^i)\}_{i=1}^{N_{\text{mol}}}$ using an accurate *ab initio* method for N_{mol} different molecules (also including different geometries of the same molecules). For the 1D implementation, we use the original training and test datasets by Li *et al.* [25] composed of hydrogen atom-based systems, e.g., H_2 , H_2H_2 , and H_4 . The data are obtained with DMRG simulations, where $N_{\text{grid}} = 513$ is fixed for all the systems. For the 3D implementation, we use the PySCF software package and the coupled cluster singles and doubles (CCSD) [75] method to generate our data for H_2 , which is equivalent to FCI (an exact solution within a basis set) in this case. When more than two electrons are present, we suggest using the singles and doubles coupled cluster method with perturbative triples (CCSD[T]), which is viewed as the “gold standard” in computational chemistry for weakly correlated systems [75]. In the presence of strong correlation, the aforementioned CC methods often fail to converge to the correct solution [76]. Consequently, we employ FCI in challenging cases, such as the square H_4 system, for which it remains feasible given the small system sizes considered in this work. A Gaussian-type orbital basis set, 6-31G, is used in all calculations. For an integration grid, we use SG-3 with a Becke scheme to combine multiatomic grids, such that for a given molecule at all geometries we obtain the same number of grid points [68]. The grid density is set to minimum to reduce the number of neural XC functional inputs, which simplifies the complexity of the problem. This entails $N_{\text{grid}} = 1192$ for H_2 and $N_{\text{grid}} = 2384$ for planar H_4 in the 3D setting. Given that we train models that allow only for a fixed number of inputs (N_{grid}), grids have to be adapted (e.g., by pruning) to have the same number of grid points for all molecules in the dataset [68]. However, this is only a limitation imposed by simple prototypical (quantum) models that are investigated in this work. See Appendix B for additional clarifications about integration grids.

To train our models, we use the loss function given by

$$\mathcal{L}(\theta) = \underbrace{\frac{1}{N_{\text{mol}}} \sum_{i=1}^{N_{\text{mol}}} \left[\sum_{j=1}^{N_{\text{grid}}} w_j (n_{\text{KS},j}^i(\theta) - n_{\text{ref},j}^i)^2 / N_e^i \right]}_{\text{density loss } \mathcal{L}_n} + \underbrace{\frac{1}{N_{\text{mol}}} \sum_{i=1}^{N_{\text{mol}}} \left[\sum_{k=1}^{N_{\text{KS}}} \eta_k (E_k^i(\theta) - E_{\text{ref}}^i)^2 / N_e^i \right]}_{\text{energy loss } \mathcal{L}_E}, \quad (23)$$

where the expectation is taken on N_{mol} molecules of the training set, η_k weights the contribution of every energy at each KS iteration, and w_j represents the integration grid weight. We use the weighting scheme introduced by Li *et al.* [25]: $\eta_k = 0.9^{N_{\text{KS}}-k} H(k-10)$, where H is the Heaviside function to tune the contribution of every loss term. This composite loss function includes two primary components: the density loss $\mathcal{L}_n$ and the energy loss $\mathcal{L}_E$. These components are essential for guiding the neural network to accurately produce electron densities and total energies, respectively. Specifically, the density loss assesses the accuracy of the Kohn-Sham electron density produced by the DFT pipeline compared to the reference electron density. Meanwhile, energy loss measures the precision of the Kohn-Sham energies generated by the neural network against the reference energies at each iteration k . The energy loss is weighted by η_k to emphasize the importance of later iterations.

All derivatives of the loss function are computed via JAX automatic differentiation in this work. On quantum hardware, exact differentiation schemes such as parameter-shift rule [39] or its generalized version [40] can be used (see Appendix D). The loss function is then minimized to obtain the optimal model parameters,

$$\theta_{\text{opt}} = \underset{\theta}{\text{argmin}} \mathcal{L}(\theta),$$

that define the optimal XC correlation functional. In the following scenario, we outline the use of a gradient-based method for training, although it should be noted that our approach is also compatible with gradient-free techniques, which would, however, prevent regularization as in Ref. [25].

At each converged SCF calculation, the gradient of loss $\mathcal{L}$ is performed via the automatic differentiation (or potentially parameter-shift rule on a quantum device). The gradient is then provided to the classical optimizer (Adam [77] or L-BFGS-B [78]). The learning rate for the Adam optimizer can be decreased from 10^{-2} to 10^{-4} as the training progresses to fine-tune the results. Otherwise, L-BFGS-B and Adam are used with the standard hyperparameters.

IV. RESULTS AND DISCUSSION

In this section, we present the outcomes of numerical experiments and comment on their performance in 1D and 3D settings. First, we evaluate all our models on H_2 using the 1D KS-DFT framework based on the implementation of Li *et al.* [25]. Next, we extend our approach to the 3D framework based on PySCFAD [31] and test the extrapolation capacity of our better models on the H_2 molecule. We assess the effect of the fractional orbital occupation on H_2 and evaluate our best

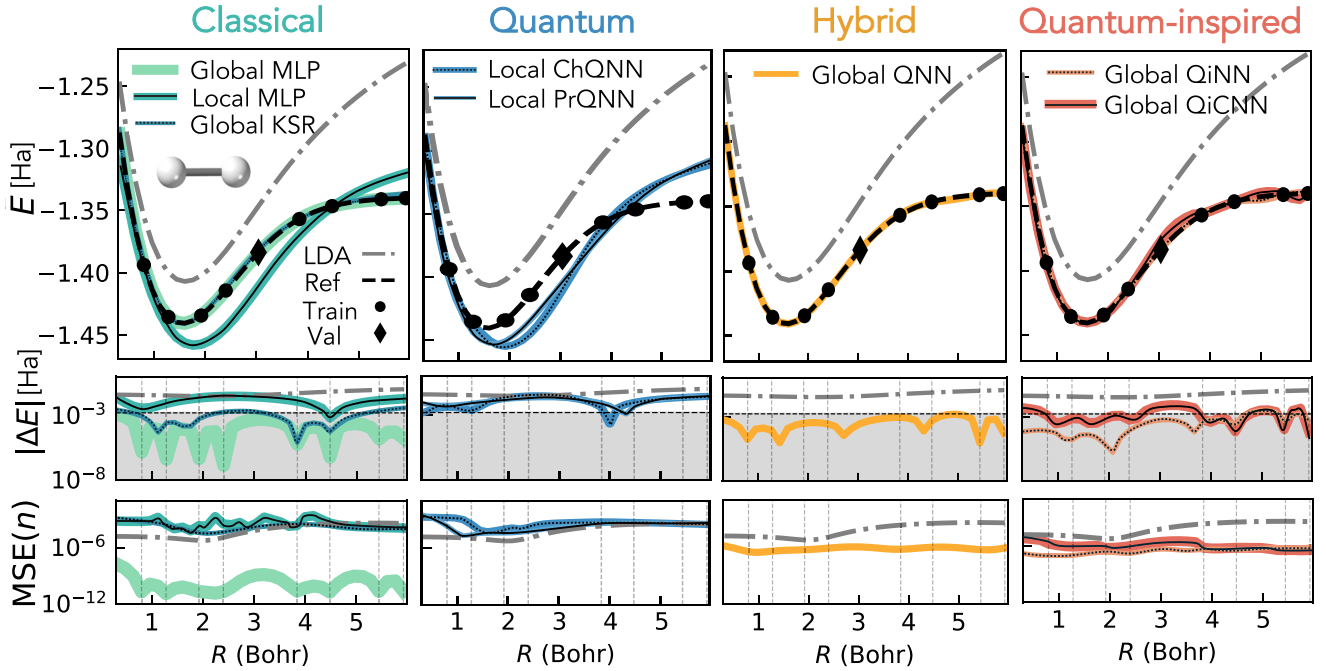


FIG. 4. Results of training neural XC functionals on different geometries of hydrogen molecule in 1D KS-DFT framework [25]. The gray area highlights results within chemical precision (≤ 1.6 mHa) and vertical dashed lines mark the training data. Local models improve upon the LDA shape but still produce a poor qualitative agreement with the reference. Best results are obtained for global models when the whole density is considered as model input. The XC models are defined in Sec. II C. All neural functionals are trained on eight training data points (circles) and the final model parameters are selected according to the lowest loss on the validation results. The reference (Ref) results are generated using DMRG by Li *et al.* [25].

model on planar H_4 . We evaluate the generalization capability on the H_2H_2 system, which was not part of the training dataset, in a 1D setting. Finally, we estimate the number of measurements required in a quantum hardware implementation. All calculations are performed on a classical computer using Qadence [79].

For all our results, we define the density error in terms of the mean-squared error (MSE) and include the local-density approximation (LDA) functional to contrast with our ML-based functionals. In addition, we analyze the results in terms of the nonparallelity error (NPE) [80], defined as the absolute difference between the largest and smallest energy errors ΔE over the whole energy profile. The average $|\Delta E|$ measures the overall closeness to the true dissociation profile, whereas the NPE measures nonparallelism after removing any constant shift. In reaction rate calculations, the absolute energy scale is irrelevant; only the relative shape of the energy profile matters, e.g., see Ref. [81]. For example, one could have average $|\Delta E| = 1$ Ha (constant shift) but $NPE < 1$ mHa, indicating that the correct chemical behavior (e.g., reaction rates) is preserved, enabling accurate predictions. Regarding the MSE on the density, in Ref. [82] it was shown that, in typical cases, density errors do not significantly affect the total energy, except in specific scenarios. Therefore, in practical applications, NPE is often the most important metric.

The hyperparameter search is performed for all proposed models and the best parameter combination (e.g., number of qubits N_q , layers N_l and depth N_d) is selected.

A. Local versus global functionals

In Fig. 4, we evaluate all our models within the 1D framework. This figure compares the prediction of 1D potential energy surfaces achieved by different models, characterizing the H-H bond dissociation in the H_2 molecule. We start with MLP, which is the prototypical expressive model that we use to evaluate the approaches of local and global embeddings. To that end, in all our MLP models, we used two hidden layers with 513 neurons, which matches the number of grid points, to achieve enough expressivity. The local MLP model has a single input and output neuron to infer the local XC energy based on corresponding density. Local MLP improves on the LDA DFT results in terms of overall energy error, but it is not sufficient to improve on the density. The restriction local information is limiting the capacity of the model to improve. Hence, we observe the best trade-off of energy-density error that this model is capable of.

In order to evaluate if local quantum models can improve on the local MLP, we propose the local ChQNN and PrQNN. These models employ the Chebyshev and product feature maps (see Appendix C), respectively. We use $N_q = 4$ qubits, $N_l = 1$ (no data reuploading) and $N_d = 8$ layers for our ansatz. The results in terms of NPE of the energy are very similar to those of the local MLP model. The energy and density profiles of local quantum and classical models are qualitatively similar, as expected due to locality of both models. It is therefore necessary to verify if including more information in the quantum models brings benefits in terms of lower energy error $|\Delta E|$ and MSE of the density profile.

For that purpose, we evaluate a global version of MLP with $N_{\text{grid}} = 513$ input neurons. This allows the model to consider the whole density to infer the XC energy via the global embedding, hence there is only one output neuron. This results in $\sim 2.6 \times 10^5$ parameters as reported in Table I. The corresponding energy profile is significantly improved (lower error with respect to the reference) in comparison to all other models. Similarly, the proposed KSR model [25] can be tailored to any locality. The global version of KSR model was shown to perform best and generalize perfectly in a single molecule even from only two data points (here 8). We used the same density initialization as for global MLP and the model struggled to generalize correctly, leading to MSE on density comparable to LDA, despite rendering chemically accurate energies. However, KSR is significantly more parameter efficient, having only $\sim 10^3$ parameters versus $\sim 2.6 \times 10^5$ for global MLP, which are generally known to overfit the training data. In addition, given good initialization, global KSR is capable of learning the dissociation profile of H_2 perfectly only from two data points, as shown in Ref. [25]. For all the remaining models with global embedding, the results perform as well or even better than local LDA-like functionals, at least for the total energies. Moreover, the global models have the lowest NPE by an order of magnitude in comparison to local models (see Table I). A significant reduction (three to five orders of magnitude) in average MSE of the electron density is achieved only by fully global models such as global QNN and global MLP.

Our results indicate that both classical and quantum models must incorporate global electron density features to achieve chemical precision.

B. Extrapolation within a single molecule

All our models provide smooth energy profiles due to regularization capabilities of the employed differentiable DFT implementation. In Fig. 4, it is clear that eight training data points, sampling low to strongly correlated regimes of the hydrogen molecule, are sufficient to generalize correctly within a single molecule, provided the model has enough expressivity. In general, all the models show characteristic dips in energy errors at the training data points. However, the density MSE profiles are rather flat along the whole dissociation curve. Similarly, in both Figs. 5 and 6, just three training points were sufficient to train models that provide qualitatively correct energy profiles, even in the presence of strong correlation. Further studies are necessary to investigate the extrapolation capabilities of quantum models on large datasets, i.e., examining whether models trained only in strongly correlated regimes can extrapolate correctly to low-correlation regimes and vice versa.

C. Hybrid and quantum-inspired quantum-classical models

Since local QNNs are not capable of correctly fitting even the training data, we explore architectures that allow us to embed L data points and hence approach nonlocal functionals. In particular, in the global QNN model, we consider the idea of embedding $L = 3$ data points in a single QNN with $N_q = 6$ qubits, and $N_l = 1$ (no data reuploading) and $N_d = 8$ layers,

TABLE I. Extrapolation results of neural XC functionals evaluated on H_2 within 1D KS-DFT framework (Fig. 4). Average energy and density errors, as well as NPE, are evaluated along the whole dissociation profile. The uncertainties were calculated using the standard deviation. For MLP-based models, we indicate the number of neurons per layer and the number of layers (e.g., 513/2). Abbreviations: No., number; Av., average.

		Global MLP	Local MLP	Global KSR	Local ChQNN	Local PrQNN	Global QNN	Global QiNN	Global QiCNN
$N_q/N_d/N_l$		513/2	513/1	—	4/8/1	4/8/1	6/8/1	9/10/1	6/8/1
No. parameters		2.64×10^5	1540	1568	96	108	316	297	596
H_2 Av. $ \Delta E $ [Ha]		$(2 \pm 2) \times 10^{-4}$	$(1.5 \pm 0.9) \times 10^{-2}$	$(1 \pm 1) \times 10^{-3}$	$(1.5 \pm 1) \times 10^{-2}$	$(1.4 \pm 0.8) \times 10^{-2}$	$(4 \pm 4) \times 10^{-4}$	$(8 \pm 8) \times 10^{-4}$	$(3 \pm 2) \times 10^{-4}$
NPE [Ha]		1.3×10^{-3}	2.9×10^{-2}	4×10^{-3}	3.1×10^{-2}	3.1×10^{-2}	1.4×10^{-3}	4×10^{-3}	7×10^{-4}
Av. MSE(n)		$(1.0 \pm 3.0) \times 10^{-9}$	$(5.6 \pm 6.0) \times 10^{-4}$	$(2.4 \pm 2.4) \times 10^{-4}$	$(6.0 \pm 5.3) \times 10^{-4}$	$(4.0 \pm 6.0) \times 10^{-4}$	$(7.2 \pm 2.4) \times 10^{-7}$	$(2.5 \pm 1.8) \times 10^{-3}$	$(2.1 \pm 1.0) \times 10^{-7}$

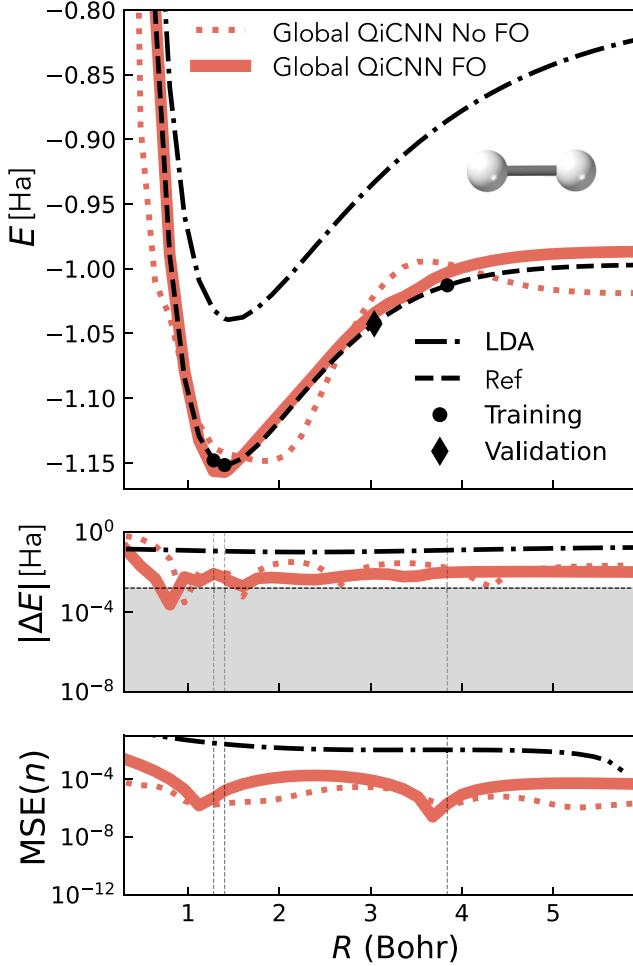


FIG. 5. Results of training neural XC functionals on different geometries of H_2 in our 3D RKS-DFT framework based on PySC-FAD. The same global QiCNN [Eq. (22)] is trained with and without fractional orbital occupation (FO). Despite the limited training data (three points), the model accurately extrapolates to distant geometries, demonstrating the strong regularization capacity of our 3D framework. The reference (Ref) results are generated using the FCI/6-31G level of theory.

and apply the same model onto N_{grid}/L different batches. The results are combined via a single-layer MLP with 171 input neurons, one bias parameter, and one output for the global embedding. This allows us to embed chunks of data into a limited quantum model with a restricted number of qubits. By combining such outputs, we obtain a fully global functional. The model has only ~ 300 trainable parameters, which is orders of magnitude smaller than for the global MLP, since both are composed of most general quantum/classical layers. Note that comparing it with the KSR model is not reasonable because the latter possesses numerous inductive biases such as the self-interaction gate and the softly imposed negativity of XC energy.

We also evaluated the amplitude embedding as a feature map, using a DQC with $N_q = 9$ qubits, $N_l = 1$, $N_d = 10$. We name this model global QiNN since, in general, it is hard to embed the classical data (electron density) exactly in a polynomial number of one-/two-qubit gates [83]. The number

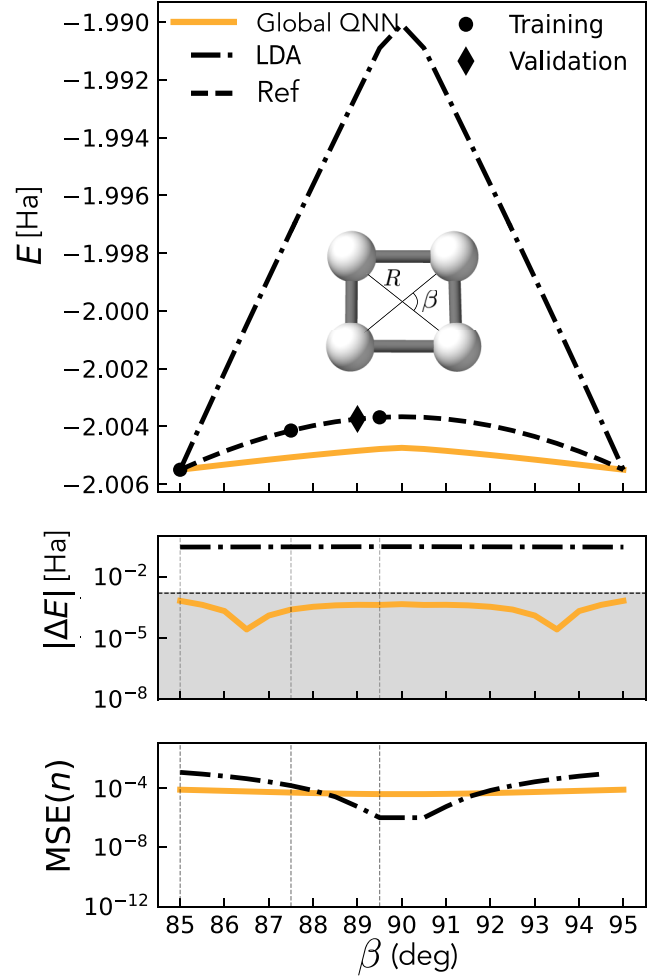


FIG. 6. Extrapolation of global QNN [Eq. (18)] on planar H_4 within the 3D framework with a fractional occupation approach [72]. Global QNN achieves chemical precision in energy, surpassing the LDA results by orders of magnitude. In the top panel, LDA and global QNN results are shifted to match the reference data at extremities to compare the shapes of the dissociation profiles. Shifts can be inferred from the $|\Delta E|$ panel. The reference (Ref) results are generated with the FCI/6-31G level of theory.

of qubits was chosen to fit all the elements of $\mathbf{n}$ exactly in a state vector. In Table I, the results for global QiNN are in good agreement with the reference: within chemical precision and overall density error that significantly improved upon LDA. Notably, the outcome is similar to global QNN, which is simple to implement on near-term devices. The global QiNN architecture, despite producing satisfactory results, does not produce a significant advantage over quantum-classical hybrid QNN. It can, however, serve for theoretical studies since only quantum layers are employed, which dissociates it completely from classical models.

Another approach is to chain quantum layers with angle embeddings, as implemented in the global QiCNN model. In that case, the DQC output becomes the input of the next DQC layer. We use three different DQCs that all employ $N_q = 6$ qubits and $N_l = 1$ and $N_d = 8$ ansatz layers. The third DQC layer feeds into an MLP layer with 19 neurons (remaining

TABLE II. Extrapolation results for global QNN evaluated on H_2 within 3D KS-DFT framework, both with and without fractional orbital occupation (FO) (Fig. 5).

	Global QiCNN No FO	Global QiCNN FO
$N_q/N_d/N_l$	4/14/1	4/14/1
No. parameters	653	653
H_2 Av. $ \Delta E $ [Ha]	$(5 \pm 13) \times 10^{-2}$	$(1 \pm 2) \times 10^{-2}$
NPE [Ha]	6.5×10^{-1}	1.3×10^{-1}
Av. MSE(n)	$(1.1 \pm 1.3) \times 10^{-5}$	$(1.5 \pm 3.7) \times 10^{-4}$

outputs of DQC at the third layer) to combine the results into the total XC energy. The results for the energy are within chemical precision. Concerning the density, the errors are significantly lower, at least by an order of magnitude, than LDA.

Overall, among the hybrid quantum models presented, the simplest, global QNN with a single quantum and classical layer, yields the best performance.

D. Extension from 1D to 3D KS-DFT framework

To demonstrate the ability to efficiently train quantum XC models in our 3D framework, we evaluate our most complex neural functional, global QiCNN, in such a setting. Similarly to the 1D version, the model consists of three chained quantum layers, each with $N_q = 4$ qubits and $N_l = 1$ and $N_d = 14$ ansatz layers. The final quantum layer connects to MLP with 149 neurons to output the total XC energy. As depicted in Fig. 5, global QiCNN suffers from energy oscillations and is unable to extrapolate correctly. This is the ideal situation to demonstrate the ability of FO methods to alleviate such problems, which are apparent in 3D due to increased complexity. To that end, we show that even with three data points (of which two are almost at the same bond distance) the complex model can be regularized in order to fit better the dissociation curve, e.g., lower NPE (see Table II). Despite not being within chemical precision along the whole energy profile, the density error is improved by orders of magnitude in comparison to the LDA result. This is achieved with a FO algorithm [72] that is discussed in Appendix A. The FO approach relaxes the constraint of the integer orbital occupation in PySCF's restricted KS-DFT implementation and allows us to approximate the true fractional ones. In addition, the neural XC functional is optimized through training, thereby compensating for errors arising from the approximation of orbital occupations. As reported in Table II, the use of the FO approach results in an approximately fivefold reduction in both the average energy error and its variance, without any extra model parameters. Note that the FO method is not limited to use only with quantum models; it can also be applied more broadly for training classical neural XC functionals.

As the ultimate test of our 3D framework, demonstrating the capability to accurately learn from challenging molecular systems, we trained our best hybrid architecture, the global QNN, on planar H_4 . This is a typical small system exhibiting strong correlation, commonly used to benchmark *ab initio* methods [76,84]. Here global QNN consists of one quantum layer with $N_q = 4$ qubits, $N_l = 1$, $N_d = 14$, and

TABLE III. Extrapolation results for global QNN evaluated on H_4 within the 3D KS-DFT framework (Fig. 6).

	Global QNN
$N_q/N_d/N_l$	4/14/1
No. parameters	764
H_4 Av. $ \Delta E $ [Ha]	$(3 \pm 2) \times 10^{-4}$
NPE [Ha]	6×10^{-4}
Av. MSE(n)	$(5.3 \pm 1.2) \times 10^{-5}$

followed by MLP with 596 neurons that outputs E_{XC} . In Fig. 6, we note that the LDA approximation greatly overestimates the energy barrier at $\beta = 90^\circ$. This quantum-classical hybrid model, given its limited expressivity, provides the best achievable fit of the training data and extrapolates smoothly to various geometries. In Table III, we note that the model achieves chemical precision over the whole energy profile (NPE $\approx 10^{-4}$). Concerning density errors, they are of the order of 10^{-5} , similarly to the H_2 density errors (Table II) that are obtained within the same 3D framework. Hence, we demonstrated that provided accurate reference data, it is possible to train quantum-enhanced functionals to learn from strongly correlated systems. In the future, it would be interesting to evaluate if such models can generalize correctly from small to other larger strongly correlated systems such as stretched N_2 .

E. Zero-shot generalization of XC functionals

In Fig. 4, we demonstrated that quantum-enhanced XC models can learn to extrapolate correctly to different geometries of a single molecule by consistently achieving chemical precision within the 1D framework. To evaluate their capacity to generalize to different molecular systems, we select our best-performing quantum-enhanced functional and test it on a molecular system that was not included in the training dataset.

Our results highlight global QNN as the simplest hybrid model, comprising a single quantum and classical layer, capable of achieving chemical precision for H_2 . For its training, the dataset only contained the hydrogen molecule at different geometries. Such a dataset does not allow the model to implicitly learn the XC energy dependency on the variable number of electrons N_e in the input density (i.e., $N_e = \sum_{i=1}^{N_{\text{grid}}} w_i n_i$) when evaluating on a different system such as H_2H_2 (separate hydrogen molecules). Hence, we enrich our train dataset with molecules such as linear H_4 at different bond distances, as done in Ref. [25]. We train global QNN using the following bond distances: For H_2 , $R \in (0.48, 1.28, 3.04, 3.84)$ bohr and for H_4 , $R \in (1.28, 2.08, 3.36, 4.48)$ bohr.

In Fig. 7, we note that chemical precision is achieved only for small values of $R < 2$ bohr and momentarily at $R = 5.3$ bohr, with the overall energy closely aligned with the reference, resulting in lower energy error compared to LDA. MSE on the density consistently remains lower than the LDA result. The XC model yields suboptimal qualitative predictions, evidenced by the energy minimum at approximately 3.5 bohr (stable configuration). Considering the utilization of only basic classical and quantum layers and a few data points, these results demonstrate the capacity of our models to

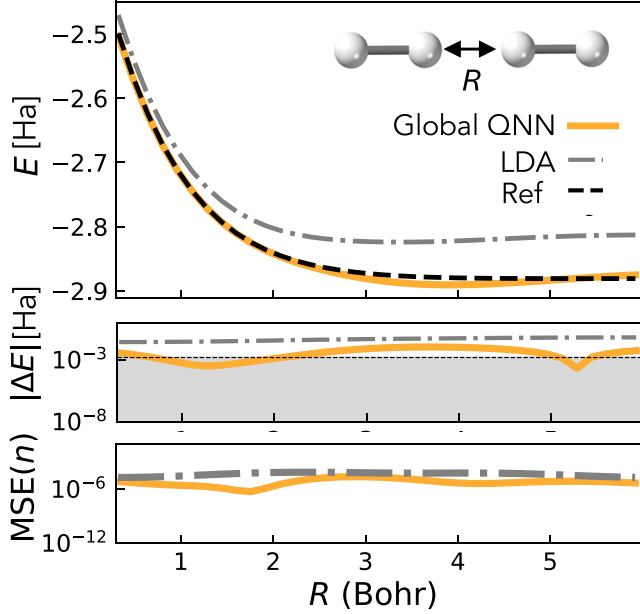


FIG. 7. Generalization of global QNN (Eq. (18)) to unseen H_2H_2 system trained on H_2 and H_4 molecules within the 1D framework. The model achieves chemical precision over some regions of the dissociation profile without being explicitly trained on it. The reference (Ref) results are generated using DMRG by Li *et al.* [25].

generalize. In fact, as stated in Table IV, our model has only 316 trainable parameters compared to the state-of-the-art global KSR model, which has ≈ 6 times more. Hence, addressing the shortcomings would require improving the architecture of our model (e.g., using molecular symmetries) and expanding the training dataset, which is beyond the scope of this study.

To summarize, we visualize all our results for the energy error of quantum (-enhanced) models in Fig. 8. Overall, the use of quantum layers can reduce the number of trainable parameters necessary to achieve a given precision in energy.

F. Performance evaluation of QNNs and MLPs in an identical backbone architecture

In the previous sections, we examined several QNN architectures. Among them, the hybrid GQNN demonstrated the best performance for the single-molecule case (Fig. 4) and exhibited promising zero-shot generalization capabilities on the H_2H_2 system (Fig. 7). However, the extent to which

TABLE IV. Zero-shot generalization results for global QNN evaluated on H_2H_2 , trained on H_2 and H_4 within the 1D KS-DFT framework (Fig. 7).

Global QNN		
	$N_q/N_d/N_l$	6/8/1
	No. parameters	316
H_2H_2	Av. $ \Delta E $ [Ha]	$(4.9 \pm 1.4) \times 10^{-2}$
	NPE [Ha]	3.9×10^{-2}
	Av. $\text{MSE}(n)$	$(5 \pm 2) \times 10^{-5}$

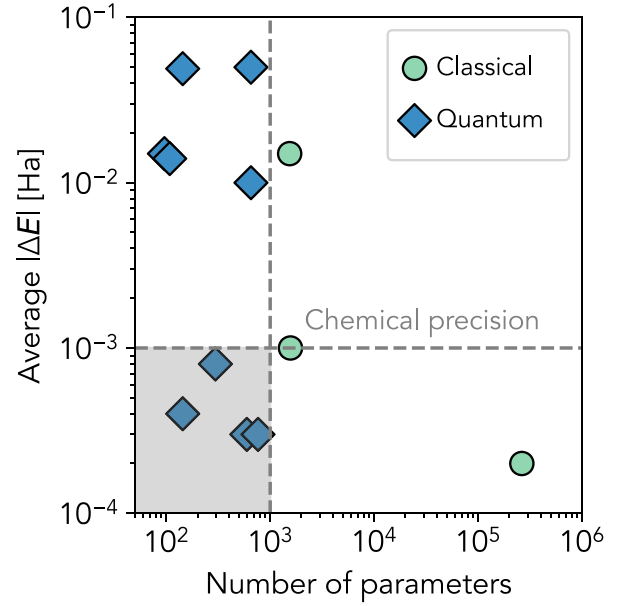


FIG. 8. Dependence of the average energy error on the size of classical and quantum (hybrid) models. Quantum models are consistently more compact ($\leq 10^3$ parameters) than classical models (e.g., GKSR with 1568 parameters) given a fixed energy error, while being still capable of achieving chemical precision.

QNNs contribute performance gains over analogous classical architectures within the same backbone remains unclear.

To address this, we substitute the quantum layers in the GQNN architecture with standard MLPs, naming that model convolutional MLP (CMLP), shown in Fig. 9. For both classical and quantum models, we preserve the overall model structure and maintain a similar parameter count to make the best possible comparison. In particular, the GQNN (CMLP)

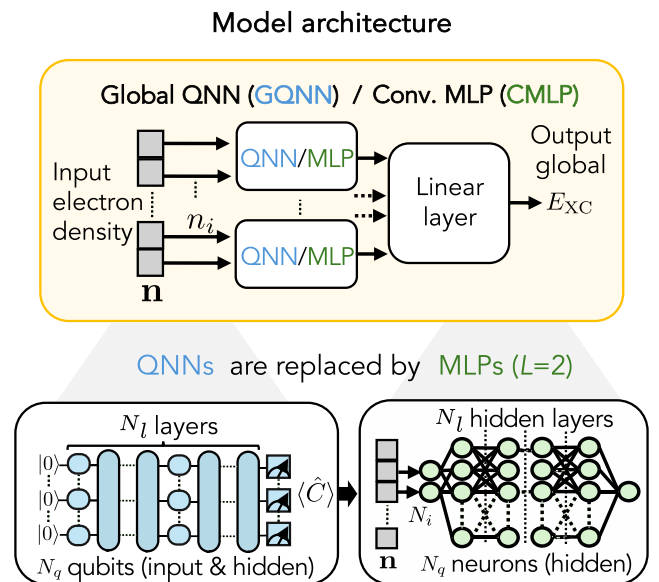


FIG. 9. Architecture of the GQNN and CMLP models. To assess the impact of quantum layers relative to classical ones, we replace the QNNs in the GQNN backbone with MLPs (CMLP).

TABLE V. Interpolation: Comparison of GQNN and CMLP (Fig. 9) in fitting the dissociation profile of the hydrogen molecule. The results are obtained within the 1D framework using the same set of molecular geometries as in Fig. 4. The best result for each metric is highlighted in bold. Abbreviations: Act., activation function; Par., number of parameters.

Model	$N_i N_q N_l $ Act.	Par.	ΔE [Ha]	Av. MSE(n)	NPE [Ha]
GQNN	3/6/8	316	$(4.0 \pm 4.0) \times 10^{-4}$	$(7.2 \pm 2.4) \times 10^{-7}$	1.40×10^{-3}
CMLP	2/2/1/tanh	521	$(1.7 \pm 1.1) \times 10^{-3}$	$(1.3 \pm 1.0) \times 10^{-5}$	3.38×10^{-3}
CMLP	2/2/1/soft	521	$(5.9 \pm 6.0) \times 10^{-3}$	$(9.1 \pm 4.6) \times 10^{-6}$	2.48×10^{-2}
CMLP	19/19/1/tanh	428	$(1.6 \pm 1.9) \times 10^{-3}$	$(5.4 \pm 1.0) \times 10^{-6}$	5.57×10^{-3}
CMLP	19/19/1/soft	428	$(7.4 \pm 34.8) \times 10^{-2}$	$(2.0 \pm 1.4) \times 10^{-4}$	2.01×10^0
CMLP	3/6/4/tanh	329	$(1.6 \pm 1.2) \times 10^{-3}$	$(5.5 \pm 2.4) \times 10^{-7}$	3.49×10^{-3}
CMLP	3/6/4/soft	329	$(9.5 \pm 11.5) \times 10^{-4}$	$(1.4 \pm 0.8) \times 10^{-7}$	4.79×10^{-3}
CMLP	4/4/1/tanh	193	$(7.5 \pm 6.5) \times 10^{-3}$	$(8.2 \pm 6.4) \times 10^{-5}$	2.14×10^{-2}
CMLP	4/4/1/soft	193	$(1.6 \pm 1.7) \times 10^{-3}$	$(7.8 \pm 1.8) \times 10^{-6}$	7.96×10^{-3}

model is composed of a QNN (MLP) module that accepts N_q (N_i) inputs, applied on N_{grid}/N_q (N_{grid}/N_i) times on the N_{grid} input density. Those outputs are combined via a linear layer with a bias, producing the total XC energy. The quantum model has N_q qubits in both the input layer (Product feature map, see Sec. II C 3) and each of the N_l hidden layers. MLP uses N_i neurons in the input layer with N_q neurons in the hidden layers.

We compare GQNN and CMLP for the interpolation and the generalization tasks. The results are assembled in Table V for the interpolation of energies across various bond distances of the hydrogen molecule, and in Table VI for generalization to a different molecular system, H_2H_2 , after training on H_2 and H_4 . We use the same datasets and hyperparameters as in previous experiments shown in Figs. 4 and 7, respectively.

The performance of CMLP-based XC functional strongly depends on the choice of activation functions. We employ the hyperbolic tangent (tanh) and softplus (soft) activation functions to ensure infinite differentiability—necessary for computing molecular properties—and to provide smooth output suitable for correct XC potential representation [28]. To control the expressivity of the XC functional, we tune the number of inputs per block—denoted N_q for GQNN and N_i for CMLP—as well as the number of layers N_l within each QNN or MLP module (see Fig. 9).

In Tables V and VI, we observe that the number of trainable parameters does not directly correlate with improved

model performance in terms of energy prediction ΔE , average density error MSE(n), or NPE. For example, the CMLP configuration (2/2/1/tanh), which has 521 parameters—nearly twice as many as the GQNN (316 parameters)—does not outperform the GQNN in either interpolation or generalization tasks. One possible explanation is the limited locality L of the MLP blocks in the CMLP model (see Fig. 9), analogous to small filter sizes in convolutional neural networks. The overall CMLP model remains global, as it ultimately processes the entire input density through repeated application of local blocks. In the configuration CMLP 2/2/1 (see Table V), the model processes only $L = 2$ inputs at a time, which may restrict its ability to capture nonlocal density correlations. Increasing the locality to $L = 19$ in a shallow configuration (CMLP 19/19/1) does not yield notable improvements in MSE on density or NPE. Interestingly, even with a modest locality of $L = 4$ (CMLP 4/4/1), the results remain of the same order of magnitude as for the $L = 19$ model. Increasing the number of layers to 4 (CMLP 3/6/4/soft) produces the best density accuracy among the CMLP variants. Nonetheless, as shown in Table V, the GQNN still achieves the lowest energy error ΔE and NPE overall. Most importantly, in the generalization setting (Table VI), GQNN outperforms all CMLP configurations across all metrics except for average energy error.

These results suggest that quantum-enhanced XC models can perform comparably to classical neural networks in the ideal noiseless setting, with marginal benefits on small

TABLE VI. Generalization: Comparison of GQNN and CMLP (Fig. 9) for zero-shot generalization. Results are evaluated within the 1D framework on H_2H_2 , trained on H_2 and H_4 , exactly as in Fig. 7. The best result for each metric is highlighted in bold. Abbreviations: Act., activation function; Par., number of parameters.

Model	$N_i N_q N_l $ Act.	Par.	ΔE [Ha]	Av. MSE(n)	NPE [Ha]
GQNN	3/6/8	316	$(4.9 \pm 1.4) \times 10^{-2}$	$(5 \pm 2) \times 10^{-5}$	3.90×10^{-2}
CMLP	2/2/1/tanh	521	$(3.8 \pm 2.8) \times 10^{-2}$	$(4.7 \pm 3.8) \times 10^{-4}$	8.00×10^{-2}
CMLP	2/2/1/soft	521	$(4.5 \pm 4.2) \times 10^{-2}$	$(8.0 \pm 4.3) \times 10^{-4}$	1.32×10^{-1}
CMLP	19/19/1/tanh	428	$(5.1 \pm 3.0) \times 10^{-2}$	$(1.7 \pm 1.1) \times 10^{-4}$	8.69×10^{-2}
CMLP	19/19/1/soft	428	$(5.7 \pm 3.5) \times 10^{-2}$	$(6.8 \pm 5.4) \times 10^{-4}$	1.05×10^{-1}
CMLP	3/6/4/tanh	329	$(3.3 \pm 2.5) \times 10^{-2}$	$(4.9 \pm 4.8) \times 10^{-4}$	7.19×10^{-2}
CMLP	3/6/4/soft	329	$(2.8 \pm 1.4) \times 10^{-2}$	$(1.0 \pm 0.5) \times 10^{-4}$	4.62×10^{-2}
CMLP	4/4/1/tanh	193	$(3.5 \pm 2.7) \times 10^{-2}$	$(5.2 \pm 5.0) \times 10^{-4}$	7.75×10^{-2}
CMLP	4/4/1/soft	193	$(3.5 \pm 1.9) \times 10^{-2}$	$(1.6 \pm 1.0) \times 10^{-4}$	5.78×10^{-2}

molecular systems, such as reduced NPE. To assess their performance under realistic conditions, we next examine the impact of gate noise and measurement (shot) noise on our results.

G. Error bounds on energy and density

Ensuring the stability of SCF cycles is essential for both successful training and reliable inference with neural XC functionals. For QNN-based functionals on quantum hardware, measurement noise and other hardware errors perturb the energy functional and thus the SCF convergence. Next, we develop a theoretical framework to characterize and analyze the propagation of such errors throughout the SCF procedure.

We denote by $n(\mathbf{r}) \in \mathcal{X}$ an electronic density in a suitable Banach space $\mathcal{X}$. We recall from Sec. II that the total energy is defined as $E[n] = T_S[n] + V[n] + E_H[n] + E_{XC}[n]$. Here $E_{XC}[n]$ is the (unknown) universal exchange-correlation functional. Let n^* be the ground-state density minimizing the total energy, and let $E^* = E[n^*]$ be its exact energy. In practice, the KS equations are solved by a fixed-point iteration. Following the convergence analysis in Refs. [85–87], we consider an SCF iteration as a map $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{X}$ such that

$$n^{(k+1)} = \mathcal{T}[n^{(k)}]. \quad (24)$$

The exact solution n^* satisfies $n^* = \mathcal{T}[n^*]$. Under typical conditions (e.g., sufficient spectral gap or use of level-shifting approach [88]), local convergence can be reached. Now suppose we replace E_{XC} by $\tilde{E}_{XC}$, obtaining an *approximate* total energy $\tilde{E}[n] = \tilde{T}_S[n] + V[n] + E_H[n] + \tilde{E}_{XC}[n]$. The iteration map becomes $\tilde{\mathcal{T}}$. We denote by $\tilde{n}$ its convergent fixed point: $\tilde{n} = \tilde{\mathcal{T}}[\tilde{n}]$, and $\tilde{E} = \tilde{E}[\tilde{n}]$ the final approximate ground-state energy. To guarantee convergence, we impose the following assumptions on the components of our framework and define $\|\cdot\| \equiv \|\cdot\|_2$.

Assumption 1 (Local Contractivity). There is a neighborhood U of n^* in $\mathcal{X}$, and a constant $0 < \kappa < 1$, such that $\|\mathcal{T}[n_1] - \mathcal{T}[n_2]\| \leq \kappa \|n_1 - n_2\|$ for all $n_1, n_2 \in U$. Moreover, $\mathcal{T}[U] \subset U$.

This ensures that n^* is the unique fixed point in U and that $\{n^{(k)}\}$ remains in U under iteration. (See also Refs. [85,89] for the role of contractivity in SCF).

Assumption 2 (Uniform XC Error). Fix $\varepsilon > 0$ and $K > 0$ such that, for all $n \in U$,

$$|\tilde{E}_{XC}[n] - E_{XC}[n]| \leq \varepsilon, \quad (25)$$

$$\left\| \frac{\delta \tilde{E}_{XC}}{\delta n}[n] - \frac{\delta E_{XC}}{\delta n}[n] \right\| \leq K \varepsilon. \quad (26)$$

That is, the approximate functional and its functional derivative (XC potential) differ from the exact ones by at most ε and $K\varepsilon$, respectively, *uniformly* over U .

Assumption 3 (Approximate Iteration in U). The modified map $\tilde{\mathcal{T}}$ also keeps iterates in U , i.e., $\tilde{\mathcal{T}}[U] \subset U$, and has a unique fixed point $\tilde{n} \in U$.

Under these assumptions, we want to show the final SCF solution $\tilde{n}$ remains $O(\varepsilon)$ close to n^* , and the energies differ by $O(\varepsilon)$. We propose the following theorem:

Theorem 1 (Iterative SCF Stability). Let $\{n^{(k)}\}$ be the SCF sequence defined by $n^{(k+1)} = \tilde{\mathcal{T}}[n^{(k)}]$, for $k = 0, 1, 2, \dots$,

with initial guess $n^{(0)} \in U$. Assume 1–3. Then for each integer $k \geq 0$,

$$\|n^{(k)} - n^*\| \leq \kappa^k \|n^{(0)} - n^*\| + \frac{\alpha \varepsilon}{1 - \kappa} (1 - \kappa^k), \quad (27)$$

for some constant $\alpha > 0$. Consequently, as $k \rightarrow \infty$, the limit $\tilde{n} = \lim_{k \rightarrow \infty} n^{(k)}$ satisfies

$$\|\tilde{n} - n^*\| \leq C\varepsilon, \quad (28)$$

with a constant $C = \frac{\alpha}{1 - \kappa}$. Furthermore, the total energy difference $|\tilde{E} - E^*|$ is also $O(\varepsilon)$:

$$|\tilde{E} - E^*| \leq C' \varepsilon,$$

where $C' = 1 + \frac{M\alpha}{1 - \kappa}$ for some constant $M > 0$.

The proof, provided in Appendix G, applies the contraction mapping argument in combination with uniform bounds on the functional derivatives. The constant κ depends on the specific SCF implementation (e.g., QEX [59]), α reflects the quality of the XC potential modeled by the neural XC functional, and M quantifies the proximity of the final converged density in our implementation to the true electron density within the chosen basis set. Under the stated assumptions, we ensure that the SCF error is bounded at every iteration, which in turn guarantees stability of the entire SCF loop under noisy conditions.

In practice, these theoretical assumptions cannot always be met, for instance in cases of vanishing gap situation, often arise and need use of damping schemes [87]. If κ is close to 1, then the factor $(1 - \kappa)^{-1}$ in C becomes large, so a small uniform error in neural XC functional might still yield a non-trivial final difference. This is consistent with the well-known difficulty of SCF in small-gap or gapless systems [85]. Often, advanced mixing or level-shifting schemes can be used to artificially increase κ -contractivity [86]. Hence, our theoretical assumptions can be met through the following mechanisms:

(1) Algorithmic stabilization. To help guarantee Assumption 1, if the iteration is not strictly contractive, then convergence can be promoted by standard acceleration schemes, such as density mixing (DIIS [90] and EDIIS [91]), preconditioning [92], and level shifting [93], among others.

(2) Regularity and bounded error of the functional. To help guarantee Assumption 2, when the XC functional is represented by a classical MLP or by a QNN, uniform bounds can be established for the total energy. In the QNN case, the Fourier representation of variational circuits [44] ensures smoothness and local Lipschitz continuity of the learned functional, thereby controlling error propagation in the SCF cycle; see Appendix H, Lemma H1, for the proof.

(3) Physically motivated initialization. To help guarantee Assumption 3, the SCF cycle can be initialized with a non-interacting electron density at $n^{(0)}$ that yields a physically reasonable starting point, as implemented for example in PySCF [71]. Such initialization increases the likelihood that the Kohn-Sham map $\mathcal{T}$ begins in a locally contractive regime, although this is not universally guaranteed [87]. In addition, during the training of neural XC functional, the learning rate can be tuned to prevent the model from driving the SCF procedure outside the contractive region.

These arguments are supported by concrete realizations in Refs. [25,28], where classical neural XC functionals were

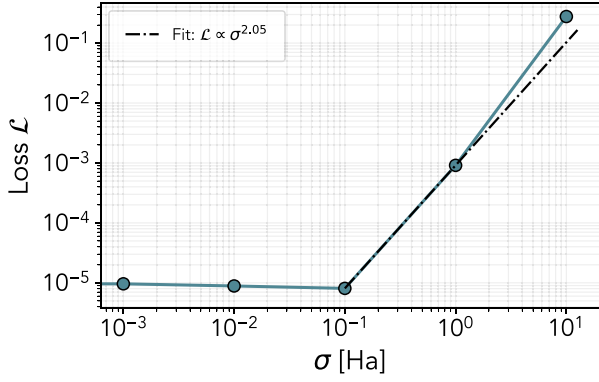


FIG. 10. Dependence of the converged loss [see Eq. (23)] on the noise level σ when training the global KSR model. The noise is sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ with standard deviation σ denoting the noise level. The results are obtained within the 1D framework using the same set of molecular geometries and hyperparameters as in Fig. 4. The fit is performed using the last three points ($\sigma \in \{0.1, 1, 10\}$ Ha), and all data points correspond to the mean converged loss over four different random initializations of the model parameters.

successfully trained and, more recently, by the neural XC functional Skala [94], which competes with standard functionals. In addition, Refs. [95,96] provide rigorous studies of stochastic SCF and its convergence, which, although not directly focused on our machine-learning task, are complementary to this work and motivate further investigation of the convergence behavior with respect to ε when scaling to larger systems.

The validation of the theory lies in its verification within a concrete implementation. To this end, we quantify the effect of typical noise sources on the training capacity of our models. For this purpose, we select the most sophisticated classical model used in this work (global KSR) without modifying its hyperparameters from Ref. [25]. Noise is modeled by sampling from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where the standard deviation σ specifies the noise level and mimics typical shot noise in QNNs. The sampled noise is added to the output of the global KSR model. In this setting, Fig. 10 reports the actual errors (loss) produced by our KS-DFT implementation using global KSR. This experiment isolates the effect of noise and directly tests whether the expected linear scaling is recovered in a full SCF implementation. We observe that our KS-DFT framework is robust up to significant levels of Gaussian noise ($\sigma = 0.1$ Ha) in the KSR model and, for $\sigma > 0.1$ Ha, the (converged) energy and density squared error [see Eq. (23)] scale with $\mathcal{O}(\sigma^{2.05})$ as expected. Hence, the total error ε scales approximately linearly with σ , consistent with Theorem 1. The slight deviation from exact quadratic scaling in the fit can be attributed to the instability in training at very large noise levels ($\sigma = 10$ Ha). This shows that the global KSR model, is expressive enough to compensate for the added noise and still converge to a similar loss in magnitude up to $\sigma = 0.1$ Ha.

For an XC functional trained in a noisy setting, an important question is whether it can generalize to unseen data and maintain a similar error. We test this hypothesis in Table

TABLE VII. Performance of global KSR within our differentiable 1D KS-DFT framework for fitting the dissociation profile of the hydrogen molecule, with additive noise drawn from a normal distribution $\mathcal{N}(0, \sigma^2)$. The results are obtained using the same hyperparameters and molecular geometries as in Table I.

σ [Ha]	ΔE [Ha]	Av. MSE(n)	NPE [Ha]
0	$(1.0 \pm 1.0) \times 10^{-3}$	$(2.4 \pm 2.4) \times 10^{-6}$	4.0×10^{-3}
10^{-3}	$(1.4 \pm 1.8) \times 10^{-3}$	$(7.5 \pm 5.5) \times 10^{-6}$	2.60×10^{-3}
10^{-2}	$(2.8 \pm 0.7) \times 10^{-3}$	$(7.5 \pm 9.7) \times 10^{-6}$	3.85×10^{-3}
10^{-1}	$(1.5 \pm 1.0) \times 10^{-2}$	$(2.7 \pm 2.1) \times 10^{-5}$	3.17×10^{-2}
10^0	$(5.2 \pm 3.1) \times 10^{-2}$	$(8.6 \pm 6.0) \times 10^{-4}$	1.15×10^{-1}
10^1	$(4.9 \pm 0.2) \times 10^0$	$(6.3 \pm 4.0) \times 10^{-3}$	5.32×10^0

VII, where the model interpolates the dissociation curve of hydrogen and is evaluated on unseen data points. All reported metrics remain stable for $\sigma < 10^{-1}$, e.g., NPE is of order of 10^{-3} as in the noiseless $\sigma = 0$ case, indicating that the noise does not impair the model's generalization capacity. Instead, it appears to act as a regularizer, mitigating overfitting, as also observed for QNNs in Ref. [97].

To isolate the effect of quantum noise on the quality of the XC functional, we consider a purely quantum model, Product QNN, without any classical postprocessing. We choose a simple configuration with $N_q = 3$ qubits and $N_l = 3$ layers, using the local XC embedding from Eq. (4), identical to that employed in the global KSR model above and commonly used in KS-DFT implementations. Thus, the only change from the previous experiment in Table VII is the model architecture. We focus on a standard noise mechanism—the combination of amplitude damping and phase damping—described in detail in Appendix F. Shot noise and gate noise are studied separately: (i) For gate noise, we fix $N_{\text{shots}} = \infty$ and vary the noise probability $p \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, error rates achievable on current quantum computers based on superconducting qubits [98]. (ii) For shot noise, we set $p = 0$ and vary $N_{\text{shots}} \in \{10, 10^2, 10^4, 10^6\}$. Table VIII summarizes the results. For all gate-noise experiments, energy and density errors remain of the same order of magnitude regardless of p , as the baseline model already exhibits relatively large but constant errors of order 10^{-2} Ha. In contrast, reducing N_{shots}

TABLE VIII. Performance of local PrQNN in our differentiable 1D KS-DFT framework for fitting the dissociation profile of the hydrogen molecule, with gate (amplitude and phase damping channels) and measurement noise. The results are obtained using the same hyperparameters and molecular geometries as in Table I.

p	N_{shots}	ΔE [Ha]	Av. MSE(n)	NPE [Ha]
0	∞	$(1.6 \pm 1.6) \times 10^{-2}$	$(1.5 \pm 1.5) \times 10^{-4}$	5.35×10^{-2}
10^{-3}	∞	$(1.7 \pm 1.5) \times 10^{-2}$	$(2.3 \pm 2.4) \times 10^{-4}$	5.27×10^{-2}
10^{-2}	∞	$(1.6 \pm 1.6) \times 10^{-2}$	$(1.5 \pm 1.5) \times 10^{-4}$	5.35×10^{-2}
10^{-1}	∞	$(1.7 \pm 1.5) \times 10^{-2}$	$(1.5 \pm 1.4) \times 10^{-4}$	5.27×10^{-2}
0	10^6	$(1.6 \pm 1.6) \times 10^{-2}$	$(1.5 \pm 1.5) \times 10^{-4}$	5.36×10^{-2}
0	10^4	$(1.6 \pm 1.6) \times 10^{-2}$	$(1.5 \pm 1.5) \times 10^{-4}$	5.42×10^{-2}
0	10^2	$(5.5 \pm 11.9) \times 10^{-2}$	$(1.6 \pm 4.1) \times 10^{-3}$	3.92×10^{-1}
0	10^1	$(1.3 \pm 1.7) \times 10^{-1}$	$(4.0 \pm 6.3) \times 10^{-3}$	5.02×10^{-1}

below 10^4 significantly degrades generalization performance, with errors increasing by roughly an order of magnitude. As shown in Appendix E, the typical statistical error scales as $\varepsilon = \mathcal{O}(N_{\text{shots}}^{-1/2})$, which is clearly observed for $N_{\text{shots}} = 10^2$, yielding $\text{NPE} \approx 10^{-1}$. This demonstrates that within the differentiable KS-DFT framework, at least $N_{\text{shots}} > 10^4$ are required to match the noiseless performance in case of local PrQNN, and achieving chemical precision of 1 mHa would likely require $N_{\text{shots}} \approx 10^6$ with other, more expressive, models. Next, we discuss in detail the resource estimates for hardware implementation based on these estimates.

H. Resource estimations for quantum hardware implementation

A robust evaluation of a competitive quantum model ideally involves assessing its time-to-solution, generalization capacity, parameter efficiency, and expressive power. In this context, the resources required for evaluating quantum XC models align with those conventionally utilized for DQC-based QNNs [36]. We focus on the time-to-solution, which is contingent on the total number of measurements (shots) required for an accurate output evaluation of quantum models. To that end, we examine the number of shots $N_{\text{shots}}^{\text{tot}}$ needed for the gradients $\nabla_{\theta} E_{\text{XC}}$, and $v_{\text{XC}}[n](\mathbf{r}) = \delta E_{\text{XC}}[n]/\delta n(\mathbf{r})$ and energy E_{XC} evaluations during the training phase.

Assumptions. To reasonably assess the resource requirements, it is important to note that, to achieve a desired precision ε on the total energy, the required number of epochs N_{epochs} and the number of model parameters N_{param} are likely to depend on the problem size, i.e., the number of atoms N_{atoms} . Larger and more diverse datasets may require a more expressive XC model—thus a greater number of parameters—as well as adjustments to the batch size and, consequently, an increase in the number of epochs. In such situations, with large N_{param} (i.e., large depth), the barren plateau problem may arise, necessitating the use of barren-plateau-resistant architectures [99] and appropriate mitigation techniques [100]. For the sake of providing an approximate estimate, we assume that N_{epochs} and N_{param} do not depend on the problem size. The exact scaling would depend on the chosen model and other hyperparameters of the framework. Since neural XC functionals in our approach can be designed to be independent of the problem size, they do not need to change as N_{atoms} increases, although they may no longer achieve accuracy ε without architectural adjustment. Thus, our estimates do not guarantee achieving accuracy ε but instead provide an approximate minimal shot budget under the aforementioned assumptions.

Energy evaluation. For a single evaluation of the XC functional, the cost operator $\hat{C} = \sum_{i=1}^{N_q} \hat{Z}_i$, in the loss function is diagonal by design. It allows for the simultaneous measurement of all its terms (due to mutual commutation), thereby mitigating associated errors on noisy quantum processors. This drastically reduces the number of shots in comparison to common *ab initio* variational approaches [101,102] that require the cost operator to be the electronic Hamiltonian of given system, which typically has many noncommuting terms to be measured [i.e., $\mathcal{O}(N^4)$ terms, where N is the system size]. In Appendix E, we establish that $N_{\text{shots}} = \mathcal{O}(\varepsilon^{-2})$ measurements per expectation value are required to achieve an

accuracy ε on the XC energy E_{XC} when using total magnetization cost operator.

Gradient evaluation. To evaluate $\nabla_{\theta} E_{\text{XC}}$ or v_{XC} by using the parameter-shift rule [39], two expectation values need to be computed at each epoch during the training (see Appendix D). In practical terms, considering that we have N_{epochs} during training, we require N_{shots} of order $\mathcal{O}(\varepsilon^{-2})$ per expectation value to achieve the desired accuracy ε . Consequently, the total number of shots during training performed for $\nabla_{\theta} E_{\text{XC}}$, taking into account N_{param} training parameters, amounts to $2N_{\text{shots}}N_{\text{param}}N_{\text{epochs}}$. Similarly for v_{XC} , the gradient $\delta E_{\text{XC}}[n]/\delta n(\mathbf{r})$ is evaluated at N_{grid} points at N_{KS} iterations, resulting in $2N_{\text{shots}}N_{\text{KS}}N_{\text{grid}}N_{\text{epochs}}$ shots.

Global embedding. In addition, E_{XC} needs to be inferred at every KS iteration. In our experience, the number of KS iterations can be set at $N_{\text{KS}} \in [15 - 100]$ depending on the size and complexity (strongly correlated regime) of the system. See Ref. [103], where large-scale KS-DFT calculations required up to 30 KS iterations. Therefore, the total cost in terms of measurements in our framework is given by

$$N_{\text{shots}}^{\text{tot},g} = (2N_{\text{param}} + 2N_{\text{grid}}N_{\text{KS}} + N_{\text{KS}})N_{\text{shots}}N_{\text{epochs}}.$$

For the scaling approximation, we consider the number of trainable parameters N_{param} and the number of grid points N_{grid} to be significantly larger than N_{KS} . In the following, we consider N_{KS} constant. Hence, the overall cost in terms of measurements is generally expected to scale as

$$N_{\text{shots}}^{\text{tot},g} \sim \mathcal{O}((N_{\text{param}} + N_{\text{grid}})N_{\text{shots}}N_{\text{epochs}}).$$

This scaling is valid for the global energy embedding [Eq. (5)], where a singular energy evaluation suffices ($E_{\text{XC}}[n] \approx F_{\theta}[n]$).

Local embedding. In a local embedding scenario, the XC energy E_{XC} is approximated through N_{grid} model evaluations, where the total error is the weighted sum of individual errors, leading to number of shots much greater than with the global energy embedding, $N_{\text{grid}}N_{\text{KS}}N_{\text{shots}}N_{\text{epochs}} \gg N_{\text{KS}}N_{\text{shots}}N_{\text{epochs}}$. The total number of shots required is

$$N_{\text{shots}}^{\text{tot},l} = (2N_{\text{param}} + 2N_{\text{grid}}N_{\text{KS}} + N_{\text{grid}}N_{\text{KS}})N_{\text{shots}}N_{\text{epochs}}.$$

Hence, the overall total measurement scaling for local embedding can be estimated as

$$N_{\text{shots}}^{\text{tot},l} \sim \mathcal{O}((N_{\text{param}} + N_{\text{grid}})N_{\text{shots}}N_{\text{epochs}}).$$

In addition, the grid size N_{grid} scales linearly with the number of atoms $N_{\text{grid}} \sim \mathcal{O}(N_{\text{atoms}})$ [104]. Thus, under our assumptions, the total cost for training with local or global energy embedding scales linearly with the system size in terms of shots due to the need to evaluate the local XC potential on the grid. The aforementioned costs are summarized in Table IX.

Concrete estimations. To estimate the shot budget required for an actual implementation on quantum hardware, we assume the typical values used in this work: $N_{\text{epochs}} = 10^3$, $N_{\text{KS}} = 10$, $N_{\text{grid}} = 10^3$, and $N_{\text{param}} = 10^2$ (e.g., global QNN). For the N_{shots} shots per energy evaluation, Sec. IV G shows, for a QNN-based functional, that deterioration from the noiseless result occurs as soon as $N_{\text{shots}} < 10^4$, consistent with the derivations in Appendix E. Based on these estimates, achieving $\varepsilon \leq 1$ mHa (chemical precision) is expected to

TABLE IX. Scaling of the number of measurements for the training with the proposed local [Eq. (4)] and global [Eq. (5)] energy embeddings in quantum-enhanced KS-DFT. We assume that N_{KS} remains constant [103], and define $N_{\text{param}}^{\text{grid}} = N_{\text{grid}} + N_{\text{param}}$.

	Local	Global
E_{XC}	$\mathcal{O}(N_{\text{grid}} N_{\text{shots}} N_{\text{epochs}})$	$\mathcal{O}(N_{\text{shots}} N_{\text{epochs}})$
$\nabla_{\theta} E_{\text{XC}}$	$\mathcal{O}(N_{\text{param}} N_{\text{shots}} N_{\text{epochs}})$	$\mathcal{O}(N_{\text{param}} N_{\text{shots}} N_{\text{epochs}})$
v_{XC}	$\mathcal{O}(N_{\text{grid}} N_{\text{shots}} N_{\text{epochs}})$	$\mathcal{O}(N_{\text{grid}} N_{\text{shots}} N_{\text{epochs}})$
Total	$\mathcal{O}(N_{\text{param}}^{\text{grid}} N_{\text{shots}} N_{\text{epochs}})$	$\mathcal{O}(N_{\text{param}}^{\text{grid}} N_{\text{shots}} N_{\text{epochs}})$

require at least $N_{\text{shots}} \approx 10^4$ – 10^6 , since the error scales as $\varepsilon = \mathcal{O}(N_{\text{shots}}^{-1/2})$. With the global embedding, the total number of shots is $N_{\text{shots}}^{\text{tot},g}$. Hence, per epoch, we require $\approx 10^8$ – 10^{10} shots, and, for the entire training, $N_{\text{shots}}^{\text{tot},g} \approx 10^{11}$ – 10^{13} shots.

The above assumes sequential measurements; however, only the epochs and KS iterations are inherently sequential. The gradients $\nabla_{\theta} E_{\text{XC}}$ and v_{XC} can be fully parallelized on quantum hardware (i.e., we set $N_{\text{param}} = 1$ and $N_{\text{grid}} = 1$), reducing the shot requirement to

$$N_{\text{shots}}^{\text{tot},g} \approx N_{\text{KS}} N_{\text{shots}} N_{\text{epochs}},$$

which yields 10^8 – 10^{10} shots for the entire training. On quantum hardware, the overall time required for training quantum models depends on the single-shot rate, highly dependent on the hardware type. This consideration ultimately determines the computational resources needed to attain a practical advantage (e.g., lower inference time) over classical XC models. Assuming an optimistic measurement time on a superconducting device of 1 ms per shot (e.g., ≈ 1 kHz shot rate reported in Ref. [105]) and 10^8 total shots, complete training could, in principle, be done in about one day. Additional measurement schemes, such as neural network estimators [106], could be employed to reduce variance, which would otherwise prevent achieving chemical precision.

Given the imperative to consistently achieve chemical precision for meaningful outcomes (e.g., accurate energy barriers), global embedding of the quantum XC functional [see Eq. (5)] emerges as the most promising near-term approach (i.e., especially in situations with small grid size, $N_{\text{grid}} \ll N_{\text{shots}}$ and $N_{\text{grid}} \ll N_{\text{epochs}}$). In our experiments, as our loss function is defined against energies, the fact that the global energy embedding does not scale with grid size significantly reduces the energy error of our models in comparison to the local embedding, leading to the best outcomes.

V. CONCLUSIONS AND OUTLOOK

In this work, we developed various (Q)NN-based architectures and evaluated their capabilities to represent XC functionals within KS-DFT. To that end, we extended existing differentiable KS-DFT codes based on JAX and PySCF frameworks, in both 1D and 3D settings, to accommodate quantum models. Next, we defined four classes of models capable of incorporating an arbitrary number of quantum and/or classical layers. As in other works, [25,28] we classified the models by their locality (i.e., the fraction of density

points considered as the input). To integrate these architectures into KS-DFT, we introduced a global embedding strategy, well suited for quantum architectures due to its minimal measurement requirements on quantum hardware. In terms of scalability, both local and global approaches are influenced by the system size due to the need to compute the local XC potential. However, a significant advantage of our model is that the global energy embedding does not scale with the grid size. This feature substantially improved the accuracy of our model, yielding the best results in our experiments. Finally, we performed numerical simulations of the entire training process on a classical computer.

The results of our experiments demonstrate the promising capabilities of QNNs in efficiently modeling the XC energy functionals. In our proof-of-concept evaluation, we showed that quantum models can achieve performance comparable to classical models with a similar number of parameters (Sec. IV F), offering small improvements in generalization. Moreover, the potentially greater expressivity of QNNs compared to classical models [37] (e.g., exponential number of Fourier frequencies [44] as discussed in Sec. II C 3) may not be strictly necessary for approximating the exact XC functional. Hence, the metrics typically adopted in analyzing quantum ML models (such as expressivity and entangling capability) may not be directly related to the ability of those models to perform well on a specific classical dataset. This observation motivates further investigations on larger datasets and more complex instances of quantum models, a nontrivial task given the inherent challenges of simulating quantum mechanics on classical computers.

Only fully global (quantum-enhanced) models, where the whole density is considered, produced the best results. In particular, global QNN provided the simplest, yet sufficiently expressive, architecture to model with chemical precision the small molecular systems considered in this work, both in 1D (H_2) and 3D (planar H_4), even in the presence of strong correlation. With a mere 316 trainable parameters, global QNN produced similar results to the state-of-the-art global KSR model, which requires 10 times more parameters. The same modest model demonstrated satisfactory generalization capability on the H_2 - H_2 system (two separate dihydrogen molecules), despite being trained solely on H_2 and H_4 . This offers a promising perspective on the potential of QNNs for generalization within this framework.

Moreover, we incorporated a fractional orbital occupation approach into our framework, which can be also employed to train purely classical models. This FO approach was instrumental in enhancing the accuracy of our models in 3D. By employing the FO method, we reduced both the average energy error and its variance by approximately sixfold. In that way, we effectively mitigated energy oscillations and substantially improved the extrapolation accuracy of global QiCNN.

To ensure the stability of our framework in realistic noisy conditions present in current quantum hardware—specifically, the SCF loop, which is the core subroutine of KS-DFT—we developed a theoretical framework in Sec. IV G that establishes error bounds and specifies the conditions the model and implementation must satisfy to guarantee well-controlled errors. We verified this theory in a concrete implementation under various noise settings, including typical hardware

noise, and recovered the predicted linear scaling. Furthermore, we provided resource estimates indicating the number of measurements required for a hardware implementation, demonstrating the feasibility of our approach when observables are measured in parallel.

Additionally, in our QML approach, a quantum-enhanced neural functional can be initially trained and subsequently fine-tuned with additional data as needed. Once trained, this functional can be directly applied to any system to infer its properties. This stands in contrast to variational quantum eigensolver [101] and imaginary time [102] approaches, which necessitate full optimization for each system. The latter methods evolve the quantum state using the exact electronic Hamiltonian. In contrast, QML methods depend on the quality of the training data to ensure qualitatively correct predictions.

As an outlook, in this study, we provided an initial evaluation of proposed techniques on quantum and quantum-inspired neural architectures and molecules. In future works, it would be necessary to expand our training dataset substantially to understand the practical efficiency of quantum-enhanced models. Efforts to construct such datasets are currently ongoing [32]. The capability of our models to generalize to significantly larger systems is a critical aspect to investigate. Specifically, can these models effectively learn from small, strongly correlated systems and subsequently generalize accurately to larger systems?

Additionally, the FO approach [72] employed in this study has been evaluated solely on dimers. It remains to be verified whether a single value of the temperature parameter γ , as detailed in Appendix A, is adequate for training on larger datasets. Alternatively, parameter-free FO approaches, such as the quadratically convergent method can be employed instead [73]. In this work, we opted for the simplest-to-implement method.

An essential direction for future work is to establish rigorous criteria under which QNNs provide genuine advantages over classical models, moving beyond general expressivity arguments. While recent advances in ML-based XC design [25,28,94] have yielded promising results, the quest for a universal functional remains unresolved. To systematically explore this challenge, we have released QEX [59], a flexible framework that supports arbitrary classical-quantum hybrids. Our approach is inherently compatible with barren plateau mitigation techniques (e.g., engineered dissipation [100]) and with architectures naturally resilient to trainability issues (e.g., QCNNs [42] and permutation-equivariant models [99]). Looking ahead, inspired by quantum-enhanced DFT approaches that leverage QPU-prepared ground states [107,108], we propose using QNNs as representation learners that extract task-relevant features from approximate ground-state wave functions. Analogously to QCNNs on quantum data [42], such models could provide inductive biases tailored to electronic-structure tasks. This perspective positions QNNs not merely as end-to-end regressors for the XC functional but as modules within a hybrid pipeline for learning more expressive and accurate approximations.

The advent of early fault-tolerant quantum computing (FTQC) architectures is expected to significantly broaden the range of feasible quantum algorithms for chemistry. While

fully fledged protocols such as quantum phase estimation (QPE) will ultimately become the method of choice as devices mature, recent studies already show that logical-qubit architectures can outperform variational methods on near-term problems [109]. Such approaches are fully compatible with our framework and may serve as a natural stepping stone prior to large-scale QPE. In the broader landscape, we anticipate that FTQC-based methods and DFT will coexist as complementary paradigms, with hybrid strategies likely to play a central role in the foreseeable future.

Overall, our findings emphasize the promising potential of QNN-based XC functionals within DFT. With this contribution, we provided an initial blueprint that would serve as a stepping stone towards the development of more optimized frameworks and quantum-enhanced neural architectures that can tackle even more challenging chemical problems with greater accuracy and efficiency.

ACKNOWLEDGMENT

The authors thank the Pasqal team for fruitful discussions, especially Atiyo Ghosh, Andrea A. Gentile, John West, and Louis-Paul Henry for their valuable insights.

APPENDIX A: INCORPORATING FRACTIONAL OCCUPATION IN RKS-DFT

Chai's algorithm for fractional occupation is based on the argument that the distribution of occupation numbers is independent of the specific many-body system and adheres to a universal form described by the Fermi-Dirac distribution [72]. Hence, the density can be expressed as $n(\mathbf{r}) = \sum_{i=1}^{\infty} g_i |\psi_i(\mathbf{r})|^2$, where the occupation number g_i is the Fermi-Dirac function $g_i = \{1 + \exp[(\epsilon_i - \mu)/\gamma]\}^{-1}$, which obeys the following two conditions, $\sum_{i=1}^{\infty} g_i = N$, and $0 \leq g_i \leq 1$, where ϵ_i is the orbital energy of the i th orbital $\psi_i(\mathbf{r})$ and μ is the chemical potential chosen to maintain a constant number of electrons N . To account for this modification, the expression for the energy is then given by

$$E[n] = T_s^\gamma[\{g_i, \psi_i\}] - \frac{\gamma}{k_B} S[\{g_i\}] + \int n(\mathbf{r})v(\mathbf{r})d\mathbf{r} + E_H[n] + E_{XC}^\gamma[n]. \quad (\text{A1})$$

with γ -dependent XC energy given by $E_{XC}^\gamma[n] = E_{XC}[n] + E_\gamma[n]$, where $E_\gamma[n]$ represents the difference between the noninteracting kinetic free energy at zero temperature and that at temperature γ . The entropy contribution is given by $-\frac{\gamma}{k_B} S[\{g_i\}] = \gamma \sum_{i=1}^{\infty} [g_i \ln(g_i) + (1 - g_i) \ln(1 - g_i)]$. Given a suitable electronic temperature γ , the chemical potential μ is fixed by $\sum_{i=1}^{\infty} \{1 + \exp[(\epsilon_i - \mu)/\gamma]\}^{-1} = N$ to conserve the number of particles. In essence, only the entropy contribution needs to be added to an existing RKS-DFT implementation and $E_{XC}^\gamma[n]$ is learned by the neural XC. This approach does not impact the overall computational scaling. The trade-off for improved density representability is the inclusion of a single free parameter, the temperature γ , which may require initial tuning for consistency across systems, a subject that warrants further investigation.

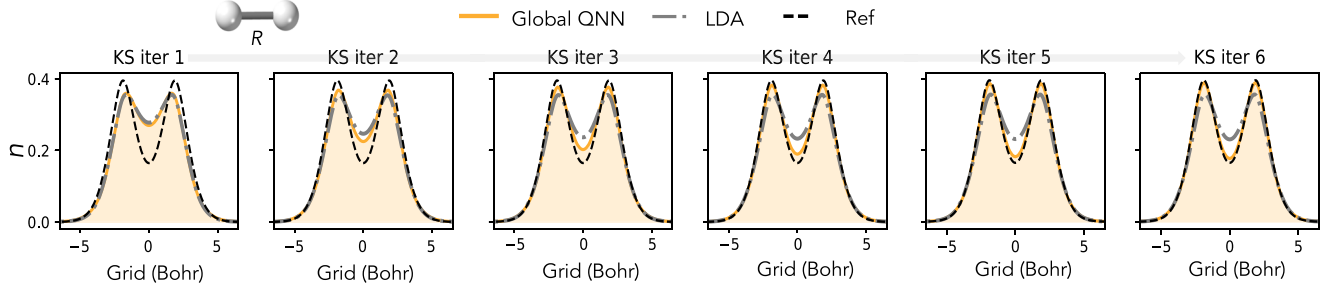


FIG. 11. Electron density of a hydrogen molecule ($R = 3.84$ bohr) at first six KS iterations using global QNN within the 1D KS-DFT framework. At every KS iteration, the electron density is improved, eventually reaching the reference one. The reference (Ref) result is generated using DMRG by Li *et al.* [25].

APPENDIX B: INTEGRATION GRIDS AND NEURAL XC FUNCTIONALS

A neural XC model needs to implicitly learn a function f_θ that is adapted to the $\{w_i\}_{i=1}^{N_{\text{grid}}}$ grid points, distributed according to a fixed grid type. There are multiple choices of grid types and densities (number of grid points N_{grid}), with a finer grid mesh yielding more accurate results at a higher computational cost. Ideally, a neural XC model needs to be trained/constrained to be invariant to finer mesh grids of the same type. See Fig. 11, where a trained global QNN model is used to compute the electron density throughout KS iterations on a fixed grid of 513 points. This adds to the challenge of making generalizable neural XC functionals where grid (hence the number of model inputs) needs to be adapted and neural XC functional made consistent with the changes. For instance, a potential solution is to design a grid invariant embedding or add constraints in the loss to make the NN f_θ invariant to grid density changes. Another approach would be to use also the coordinates $\{\mathbf{r}_i\}_{i=1}^{N_{\text{grid}}}$ as inputs of the model f_θ to explicitly include the information about the distribution of grid points in the Cartesian space. However, this approach would quadruple the number of inputs, so it is not currently considered for simplicity. In our work, we minimize grid variations by using a single grid type (e.g., of Lebedev type [110]) and the same number of grid points for all molecules whenever possible (our models have a fixed number of N_{grid} inputs) in order to facilitate the training.

APPENDIX C: BUILDING BLOCKS OF QUANTUM NEURAL NETWORKS

In this Appendix, we present typical DQCs that are used in this work.

Quantum circuit ansatz

As an ansatz U_a layer, we employ the popular hardware-efficient ansatz introduced in the context of VQE for chemistry [101]. The heuristic ansatz structure is not specifically tailored to the problem; thus, it represents the most general quantum model, closely resembling a classical MLP. It is composed of selected single qubit rotations [i.e., $R_z(\theta) = e^{-\frac{i}{2}\theta\hat{Z}}$] that are placed on all available qubits with distinct

angles θ_a ,

$$U_{\text{sqr}}(\theta_a) = \prod_{i=1}^{N_q} e^{-\frac{i}{2}\theta_{a,x}^i \hat{Z}} e^{-\frac{i}{2}\theta_{a,y}^i \hat{X}} e^{-\frac{i}{2}\theta_{a,z}^i \hat{Z}}, \quad (\text{C1})$$

which enables the representation of the most general rotation on a Bloch sphere. The entanglement block consists of an arbitrary pattern of (parameterized) multiqubit gates that introduce interactions between qubits. For example, considering the two-qubit case, a unitary is given by

$$U_{\text{ent}} = \prod_{(i,j) \in S} \text{CNOT}(i, j), \quad (\text{C2})$$

where $\text{CNOT}(i, j) = e^{-\frac{i}{4}(I_i - \hat{Z}_i)(I_j - \hat{X}_j)}$ and S represents an ordered set of pairs of qubit indices. We will employ an “alternate ladder” set $S_{\text{al}} = S_1 \cup S_2$ with $S_1 = ((0, 1), (2, 3), \dots, (N_q - 1, N_q))$ and $S_2 = ((1, 2), (3, 4), \dots, (N_q - 2, N_q - 1))$ unless specified otherwise. Note that a parametrized entanglement layer can also be defined as $U_{\text{ent}}^{\text{RZZ}}(\theta_{\text{ent}}) = \prod_{(i,j) \in S} R_{\text{ZZ}}^{i,j}(\theta_{\text{ent}}^{i,j}) = \prod_{(i,j) \in S} \text{CNOT}(i, j) e^{-\frac{i}{2}\theta_{\text{ent}}^{i,j} \hat{Z}} \text{CNOT}(i, j)$, where $R_{\text{ZZ}}^{i,j}$ represents the parametrized ZZ gate between qubits. This allows the model to cancel the interaction ($\theta_{\text{ent}}^{i,j} = 0$) between qubits if necessary. The final quantum state $|\psi\rangle$ prepared by repetitive applications of the feature map and ansatz is given by

$$|\psi(n, \theta)\rangle = \prod_{l=0}^{N_l} \left(\prod_{d=0}^{N_d} U_{\text{ent}}^{l,d} U_{\text{sqr}}^{l,d}(\theta_a^{l,d}) U_f^l(n) \right) |0\rangle, \quad (\text{C3})$$

and subsequently used in Eq. (13) as $f_\theta(n) = \langle \psi(n, \theta) | \hat{C} | \psi(n, \theta) \rangle$. Throughout this work, all the quantum layers will be evaluated in this manner. Next, we discuss how classical data can be embedded in quantum circuits $U_f(n)$.

APPENDIX D: DIFFERENTIATION OF QUANTUM MODELS

Gradient evaluation is an essential component for training NNs in general. Since (quantum) NNs can be thought of as a function f_θ , there are various approaches for differentiation. The standard approach in classical ML is the *automatic differentiation* that exploits the chain rule and stores the intermediate values of subfunctions (layers) of ML models to

evaluate the gradients of f_θ . This approach has an advantageous scaling that enables training large classical models and, hence, is always used for the differentiation of classical layers in this work. In the context of quantum computing, there has been a need to develop an efficient approach, beyond finite differences, for gradient evaluation. Schuld *et al.* showed that due to the trigonometric relations within the unitaries associated with single qubit rotations, $U_{\text{sqf}} = e^{-\frac{i}{2}\theta\hat{O}}$ (that include single-qubit Pauli generators $\hat{O}$), it is possible to derive a simple and exact derivative formula which highly resembles finite differences, the so-called *parameter-shift rule* (PSR) [39]:

$$\nabla_\theta f_\theta(n) = \frac{f_{\theta+\pi/2}(n) - f_{\theta-\pi/2}(n)}{2}. \quad (\text{D1})$$

The generalized PSR rule [40] is needed for general multi-qubit generators.

APPENDIX E: NUMBER OF MEASUREMENTS

To determine the measurement budget to reach an accuracy ε with QNNs presented in this work, we follow the standard frequentist approach [111,112]. The output of QNNs is obtained by measuring the total magnetization cost operator, which can be expressed as the sum of the expectation values of Pauli-Z operators over all N_q qubits, $M = \sum_{i=1}^{N_q} \langle \hat{Z}_i \rangle$. Assume each term $\hat{Z}_i$ is measured m_i times. The statistical error on each term, ε_i , is given by $\varepsilon_i^2 = \frac{\text{Var}[\langle \hat{Z}_i \rangle]}{m_i}$. Since $\langle \hat{Z}_i \rangle \in [-1, 1]$, its variance is upper bounded by 1, $\text{Var}[\langle \hat{Z}_i \rangle] \leq 1$. To achieve a total error ε , we can distribute the per-term errors such that $\varepsilon_i^2 = \frac{1}{N_q} \varepsilon^2$. Then, the total number of measurements N_{shots} across all qubits becomes

$$N_{\text{shots}} = \sum_{i=1}^{N_q} m_i = \sum_{i=1}^{N_q} \frac{\text{Var}[\langle \hat{Z}_i \rangle]}{\varepsilon_i^2} \leq \sum_{i=1}^{N_q} \frac{1}{\varepsilon^2/N_q} = \frac{N_q^2}{\varepsilon^2}. \quad (\text{E1})$$

Since all the operators are diagonal, they can be measured simultaneously, resulting in a total shot cost of $N_{\text{shots}} = \mathcal{O}(\varepsilon^{-2})$.

APPENDIX F: NOISE CHANNELS

In Appendix E, we discussed the error arising from finite sampling. In modeling realistic quantum devices, additional noise processes are commonly represented by completely positive trace-preserving (CPTP) maps acting on a qubit's density matrix $\rho = |\Psi\rangle\langle\Psi|$. Any CPTP map can be expressed in terms of a set of Kraus operators $\{\hat{K}_i\}$ satisfying $\sum_i \hat{K}_i^\dagger \hat{K}_i = \mathbb{I}$. Two common single-qubit noise channels are *dephasing* and *amplitude damping*, which we focused on in our experiment in Table VIII implemented in QEX [59].

Dephasing (phase-damping) channel. The dephasing channel describes the decay of quantum coherences without altering computational-basis populations. A Kraus representation is

$$\hat{K}_0 = \sqrt{1-p} \mathbb{I}, \quad \hat{K}_1 = \sqrt{p} \hat{Z}, \quad (\text{F1})$$

where $0 \leq p \leq 1$ is the dephasing probability and $\hat{Z}$ is the Pauli-Z operator. The channel acts as

$$\mathcal{E}_{\text{deph}}(\rho) = \hat{K}_0 \rho \hat{K}_0^\dagger + \hat{K}_1 \rho \hat{K}_1^\dagger. \quad (\text{F2})$$

Amplitude-damping channel. The amplitude-damping channel models energy relaxation from $|1\rangle$ to $|0\rangle$, as in spontaneous emission. A standard Kraus representation is

$$\hat{K}_2 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}, \quad \hat{K}_3 = \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix}, \quad (\text{F3})$$

where p is the probability of decay during the noise process. The map is defined as

$$\mathcal{E}_{\text{amp}}(\rho) = \hat{K}_2 \rho \hat{K}_2^\dagger + \hat{K}_3 \rho \hat{K}_3^\dagger. \quad (\text{F4})$$

In Sec. IV G, we analyze the performance of a typical QNN in our framework by applying these noise channels and tuning their strength via the parameter p .

APPENDIX G: PROOFS OF ERROR BOUNDS

Proof. We prove Theorem 1 under Assumptions 1–3, as stated in Sec. IV G. We aim to show that, under an error ε , the exact self-consistent density n^* obtained from the exact SCF map $\mathcal{T}$ (with the exact XC functional) differs from the approximate self-consistent density $\tilde{n}$ obtained from the approximate SCF map $\tilde{\mathcal{T}}$ by at most $\mathcal{O}(\varepsilon)$.

Step 1: We define iteration errors. Let $e_k = \|n^{(k)} - n^*\|$, with $\|\cdot\| \equiv \|\cdot\|_2$. Then by definition of $\mathcal{T}$,

$$n^{(k+1)} - n^* = \tilde{\mathcal{T}}[n^{(k)}] - \mathcal{T}[n^*].$$

We split and add $\mathcal{T}[n^{(k)}]$:

$$n^{(k+1)} - n^* = (\tilde{\mathcal{T}} - \mathcal{T})[n^{(k)}] + (\mathcal{T}[n^{(k)}] - \mathcal{T}[n^*]).$$

Hence,

$$\begin{aligned} e_{k+1} &= \|n^{(k+1)} - n^*\| \leq \|(\tilde{\mathcal{T}} - \mathcal{T})[n^{(k)}]\| \\ &\quad + \|\mathcal{T}[n^{(k)}] - \mathcal{T}[n^*]\| \\ &\leq \underbrace{\|\tilde{\mathcal{T}}[n^{(k)}] - \mathcal{T}[n^{(k)}]\|}_{\text{Term (a)}} + \underbrace{\|\mathcal{T}[n^{(k)}] - \mathcal{T}[n^*]\|}_{\text{Term (b)}}. \end{aligned} \quad (\text{G1})$$

Step 2: We bound Term (b) by contractivity. By Assumption 1, for $n_1, n_2 \in U$,

$$\|\mathcal{T}[n_1] - \mathcal{T}[n_2]\| \leq \kappa \|n_1 - n_2\|.$$

Since $n^{(k)}$ and n^* lie in U , we get

$$\|\mathcal{T}[n^{(k)}] - \mathcal{T}[n^*]\| \leq \kappa \|n^{(k)} - n^*\| = \kappa e_k.$$

Step 3: We bound Term (a) by uniform XC error. From Sec. IV G, we recall that $\tilde{\mathcal{T}}$ and $\mathcal{T}$ differ only in the XC potential part, and by Assumption 2, that potential difference is at most $K\varepsilon$. To convince yourself, see the KS equations in Sec. II, where Hamiltonian operator $\hat{H}[n] = -\nabla^2/2 + v_{\text{KS}}[n]$ ultimately defines the SCF map $\mathcal{T}$ through XC potential v_{XC} . In the local region U , changing the potential by at most $K\varepsilon$ changes the resulting density by at most some constant factor $\alpha\varepsilon$. Hence we have

$$\|\tilde{\mathcal{T}}[n] - \mathcal{T}[n]\| \leq \alpha\varepsilon \quad \text{for all } n \in U. \quad (\text{G2})$$

Thus, for $n^{(k)} \in U$,

$$\|\tilde{\mathcal{T}}[n^{(k)}] - \mathcal{T}[n^{(k)}]\| \leq \alpha\varepsilon.$$

Step 4: We combine the previous statements into Eq. (G1):

$$\begin{aligned} e_{k+1} &= \|n^{(k+1)} - n^*\| \\ &\leq \|\tilde{\mathcal{T}}[n^{(k)}] - \mathcal{T}[n^{(k)}]\| + \|\mathcal{T}[n^{(k)}] - \mathcal{T}[n^*]\| \\ &\leq \alpha \varepsilon + \kappa e_k. \end{aligned} \quad (\text{G3})$$

Step 5: The density $n^{(k)}$ remains in U by Assumption 3, we solve the recurrence by rewriting

$$e_{k+1} \leq \kappa e_k + \alpha \varepsilon,$$

as

$$e_k \leq \kappa^k e_0 + \alpha \varepsilon \sum_{j=0}^{k-1} \kappa^j = \kappa^k e_0 + \alpha \varepsilon \frac{1 - \kappa^k}{1 - \kappa}.$$

Set $n^{(0)}$ as the initial guess, so $e_0 = \|n^{(0)} - n^*\|$. This proves the finite- k bound (27). Taking $k \rightarrow \infty$, we get

$$e_\infty = \|\tilde{n} - n^*\| \leq \frac{\alpha \varepsilon}{1 - \kappa}. \quad (\text{G4})$$

Hence the approximate self-consistent density $\tilde{n}$ is within $\mathcal{O}(\varepsilon)$ of the true electron density n^* .

Step 6: Energy difference is also $\mathcal{O}(\varepsilon)$. Finally, define $\tilde{E} \equiv \tilde{E}[\tilde{n}]$ and $E^* \equiv E[n^*]$. Write

$$\tilde{E} - E^* = (\tilde{E}[\tilde{n}] - E[\tilde{n}]) + (E[\tilde{n}] - E[n^*]). \quad (\text{G5})$$

Since $\tilde{E}_{\text{XC}}$ differs from E_{XC} by at most ε in value, $\tilde{E}$ differs from E by at most ε at the *same* density $\tilde{n}$. Meanwhile, we assume that $|E[\tilde{n}] - E[n^*]| \leq M \|\tilde{n} - n^*\|$ for some Lipschitz constant $M > 0$ near n^* since according to Eq. (G4), $\tilde{n}$ and n^* are order ε close in the region U by Assumption 3. Hence,

$$|\tilde{E} - E^*| \leq \varepsilon + M \|\tilde{n} - n^*\| \leq \varepsilon + M \frac{\alpha}{1 - \kappa} \varepsilon = C' \varepsilon,$$

where $C' = 1 + \frac{M\alpha}{1-\kappa}$. This completes the proof. ■

Thus, the final approximate energy is $\mathcal{O}(\varepsilon)$ -close to the true energy, demonstrating that, under our assumptions, catastrophic error accumulation in the SCF loop is avoided.

APPENDIX H: LIPSCHITZ CONTINUITY OF QNNS

In this Appendix, we establish the Lipschitz continuity of QNNs employed in this work, following the derivations and notation from Ref. [39] for clarity. Recall the definition of quantum models, Eq. (13), here denoted as

$$f_\theta(n) = \langle 0 | U^\dagger(n, \theta) \hat{C} U(n, \theta) | 0 \rangle,$$

which forms the basis of all our N_q -qubit models. We propose the following Lemma to support our arguments from Sec. IV G.

Lemma H1 (QNN Lipschitz continuity). Let QNN be represented by $f_\theta(n) = \langle 0 | U^\dagger(n, \theta) \hat{C} U(n, \theta) | 0 \rangle$ with

$$\begin{aligned} U(n) &= W^{(L+1)} S(n) W^{(L)} \dots W^{(2)} S(n) W^{(1)}, \\ S(n) &= e^{-inH}, \quad \hat{C} = \sum_{i=1}^{N_q} \hat{Z}_i, \end{aligned}$$

where $W^k \equiv W^k(\theta)$ are x -independent unitaries (representing ansatz), $S(n)$ is the feature map with Hermitian operator H with eigenvalues $\{\lambda_m\}_{m=1}^d$, $d = 2^{N_q}$, and $L = N_l$ is the number of layers. Then f_θ is globally Lipschitz in x , defined according to Ref. [39] as

$$f(n) = \sum_{k,j \in [d]^L} e^{i(\Lambda_k - \Lambda_j)n} a_{k,j}, \quad \Lambda_j = \lambda_{j_1} + \dots + \lambda_{j_L},$$

Consequently,

$$\sup_{x \in \mathbb{R}} |f'_\theta(n)| \leq N_q \sum_{k,j \in [d]^L} |\Lambda_k - \Lambda_j| < \infty,$$

where we use multi-index $j = \{j_1, \dots, j_L\} \in [d]^L$, $[d]^L$ denotes any set of L integers between 1 and d , and N_q is the number of qubits. Hence, f_θ is globally Lipschitz on $\mathbb{R}$ (for any density n) with a θ -independent constant.

Proof. We follow the Ref. [39] and recall that:

$$f(n) = \sum_{k,j \in [d]^L} e^{i(\Lambda_k - \Lambda_j)n} a_{k,j},$$

where

$$\begin{aligned} a_{k,j} &= \sum_{i,i'} (W^*)_{ik_1}^{(1)} (W^*)_{ji_2}^{(2)} \dots (W^*)_{ji_L}^{(L+1)} C_{i,i'} \\ &\times W_{i'j_L}^{(L+1)} \dots W_{j_2j_1}^{(2)} W_{j_11}^{(1)}. \end{aligned}$$

Since $\|\hat{C}\| = N_q$, W^k are unitary, and $\sup_n |f_\theta(n)| \leq N_q$, consequently the elements $|a_{k,j}| \leq \|\hat{C}\| = N_q$. We can then write the derivative of our model,

$$f'_\theta(n) = \sum_{k,j \in [d]^L} i(\Lambda_k - \Lambda_j) e^{i(\Lambda_k - \Lambda_j)n} a_{k,j}.$$

So $\sup_n |f'_\theta(n)| \leq \sum_{k,j \in [d]^L} |\Lambda_k - \Lambda_j| |a_{k,j}| \leq N_q \sum_{k,j \in [d]^L} |\Lambda_k - \Lambda_j| < \infty$ since the spectrum is finite. The mean-value theorem implies the stated Lipschitz continuity of $f_\theta(n)$. ■

- [1] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, International Series of Monographs on Chemistry (Oxford University Press, Oxford, 1994).
- [2] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley, New York, 2015).
- [3] E. Engel and R. M. Dreizler, *Density Functional Theory: An Advanced Course*, Theoretical and Mathematical Physics (Springer, Berlin, 2011).

- [4] A. D. Becke, *J. Chem. Phys.* **140**, 18A301 (2014).
- [5] R. O. Jones, *Rev. Mod. Phys.* **87**, 897 (2015).
- [6] A. J. Cohen, P. Mori-Sánchez, and W. Yang, *Chem. Rev.* **112**, 289 (2012).
- [7] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, *Science* **355**, 49 (2017).
- [8] K. P. Kepp, *Science* **356**, 496 (2017).

- [9] A. J. Cohen, P. Mori-Sánchez, and W. Yang, *Science* **321**, 792 (2008).
- [10] A. J. Cohen, P. Mori-Sánchez, and W. Yang, *J. Chem. Phys.* **129**, 121104 (2008).
- [11] N. Mardirossian and M. Head-Gordon, *Mol. Phys.* **115**, 2315 (2017).
- [12] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, *Phys. Chem. Chem. Phys.* **19**, 32184 (2017).
- [13] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, *J. Chem. Phys.* **150**, 154122 (2019).
- [14] P. Verma and D. G. Truhlar, *Trends Chem.* **2**, 302 (2020).
- [15] M. A. Rohrdanz, K. M. Martins, and J. M. Herbert, *J. Chem. Phys.* **130**, 054112 (2009).
- [16] S. Mohr, L. E. Ratcliff, L. Genovese, D. Caliste, P. Boulanger, S. Goedecker, and T. Deutsch, *Phys. Chem. Chem. Phys.* **17**, 31360 (2015).
- [17] M. Bursch, J.-M. Mewes, A. Hansen, and S. Grimme, *Angew. Chem. Int. Ed.* **61**, e202205735 (2022).
- [18] D. J. Tozer, V. E. Ingamells, and N. C. Handy, *J. Chem. Phys.* **105**, 9200 (1996).
- [19] D. J. Tozer and N. C. Handy, *Recent Advances In Density Functional Methods: (Part II)* (World Scientific, Singapore, 1997), pp. 1–14.
- [20] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, *Nat. Commun.* **8**, 872 (2017).
- [21] R. Nagai, R. Akashi, and O. Sugino, *npj Comput. Mater.* **6**, 43 (2020).
- [22] S. Dick and M. Fernandez-Serra, *Nat. Commun.* **11**, 3509 (2020).
- [23] B. Kalita, L. Li, R. J. McCarty, and K. Burke, *Acc. Chem. Res.* **54**, 818 (2021).
- [24] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, *Science* **374**, 1385 (2021).
- [25] L. Li, S. Hoyer, R. Pederson, R. Sun, E. D. Cubuk, P. Riley, and K. Burke, *Phys. Rev. Lett.* **126**, 036401 (2021).
- [26] E. Cuierrier, P.-O. Roy, R. Wang, and M. Ernzerhof, *J. Chem. Phys.* **157**, 171103 (2022).
- [27] S. Riemelmoser, C. Verdi, M. Kaltak, and G. Kresse, *J. Chem. Theory Comput.* **19**, 7287 (2023).
- [28] M. F. Kasim and S. M. Vinko, *Phys. Rev. Lett.* **127**, 126403 (2021).
- [29] R. Nagai, R. Akashi, and O. Sugino, *Phys. Rev. Res.* **4**, 013106 (2022).
- [30] J. Wu, S.-M. Pun, X. Zheng, and G. Chen, *J. Chem. Phys.* **159**, 090901 (2023).
- [31] X. Zhang and G. K.-L. Chan, *J. Chem. Phys.* **157**, 204801 (2022).
- [32] P. A. M. Casares, J. S. Baker, M. Medvidović, R. d. Reis, and J. M. Arrazola, *J. Chem. Phys.* **160**, 062501 (2024).
- [33] B. Kalita, R. Pederson, J. Chen, L. Li, and K. Burke, *J. Phys. Chem. Lett.* **13**, 2540 (2022).
- [34] J. Schmidt, C. L. Benavides-Riveros, and M. A. Marques, *J. Phys. Chem. Lett.* **10**, 6425 (2019).
- [35] H. Ma, A. Narayanaswamy, P. Riley, and L. Li, *Sci. Adv.* **8**, eabq0279 (2022).
- [36] O. Kyriienko, A. E. Paine, and V. E. Elfving, *Phys. Rev. A* **103**, 052416 (2021).
- [37] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, *Nat. Comput. Sci.* **1**, 403 (2021).
- [38] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, *Phys. Rev. A* **98**, 032309 (2018).
- [39] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, *Phys. Rev. A* **99**, 032331 (2019).
- [40] O. Kyriienko and V. E. Elfving, *Phys. Rev. A* **104**, 052417 (2021).
- [41] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, *Nat. Rev. Phys.* **3**, 625 (2021).
- [42] I. Cong, S. Choi, and M. D. Lukin, *Nat. Phys.* **15**, 1273 (2019).
- [43] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, *et al.*, *Science* **376**, 1182 (2022).
- [44] M. Schuld, R. Sweke, and J. J. Meyer, *Phys. Rev. A* **103**, 032430 (2021).
- [45] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, *Quantum* **6**, 824 (2022).
- [46] A. Abbas, R. King, H.-Y. Huang, W. J. Huggins, R. Movassagh, D. Gilboa, and J. R. McClean, in *Proceedings of the 37th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2023).
- [47] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, D. Scheiermann, and R. Wolf, *Nat. Commun.* **11**, 808 (2020).
- [48] S. H. Sack, R. A. Medina, A. A. Michailidis, R. Kueng, and M. Serbyn, *PRX Quantum* **3**, 020365 (2022).
- [49] H.-Y. Huang, R. Kueng, and J. Preskill, *Nat. Phys.* **16**, 1050 (2020).
- [50] D. Miller, L. E. Fischer, K. Levi, E. J. Kuehnke, I. O. Sokolov, P. K. Barkoutsos, J. Eisert, and I. Tavernelli, *npj Quantum Inf.* **10**, 122 (2024).
- [51] E. Gil-Fuster, C. Gyurik, A. Pérez-Salinas, and V. Dunjko, in *The Thirteenth International Conference on Learning Representations* (ICLR, 2025).
- [52] B. Senjean, S. Yalouz, and M. Saubanière, *SciPost Phys.* **14**, 055 (2023).
- [53] M. Rossmannek, P. K. Barkoutsos, P. J. Ollitrault, and I. Tavernelli, *J. Chem. Phys.* **154**, 114105 (2021).
- [54] R. Schade, C. Bauer, K. Tamoev, L. Mazur, C. Pleschl, and T. D. Kühne, *Phys. Rev. Res.* **4**, 033160 (2022).
- [55] F. Gaitan and F. Nori, *Phys. Rev. B* **79**, 205117 (2009).
- [56] T. E. Baker and D. Poulin, *Phys. Rev. Res.* **2**, 043238 (2020).
- [57] P. Barkoutsos, I. Olegovich Sokolov, G.-J. Both, and V. E. Elfving, Determining exchange-correlation functionals for molecular systems using a quantum computer (2025), <https://patents.google.com/patent/EP4629143A1/en>.
- [58] M. Schuld and N. Killoran, *PRX Quantum* **3**, 030101 (2022).
- [59] I. O. Sokolov and S. Acheche, QEX: Quantum-enhanced neural exchange-correlation functionals (2025), <https://github.com/pasqal-io/qex>.
- [60] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable

- transformations of python+numpy programs (2018), <http://github.com/jax-ml/jax>.
- [61] P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).
 - [62] K. Hornik, M. Stinchcombe, and H. White, *Neural Netw.* **2**, 359 (1989).
 - [63] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, *Nat. Commun.* **12**, 1791 (2021).
 - [64] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, *Quantum* **4**, 226 (2020).
 - [65] A. Pérez-Salinas, D. López-Núñez, A. García-Sáez, P. Forn-Díaz, and J. I. Latorre, *Phys. Rev. A* **104**, 012405 (2021).
 - [66] J. Landman, S. Thabet, C. Dalyac, H. Mhiri, and E. Kashefi, in *The Eleventh International Conference on Learning Representations* (ICLR, 2023).
 - [67] K. Nakaji, S. Uno, Y. Suzuki, R. Raymond, T. Onodera, T. Tanaka, H. Tezuka, N. Mitsuda, and N. Yamamoto, *Phys. Rev. Res.* **4**, 023136 (2022).
 - [68] S. Dasgupta and J. M. Herbert, *J. Comput. Chem.* **38**, 869 (2017).
 - [69] X.-M. Zhang, M.-H. Yung, and X. Yuan, *Phys. Rev. Res.* **3**, 043200 (2021).
 - [70] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, in *NIPS 2017 Workshop on Autodiff* (ICLR, 2017).
 - [71] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, *et al.*, *WIREs Comput. Mol. Sci.* **8**, e1340 (2018).
 - [72] J.-D. Chai, *J. Chem. Phys.* **136**, 154104 (2012).
 - [73] E. Cancès, K. N. Kudin, G. E. Scuseria, and G. Turinici, *J. Chem. Phys.* **118**, 5364 (2003).
 - [74] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-structure Theory* (John Wiley & Sons, Hoboken, USA, 2013).
 - [75] R. J. Bartlett and M. Musiał, *Rev. Mod. Phys.* **79**, 291 (2007).
 - [76] I. W. Bulik, T. M. Henderson, and G. E. Scuseria, *J. Chem. Theory Comput.* **11**, 3171 (2015).
 - [77] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
 - [78] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, *ACM Trans. Math. Softw.* **23**, 550 (1997).
 - [79] D. Seitz, N. Heim, J. P. Moutinho, R. Guichard, V. Abramavicius, A. Wennersteen, G.-J. Both, A. Quelle, C. de Groot, G. V. Velikova, V. E. Elfving, and M. Dagrada, *IEEE Software*, **42**, 44 (2025).
 - [80] D. Tuna, D. Lefrancois, Ł. Wolanski, S. Gozem, I. Schapiro, T. Andruniów, A. Dreuw, and M. Olivucci, *J. Chem. Theory Comput.* **11**, 5758 (2015).
 - [81] H. Douroudgari, M. Vahedpour, and F. Khouini, *Sci. Rep.* **11**, 13220 (2021).
 - [82] M.-C. Kim, E. Sim, and K. Burke, *Phys. Rev. Lett.* **111**, 073003 (2013).
 - [83] D. K. Park, F. Petruccione, and J.-K. K. Rhee, *Sci. Rep.* **9**, 3949 (2019).
 - [84] I. O. Sokolov, P. K. Barkoutsos, P. J. Ollitrault, D. Greenberg, J. Rice, M. Pistoia, and I. Tavernelli, *J. Chem. Phys.* **152**, 124107 (2020).
 - [85] X. Liu, X. Wang, Z. Wen, and Y. Yuan, *SIAM J. Matrix Anal. Appl.* **35**, 546 (2014).
 - [86] X. Liu, Z. Wen, X. Wang, M. Ulbrich, and Y. Yuan, *SIAM J. Numer. Anal.* **53**, 1758 (2015).
 - [87] N. D. Woods, M. Payne, and P. Hasnip, *J. Phys.: Condens. Matter* **31**, 453001 (2019).
 - [88] E. Cancès and C. Le Bris, *ESAIM: M2AN* **34**, 749 (2000).
 - [89] Z. Bai, R.-C. Li, and D. Lu, [arXiv:2009.09022](https://arxiv.org/abs/2009.09022).
 - [90] P. Pulay, *Chem. Phys. Lett.* **73**, 393 (1980).
 - [91] K. N. Kudin, G. E. Scuseria, and E. Cancès, *J. Chem. Phys.* **116**, 8255 (2002).
 - [92] G. P. Kerker, *Phys. Rev. B* **23**, 3082 (1981).
 - [93] V. R. Saunders and I. H. Hillier, *Int. J. Quantum Chem.* **7**, 699 (1973).
 - [94] G. Luise, C.-W. Huang, T. Vogels, D. P. Kooi, S. Ehlert, S. Lanius, K. J. Giesbertz, A. Karton, D. Gunceler, M. Stanley, *et al.*, [arXiv:2506.14665](https://arxiv.org/abs/2506.14665).
 - [95] T. Ko and X. Li, *Math. Comp.* **92**, 1693 (2023).
 - [96] T. Ko, X. Li, and C. Wang, in *Proceedings of the IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE, Los Alamitos, CA, 2024), Vol. 1, pp. 669–680.
 - [97] W. Somogyi, E. Pankovets, V. Kuzmin, and A. Melnikov, [arXiv:2410.19921](https://arxiv.org/abs/2410.19921).
 - [98] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, *Nature (London)* **627**, 778 (2024).
 - [99] L. Schatzki, M. Larocca, Q. T. Nguyen, F. Sauvage, and M. Cerezo, *npj Quantum Inf.* **10**, 12 (2024).
 - [100] A. Sannia, F. Tacchino, I. Tavernelli, G. L. Giorgi, and R. Zambrini, *npj Quantum Inf.* **10**, 81 (2024).
 - [101] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O’Brien, *Nat. Commun.* **5**, 4213 (2014).
 - [102] S. McArdle, T. Jones, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan, *npj Quantum Inf.* **5**, 75 (2019).
 - [103] S. Das and V. Gavini, *Phys. Rev. B* **107**, 125133 (2023).
 - [104] A. D. Becke, *J. Chem. Phys.* **88**, 2547 (1988).
 - [105] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, *et al.*, *Nature (London)* **618**, 500 (2023).
 - [106] G. Torlai, G. Mazzola, G. Carleo, and A. Mezzacapo, *Phys. Rev. Res.* **2**, 022060(R) (2020).
 - [107] E. Sheridan, L. Mineh, R. A. Santos, and T. Cubitt, [arXiv:2402.18534](https://arxiv.org/abs/2402.18534).
 - [108] E. Koridon, F. Frohnert, E. Prehn, E. van Nieuwenburg, J. Tura, and S. Polla, *Mach. Learn.: Sci. Technol.* **6**, 025020 (2025).
 - [109] W. van Dam, H. Liu, G. H. Low, A. Paetznick, A. Paz, M. Silva, A. Sundaram, K. Svore, and M. Troyer, [arXiv:2409.05835](https://arxiv.org/abs/2409.05835).
 - [110] V. Lebedev, *USSR Comput. Math. Math. Phys.* **15**, 44 (1975).
 - [111] D. Wecker, M. B. Hastings, and M. Troyer, *Phys. Rev. A* **92**, 042303 (2015).
 - [112] J. Romero, R. Babbush, J. R. McClean, C. Hempel, P. J. Love, and A. Aspuru-Guzik, *Quantum Sci. Technol.* **4**, 014008 (2018).