Modelling and Simulation in Materials Science and Engineering

OPEN ACCESS

RECEIVED
5 January 2026

REVISED
17 April 2026

ACCEPTED FOR PUBLICATION
25 May 2026

PUBLISHED
29 July 2026

TOPICAL REVIEW

Recent advances in electronic structure learning

Tim S Hindges[1,2,*], **Wenhao He**[3] and **Ju Li**[4,5]

[1] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[2] Department of Chemistry, Imperial College London, London, United Kingdom
[3] The Center for Computational Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[4] Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[5] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: tim.hindges21@imperial.ac.uk

Keywords: electronic structure, machine learning, deep learning, quantum chemistry, density functional theory, neural networks, neural quantum states

Abstract

Machine learning (ML) has emerged as a transformative approach for overcoming the unfavorable scaling that limits system sizes and accessible timescales in electronic structure calculations, which underpin modern computational sciences. Herein, we provide a comprehensive examination of ML methods that learn electronic structure representations across multiple levels of theory. We introduce an 'ML Jacob's ladder' framework that categorizes methods by the completeness of electronic structure information they capture, from determining the many-body wavefunction through neural quantum states, to parameterizing exchange- correlation functionals in density functional theory, to predicting Hamiltonians and Fock matrices that bypass self-consistent-field iterations, and finally to learning electron densities from which diverse observables can be derived. Throughout, we emphasize the conceptual unification of these approaches through diverse representations and shared physical principles such as the variational theorem, Kohn–Sham formalism, and hierarchical corrections that make ML uniquely suited for this domain. We also highlight generalizable models emerging across rungs that promise transferability between chemical systems and levels of theory, alongside practical implementations for real molecular and materials systems. By examining how different architectural choices and physical constraints shape model capabilities, we map the evolving landscape where ML is reshaping the accessibility, accuracy, and interpretability of quantum mechanical predictions.

1. Introduction

Quantum chemistry (QC) techniques undertake a fundamental role in materials, chemical, physical, and biological research [1–5], allowing *ab initio* simulation of systems and the extraction of relevant properties. The refinement and improvement of computing power and algorithms have recently allowed such complex methods to become common practice in these fields, with chemistry and materials related research often taking over a third of supercomputer hours [6, 7].

The substantial amount of computational resources dedicated to these methods underscores their importance to the scientific community, but simultaneously highlights the need for improved efficiency. The unfavorable scaling of the QC methods restricts the abilities of researchers, particularly for large systems, demonstrating the value in enabling accelerated *ab initio* calculations. For instance, the exponential scaling of full configuration interaction (FCI) renders calculations infeasible for systems exceeding 20 electrons in 20 orbitals [8].

Therefore, with the recent emergence of machine learning (ML) methods as statistical frameworks for inferring complex functional relationships, they have increasingly been applied to QC and electronic-structure problems in order to make *ab initio* results accessible at a cheaper computational cost. Indeed, early quantitative structure-activity relationship models showed the ability of simple statistical methods

to predict molecular properties from their structural components [9]. With the subsequent improvement of ML architectures and the advent of neural networks (NNs), the complexity of relationships between inputs and predictions evolved accordingly. As such, models have been able to determine an expanding number of physicochemical properties, achieving accuracy that nears, and in some cases rivals, high-level quantum-chemical methods [10–15]. Relatedly, a particularly popular and growing use of NNs across the molecular and material sciences is ML interatomic potentials (MLIPs), which aim to provide reliable energies and forces for large-scale molecular and materials simulations such as molecular dynamics which are unfeasible or too costly with the above QC techniques [16–18]. These successes illustrate the potential of NNs to learn underlying physical properties of systems, and the architectures served as inspiration for many of the methods we will discuss below.

However, in this review, we focus on advances in ML models aimed at learning beyond direct property prediction. In particular, we highlight recent efforts that use ML to predict and understand electronic structure itself. We believe this to be particularly important for the community of *ab initio* modelers and ML researchers because expanding the learning task to encompass broader aspects of the underlying physics enables NNs to acquire a more fundamental and transferable representation of molecular and materials systems. By learning electronic structure directly, a model is inherently more applicable since many downstream properties can be computed without additional supervision. Moreover, this approach can enhance accuracy because models trained only on target properties may lack the complete physical information encoded in the electronic structure, resulting in degraded performance on challenging or out-of-distribution cases. In fact, increasing the electronic structure information included in ML models has been shown to lead to better accuracy [19], suggesting that directly learning electronic structure can further improve results. Indeed, the models in this review learn underlying physics with accuracy approaching that of the original QC methods, hinting towards a future with more accessible *ab initio* results that retain their informational breadth and accuracy. It should be noted that the accuracy of supervised models in this review is compared to the method the authors trained the model on (unless stated otherwise). As such, numerical results and 'chemical accuracy' claims (roughly 1 kcal mol^{-1} $\approx$43 meV per molecule) are contingent on the baseline accuracy of the method used, which is important context for the reader.

We believe ML to be particularly primed for learning electronic structure due to four key paradigms. Firstly, due to the reliance of diverse fields on QC and recent optimizations of algorithms and computational efficiency, a lot of relevant *ab initio* data is available to train models, with databases for materials [20]), small chemical systems [21, 22]) and biologically relevant systems [23] becoming available. Therefore, since these calculations often require solving the Schrödinger equation repeatedly for chemically similar systems that are built from a limited set of elements, these datasets present a high-value landscape for data-driven models.

Secondly, ML models possess the theoretical expressivity needed to represent the high-dimensional objects that define electronic structure. For example, kernel ridge regression (KRR) operates in a constructed reproducing kernel Hilbert space to represent and minimize risk over infinite-dimensional function spaces in a tractable way [24, 25]. This shares the same functional-analytic foundation as quantum mechanics with its Hilbert space. Furthermore, the universal approximation theorem [26] establishes that sufficiently expressive NNs can approximate continuous functions and functionals to arbitrary accuracy, providing a principled foundation for learning wavefunctions, exchange-correlation (XC) functionals, Hamiltonians, and densities directly from data. When combined with symmetry-preserving architectures that encode the physically required invariances, this expressive power makes NNs exceptionally well suited for modeling the structured mappings at the heart of QC.

Thirdly, electronic structure methods follow a hierarchical and multifidelity structure from semi-empirical methods to density functional theory (DFT) to coupled cluster (CC) and beyond. Thus, rather than learning expensive high-level theory directly, ML models can learn corrections $\Delta = E_{\mathrm{high}} - E_{\mathrm{low}}$ from a computationally cheap baseline method. By learning corrections to the electronic structure, the model requires far fewer training examples and enables systematic improvement over the baseline methods. This was initially applied to the chemical space by Ramakrishnan *et al* [27], and has since been a key part of certain kernel based models [28–30], and, recently, NNs [31, 32].

Lastly, ML is especially disposed for learning electronic structure due to the existence of two key theoretical foundations; the variational principle and the Kohn–Sham (KS) formulation of DFT. Indeed, KS theory guarantees the existence of an exact universal functional $E[\rho]$, but provides no prescription for finding it. Traditional functional development has progressed through increasingly sophisticated handcrafted approximations as portrayed by Jacob's ladder [33] (see figure 1), essentially manually feature fitting to known exact conditions. By nature, ML offers a complementary approach to this fitting by

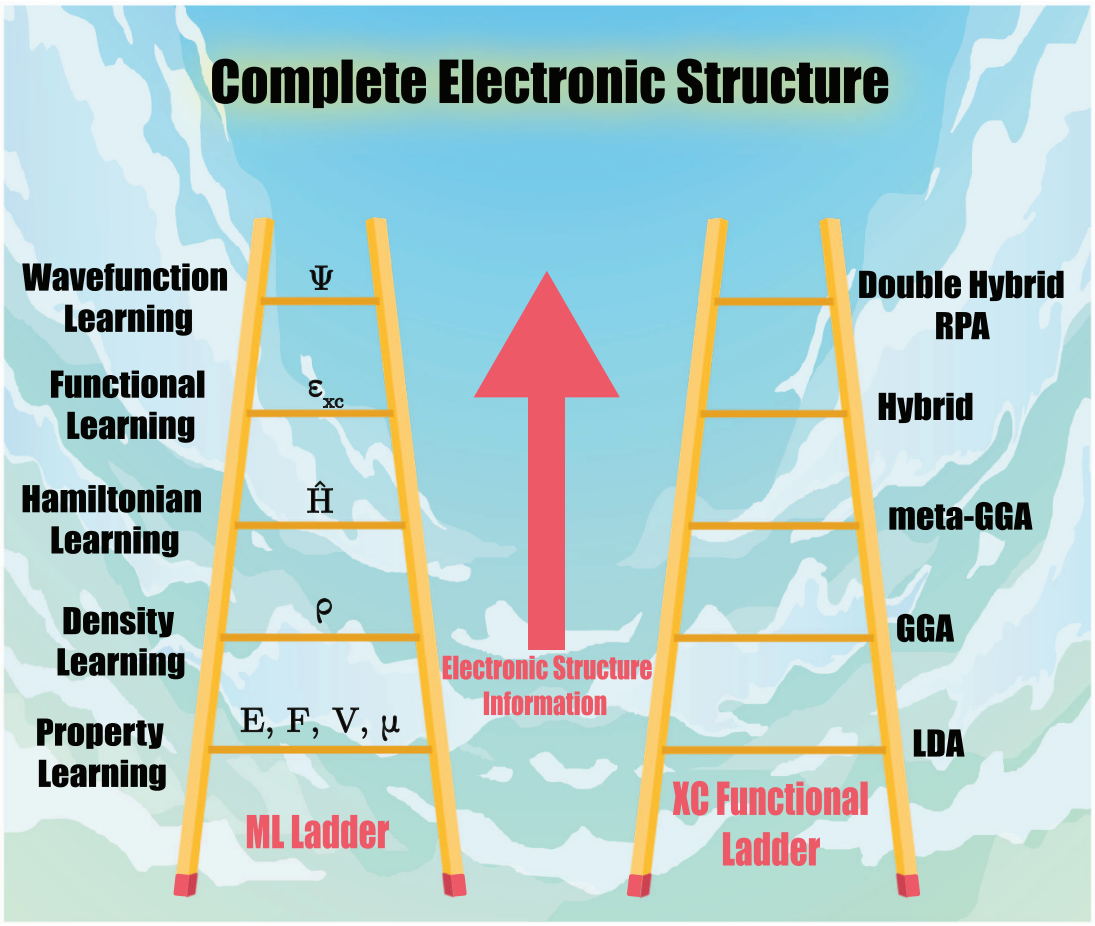

Figure 1. Right: commonly known Jacob's ladder with each rung representing a more accurate representation of the XC functional. Left: Proposed ML for electronic structure ladder in which each rung is an increase in electronic structure information that the ML model learns, with the ultimate goal of obtaining a representation of the exact electronic structure.

parameterizing XC functionals $E_{xc}[\rho]$ or related kernel functions with NNs, allowing data-driven discovery of functional forms while maintaining essential physical constraints. Similarly, the variational principle allows neural quantum states (NQSs) to represent $|\Psi\rangle$ directly through NN architectures in an unsupervised manner, as the Hamiltonian itself provides the optimization objective through the variational principle; minimize the system energy. In light of the computational expense required to generate very accurate data, the ability of networks to learn in an unsupervised fashion therefore further highlights the potential and importance of ML electronic structure.

These conditions exemplify the synergy between ML and electronic structure learning and provide the roadmap for a shift from purely physics-based calculations to hybrid methods that combine physical constraints with data-driven learning. Therefore, this review examines the theoretical foundations and practical implementations of such methods. Inspired by Jacob's ladder for XC functionals, we have constructed an 'ML Jacob's ladder' in which increasing electronic structure information is learned by models at each rung (figure 1). Climbing it thus leads to a complete representation of electronic structure, from which any downstream property can be derived. Many complementary reviews focus on specific rungs of the ladder [34–38], and thus we aim to provide a holistic and updated view of the rapidly growing field. We will structure this review by descending this ladder, going from learning the complete wavefunction to the charge density. Furthermore, we aim to focus on ML models that can be applied to real molecular and material systems, although we recognize and mention that several of these models were inspired by work on 'toy' physical systems.

Therefore, our outline is as follows:

(i) **Learning the wavefunction**, where unsupervised NNs serve as variational *ansätze* capable of representing highly correlated many-electron states.

(ii) **Learning the exchange–correlation functional**, enabling data-driven construction of the central object in KS DFT.

(iii) **Learning the Hamiltonian/Fock matrix**, which provides a direct route to bypassing costly self-consistent-field (SCF) iterations and enables efficient access to energies, forces, spectra, and response properties.

(iv) **Learning the electron density**, which also helps to circumvent SCF iterations and serves as a foundation from which many observables can be derived.

By exploring these four axes, we highlight how the landscape of electronic-structure theory is being reshaped by various ML architectures and methods, pointing toward a future in which accurate, scalable, interpretable, and general *ab initio* predictions are made accessible.

2. Theoretical background

To introduce the reader to concepts in both QC and ML, we include a short theoretical background to the works detailed in this review. Further details should be sought in more specialized QC [39–41] and ML [42–44] reviews.

2.1. QC

The basis behind such techniques is electronic structure theory, which revolves around the many-electron Schrödinger equation (S.E.), usually under the Born-Oppenheimer (BO) approximation:

$$\hat{H}\Psi\left(\mathbf{r}_1,\ldots,\mathbf{r}_{N_e}\right) = E\Psi\left(\mathbf{r}_1,\ldots,\mathbf{r}_{N_e}\right), \tag{1}$$

where $\hat{H}$ is the electronic Hamiltonian for N_e electrons. In order to solve this equation, several approximate methods have been developed, leading to a tradeoff between the accuracy of approximations and the computational scaling.

Indeed, an early ansatz is the Hartree–Fock (HF) method, which approximates (1) as a single Slater determinant. The Slater determinant is made of one-electron spin-orbitals and is a logical choice to represent a wavefunction since it ensures that the Pauli exclusion principle is respected. For N_e electrons occupying spin-orbitals $\{\phi_1,\phi_2,\ldots,\phi_{N_e}\}$, the Slater determinant, and thus the HF wavefunction is defined as:

$$\Psi_{\mathrm{HF}}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right) = \frac{1}{\sqrt{N_e!}}\begin{vmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_N(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_{N_e}) & \phi_2(\mathbf{x}_{N_e}) & \cdots & \phi_N(\mathbf{x}_{N_e}) \end{vmatrix}. \tag{2}$$

In order to best represent the system with this ansatz, Ψ_{HF} is optimized using the variational principle. This states that the most representative/best ground-state wavefunction is obtained by minimizing the Rayleigh coefficient (energy expectation value $\langle E\rangle$) since:

$$E_0 \leqslant \frac{\langle\Psi_{\mathrm{trial}}|\hat{H}|\Psi_{\mathrm{trial}}\rangle}{\langle\Psi_{\mathrm{trial}}|\Psi_{\mathrm{trial}}\rangle}, \tag{3}$$

where Ψ_{trial} is the tested/evaluated trial wavefunction and E_0 is the exact ground-state energy of the system. However, by only considering the wavefunction as a single antisymmetric product of one-electron orbitals, electron correlation beyond the mean-field level is neglected, therefore making it difficult to achieve 'chemical accuracy' for more complex systems. Nevertheless, due to its relatively simple protocol, it scales nominally with the number of basis functions as $\mathcal{O}(N^4)$ [45].

A conceptually simple extension to HF and its accuracy is FCI, where the wavefunction is expanded in a complete basis of Slater determinants. This can lead to exact solutions, but the method scales exponentially with N, illustrating the tradeoff between modeling systems at a certain level of accuracy, and being able to model larger systems. As such, methods that split the two extremes are more common in practice.

Indeed, a popular approach is CC [39], which expresses the wavefunction as an exponential of excitation operators acting on a reference determinant $|\Phi_0\rangle$:

$$|\Psi_{\mathrm{CC}}\rangle = \mathrm{e}^{\hat{T}}|\Phi_0\rangle \tag{4}$$

where $\hat{T} = \hat{T}_1 + \hat{T}_2 + \cdots$ includes single, double, and higher excitations. The most popular CC method uses the Single, Double, and pertubative Triple excitations: CCSD(T). The inclusion of these excitations allows an accurate, size-extensive treatment of electron correlation. Therefore, it is commonly dubbed the 'gold-standard' method due to its high accuracy, often reaching 'chemical accuracy' for ground state energies. It scales as $\mathcal{O}(N^7)$ with the number of particles, which, although steep, is much better than FCI, and with the progress in modern computing power/efficacy, it is usable for many industrially relevant systems.

Finally, DFT is arguably the most widely used QC method, particularly for large molecular systems. It is underpinned by a fundamentally different theoretical framework which stems from key principles established by Hohenberg and Kohn [46]. The first Hohenberg–Kohn (HK) theorem states that the ground-state electron density uniquely maps to the external potential, and the second HK theorem that there exists an energy functional $E[\rho]$ such that the total energy is minimized by the true ground-state density, according to the variational principle. However, the exact form of $E[\rho]$ is unknown. Therefore, Kohn and Sham [47] introduced an auxiliary system of non-interacting electrons that reproduces the same ground-state density as the interacting system. The corresponding orbitals $\{\phi_j\}$ satisfy

$$\left[-\frac{1}{2}\nabla^2 + v_s(\mathbf{r})\right]\phi_j(\mathbf{r}) = \epsilon_j\phi_j(\mathbf{r}), \tag{5}$$

where the KS potential $v_s(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}[n](\mathbf{r}) + v_{\text{xc}}[n](\mathbf{r})$ includes the external potential, classical Hartree, and XC contributions, respectively. In practice, the KS total energy functional is expressed as

$$E_{\text{total}}[n] = T_s[n] + E_H[n] + E_{\text{xc}}[n] + E^{\text{ei}}[n] + E^{\text{ii}}, \tag{6}$$

where $T_s[n]$ is the kinetic energy of the non-interacting reference system, $E_{\text{H}}[n]$ is the Hartree (electron-electron) term, $E^{\text{ei}}[n]$ is the electron-ion interaction energy, and E^{ii} is the classical ion-ion repulsion, and are known. However $E_{\text{xc}}[n]$, the XC functional, takes into account the errors from non-interacting kinetic energy terms and classical treatment of electron-electron interactions, but is not known completely. This unknown has been the subject of a lot of research because, according to the HK theorems, KS-DFT is exact in nature, yet in practice its accuracy is limited to the approximation made for the XC functional. The various approximations to this functional can be organized by accuracy, a hierarchy commonly called 'Jacob's ladder'. These levels are shown in figure 1. Nevertheless, with its self-consistent loop, DFT often scales as $\mathcal{O}(N^3)$ with the number of particles [48], leading to its popularity. For the study of excited states, time-dependent density functional theory (TDDFT) extends DFT to time-dependent (TD) systems through the Runge–Gross theorem [49], an extension of the HK theorems which establishes a one-to-one correspondence between the TD density $\rho(\mathbf{r}, t)$ and the TD potential $v_{\text{ext}}(\mathbf{r}, t)$.

2.2. ML

We briefly distinguish the two learning settings recurring in this review (and developed further in subsequent sections). *Supervised learning* fits models to labeled input–output pairs—for example, molecular or crystalline structures paired with energies, Hamiltonian matrix elements, or electron densities from reference *ab initio* calculations—by minimizing a loss against those labels. *Unsupervised learning* optimizes objectives without paired outputs, as in variational Monte Carlo (VMC) for NQSs (section 3), where the energy expectation supplies the training signal. Many methods later in the review combine both ideas (e.g. physics-informed losses alongside supervised fits).

2.2.1. Kernel methods

Kernel methods [24] have been central to early ML applications in QC. Given training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, KRR predicts at a test point $\mathbf{x}_*$ via:

$$\hat{f}(\mathbf{x}_*) = \mathbf{k}_*^T(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{7}$$

where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel (Gram) matrix, $\mathbf{k}_* = [k(\mathbf{x}_*, \mathbf{x}_i)]_{i=1}^N$, and $\lambda > 0$ is a regularization parameter. Molecular applications usually represent each atomic environment with a symmetry-aware descriptor. SOAP (smooth overlap of atomic positions) [50] is representative: it respects permutations among equivalent atoms and retains a local cutoff. Chemistry papers sometimes describe SOAP itself as a "kernel," but it is more precisely a fixed featurization. The kernel k in equation (7) is specified separately and defines similarity between those feature vectors in the usual way.

Equation (7) is often presented within the Gaussian process (GP) framework, where it arises as the posterior mean of a probabilistic model that additionally provides predictive variance:

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{k}_* \tag{8}$$

with λ identified as noise variance σ_n^2. However, interpreting σ_*^2 as a calibrated uncertainty estimate requires that the kernel correspond to a valid covariance function of a stochastic process—i.e. that it is bounded and positive semi-definite in a way consistent with the GP prior assumptions. When stationary kernels such as the Matérn family are used on SOAP or related descriptor vectors—the usual pairing in GP/KRR models built on these featurizations—this interpretation is well-justified [50]. However, polynomial (dot-product) kernels are also widely employed with the same descriptors, and these are unbounded and do not define a proper stationary process, rendering the uncertainty estimates unreliable as confidence intervals. In such cases, KRR is better understood as regularized linear regression in the feature space induced by the kernel, without attaching probabilistic meaning to the predictive variance.

Despite this caveat, kernel methods remain valuable for small-data regimes ($N \lesssim 10^4$) commonly encountered in high-accuracy QC, where their closed-form solution avoids the optimization challenges of NNs. Exact kernel and Gaussian-process regression require an $\mathcal{O}(N^3)$ decomposition or inversion of the $N \times N$ Gram matrix [24, 42], which in practice caps feasible training set sizes at roughly $N \sim 10^4$ on typical hardware before time and memory become prohibitive. For substantially larger datasets, NNs trained with stochastic optimization process mini-batches of size $m \ll N$ per step, so that the cost per epoch scales as $\mathcal{O}(N)$ in N at fixed architecture and batch size, and they therefore typically offer better computational efficiency in the large-N regime [43, 44]. Their principal limitation for equation (7) remains the $\mathcal{O}(N^3)$ linear solve, which motivates the NN architectures discussed below.

2.2.2. Multilayer perceptrons (MLPs)

The MLP, a standard feed-forward NN architecture [51], consists of multiple layers, each applying an affine transformation followed by a nonlinear activation function:

$$\mathbf{h}^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right) \tag{9}$$

where $\mathbf{h}^{(l)} \in \mathbb{R}^{d_l}$ is the hidden representation at layer l, $\mathbf{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$ is the weight matrix, $\mathbf{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$ is the bias vector, and σ is a nonlinear activation (commonly ReLU, tanh, or SiLU/swish). While the universal approximation theorem guarantees that even sufficiently wide one-layer MLPs can approximate any continuous function on a compact domain as mentioned above, it does not mean that a given function can be learned efficiently and, often, obtaining enough data to allow the MLP to learn such complex functions is the bottleneck. This is why the field has moved toward architectures with stronger inductive biases—graph NNs, equivariant networks, and transformers—that encode known physical structure (molecular topology, spatial locality, rotational symmetry) directly into the model. By building in constraints that an MLP would have to discover from data, these architectures effectively reduce the dimensionality of the function space that must be searched, achieving better accuracy with fewer parameters and, crucially, less training data. The following sections describe these architectures in order of increasing geometric sophistication.

Despite these limitation, MLPs remain widely used for learning atomic contributions in local chemical environments (e.g. in the Behler–Parrinello NN [52]) or as components within more complex architectures where they process aggregated features.

2.2.3. Graph neural networks (GNNs) and message passing

Molecular systems naturally lend themselves to graphical representations $G = (V, E)$, where atoms form the node set V and chemical bonds (or spatial proximity) define the edge set E. Often, GNNs [53] exploit this representation by iteratively updating node features through message passing—aggregating information from neighboring nodes while respecting the graph topology.

The general message passing framework, formalized by Gilmer *et al* [54], updates node features $\mathbf{h}_i$ for atom i through two stages

$$\mathbf{m}_i^{(t+1)} = \sum_{j \in \mathcal{N}(i)} M\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)}, \mathbf{e}_{ij}\right) \tag{10}$$

$$\mathbf{h}_i^{(t+1)} = U\left(\mathbf{h}_i^{(t)}, \mathbf{m}_i^{(t+1)}\right) \tag{11}$$

where $\mathbf{h}_i^{(t)} \in \mathbb{R}^d$ represents the feature vector for atom i at iteration t, $\mathcal{N}(i)$ denotes the set of neighboring atoms of node i (defined by bonding or a distance cutoff), $\mathbf{m}_i^{(t+1)}$ is the aggregated message received by node i, $\mathbf{e}_{ij} \in \mathbb{R}^{d_e}$ represents edge features (typically including interatomic distance $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$, and sometimes bond type or directional information), M is a learnable message function (often implemented as an MLP), and U is a learnable update function (also typically an MLP or GRU). The summation operator $\sum$ ensures permutation invariance—the order of neighbors does not affect the result.

After T message passing iterations, the receptive field of each node extends to atoms within T bonds or distance steps. Global molecular properties are obtained by aggregating node features: $\mathbf{h}_{\text{mol}} = \rho(\{\mathbf{h}_i^{(T)}\}_{i \in V})$, where ρ is a permutation-invariant readout function such as summation or mean pooling.

Architectures like SchNet [55] introduce continuous-filter convolutions that weight messages by Gaussian basis expansions of interatomic distances. DimeNet [56] extends this by incorporating directional information through bond angles θ_{ijk} between atom triplets. PaiNN (polarizable atom interaction NN) [57] maintains both scalar features $\mathbf{s}_i$ (invariant to rotations) and vector features $\mathbf{v}_i$ (equivariant under rotations), enabling the network to represent directional properties like dipole moments while maintaining appropriate geometric transformation properties.

A practical consideration often underemphasized in the literature is the choice of cutoff radius and graph construction, which implicitly encodes assumptions about locality. Too small a cutoff misses relevant interactions (e.g. long-range electrostatics or dispersion); too large a cutoff inflates computational cost quadratically and may introduce noise. For condensed-phase or charged systems, the strictly local message-passing paradigm can be fundamentally limiting, motivating the long-range approaches and transformer architectures discussed below.

2.2.4. Transformer architectures

Originally developed for natural language processing (NLP) [58], transformers have recently been adapted for molecular systems. The core innovation is the self-attention mechanism, which computes pairwise interactions between all atoms simultaneously rather than relying on iterative message passing over local neighborhoods.

For a molecule with N atoms, each atom i is first embedded into three representations: a query $\mathbf{q}_i = \mathbf{W}_Q \mathbf{h}_i$, key $\mathbf{k}_i = \mathbf{W}_K \mathbf{h}_i$, and value $\mathbf{v}_i = \mathbf{W}_V \mathbf{h}_i$, where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_k \times d}$ are learnable projection matrices and d_k is the key/query dimension. The attention mechanism computes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{12}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_k}$ are matrices stacking all queries, keys, and values respectively, $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{N \times N}$ computes dot products measuring similarity between all atom pairs, the scaling factor $1/\sqrt{d_k}$ prevents gradients from vanishing for large d_k, and softmax is applied row-wise to produce attention weights $\alpha_{ij} \in [0, 1]$ with $\sum_j \alpha_{ij} = 1$. The output for atom i is thus $\mathbf{h}_i^{\text{out}} = \sum_j \alpha_{ij} \mathbf{v}_j$, a weighted combination of all atom features.

Multi-head attention runs H parallel attention operations with different projection matrices and concatenates results: $\text{MultiHead} = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\mathbf{W}^O$, allowing the model to attend to different aspects of atomic relationships simultaneously.

In the *dense* self-attention used in standard transformers, each node can in principle attend to all other nodes within a layer, yielding a global interaction pattern at $O(N^2)$ cost. However, vanilla transformers lack inherent awareness of 3D geometry, and many molecular adaptations either inject geometric priors into attention or restrict interactions to a molecular graph/neighborhood. Common strategies include: (1) distance- or geometry-dependent attention biases that modify α_{ij} based on r_{ij}, (2) geometric/relative positional embeddings, or (3) hybrid designs that combine attention with geometric message passing. The Molecule Attention Transformer [59] augments the attention matrix with distance and graph-structure terms, while Equiformer/EquiformerV2 [60] implement an *equivariant graph attention* that aggregates messages over neighbors $j \in \mathcal{N}(i)$ rather than unconstrained all-pairs attention.

2.2.5. Equivariant neural networks (E3NN)

A fundamental requirement for molecular property prediction is appropriate transformation under spatial symmetries [61]. Molecular energies must be invariant (unchanged) under rotations and translations, while forces and dipole moments must be equivariant (transform predictably). Formally, for any rotation $R \in \text{SO}(3)$ and translation $\mathbf{t}$, a function f is equivariant if $f(R\mathbf{r} + \mathbf{t}) = Rf(\mathbf{r})$.

E3NN (Euclidean neural networks) [62] achieve equivariance by construction through operating on irreducible representations (irreps) of the rotation group SO(3). Rather than representing atom features

as simple vectors in $\mathbb{R}^d$, features are decomposed into spherical harmonic components with well-defined transformation properties.

A feature of type l (where $l = 0, 1, 2, \ldots$ corresponds to scalar, vector, tensor, ... representations) transforms as a spherical harmonic Y_l^m with multiplicity (degeneracy) $2l+1$ (for $m = -l, \ldots, +l$). The feature representation for atom i becomes:

$$\mathbf{h}_i = \bigoplus_{l=0}^{L_{\max}} \mathbf{h}_i^{(l)} \tag{13}$$

where $\mathbf{h}_i^{(l)} \in \mathbb{R}^{m_l \times (2l+1)}$ contains m_l channels of type-l features (m_l is the multiplicity or number of l-type features), and $\bigoplus$ denotes direct sum (concatenation of different irrep types). For example, $\mathbf{h}_i^{(0)}$ contains scalar features (invariants like atomic number, partial charge), $\mathbf{h}_i^{(1)}$ contains vector features (equivariants like displacement vectors, forces), and $\mathbf{h}_i^{(2)}$ contains second-order tensor features.

The key operation in E3NN is the tensor product between irreps, which combines features while preserving equivariance:

$$\mathbf{h}^{(l_1)} \otimes \mathbf{h}^{(l_2)} = \bigoplus_{l_3=|l_1-l_2|}^{l_1+l_2} C_{l_1,l_2}^{l_3} \mathbf{h}^{(l_3)} \tag{14}$$

where $C_{l_1,l_2}^{l_3}$ are Clebsch–Gordan coefficients that define how spherical harmonics couple under rotation group representation theory. This ensures the output transforms correctly—for instance, combining two vectors ($l = 1$) produces a scalar ($l = 0$), a vector ($l = 1$), and a second-rank tensor ($l = 2$).

In practice, E3NN message passing operates as:

1. **Edge features**: compute the relative position vector $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ and its spherical harmonic expansion $Y_l^m(\hat{\mathbf{r}}_{ij})$, where $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/\|\mathbf{r}_{ij}\|$ is the unit direction vector.
2. **Tensor product**: Combine node features with edge spherical harmonics via the tensor product, weighted by learnable radial functions that depend on distance r_{ij}.
3. **Message aggregation**: sum messages from all neighbors, where the spherical harmonic representation ensures rotational equivariance is maintained.

Architectures like NequIP [63] explicitly construct features as irreps and use tensor products for all operations. multi-atomic cluster expansion [64] extends this by incorporating higher-order body interactions through products of pairwise features, achieving many-body equivariance. These architectures achieve state-of-the-art accuracy on molecular force fields and properties requiring geometric precision (forces, stress tensors, polarizabilities), though the spherical harmonic transformations and Clebsch–Gordan coupling incur ∼2–5× computational cost compared to simpler GNNs like SchNet.

The fundamental advantage of E3NN is thus guaranteed equivariance by mathematical construction rather than approximate equivariance learned through data augmentation, leading to better generalization with less training data and more physically meaningful learned representations.

Nevertheless there is a computational trade-off : the Clebsch–Gordan tensor products and higher-order spherical harmonic operations incur overhead compared to invariant GNNs such as SchNet. Whether this cost is justified depends on the application: for scalar energy prediction with abundant training data, simpler invariant models may suffice, but for tensorial properties (polarizability, stress, NMR shielding) or data-scarce settings, the inductive bias of exact equivariance is often decisive.

3. Learning wavefunctions

In this section, we will review the development of ML techniques for the unsupervised learning of the wavefunction, dubbed NQSs.

3.1. Introduction and early models

Essential to the appearance and function all NQSs is the VMC method [65]. As the name suggests, this method utilizes the variational principle (3) to approximate the ground state energy by minimizing the expectation value over a trial wavefunction.

However, this involves evaluating a $3N$-dimensional integral over all electronic coordinates, which is infeasible deterministically. VMC, instead, uses the Monte Carlo (MC) algorithm to sample electronic configurations from the probability density $\pi(\mathbf{R})$:

$$\pi(\mathbf{R}) = \frac{|\Psi_\theta(\mathbf{R})|^2}{\int d\mathbf{R}\,|\Psi_\theta(\mathbf{R})|^2} \tag{15}$$

(3) thus becomes:

$$\langle E\rangle_\theta = \frac{\langle\Psi_\theta|\hat{H}|\Psi_\theta\rangle}{\langle\Psi_\theta|\Psi_\theta\rangle} = \int \mathrm{d}\mathbf{R}\,\pi_\theta(\mathbf{R})\,E_L(\mathbf{R}) \tag{16}$$

where E_L is the local energy and is defined as:

$$E_L(\mathbf{R}) \equiv \frac{\hat{H}\Psi_\theta(\mathbf{R})}{\Psi_\theta(\mathbf{R})}. \tag{17}$$

Therefore, using walkers to stochastically estimate $\langle E\rangle_\theta$, then taking enough samples M of the electronic structure $\mathbf{R}$, we get:

$$\langle E\rangle_\theta = \frac{1}{M}\sum_{k=1}^{M} E_L\left(\mathbf{R}^{(k)}\right). \tag{18}$$

A common trial wavefunction has a part represented by product of separate ↑ and ↓ Slater determinants (since $\hat{H}$ is spin independent) with single particle orbitals to ensure antisymmetry under exchange of same-spin orbitals. This product is multiplied by a symmetric Jastrow factor [66, 67] that accounts for the Kato cusp conditions by enforcing the correct short-range behavior of the wavefunction as electrons approach each other or a nucleus. This combination captures both exchange and correlation effects and gives rise to a trial wavefunction of the form:

$$\Psi_T(\mathbf{R}) = \mathrm{e}^{J(\mathbf{R})}\,\det\left[\phi_i^{\uparrow}\left(\mathbf{r}_j^{\uparrow}\right)\right]\det\left[\phi_k^{\downarrow}\left(\mathbf{r}_l^{\downarrow}\right)\right], \tag{19}$$

Nevertheless, traditional VMC and other quantum Monte Carlo (QMC) methods are not as widely used as those mentioned in the introduction largely because of the complexity of the trial wavefunction required to reach accurate energies, leading to a compromise between compute needed and accuracy [68].

However, with the advances in NN performance, and the natural link between stochastic sampling of wavefunctions and the stochastic estimate of gradients popularized in ML, an NQS was presented by Carleo and Troyer [69]. They exploit the capacity of NNs to flexibly represent high-dimensional probability distributions, enabling more accurate wavefunction *ansätze* at reduced computational cost compared to traditional VMC. Carleo and Troyer showed that a restricted Boltzmann machine (RBM) [70] can learn the ground state of lattice quantum spin models to high accuracy.

Their RBM takes a configuration as input and returns a phase and amplitude of the wavefunction. This is achieved through visible nodes for the physical spin variables and hidden nodes to learn correlations. Learning is achieved through the same VMC workflow where Ψ_T in (19) becomes Ψ_θ and NN parameters are updated to minimize $\langle E\rangle_\theta$. This optimization is done through stochastic reconfiguration (SR) which ensures that the parameters are updated whilst taking the deformation caused to Ψ_θ into account to ensure stable training.

This led to high accuracy when compared with 1D, 2D Ising and Heisenberg models, and they even show promise in short time evolution simulations by projecting the time dependent S.E. on the RBM. As such, the concept of an NQS that can capture correlation and entanglement was established, which motivated further research in the field. This motivation, coupled with increasingly expressive ML architectures, led to rapid development. Particularly impactful work was the use of backflow orbitals that take electron position and correlation into account instead of using mean field Slater determinants [71, 72]. These were applied to NQSs as neural network backflow (NNB) wavefunctions that showed improved accuracy in Hubbard models [73].

Landmark models FermiNet [68] and PauliNet [74] were also released. These established NQS as viable *ansatze* for continuous electronic systems such as molecules.

Firstly PauliNet, inspired by traditional VMC, takes Slater determinants of precomputed single electron orbitals (HF), and trains a NN to parametrize the ground state wavefunction through VMC. This

is done through a combination of SchNet-inspired [55] continuous-filter convolutions of electron coordinates and deep NNs that learn the Jastrow term and incorporate backflow considerations. These NNs use the weighted adaptive moment estimation (Adam) [75] optimizer and stochastic gradient descent (since SR is very expensive for continuous systems). The continuity of the electronic structure of the system also requires that the cusp conditions be taken into account in the ansatz, which is done with fixed explicit terms. The fixed nature of these physical conditions helps to accelerate convergence, and the general architecture leads to impressive accuracy. Indeed, when summing multiple learned determinants to better account for electron correlation, PauliNet achieved 'chemical accuracy' (< 1 kcal mol^{-1}) across small molecules and atoms up to Ne. It also outperformed CCSD(T) on cyclobutadiene, a molecule with multi-reference (MR) character, showing its ability to learn electron correlation whilst scaling only as $\mathcal{O}(N_e^4)$ with the number of electrons.

By contrast, FermiNet (figure 2) does not explicitly enforce cusp conditions and, instead, lets the NN learn these. In order to achieve this, the network additionally takes features of electron pairs as input, including the distances between these $|\mathbf{r_i} - \mathbf{r_j}|$ and relative displacements $(\mathbf{r_i} - \mathbf{r_j})$, allowing the permutation equivariant NN to learn the cusp conditions and removing the need for the Jastrow factor. Moreover, in the final layer, an exponentially decaying envelope function ensures the wavefunction decays to 0 as electrons get further from nuclei. Their *anstatz* is:

$$\Psi_\theta(\mathbf{R}) = \sum_{k=1}^{K} \det\left[\phi_{i\uparrow}^{(k)}\left(\mathbf{r}_j^{\uparrow}, \mathbf{R}\right)\right] \det\left[\phi_{i\downarrow}^{(k)}\left(\mathbf{r}_j^{\downarrow}, \mathbf{R}\right)\right], \quad \mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_N). \tag{20}$$

$$\phi_{i\sigma}^{(k)}\left(\mathbf{r}_{j\sigma}, \mathbf{R}\right) = f_{i\sigma}^{(k)}\Big(\mathbf{r}_{j\sigma}, \left\{\mathbf{r}_{\ell\sigma} - \mathbf{r}_{j\sigma}\right\}_{\ell\neq j}, \left\{\mathbf{r}_{\ell\bar{\sigma}} - \mathbf{r}_{j\sigma}\right\}_{\ell}, \left\{\mathbf{R}_I - \mathbf{r}_{j\sigma}\right\}_I\Big), \quad \sigma \in \{\uparrow, \downarrow\} \tag{21}$$

where $f_{i\sigma}^{(k)}$ denotes the NN that maps the input features. The inputs to $f_{i\sigma}^{(k)}$ include: $\mathbf{r}_{j\sigma}$ (electron position), $\{\mathbf{r}_{\ell\sigma} - \mathbf{r}_{j\sigma}\}_{\ell\neq j}$ (relative same-spin displacements), $\{\mathbf{r}_{\ell\bar{\sigma}} - \mathbf{r}_{j\sigma}\}_{\ell}$ (opposite-spin displacements), and $\{\mathbf{R}_I - \mathbf{r}_{j\sigma}\}_I$ (electron-nucleus vectors).

This ansatz has $\approx$ 700 000 parameters on average, which are optimized using a modified version of Kronecker-factored approximate curvature [76]. Although this led to better accuracy than PauliNet (and pretraining against HF orbitals is possible), a 30 electron system takes 1 month on 16 GPUs. This was later improved with an optimization of the wavefunction form and better GPU utilization [77]. The *ansatze* mentioned above have also been used in diffusion Monte Carlo (DMC), which shows their transferability to other MC methods [78].

These developments and further considerations have also been comprehensively reviewed by Hermann *et al* [37]. As such, the present review will focus on more recent advances that have sought to broaden the applicability of NQSs. In particular, we highlight progress in refining foundational architectures such as FermiNet and PauliNet, extending their scalability to larger systems, incorporating transformer-based representations inspired by general ML research, expanding their scope beyond the electronic ground state, and making NQSs generalizable.

3.2. Architectural refinements and expansions to larger systems

With the aim to render FermiNet accessible to larger atoms and systems, Li *et al* [79] implemented an effective core potential (ECP), where core electrons are accounted for using ccECP, allowing FermiNet to optimize the wavefunction depending only on valence electrons. This allowed ground state energies of 3*d* transition state metals to be simulated 10–40 mHa from CCSD(T)/CBS. Similarly, to increase the size of systems that can be simulated by FermiNet, Gérard *et al* [80] implemented a SchNet-like convolutional layer embedding (as in PauliNet), which sped up calculations and increased accuracy along with their other modifications to FermiNet. Indeed, after pretraining to initialize the model with HF orbitals, glycine was able to be simulated at a higher accuracy than multireference configuration interation-F12(Q) using the new NQS ansatz.

Another expansion is the extension of neural-network wavefunction methods to periodic systems. Yoshioka *et al* [81] formulate the many-electron problem in periodic solids via a second-quantized fermionic Hamiltonian expressed in an orthonormal crystalline-orbital basis (from HF). This allows them to compute potential energy curves of graphene and rocksalt-structured LiH with CCSD(T) and FCI-like accuracy.

Nevertheless, an alternative approach to treating periodicity arises in real-space neural wavefunctions, where periodic boundary conditions must be imposed explicitly rather than being embedded in the basis. Pescia *et al* [82] establish PBCs for bosonic systems by mapping the position of a particle to a periodic Fourier basis:

$$\mathbf{x}_i \mapsto \left(\sin\left(\frac{2\pi k}{L}\mathbf{x}_i\right), \cos\left(\frac{2\pi k}{L}\mathbf{x}_i\right)\right)_{k=1}^{K} = \mathbf{x}_i^{(K)} \tag{22}$$

where L is the representative length of the periodic lattice.

As such, Cassella *et al* [83] expand on this and establish PBCs for fermionic systems, which require the inclusion of long-range Coulomb interactions, using Ewald summation [84]. These adaptations are designed such that periodicity is also respected for non-cubic boundaries, and are applied to FermiNet to simulate homogeneous electron gas (HEG). Likewise, Wilson *et al* [85] dedicate a specific ansatz to enforce periodicity named Wavefunction *Ansatze* but Periodic (WAP)-Net to accurately simulate the HEG. An interesting approach that enforces PBCs, but differs from previous examples, is Pescia *et al*'s [86] use of an equivariant message passing neural network (MPNN) and a particle attention mechanism to generate backflow orbitals. This increases the expressivity of the backflow transformations and reduces the number of optimizable parameters in the wavefunction, allowing more electrons in a HEG to be simulated.

However, within the scope of real solids, Li *et al* [87] embed translational symmetry directly within the orbital construction itself. Indeed, they include a Bloch-phase factor in the Slater determinants of their ansatz such that:

$$\Psi(\mathbf{r}) = \det{}_{\uparrow}\left[e^{i\mathbf{k}\cdot\mathbf{r}_{\uparrow}} u_{\mathrm{mol}}^{\uparrow}(d)\right] \det{}_{\downarrow}\left[e^{i\mathbf{k}\cdot\mathbf{r}_{\downarrow}} u_{\mathrm{mol}}^{\downarrow}(d)\right], \tag{23}$$

where $u_{\mathrm{mol}}^{\sigma}(d)$ are neural orbitals that depend on a periodic distance feature $d(\mathbf{r})$ defined by the crystal lattice, which encodes the necessary correlation and physical conditions. The inclusion of Bloch-phase factors $e^{i\mathbf{k}\cdot\mathbf{r}}$ in (23) ensures that $\Psi(\mathbf{r}+\mathbf{L}) = e^{i\mathbf{k}\cdot\mathbf{L}}\Psi(\mathbf{r})$, thereby satisfying the Born-von Kármán boundary condition analytically, allowing it to describe real crystalline solids in continuous space. This is exemplified by its accurate results when applied to hydrogen chains, graphene, and LiH crystals (1, 2, and 3D systems, respectively). Impressively, with HF calculated finite-cell corrections, the cohesive energy of a 108 electron supercell of LiH is only 0.011 eV from the experimental value, showing the ability of NQS to be used to simulate real periodic systems at high accuracy.

Therefore, with the growth in use cases and the gained interest in the method, algorithmic refinements emerged in order to aid the continuing development. To facilitate broader adoption, open-source suites were created that unify NQSs in modular and extensible code frameworks, such as NetKet [88], DeepQMC [89], and QMCTorch [90].

Although all the methods so far have relied on expensive MCMC sampling, Sharir *et al* [91] implemented autoregressive sampling for NQSs coined neural autoregressive quantum states, constituting an algorithmic change. These are defined by a product of conditional wavefunctions:

$$\Psi(\mathbf{s}) = \prod_{i=1}^{N} \psi_\theta(s_i \,|\, s_1, s_2, \ldots, s_{i-1}), \tag{24}$$

where normalization is enforced for Ψ_θ, and s_i are degrees of freedom the model can have. Thus, Ψ_θ are the outputs of NNs that give the probability of s_i to take a specific value, conditional on $s_1, s_2, \ldots, s_{i-1}$ (input to NN). This approach allows them to train a 20 layer convolutional neural network (CNN) with $\approx$1 million parameters with this independent sampling, showing great applicability for discretized systems, particularly in 2D. However, Barrett *et al* [92] then managed to apply it to molecules through second-quantized discretization of spin orbitals, where they obtained similar results to FCI on real molecules with up to 20 electrons, showing the potential of this sampling method for second quantized systems. Nevertheless, this sampling is reliant on discretization and is thus restricted to these second-quantized systems.

More recent architectural modifications include Li *et al*'s LapNet [93], in which the bottleneck of calculating the NN's Laplacian is greatly sped up. Indeed, a large portion of the computational expense for optimizing the NQS is in calculating the Laplacian by first determining the Hessian matrix. In this novel method, the Laplacian is able to be exactly calculated without building the Hessian, dubbing the process 'Forward Laplacian', which was subsequently made applicable to other NQSs. Moreover, they make use of the sparcity in block diagonal matrices that LapNet and other NNs compute, to speed up calculations

further by dropping the zeroes. As such, they achieve orders of magnitude more efficient calculations than FermiNet whilst being more accurate (due to the inclusion of transformer heads discussed in later parts). This enables LapNet to achieve near chemical accuracy relative to CCSD(T)/CBS accuracy for the uracil dimer, a 116 electron system.

Lastly, a key development in a similar regard is Scherbela *et al*'s finite range embeddings (FiRE) [94]. As the name suggests, the electron embeddings from previous NQSs (like FermiNet) are modified such that, instead of depending on all electron positions, the electron-electron field of vision is truncated:

$$h_i(\mathbf{r}) = h\left(\mathbf{r}_i, \left\{\mathbf{r}_j \mid j \neq i \wedge \|\mathbf{r}_i - \mathbf{r}_j\| \leqslant c\right\}\right). \tag{25}$$

Thus, h_i only 'sees' neighbors within radius c. This sparcifies the Jacobians obtained, further speeding up Laplacian calculation, and also allows for low-rank updates to be performed to the wavefunction. In addition, to ensure the long-range correlation is maintained even with the cutoff, a global attention Jastrow is included in the model's Jastrow factor. The combination of these properties and improved MCMC sampling through global single-electron jumps allows the FiRE model to outperform CCSD(T)/CBS on a 180 electron molecule.

As such, this subsection has summarized the numerous recent developments that optimize and build on previous *ansatze*. This thus illustrates the potential NQSs to become viable for large(r) systems with more electrons, via ECP, PBCs, or architectural refinements as seen in recent trends.

3.3. Transformers and attention mechanisms

As alluded to in certain examples of the previous section, the emergence of transformer and attention-based architectures in NLP [58] to determine the relation between words has had a significant effect on NQSs. Indeed, von Glehn *et al* [95] applied this concept to represent the correlation between electrons, creating Psiformer (figure 2). They employ self-attention layers which are permutation invariant by default, allowing the embeddings of electrons to be updated as:

$$\mathbf{A}_i = \sum_j \mathrm{softmax}_j\left(\frac{(\mathbf{q}_i)^\top \mathbf{k}_j}{\sqrt{d_k}}\right)\mathbf{v}_j, \tag{26}$$

Where $\mathbf{A}_i$ is the updated embedding for electron i after attending to all other electrons and the vectors seen in (12) correspond to electrons i and j, accordingly.

These do not take pairwise electron features as inputs and thus a Jastrow factor is included to account for the electron-electron cusps. Another deviation from FermiNet is the logarithmic rescaling of electron-nuclei distances for more stable training, and mapping spin $\uparrow$ to 1 and $\downarrow$ to -1. This type of architecture allows the ansatz to be minimized with fewer parameters than FermiNet whilst maintaining high accuracy predictions for small molecules. However, impressively, due to the learned correlations and lesser number of parameters, Psiformer performs better on larger molecules than FermiNet (figure 3), and was able to surpass classical DMC for the benzene dimer. Similarly, Viteritti *et al* [96] use a vision transformer-like ansatz to find accurate ground state energies for frustrated 1D Heisenberg spin systems.

These concretized the concept of using transformer-based architectures in NQS development. Therefore, Gu *et al* [97], using previous work on PBCs and an attention mechanism, learn an NQS for the 2D doped Hubbard model:

$$\begin{aligned}\hat{H} = &-t\sum_{\langle i,j\rangle,\sigma}\left(\hat{c}^\dagger_{i\sigma}\hat{c}_{j\sigma} + \hat{c}^\dagger_{j\sigma}\hat{c}_{i\sigma}\right) - t'\sum_{\langle\langle i,j\rangle\rangle,\sigma}\left(\hat{c}^\dagger_{i\sigma}\hat{c}_{j\sigma} + \hat{c}^\dagger_{j\sigma}\hat{c}_{i\sigma}\right)\\ &+ U\sum_i \hat{n}_{i\uparrow}\hat{n}_{i\downarrow} - \mu\sum_{i,\sigma}\hat{n}_{i\sigma}.\end{aligned} \tag{27}$$

Here, t and t' are the nearest- and next-nearest-neighbor hopping amplitudes, U is the on-site Coulomb repulsion, and μ is the chemical potential controlling the doping level. The operators $\hat{c}^\dagger_{i\sigma}$ and $\hat{c}_{i\sigma}$ create and annihilate an electron with spin σ on site i, and $\hat{n}_{i\sigma} = \hat{c}^\dagger_{i\sigma}\hat{c}_{i\sigma}$ is the corresponding number operator. The sums $\langle i,j\rangle$ and $\langle\langle i,j\rangle\rangle$ run over nearest- and next-nearest-neighbor pairs on the 2D lattice.

Since it is a lattice model, they introduce positional encoding that differentiates between sites and revert to a sum of spin up and spin down determinants (without a Jastrow factor since it is a lattice). The attention mechanism is used to generate the backflow orbitals which are pretrained by minimizing the difference between their orbitals and NNB orbitals. Furthermore, due to the distretization of the system, the focus of each attention head's learned correlations can be clearly deduced by visualizing the normalized attention scores. Indeed, figure 4 illustrates the specialization learned by each attention

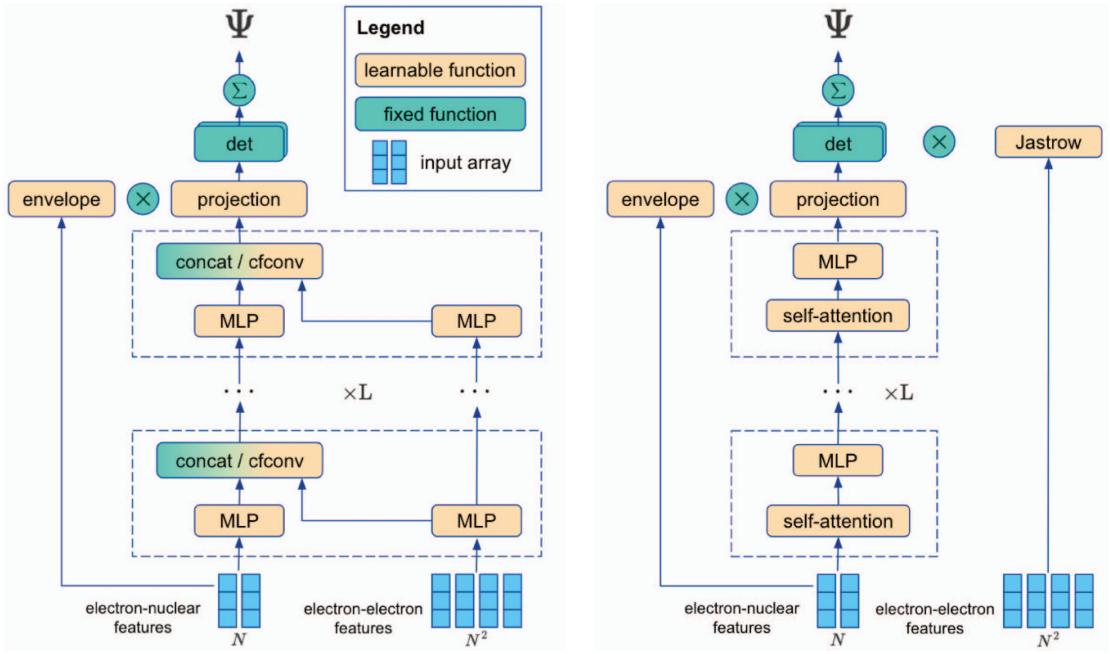

Figure 2. Left: Architecture of FermiNet and a SchNet-FermiNet hybrid. The latter's nuclear embedding stream and separate spin-dependant electron-electron streams are not pictured here. Right: Architecture of Psiformer. Illustrating the inclusion of the self-attention mechanism and different treatment of electron-electron features with the inclusion of a Jastrow factor. Reproduced from [95]. CC BY 4.0.

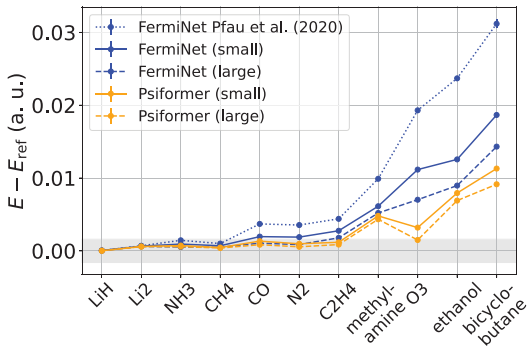

Figure 3. Results from FermiNet and Psiformer on a set of molecules compared to a computed CCSD(T)/CBS reference, where the gray region is 'chemical accuracy' compared to the baseline. The *'(small)'* is the original size of the FermiNet architecture and *'(large)'* corresponds to a model with double the number determinants and wider layers. We can see how, for the larger systems, the attention mechanism and updated architecture leads to more accurate results. Reproduced from [95]. CC BY 4.0.

head in the trained transformer. Analysis of the attention weights reveals that individual heads focus on distinct physical features of the system—such as electron hopping or local correlation environments—rather than distributing attention uniformly. This emergent specialization suggests that the model captures physically meaningful structure, and is not simply exploiting spurious correlations in the training data.

This further exemplifies a benefit over property learning as the model learns the physics through training, allowing it to achieve state of the art accuracy (99.98%) for a 12 x 12 lattice and correctly predict experimentally observed stripes in cuprates.

Attention mechanisms are also central to Geier *et al*'s [98] N_e^2 scaling with the number of electrons for simulating Fermi Liquid to Wigner Crystal phase transitions, and Teng *et al*'s [99] accurate ground state fractional quantum Hall wavefunctions.

Recently, with a focus on molecular systems and bond dissociation behavior, Forster *et al* [100] introduced another transformer-based architecture, Orbformer. In this approach, nuclear coordinates are first processed by an MPNN to produce nucleus embeddings, which are then used to generate filtered

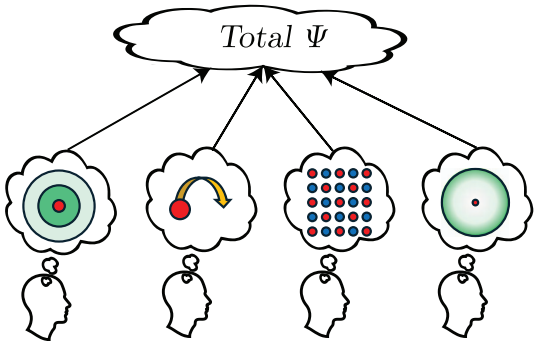

Figure 4. Cartoon representation of the focus of each attention 'head' in [97]. Each head learns fundamental physical correlations that are important to the representation of the total wavefunction. The first head learns local correlations, the second head learns information relevant to electron hopping (nearest neighbors), the third head pays attention electrons important to the antiferromagnetic character of the system, and the fourth learns more delocalized and complex interactions. Reproduced with permission from [97].

electron-nucleus features. These features encode short-range electron-nuclear correlations and help the network learn the correct cusp behavior of the wavefunction near nuclei. Unlike PsiFormer, Orbformer incorporates an explicit electron-electron distance-dependent bias term into the self-attention mechanism, which smoothly attenuates information exchange between electrons that are spatially distant. This localization bias allows each attention head to adapt its effective 'field of view'. These refinements lead to state of the art accuracy, as Orbformer converges to chemical accuracy for multiple bond dissociations and predicts the correct Diels–Alder reaction mechanism. Moreover, key elements of this model's architecture and results will be expanded upon in the 'transferability' section of the review since this model was also designed to be fine-tuned after a large pretraining.

In summary, this section has highlighted the effectiveness of transformer-based NQSs in modeling a wide range of physical systems, illustrating how advances in general ML research continue to inspire progress in quantum many-body modeling.

3.4. Beyond the ground state

As highlighted by Carleo and Troyer in their initial paper, NQSs are not limited to ground state properties. Indeed, their model can describe the time evolution of certain Heisenberg models. This led to a branch of NQSs aiming beyond the ground state. Many have incorporated time dependence into NQSs for simulating real-time quantum dynamics, including studies of spin models such as the transverse-field Ising model [101–103], bosonic systems such as the one-dimensional Bose gas [104], and continuous-variable rotor models [105]. Nys *et al* [106] extend these methods to the scope of our review by creating time dependent parameters for Jastrow and particle-attention backflow transformations inspired by PauliNet [74] and Pescia *et al*'s [86] MPNN for HEGs. This allows for the application of time dependence to continuous systems in first quantization. Furthermore, when operating in second quantization, they take inspiration from Carleo and Troyer to create a TD RBM. Furthermore, they employ TD variational principles and an expansion upon time dependent VMC [107]. As a result, the ansatz gives accurate predictions of TD properties for a Harmonic Interaction model, diatomics in an intense laser field, and a quenched quantum dot.

Moreover, another axis of research has instead focused on modeling excited states [81]. Indeed, Choo *et al* [108] have shown that it is possible to characterize excited states of the one-dimensional spin-$\frac{1}{2}$ Heisenberg and Bose–Hubbard models. This was achieved either by exploiting Abelian spatial symmetries to target specific excited eigenstates, or by minimizing the energy expectation values of multiple neural wavefunctions simultaneously, with orthogonality between states enforced through a Gram–Schmidt procedure. Similarly, Entwistle *et al* [109] use a PauliNet architecture, but modify the loss terms such that they can minimize the energy expectation value for a set of orthogonal states by also minimizing the overlap between the ground state and the excited state(s) [110]. Their loss term is thus:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \mathbb{E}_i \left[E_{\mathrm{loc}}[\psi_{\boldsymbol{\theta},i}](\mathbf{r}) \right] + \alpha \sum_{i>j} \left(\frac{1}{1-|S_{ij}|} - 1 \right). \tag{28}$$

The second term introduces a differentiable penalty based on the pairwise overlaps $S_{ij} = \langle \psi_{\theta,i} | \psi_{\theta,j} \rangle$ between states, scaled by a hyperparameter α.

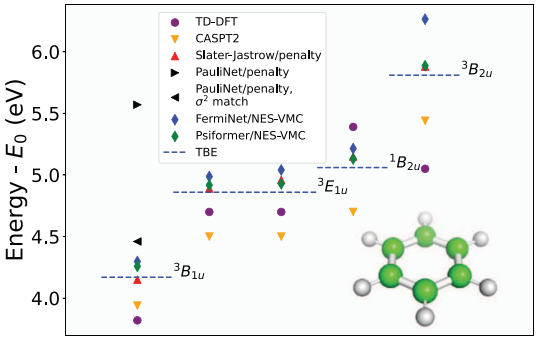

Figure 5. Prediction of excited state energies of benzene from multiple non-ML, penalty-based, and NES-VMC methods. Clearly, NES-VMC (especially when applied to the Psiformer model) is the most accurate model, as it nears theoretical best estimates. From [112]. Reprinted with permission from AAAS.

This method gives accurate excitation energies for small molecules, matching, and sometimes outperforming, results obtained through MR-CC and TD-DFT for small molecules and benzene. An extension of Entwistle *et al*'s work is Lu and Fu's [111] auxillary wavefunction (AW) method. Here, they also minimize overlap and converge to orthogonal states, but use an AW instead of a penalty term. This was found to be favorable for low-lying excited states, and was applied to small atoms and molecules.

However, these approaches introduce hyperparameters and weight choices that can make convergence sensitive to MC noise [112, 113]. To address this, Pfau *et al* developed the natural excited state (NES)-VMC method [112]. For the K lowest-energy states, they define a composite wavefunction ansatz as:

$$\Psi\left(\mathbf{x}^1,\ldots,\mathbf{x}^K\right) \triangleq \det\begin{bmatrix}\psi_1\left(\mathbf{x}^1\right) & \cdots & \psi_K\left(\mathbf{x}^1\right)\\ \vdots & \ddots & \vdots\\ \psi_1\left(\mathbf{x}^K\right) & \cdots & \psi_K\left(\mathbf{x}^K\right)\end{bmatrix}, \tag{29}$$

where each ψ_i represents an independent NN wavefunction corresponding to one of the K excited states, and $\mathbf{x}^i$ is a particle state. MC sampling is then taken from the distribution of all states and they can be optimized through VMC. Moreover, there is no need to enforce orthogonality here, since if two ψ_i become identical, the determinant (and thus the wavefunction) goes to 0. This method can be applied to different *ansatze*: FermiNet and Psiformer both give chemically accurate results on excited state energies and oscillator strengths for first row atoms. Furthermore, on larger systems such as benzene, NES-VMC (particularly the Psiformer version) outperforms competing methods and approaches theoretical best estimates as seen in figure 5).

However, essentially restructuring the excited state as an expanded system leads to higher computational costs. Thus, Szabó *et al* [113] further extended penalty-based methods by switching to a Psiformer architecture and autotuning the penalty scaling to cut gradient noise. A spin penalty term was also added to allow better convergence to a specific spin excited state, and a loss term with provable convergence without biasing was developed. Together, this enables chemically accurate prediction of many excited states and the generation of accurate potential energy surfaces (PESs) of certain excited states, which NES-VMC was unable to attain, by targeting a specific spin state. Li *et al* [114] introduce a low-scaling penalty based on the spin raising operator, allowing them to enforce spin symmetry. When using a LapNet architecture and enforcing an overlap penalty, this method can even obtain chemically accurate singlet-triplet energy gaps for biradicals. Lastly, Shätzle *et al* [115] simulate excited state PES using similar overlap and spin penalties. However, a significant part of their innovation is the transferability of the method, which will be expanded upon after contextualization in the final axis this review will cover on NQS; transferability.

3.5. Transferability

As mentioned above, the development of algorithmic refinements has greatly helped scale calculations. However, another approach taken to improve the efficiency (and thus usability) of NQSs has been to make their training transferable to other molecules and their models generalizable. In this context, transferability broadly refers to the ability of a pretrained or partially optimized NQS to be reused, without training from scratch, across different systems. This encompasses a spectrum of approaches: from sharing weights across geometries of a single molecule, to generalizing across molecules as this section will detail.

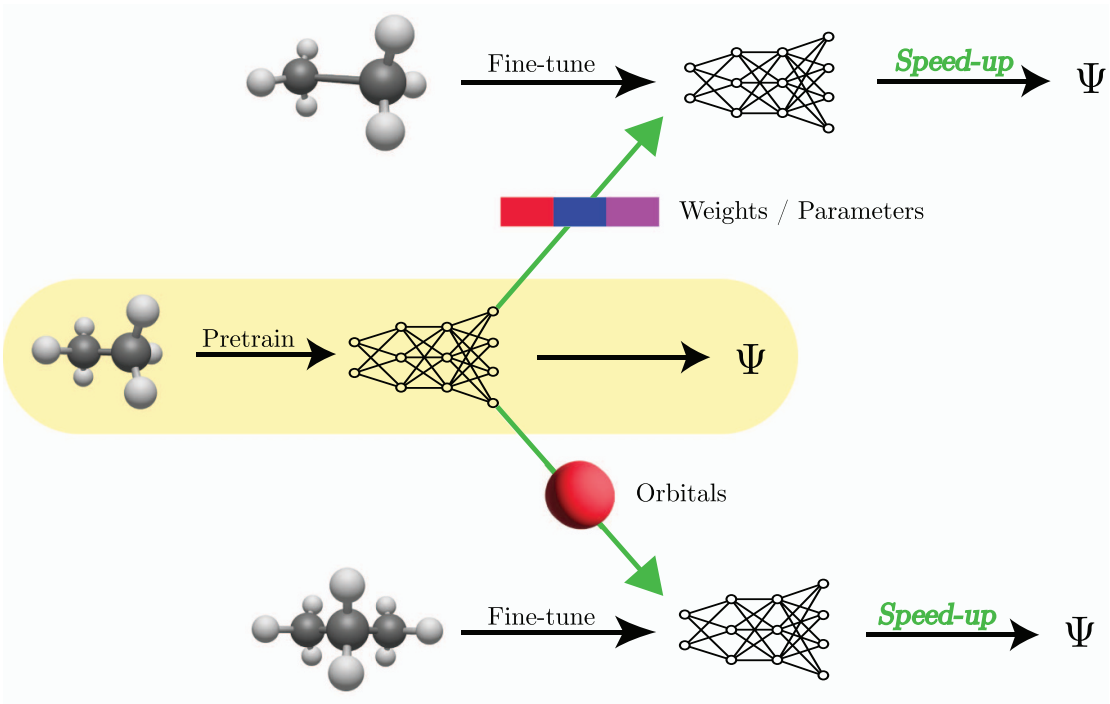

Figure 6. Illustration of the parameter/weight sharing mechanisms which speed-up the generation of PES and transferable orbital creation which allow few-shot or zero-shot learning of the wavefunction for a chemically similar molecule (bottom).

An early development in this regard was Scherbela *et al*'s DeepErwin [116]. They were motivated by the presence of fundamental/core orbitals in molecules and engineered a system in which certain weights /parameters in a PauliNet-like ansatz could be shared across similar geometries (figure 6). Indeed, the embedding of electron coordinates and parts of the backflow factor were found to be most apt for the sharing of parameters across the VMC optimization. As a result, PES of H chains with up to 112 configurations could be calculated simultaneously to chemical accuracy 6–13 times faster than previously possible. Furthermore, pretraining a model and sharing weights with *ansatze* of unseen configurations also converged, showing promise for generalizable NQSs. Gao and Günnemann [117] also performed work towards this goal by using a GNN to optimize an NQS that solves the S.E. for multiple geometries across the PES. This GNN uses message passing and an equivariant coordinate system to convey the information for each geometry to a FermiNet model with nuclei reindexing invariance built into it, allowing the GNN to reparametrize the latter part of the model. These advancements, coined PESnet, led to an accuracy comparable to FermiNet (thus slightly better than PauliNet and DeepErwin), whilst being up to 47 times faster.

In fact, it is the parameter sharing first demonstrated by DeepErwin, where specific network parameters (namely the weights of the electron embedding and backflow networks) are reused across geometries, that enabled Schätzle *et al* to generalize their excited PES model [115]. Indeed, modifying Psiformer's architecture to be causal, inhibiting information flow between electrons and nuclei, their model can accurately and efficiently generate the excited PES across 23 geometries of ethylene, which is known to have MR characteristics that make it difficult to model properly.

However, previous approaches were restricted to exploring different geometries of a single molecule to generate PESs. To enable generalization across molecules, Gao and Günnemann [118] introduced the Molecular Orbital Network (MOON), an MPNN-based Slater–Jastrow ansatz that learns localized molecular-orbital embeddings and satisfies the size-consistency condition. Building on this, they developed graph-learned orbital embeddings (GLOBE), which learns transferable orbital embeddings through atom-atom message passing and nucleus-to-orbital unidirectional message passing. These embeddings encode spatial information (e.g. core or bonding character) and are used to condition the downstream wavefunction parameters of the size-consistent ansatz. Using gradient rescaling to prevent larger systems from dominating the optimization and pretraining to HF orbitals before variational fine-tuning, this method achieves accurate energies for small molecules, which are all optimized at once (instead of independently with weight-sharing like in DeepErwin).

Another sub-direction is inspired by foundation style models in language and vision [77, 119, 120]. Indeed, Zhang *et al* [121] generated transformer quantum states that could fine tune a pretrained model to unseen and larger examples of the 1D transverse field Ising model. Scherbela *et al* [22] expand this concept to first quantized continuous space, where they define Slater-type multi-compound *ansatze*. These are innovatively based on a new type of single electron orbital: transferable atomic orbitals (TAOs) as seen in figure 6. TAOs are constructed via per-orbital descriptors extracted from HF orbitals using a graph convolutional network (GCN). These descriptors are then passed through trainable envelope and backflow MLPs, which are then combined with learned electron embeddings from an MPNN to complete the expression for a TAO. The per-atom sum of orbitals and inclusion of an envelope function ensure size consistency, which is shown by its ability to outperform HF on H_{20} as a zero-shot model trained on H_6 and H_{10}. This performance also gets better with few-shot finetuning, and with 1000 steps of finetuning, CCSD(T)/ccpVTZ accuracy is achieved, 20 times quicker than training from scratch. Zero-shot performance was further improved by incorporating a PhisNet style SE(3) equivariant orbital model to obtain better orbital and nuclear embeddings [122].

Although previously mentioned in the context of transformer architectures, Orbformer is particularly state of the art due to its transferability [100]. Indeed, it also contains generalized orbitals learned by a sub-network. This sub-network learns the coefficients to construct the electron envelopes, whose exponents are learned per atom type. Moreover, the electron embeddings generated by the transformer are projected into Slater columns using weights generated by the sub-network too. Thus, with a cusp enforcing Jastrow factor, it can be pretrained on multiple molecules and later finetuned. This transferability is also attributed to receiver gates in the attention heads, and architectural advancements of Forward Laplacian and Flash Attention which make large pretraining possible. Moreover, inspired by diffusion models, unadjusted Langevin algorithm [123] is used for initial equilibration, which also speeds up pretraining. Here, Forster *et al* pretrain Orbformer on 22 000 light atom molecules. This led to leading performance on the TinyMol dataset. Moreover, when optimizing an *ansatze* jointly, this also allowed up to a 20 times speedup over single-point calculations when predicting a 20 point dissociation curve, illustrating the importance of transferring. Furthermore, an important sign of progress towards a general model is the fact that the NN was shown to learn correct physics. For instance, Orbformer learns to localize two 'core' orbitals for carbon atoms through its training, which represent physically accurate 1 s, 2 s core C orbitals.

However, in cases where researchers want to generalize their models with Slater-determinant *ansatze* as those mentioned above, the number of ↑ and ↓ orbitals must be discretized, which can limit the scope of generalizability. As such, adapted from Kim*et al*'s work on ultra cold Fermi Gas, Gao and Günnemann presented a fully learnable Jastrow-Pfaffian ansatz: NeurPf. Their expression is

$$\Psi_{\boldsymbol{\theta}}(\mathbf{r}) = \exp[J_{\boldsymbol{\theta}}(\mathbf{r})]\ \mathrm{Pf}\left(\hat{\Phi}_{\boldsymbol{\theta}}(\mathbf{r})\, A_{\boldsymbol{\theta}}\, \hat{\Phi}_{\boldsymbol{\theta}}(\mathbf{r})^{T}\right). \tag{30}$$

The symmetric Jastrow factor $J_{\boldsymbol{\theta}}(\mathbf{r})$ captures smooth electron–electron and electron–nucleus correlations, while the Pfaffian $\mathrm{Pf}\left(\hat{\Phi}_{\boldsymbol{\theta}}(\mathbf{r}) A_{\boldsymbol{\theta}} \hat{\Phi}_{\boldsymbol{\theta}}(\mathbf{r})^{\mathrm{T}}\right)$ encodes antisymmetric pairing between electrons through a learnable skew-symmetric matrix $A_{\boldsymbol{\theta}}$. The orbital feature map $\hat{\Phi}_{\boldsymbol{\theta}}(\mathbf{r})$ is implemented as a permutation-equivariant NN and is a function of MOON-style embeddings. These conditions allow the model to be overparameterized, with $N_o \geqslant \{N_\uparrow, N_\downarrow\}$ learnable orbital functions, which enables it to generalize to multiple molecules with previously unattainable charge and/or spin configurations. To generalize, they used an architecture similar to PESnet, but using their Pfaffian instead of a Slater determinant. Once initialized with HF pretraining, wavefunctions for second row elements were simultaneously optimized and chemically accurate electron affinities and ionization potentials were obtained, which was not previously generalizable. It also managed to outperform GLOBE and TAO's for small and large molecules in the TinyMol dataset, even outperforming CCSD(T)/CBS for alkanes up to 106 electrons after 180 000 fine-tuning steps.

Just beyond the scope of this review, Rende *et al* [124] introduced the foundation neural quantum states (FNQS) framework for J_1–J_2 Heisenberg models. Unlike previous approaches, FNQS expands the network inputs to include the Hamiltonian coupling parameters, allowing sampling over a distribution of (J_1, J_2) values. This induces a multimodal representation that enables the simultaneous optimization of multiple ground states within a single ansatz across different Hamiltonians. Such foundation-style learning represents a promising direction for extending NQSs to more general many-body lattice systems.

3.6. Summary

Therefore, with continuous advances in accuracy, generalizability to excited and TD states, transferability, and computational scaling, the interest in NQSs is unlikely to diminish. The field is poised to further broaden the applicability of these ansätze by integrating recent methodological innovations, leveraging advances in ML across other domains, and developing deeper connections between QC and neural-network representations.

4. ML the XC functional

4.1. Introduction

DFT remains the most widely used method in material science. The performance of DFT (cost and accuracy) is determined by the choice of XC functional. While a vast number of traditional XC functionals are designed across the Jacob ladder as seen in figure 1, they mostly rely on analytical constraints and manual design of parameters.

In recent years, ML has been introduced to design XC functional in a data-driven way. Typically a high-level electronic structure dataset (like coupled-cluster results) will be used as label. The functional will be trained to achieve a higher accuracy within the dataset domain.

In this section, we examine how the machine-learning community has made use of this framework (or generalizations thereof) to develop XC functionals: what descriptor sets are used, how physical constraints are enforced, how training data are selected, and what performance has been achieved. We categorize the literature into three classes: Local 'enhancement-factor' ML functionals, Non-local / neural-operator functionals, and Residual or hybrid ML functionals.

4.2. Semi-local 'enhancement-factor' ML XC functionals

Inspired by classical local density approximation (LDA) and generalized gradient approximation (GGA) functionals, many modern ML-XC approaches adopt a transparent semi-local parameterization of the exchange–correlation energy: $E_{\mathrm{xc}}[\rho]$ is written through the electron density ρ and local density invariants $\mathbf{x}[\rho](\mathbf{r})$ at each point $\mathbf{r}$ (e.g. $|\nabla\rho|$, $\nabla^2\rho$, kinetic-energy density τ), i.e. the usual DFT 'ingredients' on a real-space grid—not a learned molecular fingerprint or SOAP-style descriptor of atomic environments built on the Dirac exchange form,

$$E_{\mathrm{xc}}^{\theta}[\rho] = -\frac{3}{4}\left(\frac{6}{\pi}\right)^{1/3}\int\left(\rho^{\uparrow}(\mathbf{r})^{4/3}+\rho^{\downarrow}(\mathbf{r})^{4/3}\right)\times f_{\theta}\left[\mathbf{x}[\rho](\mathbf{r})\right]d\mathbf{r}, \tag{31}$$

where $\rho^{\uparrow}$ and $\rho^{\downarrow}$ denote spin densities, f_θ is a learnable enhancement factor with parameters θ, and $\mathbf{x}[\rho](\mathbf{r})$ denotes local density descriptors (e.g. ρ, $|\nabla\rho|$, $\nabla^2\rho$, τ). When f_θ is replaced by a hand-crafted analytic form, this reduces to traditional LDA, GGA, or meta-GGA functionals. Early machine-learning (XC) functionals primarily focused on learning such local enhancement factors f_θ within the semi-local framework. Aldegunde *et al* [125] introduced a Bayesian meta-GGA that parametrized an enhancement factor F_θ using basis functions of local descriptors, fitted to atomization and bulk property data with built-in uncertainty quantification. Their relevance vector machine approach achieved a mean absolute error of 0.116 eV on atomization energies for test molecules in the G2/97 set, though with larger errors of 0.314 eV for test solids. Ryabov *et al* [126] implemented a neural-network interpolation of LDA/GGA functionals, learning a direct mapping from local density features to the XC energy density $\varepsilon_{\mathrm{xc}}(\mathbf{r})$. Lei and Medford [127] developed rotationally invariant convolutional descriptors, applying NNs to locally convolved densities to predict $\varepsilon_{\mathrm{xc}}(\mathbf{r})$, effectively learning a localized enhancement factor. Nagai *et al* [128] extended this idea by embedding physical asymptotic constraints into a meta-GGA-style $F_\theta(r_s, s, \alpha, \ldots)$ that multiplies a standard reference energy density. These methods retain the favorable cost and interpretability of meta-GGA-level functionals while improving flexibility through data-driven fitting. However, their local character limits their ability to capture genuinely non-local exchange and correlation effects, long-range dispersion, or static correlation in strongly correlated systems.

4.3. Non-local and operator-like ML XC functionals

Beyond semi-local density invariants (features of ρ and its derivatives at $\mathbf{r}$), several recent approaches explicitly endow the enhancement factor with non-local dependence on the electronic density (or orbitals), thereby capturing long-range exchange, dispersion-like correlation, and multi-center effects that local maps cannot. Here we highlight five representative lines of work—*CIDER* [129], *NeuralXC* [130], *DM21* [131], *Skala* [132], and *EG-XC* [133] . These methods pursue distinct routes to non-locality:

CIDER employs feature-space convolutions with atom-centered kernels; NeuralXC uses angular projections onto localized basis sets; DM21 incorporates orbital-dependent exact exchange via learned mixing; Skala implements neural operators coupling multi-resolution grids; and EG-XC leverages equivariant message passing on atomic graphs. Together, these strategies span from convolutions and basis projections to explicit orbital dependence and graph aggregation, each balancing expressiveness, cost, and exact constraints.

CIDER realizes non-locality through a feature-space convolution framework that constructs non-local descriptors by convolving semi-local features (density, gradient, kinetic energy density) with parameterized atom-centered kernel functions. This convolution integrates density information over finite spatial regions, capturing multi-center correlations through spatially extended descriptors built from radial basis functions and spherical harmonics. Rather than employing a NN, CIDER leverages Gaussian process regression (GPR) to map these non-local descriptors to the enhancement factor, enabling rigorous uncertainty quantification, data-efficient learning, and natural enforcement of smoothness. Exact constraints such as uniform-density limits, slowly varying behavior, size consistency, and uniform coordinate scaling are satisfied through careful kernel design and feature normalization. The resulting functional retains near meta-GGA computational cost (via finite-range atomic neighborhoods) while accessing genuinely non-local information. The CIDER exchange functional trained on molecular data achieves mean absolute errors well below those of semilocal functionals on validation sets. When incorporated as PBE0/NL-MGGA-DTR (a surrogate hybrid combining CIDER with PBE correlation), the functional demonstrates excellent performance on the GMTKN55 benchmark, with mean deviations from PBE0 reference values of only a few kcal mol^{-1} across diverse reaction types. For solid-state calculations in plane-wave (PW) codes, CIDER exhibits computational scaling comparable to meta-GGAs while approaching hybrid-functional accuracy. On silicon defect calculations, CIDER correctly predicts transition levels for vacancy, boron substitutional, and phosphorus substitutional defects, demonstrating its applicability to materials with strong correlation effects. The method shows significant improvements over semi-local functionals for reaction barriers, non-covalent interactions, and MR systems [129, 134].

NeuralXC constructs atom-centered, rotationally invariant density descriptors by projecting $\rho(\mathbf{r})$ onto localized, smooth basis functions around each nucleus and pooling over angular channels. A NN predicts an additive correction $\Delta\varepsilon_{\mathrm{xc}}(\mathbf{r})$ which is integrated and used self-consistently. Although the descriptors are assembled locally, their finite spatial support (with multiple radial shells and angular moments) yields an *effective* non-local receptive field extending over each atomic environment; further, self-consistency propagates these corrections non-locally through the KS response. In practice, this delivers mid-range non-locality while preserving rotational invariance and near meta-GGA cost. A NeuralXC functional trained on ethane and propane achieves MAE of 6.6 meV and 6.1 meV for n-butane and isobutane test structures, respectively, demonstrating strong transferability across alkane chain lengths. In BO molecular dynamics of liquid water at 300 K, NXC-W01 reproduces radial distribution functions in close agreement with MB-pol classical molecular dynamics (MD) and experimental measurements, significantly outperforming SCAN and ωB97M-V functionals which produce overstructured liquids [130].

DM21 is a data-driven local-hybrid meta-GGA in which a neural map modulates, pointwise, the mixing between semi-local exchange and exact (Fock) exchange (figure 7). The explicit inclusion of HF exchange renders the energy functional orbital dependent and therefore non-local by construction (since Fock exchange is a non-local operator). The learned, position-dependent mixing function is trained under fractional-charge/spin constraints to mitigate delocalization error; as a result, non-local effects are introduced via the exchange operator while the neural component remains tied to semi-local density invariants (the same local meta-GGA ingredients as above, not molecular environment vectors). On the GMTKN55 benchmark, DM21 achieves state-of-the-art performance with mean-of-means (MoM) errors competitive superior to the best hybrid and double functionals. On the bond-breaking benchmark measuring mean absolute errors for first- and second-row diatomics, DM21 demonstrates sub-5 kcal mol^{-1} accuracy. The functional correctly captures piecewise linear energies for fractionally charged atoms and maintains energy constancy for spin-interpolated states, addressing fundamental delocalization and spin-symmetry-breaking errors that plague conventional functionals. On the QM9 dataset of 134k organic molecules, DM21 achieves errors of 1.66 kcal mol^{-1}, demonstrating impressive MoM performance across GMTKN55 and QM9 datasets [131].

EG-XC formulates non-locality through an SE(3)-equivariant graph NN built on a nuclei-centered representation of the density. Nodes carry density-derived features; edges encode interatomic geometry; message passing aggregates information over multiple hops, thereby realizing a controllable non-local receptive field while rigorously preserving rotational equivariance. The network outputs a pointwise (or atom-resolved) contribution to $\varepsilon_{\mathrm{xc}}$, which is integrated self-consistently. Equivariance ensures correct tensorial transformation properties and sample efficiency, and message passing supplies the non-local

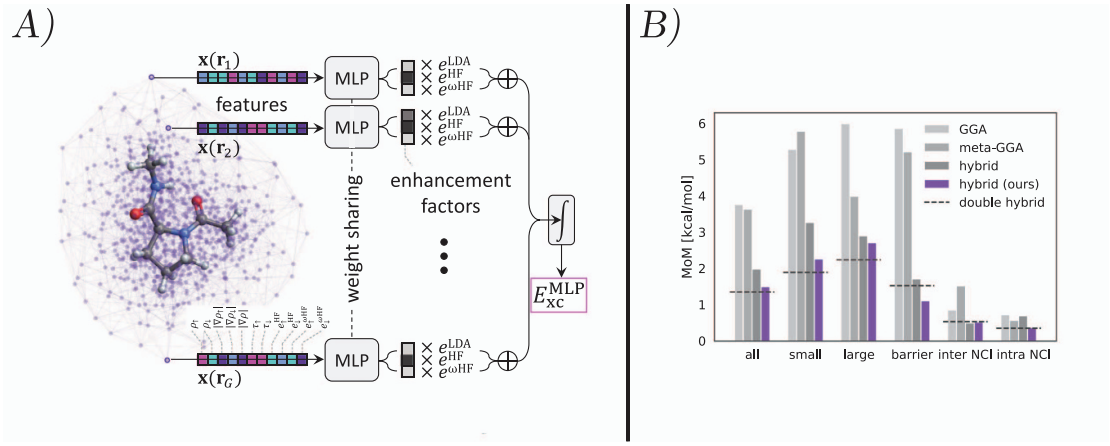

Figure 7. (A) Architecture of DM21. (B) Results of the DM21 functional compared to other popular functionals, illustrating its ability to outperform other hybrid methods and even surpass double-hybrid methods in some cases. From [131]. Reprinted with permission from AAAS.

coupling absent in purely local DFAs. On the MD17 molecular dynamics benchmark, EG-XC accurately reconstructs CCSD(T) energies with high fidelity. For out-of-distribution conformations of the 3BPA (bisphenol A) molecule, EG-XC reduces relative mean absolute error by 35%–50% compared to baseline methods. Remarkably, on the QM9 dataset, EG-XC demonstrates exceptional data efficiency and size extrapolation: when trained on identical datasets, it achieves 51% lower MAE on average compared to the SchNet force field [55], and matches the performance of force fields trained on 5 times more data and on larger molecules, demonstrating superior transferability and sample efficiency [133].

Most recently, Skala implements a genuinely non-local neural *operator* over the standard DFT quadrature grid. Meta-GGA features (spin densities, gradients, kinetic energy densities) are computed on a fine grid, while a coarse auxiliary grid communicates global information via spherical-harmonic and radial-basis expansions. A learned coupling between coarse and fine representations produces an enhancement factor $f_\theta[\mathbf{x}[\rho]](\mathbf{r})$ that depends on the density field *beyond* the local neighborhood. Architectural constraints (e.g. bounded f_θ) enforce exact conditions such as uniform scaling, size consistency, and Lieb–Oxford bounds, enabling long-range non-local interactions at roughly meta-GGA computational cost. The Skala functional achieves state-of-the-art performance on the W4-17 benchmark for atomization energies, outperforming popular hybrid functionals and achieving near-chemical accuracy. The multi-resolution grid approach enables the functional to capture both local and global density features efficiently, with computational costs scaling favorably compared to hybrid DFT while maintaining accuracy competitive with more expensive double-hybrid methods. The architectural design ensures physical constraints are satisfied automatically through the network structure rather than explicit post-processing [132].

This line of work targets an explicitly kernel-type non-local exchange functional, in which the exchange energy is expressed through finite-range pair integrals of density-derived descriptors. A trainable kernel (parametrized by an ML model) mediates interactions between spatially separated points,

$$E_{\mathrm{x}}^{\mathrm{ML}}[\rho] = \frac{1}{2}\int\int \rho(\mathbf{r})\, K_\theta(\mathbf{r},\mathbf{r}';\Phi[\rho])\, \rho(\mathbf{r}')\, \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \tag{32}$$

with $\Phi[\rho]$ denoting auxiliary local features (e.g. reduced gradients or kinetic-energy proxies). By construction, the energy depends on *pairs* of points, furnishing true non-local behavior appropriate for both molecular and periodic systems. Practical implementations restrict the kernel to a finite range and impose symmetry/decay constraints to ensure accuracy, transferability, and tractable scaling.

4.4. Learning of kinetic energy

A parallel line of research to ML-based XC functionals focuses on learning the non-interacting kinetic energy functional $T_{\mathrm{s}}[\rho]$, the central quantity in orbital-free density functional theory (OF-DFT). In OF-DFT, the total electronic energy is expressed entirely as a functional of the density,

$$E[\rho] = T_{\mathrm{s}}[\rho] + E_{\mathrm{H}}[\rho] + E_{\mathrm{xc}}[\rho] + \int v_{\mathrm{ext}}(\mathbf{r})\,\rho(\mathbf{r})\,d\mathbf{r}, \tag{33}$$

thus bypassing KS orbitals and potentially reducing computational scaling from $O(N^3)$ to nearly linear in system size. The main difficulty lies in constructing an accurate and transferable approximation for $T_s[\rho]$, which must recover both the local Thomas-Fermi limit and the non-local Pauli contributions governing shell structure and chemical bonding.

Early orbital-free formulations relied on analytic gradient expansions or density-dependent kernels such as the Wang–Teter and Wang-Govind-Carter functionals, which introduce non-local convolution integrals between densities and their gradients. The first machine-learning approach to the kinetic functional was introduced by Snyder *et al* [135], who used kernel-ridge regression to learn the functional derivative $\delta T_s[\rho]/\delta\rho(\mathbf{r})$ directly from reference KS data. Working in one dimension, they demonstrated that a kernel model can accurately capture shell and bonding effects absent in conventional gradient expansions, thereby establishing ML as a viable path toward non-local kinetic functionals. On their one-dimensional model problem of noninteracting Fermions in a box, Snyder *et al* achieved mean absolute errors below 1 kcal mol^{-1} on test densities using fewer than 100 training densities, demonstrating remarkable data efficiency [135]. Via principal component analysis, their projected functional derivative method produced highly accurate self-consistent densities, addressing the challenge of iterative density optimization without requiring explicit calculation of poorly-behaved functional derivatives in data-sparse regions. In follow-up work [136], they extended this approach to one-dimensional diatomic bond breaking, showing that the machine-learned functional could (i) accurately dissociate diatomics, (ii) be systematically improved with increased reference data, and (iii) generate accurate self-consistent densities via the projection method. With relatively few training densities (on the order of 20 data points for bond-breaking problems), the error due to machine-learning interpolation was demonstrated to be smaller than typical errors in standard XC functionals, establishing that ML-based kinetic functionals could surpass conventional analytical approximations in accuracy [136].

Symbolic regression has also been explored as a route to analytic (semi-local) kinetic-energy density functionals with tractable functional derivatives. Mitchell *et al* [137] used symbolic regression in a one-dimensional KEDF setting to obtain expressions that track the transition from the von Weizsäcker limit (exact for one-electron cases) toward the Thomas-Fermi limit (exact for the HEG). Complementarily, Manzhos *et al* constructed explicit analytic kinetic-energy functionals guided by interpretative ML [138].

Beyond model accuracy, the quality of the functional derivative matters for orbital-free density optimization. Meyer, Weichselbaum, and Hauser showed that simultaneously training on the kinetic energy density functional *and* its functional derivative improves the derivative accuracy compared to training on the functional alone [139].

Kernel methods, in particular GPR, have been used to obtain robust kinetic-energy density models from highly uneven descriptor sets. Manzhos *et al* [140] introduced target and feature averaging for GPR-based kinetic energy density learning, achieving accurate and stable models from as few as $\sim$2000 training samples across multiple light materials.

Small NNs can be effective when trained on physically motivated inputs. Golub and Manzhos demonstrated that small (single-hidden-layer) NNs trained using fourth-order gradient expansion terms can achieve ultra-low kinetic-energy density fitting errors without overfitting [141].

More recently, physically constrained nonlocal KEDFs have been built with machine-learning architectures designed for stability and exact constraint satisfaction in self-consistent OF-DFT. Sun and Chen introduced a physical-constrained nonlocal kinetic energy density functional for simple metals (implemented using the ABACUS package) [142], and extended it to a multi-channel formulation for semiconductors using multiple length scales [143].

Subsequent developments extended these ideas to three-dimensional densities using more expressive neural architectures. Remme *et al* [144] further incorporated rotational and translational equivariance into neural architectures, ensuring that the learned $T_s[\rho]$ satisfies basic symmetry requirements while remaining transferable across atomic numbers and chemical compositions. These neural models can produce accurate kinetic potentials for self-consistent OF-DFT, reproducing both shell structure and bonding without the need for explicit orbitals. The KineticNet architecture introduced by Remme *et al* represents a major advancement: it employs an equivariant deep NN based on point convolutions adapted to molecular quadrature grids, with convolution filters providing sufficient spatial resolution near nuclear cusps and an atom-centric sparse architecture that relays information across multiple bond lengths. Crucially, they developed a novel training strategy that generates diverse training data by finding ground-state densities perturbed by random external potentials, enabling the functional to handle iterative density optimization even from poor initial guesses. KineticNet achieved, for the first time, chemical accuracy of the learned functionals across input densities and geometries of small molecules (MAE $<$1 kcal/mol), representing a qualitative leap beyond earlier 1D demonstrations [144]. For two-electron

systems, they additionally demonstrated OF-DFT density optimization with chemical accuracy, showing that the learned kinetic energy functional produces converged self-consistent densities that closely match KS reference densities [144]. The ability of KineticNet to maintain accuracy across diverse molecular geometries and chemical compositions, combined with its equivariance properties, demonstrates that neural functionals can generalize substantially beyond their training distributions.

These advances demonstrate that ML-based OF-DFT can achieve near-KS accuracy for small molecules and light metals while retaining linear scaling and smooth density dependence. The progression from Snyder *et al*'s pioneering 1D work (MAE <1 kcal mol^{-1} with <100 training points) to Remme *et al*'s 3D chemical accuracy on diverse molecules illustrates the rapid maturation of machine-learned kinetic functionals. The ability to represent long-range coupling through learned non-local kernels or neural operators makes these methods promising for large-scale materials simulations involving thousands to millions of atoms, where conventional orbital-based approaches become prohibitively expensive. Recent work has extended these methods to periodic systems and demonstrated stable self-consistent optimization on increasingly complex chemical systems, bringing the original HK vision of a purely density-based electronic structure theory closer to practical realization.

4.5. Back propogation of self-consistensy

A fundamental technical challenge in training machine-learned density functionals lies in the non-trivial dependence of the total energy on the model parameters through the self-consistent KS (SCF) equations. Because the electronic density $\rho^*[\theta]$ is obtained implicitly from the stationary condition

$$\frac{\delta E_\theta[\rho]}{\delta \rho(\mathbf{r})}\Big|_{\rho=\rho^*} = 0, \tag{34}$$

the gradient $\partial E_\theta[\rho^*]/\partial\theta$ involves implicit derivatives of the converged density with respect to the functional parameters θ. Traditional back-propagation cannot be applied directly, as the SCF iteration constitutes a fixed-point solver rather than an explicit computational graph.

Early studies circumvented this issue by optimizing functional parameters using gradient-free or stochastic schemes. In particular, several early neural exchange–correlation functionals employed simulated annealing or evolutionary algorithms to minimize energy errors across a training set, treating the SCF solver as a black box. While conceptually simple, such methods scale poorly and provide limited insight into the sensitivity of the self-consistent density to functional parameters.

More recent work has introduced explicit differentiation through the SCF cycle using techniques from implicit differentiation and differentiable programming. Kasim and Vinko [145] demonstrated a fully differentiable DFT implementation in which both the SCF iterations and the underlying Poisson and KS solvers are embedded within an automatic-differentiation framework, allowing end-to-end back-propagation of energy gradients with respect to the neural functional parameters. Related approaches employ implicit-function theorem differentiation, expressing the SCF solution as a fixed point of a differentiable mapping and solving the associated linear system for $\partial\rho^*/\partial\theta$. These developments enable efficient and numerically stable gradient-based optimization of ML functionals, dramatically improving convergence and scalability compared to stochastic search methods.

The emergence of differentiable DFT frameworks thus marks a crucial advance for functional learning: it allows model parameters to be optimized directly with respect to total energy and density errors, enforces self-consistency during training, and provides a unified route to differentiate through other electronic-structure components such as response functions, band gaps, or atomic forces. Combined with modern automatic-differentiation libraries, these methods open the door to fully differentiable QC pipelines capable of learning physically constrained, self-consistent density functionals end to end.

4.6. Summary

DFT has a long and successful history, with the KS theorem guaranteeing the existence of an exact functional but not its explicit form. Traditional XC and kinetic energy functionals rely on analytic constructions, yet ML now allows these functionals to be learned directly from high-level electronic structure data. Semi-local ML functionals improve accuracy while retaining efficiency, whereas non-local and operator-based approaches capture long-range correlations and multi-center effects. Advances in differentiable, self-consistent frameworks further enable end-to-end training of neural functionals. Together, these developments push DFT forward, bridging decades of empirical design with modern neural-network-based, physically informed functional approximations.

5. ML Hamiltonians

Beyond accuracy improvements, the computational cost of SCF loops in DFT fundamentally limits its scalability. Indeed, SCF iterations hinder simulations of large systems, and they incentivize the use of inexpensive semi-local exchange–correlation functionals (LDA/GGA/meta-GGA rungs of Jacob's ladder) over more accurate but costly hybrid functionals. To address these bottlenecks, machine-learning Hamiltonians have been introduced as a promising alternative. Recent approaches focus on directly predicting the Hamiltonian matrix without performing SCF iterations, enabling DFT-level accuracy for system sizes that were previously inaccessible.

5.1. Core architectures

5.1.1. Message-passing GNNs

As mentioned in the technical background, MPNNs constitute one of the earliest and most widely adopted architectures for predicting quantum-mechanical operators from atomic configurations.

For Hamiltonian learning, MPNNs are commonly trained to predict matrix elements such as H_{ij} or Fock matrix entries directly from local coordination environments. DeepH [146] and related models apply localized message-passing and basis-aware embeddings to map atomic neighborhoods onto Hamiltonian blocks expressed in atom-centered Gaussian or NAO basis sets. These MPNN-based Hamiltonians have demonstrated strong performance and favorable scaling for molecules ranging from hundreds to thousands of atoms, making them a practical data-driven alternative to traditional SCF-based DFT in large-scale molecular modeling.

Despite these successes, conventional message-passing models remain limited by their reliance on primarily scalar features and locality assumptions. As a result, they do not fully capture the tensorial structure of Hamiltonians—particularly the rotational covariance of off-diagonal *p*–*p*, *d*–*d*, or multipole-coupled elements—which motivates the use of symmetry-preserving neural architectures. Recent work has also identified limitations in standard MPNN aggregation schemes for Hamiltonian learning, proposing refined strategies that better preserve local electronic structure information during message accumulation [147].

5.1.2. E3NNs

Similarly, as mentioned in the background, E3NNs enforce the symmetries of the Euclidean group E(3)—translations, rotations, and reflections—directly in the architecture.

Applied to Hamiltonian prediction, equivariant models naturally encode the rotational transformation laws of orbital overlaps and Hamiltonian matrix elements. Since *p*-, *d*-, and higher angular-momentum channels transform as vector and rank-2 tensors under SO(3), equivariant architectures ensure that predicted H_{ij} elements rotate covariantly with the molecular geometry, an essential requirement for physical consistency. This leads to improved generalization across orientations, conformers, and chemical environments compared to purely scalar models. Moreover, because the basis functions used in Gaussian, NAO, or PAW frameworks have known symmetry properties, equivariant Hamiltonian learners achieve higher accuracy for angular-dependent couplings and multipole interactions. Recently Gong *et al* combined equivariant message passing with localized basis-projection schemes in DeepH-E3 [148] (figure 8), which can reproduce DFT-quality Hamiltonians across wide chemical spaces and can scale to supercells with thousands of atoms. In particular, for twisted bilayer graphene, DeepH-E3 achieves MAEs of 0.4–0.5 meV for such supercells when trained on small, non-twisted, and randomly perturbed structures of this material (figure 8).

While these architectural advances establish the foundation for accurate Hamiltonian prediction, recent work has demonstrated that learned Hamiltonians enable calculations far beyond ground-state energy predictions, opening new avenues for materials simulation and dynamic phenomena.

5.2. Beyond ground-state properties

Once Hamiltonians are learned, they serve as differentiable, queryable surrogates for full DFT calculations. This capability extends far beyond predicting ground-state energies and atomic forces—learned Hamiltonians can be perturbed, differentiated, or propagated in time to access response properties, excited states, and dynamic processes that traditionally require expensive recalculations of the electronic structure at each configuration.

5.2.1. Response properties and phonons

Response properties such as phonon spectra, dielectric constants, and Born effective charges are critical for understanding lattice dynamics, optical properties, and thermal transport. These quantities are

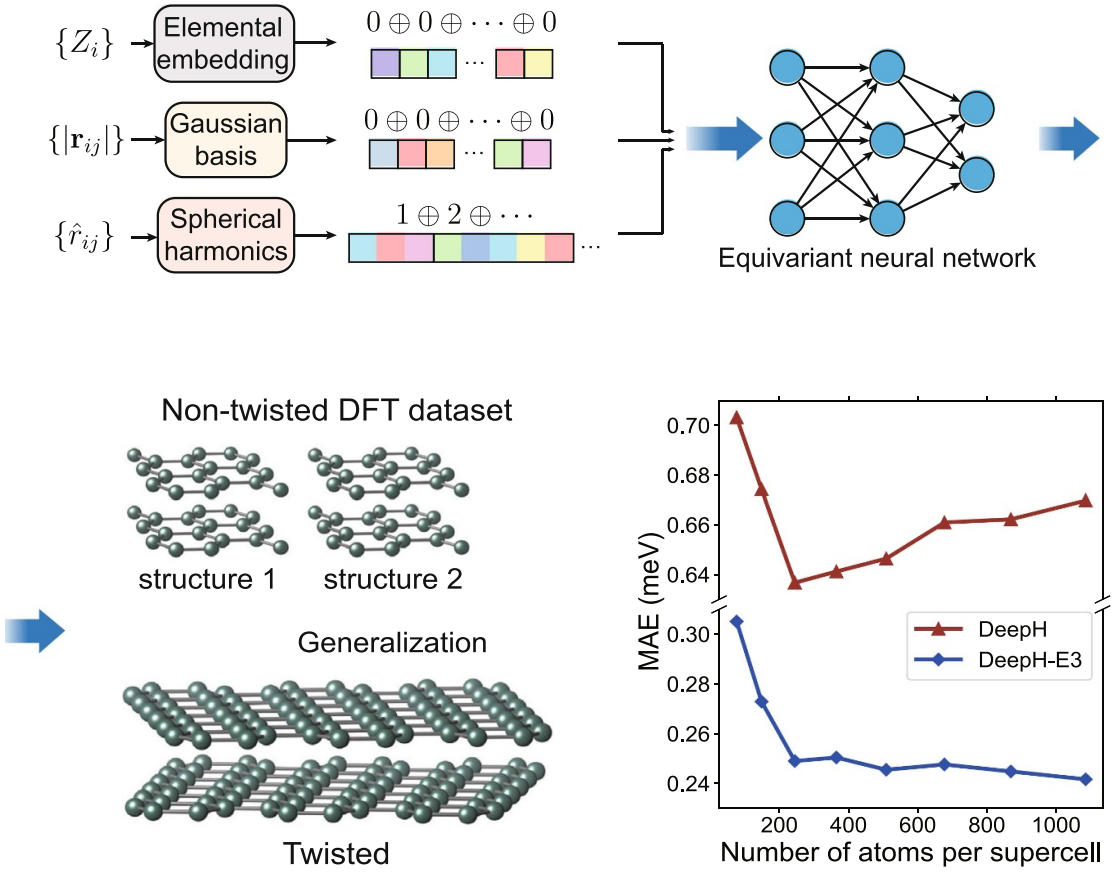

Figure 8. Top: Construction of an equivariant architecture from input features: atomic numbers, interatomic distances, and the unit vectors of relative positions. Bottom: training on twisted bilayer graphene and generalization to larger supercells for a chosen twist angle. Comparison with DeepH [146], shows importance of including equivariance. Reproduced from [148]. CC BY 4.0.

typically computed using density functional perturbation theory (DFPT), which requires calculating the derivative of the Hamiltonian with respect to atomic displacements or external fields—a process that scales poorly for large systems. Li *et al* [149] demonstrated that learned Hamiltonians can be incorporated directly into DFPT workflows through automatic differentiation, enabling efficient calculation of phonon spectra and dielectric responses without repeating costly SCF iterations for each perturbation. By training NNs to predict the KS potential matrix and applying automatic differentiation to obtain electron–phonon coupling matrices, their deep-learning DFPT framework achieves computational speedups of several orders of magnitude while maintaining near-perfect accuracy. The framework reproduces *ab initio* electron–phonon coupling properties with high fidelity, achieving low mean absolute errors (approximately 0.2–0.5 meV for Hamiltonian elements and 0.3–0.5 meV/Å^{-1} for Pulay terms) and quantitatively matching phonon linewidths, Eliashberg spectral functions, electron–phonon coupling (EPC) strength λ (e.g. 0.085 vs 0.089 for a carbon nanotube), and superconducting transition temperatures across graphene, MoS_2, and carbon nanotube benchmarks.

Extending this capability further, EPC—which governs phenomena such as superconductivity, charge transport, and polaron formation—requires computing matrix elements of the electron–phonon interaction across the Brillouin zone. This calculation is notoriously expensive, often requiring millions of matrix element evaluations. Zhong *et al* [150] showed that Hamiltonian learning accelerates EPC calculations by orders of magnitude while maintaining first-principles accuracy. Their approach combines learned Hamiltonians with Wannier interpolation schemes, enabling high-resolution sampling of electron–phonon coupling constants for materials ranging from conventional metals to complex topological semimetals. For GaAs, their ML framework using the Heyd–Scuseria–Ernzerhof hybrid functional predicted room-temperature electron mobilities of 8500 cm^2/(V·s), within 15% of experimental values, while calculations using the standard PBE functional severely underestimated mobility by a factor of three. For the Kagome superconductor CsV_3Sb_5, the method successfully reproduced the experimentally observed double-dome structure in the pressure-induced superconducting phase diagram with transition temperatures agreeing to within 1 K, revealing that phonon softening drives the

pressure-dependent superconductivity. This breakthrough opens pathways to predict mobility, superconducting transition temperatures, and thermal conductivities for materials where such calculations were previously intractable.

5.2.2. Excited-state dynamics and magnetic effects

Nonadiabatic molecular dynamics (NAMD)—where electronic excitations couple to nuclear motion—is essential for understanding photochemistry, charge transfer, and energy relaxation in molecules and materials. However, NAMD requires computing the electronic structure at every nuclear time step (typically femtosecond resolution), making it prohibitively expensive for systems larger than a few dozen atoms or for trajectories longer than picoseconds. Zhang *et al* [151] demonstrated that learned Hamiltonians can be propagated on-the-fly to enable NAMD simulations at *ab initio* quality but with computational costs approaching classical force fields. Their N2AMD (neural-network non-adiabatic molecular dynamics) framework employs an E(3)-equivariant GNN to predict Hamiltonians from which nonadiabatic couplings and orbital energies are computed directly, achieving mean absolute errors below 2.5 meV for band energies and 0.017–0.038 meV for nonadiabatic couplings while accelerating calculations by four orders of magnitude compared to hybrid-functional DFT. By training models to predict TD Hamiltonians under external perturbations (such as laser fields or ionic motion), they achieved femtosecond-scale excited-state dynamics for systems previously inaccessible to traditional methods. Their approach successfully reproduced ultrafast charge separation in organic photovoltaic interfaces and coherent phonon generation in two-dimensional materials, showcasing the potential of learned Hamiltonians to bridge the gap between *ab initio* accuracy and the spatiotemporal scales required for realistic simulations.

Beyond the learned-Hamiltonian perspective, classic NAMD development by Barbatti, Westermayr, Dral, and others established coupled-trajectory formalisms (e.g. Ehrenfest dynamics and surface hopping) [152–155] as the baseline for defining nonadiabatic couplings and electronic-state branching around avoided crossings and conical intersections [154–163]. From the ML point of view, the key practical step is to match this workflow with data-efficient on-the-fly refinement: uncertainty-driven active-learning loops decide when additional *ab initio* evaluations are needed to keep the learned surrogate accurate throughout a trajectory. Bayesian active-learning protocols of this general form have been demonstrated to reduce labeled data requirements in electronic-structure learning [164]; an analogous strategy is therefore natural for NAMD. Spin-polarized systems, where magnetism plays a central role in determining electronic structure, present unique challenges for Hamiltonian learning. In magnetic materials, separate spin-up and spin-down Hamiltonian channels must be predicted, and the model must respect magnetic symmetries such as time-reversal breaking and spin rotation. Li *et al* [165] developed xDeepH, which is an extension of the deepH framework to spin-polarized DFT, enabling prediction of magnetic ground states, spin textures, and magnon spectra without SCF iterations. Their model incorporates spin-dependent embeddings and enforces proper transformation laws under spin rotations, achieving accurate predictions for ferromagnets, antiferromagnets, and systems with complex noncollinear magnetic order. The spin-polarized xDeepH model achieves sub-meV accuracy for magnetic Hamiltonian matrix elements, with mean absolute errors of 0.56 meV for spin-spiral monolayer $NiBr_2$ and 0.36 meV for monolayer CrI_3, while accurately reproducing band structures for curved CrI_3 nanotubes and moiré-twisted bilayer CrI_3 hosting magnetic skyrmions. The model generalizes reliably to unseen magnetic textures, lattice curvatures, and twist angles, matching DFT band structures and response properties at a fraction of the computational cost and enabling high-throughput studies of magnetic quantum materials.

These applications demonstrate that learned Hamiltonians function as versatile computational engines for multi-scale simulations, enabling coupling of electronic structure to nuclear motion, electromagnetic fields, and thermal fluctuations—domains that have traditionally required prohibitively expensive *ab initio* methods or uncontrolled approximations.

5.3. Extending chemical and basis coverage

Early Hamiltonian learners were trained on narrow chemical domains, such as organic molecules or specific crystalline systems, limiting their applicability to broader materials discovery efforts. Recent work has pursued two complementary strategies to expand coverage: developing universal models that span diverse chemical spaces, and generalizing Hamiltonian representations beyond localized atomic orbital bases.

5.3.1. Toward universal models

The promise of a single, universal Hamiltonian predictor, analogous to a universal force field, has motivated efforts to train models across the periodic table [166–168]. However, this ambition faces significant challenges: different elements exhibit vastly different electronic structures (s/p-block main-group chemistry versus d/f-block transition metals and lanthanides), coordination environments vary widely (from molecular clusters to extended solids), and charge-transfer effects complicate orbital interactions in heterogeneous systems.

Several recent efforts have made progress toward this goal. Wang *et al* [166] trained a universal deepH model spanning first 4 rows of elements, from hydrogen to bismuth, using a diverse training set of over 10 000 solid materials. The model achieved competitive accuracy across main-group compounds, transition metal oxides, and intermetallic alloys, demonstrating that sufficiently large and diverse datasets enable transferable Hamiltonian prediction. Key to their success was careful dataset balancing—oversampling underrepresented chemical environments (e.g. f-block elements, high-pressure phases) and augmenting training with configurations generated from molecular dynamics trajectories to capture thermal disorder.

Zhong *et al* [167] demonstrated that a single equivariant GNN trained on Hamiltonians spanning most of the periodic table can achieve strong transferability across elements, including those with relatively sparse training data. This generalization arises from large-scale diverse training and a two-stage loss incorporating both Hamiltonian matrix elements and band eigenvalues, rather than from explicit alchemical interpolation between atomic species. More recently, Kaniselvan *et al* [168] investigated scaling laws for foundation models of electronic Hamiltonians, training large models on heterogeneous datasets spanning molecules, bulk crystals, and surfaces. They showed that pretraining on abundant small-molecule data systematically improves transfer to chemically and structurally distinct, data-scarce systems-including large inorganic-organic frameworks-highlighting the importance of scale and dataset diversity for transferable Hamiltonian learning.

Despite these advances, challenges remain. Charge-transfer systems—where electrons redistribute substantially between atoms—remain difficult to predict accurately, as standard NAO basis sets struggle to represent delocalized charge distributions. Additionally, heavy elements with relativistic effects (e.g. gold, mercury, actinides) require specialized treatment, and current universal models have not yet incorporated spin–orbit coupling or scalar-relativistic corrections systematically. Future work will likely require multi-fidelity training strategies, where inexpensive semi-local DFT data provides broad coverage while targeted hybrid-functional or many-body perturbation theory calculations refine predictions for challenging cases.

5.3.2. PW basis set

Most Hamiltonian learners operate in atom-centered Gaussian or numerical atomic orbital (NAO) bases, which are chemically intuitive and sparse but limit applicability to extended solids where PW representations dominate. PW DFT—used ubiquitously in solid-state physics for its compatibility with periodic boundary conditions, pseudopotentials, and fast Fourier transforms—represents electronic states in momentum space as $\psi(\mathbf{k})$, with Hamiltonians expressed as $H(\mathbf{k})$ matrices at each crystal momentum $\mathbf{k}$.

Gong *et al* [169] developed a method to generalize deep-learning models of DFT Hamiltonians to results from PW DFT. They do this by reconstructing atomic-orbital Hamiltonians in real space from PW DFT output, allowing NNs trained on these matrices to reproduce electronic structure properties efficiently and with accuracy comparable to standard DFT. Their reconstruction method is orders of magnitude faster than traditional projection approaches and scales linearly with system size, enabling generalization from small to large systems.

5.4. Improving accuracy and efficiency

Beyond foundational architectures and expanded coverage, recent research has focused on improving both the accuracy ceiling of learned Hamiltonians and the efficiency with which they can be trained and deployed. These efforts span novel neural architectures, training strategies that go beyond standard supervised learning, and methods for learning electronic structure at levels of theory more accurate than standard DFT.

While MPNNs and equivariant NNs have proven effective for Hamiltonian learning, alternative architectures continue to emerge. Transformer-based models, which replace local message-passing with global self-attention mechanisms, have shown promise in capturing long-range electronic correlations that decay slowly with distance. Wang *et al* [170] introduced DeepH-2, with a local coordinate transformer backbone for Hamiltonian prediction where atomic and orbital embeddings are updated through multi-head attention layers rather than iterative neighborhood aggregation. This design allows the model to

directly capture interactions between distant atoms—critical for metallic systems, charge-transfer complexes, and conjugated molecules where electronic delocalization spans many atoms.

Their results demonstrate that transformers achieve about two times of better accuracy than DeepH-E3. While computational cost is reduced from $O(L^6)$ to $O(L^3)$ where L denotes the highest angular momentum quantum number.

Rather than relying solely on supervised learning from pre-computed DFT data, Li *et al* [171] developed a physics-informed approach that integrates NN optimization with variational DFT. The method uses a generalized energy functional as the loss function: $\tilde{E}[H_\theta] = E[H_\theta] + \lambda(H_\theta - \tilde{H})$, where H_θ is the NN-predicted Hamiltonian, $E[H_\theta]$ is the DFT total energy, and $\tilde{H}$ is a reconstructed Hamiltonian from the density matrix. This hybrid loss combines physics-based energy minimization (unsupervised) with structural guidance from Hamiltonian reconstruction (supervised).

The approach requires differentiable DFT implementation to compute gradients $\nabla_H E$ via automatic differentiation. The authors developed AI2DFT, a Julia-based code enabling backpropagation through the entire DFT workflow: $H \to \rho \to n \to E$. While models are initialized from supervised pre-training on DeepH-E3, the variational fine-tuning provides substantial improvements in predicting derived quantities. For H_2O molecules, energy predictions achieve 0.013 μeV accuracy compared to 0.83 μeV for pure supervised learning—a 60-fold improvement. Similar gains are observed for graphene and other test systems.

The key advantage is that the energy gradient term $\nabla_H E$ acts as a physical regularizer, filtering unphysical high-energy configurations that arise from fitting noise in training data. This yields Hamiltonians that, despite similar matrix element errors, produce significantly more accurate total energies, density matrices, and charge densities. The method remains within the standard KS DFT framework using conventional functionals (PBE), achieving sub-meV Hamiltonian accuracy and μeV-level total energy accuracy across molecular and periodic systems (H_2O, graphene, MoS_2, bcc Na). While computationally expensive during training due to SCF iterations within the optimization loop, this physics-informed strategy demonstrates that incorporating variational principles can substantially enhance the quality of learned electronic structure beyond conventional data-driven approaches.

Most ML models for molecular electronic structures use DFT databases as ground truth in training, and their prediction accuracy cannot surpass that of DFT [32]. To overcome this fundamental limitation, Tang *et al* developed a unified multi-task ML method (MEHnet) for molecular electronic structures using the gold-standard CCSD(T) calculations as training data.

Rather than directly fitting the single-configurational representation of DFT calculations, MEHnet uses a NN to simulate the non-local XC interactions of a many-body system. The method constructs an effective single-body Hamiltonian by adding a machine-learning correction term V_θ to a fast-to-evaluate mean-field Hamiltonian F' obtained from local DFT. The correction term accounts for non-local XC effects captured in CCSD(T) calculations, but with computational cost that scales linearly with system size.

MEHnet employs E3-equivariant NNs and multi-task learning to predict multiple quantum chemical properties including energy, electric multipole moments, atomic charges, bond orders, energy gaps, and polarizabilities from a shared electronic structure representation (figure 9). Critically, custom backpropagation schemes based on quantum perturbation theory enable numerically stable gradient computation through the Hamiltonian diagonalization step.

Benchmarked on hydrocarbon molecules, MEHnet predicts molecular energies within approximately 0.1 kcal mol^{-1} per atom and consistently exhibits smaller errors than widely used hybrid and double-hybrid functionals across various properties. Remarkably, the model demonstrates robust generalization from small training molecules to larger systems including aromatic compounds and semiconducting polymers with hundreds of atoms—systems for which CCSD(T) calculations are computationally prohibitive (figure 9). By learning directly from coupled-cluster data, this approach demonstrates that ML can transcend the accuracy limitations of conventional density functionals while maintaining computational efficiency comparable to semi-local DFT.

5.5. Summary

ML Hamiltonians have matured from proof-of-concept demonstrations to practical tools enabling simulations previously inaccessible to *ab initio* methods. The field has progressed along multiple fronts: architectures now span from message-passing graphs to equivariant networks and transformers; applications extend from ground-state energies to phonons, excited-state dynamics, and magnetic properties; chemical coverage approaches universality across the periodic table; and representations generalize from localized orbitals to plane waves and reduced dimensions. Meanwhile, learning strategies increasingly target levels

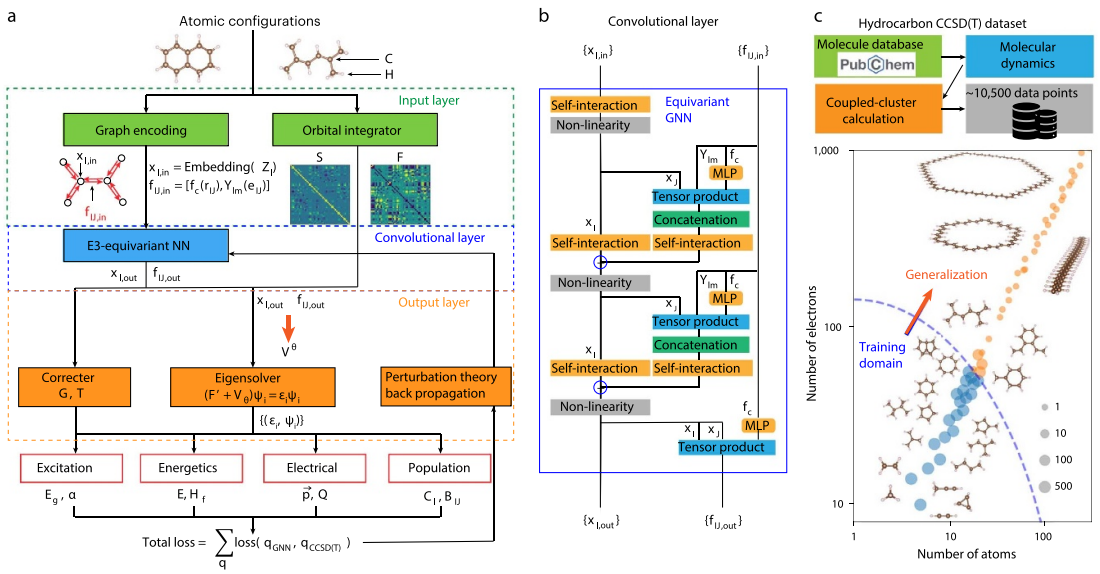

Figure 9. Workflow of MEHnet. (a) and (b) Architecture of the multi-task model, combining property prediction and learning corrections to the Fock matrix. **c**: Training workflow and ability of MEHnet to generalize predictions to molecules far outside its training distribution. Reproduced from [32], with permission from Springer Nature.

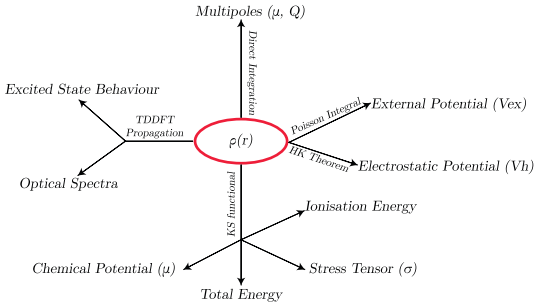

Figure 10. Diagram of commonly calculated (exactly or approximately) properties from the electronic density.

of theory beyond standard DFT, whether through transfer learning to hybrid functionals or variational training that transcends fixed reference data.

Critical challenges remain. Systematic errors persist for strongly correlated materials, charge-transfer complexes, and systems with significant relativistic effects. Dataset curation—particularly balancing chemical diversity against computational cost—continues to limit transferability. And fundamental questions about the optimal inductive biases for electronic structure prediction remain open: when should we enforce physical constraints explicitly versus learning them from data? How can we efficiently incorporate multi-fidelity information spanning semi-local DFT to CCSD(T)?

6. Learning electronic density

6.1. Introduction

This section reviews how ML can be used to predict electron density directly, bypassing SCF calculations. Beyond its intrinsic importance, electron density enables the calculation of many molecular properties (figure 10).

Here, the methods outlined mainly employ supervised learning, aiming to reduce the difference between a true density (usually obtained by DFT) and their predicted density. As such, the error reported is comparing predictions to DFT-level accuracy unless specified otherwise. Moreover, while this review focuses on methods that explicitly learn spatially resolved electronic densities, several studies have extended machine-learning approaches to the prediction of energy-resolved electronic quantities such as the density of states (DOS) and the local density of states (LDOS). These works aim to reconstruct the

electronic spectrum rather than the ground-state charge distribution, effectively learning how states are distributed across energy levels [172–174]. As these studies focus on learning the energy spectra rather than the spatial charge distribution, they fall outside the present scope but remain closely related to the broader goal of data-driven electronic-structure learning, especially since the density can be recovered accurately from these LDOS predictions [175–177]. As such, we do include approaches that predict the LDOS at each grid point as an intermediate representation from which the charge density and total energy are recovered via integration and post-processing, since these are functionally electron density prediction methods and the density remains a central output.

6.2. Learning the HK Map

Brockherde and Bogojeski *et al* [178, 179] proposed to learn the HK map between the external potential $v(\mathbf{r})$ and the ground-state electron density $\rho(\mathbf{r})$, thereby bypassing the explicit solution of the KS equations. Drawing on semiclassical theory, which demonstrates that this map can frequently be approximated smoothly without computing explicit kinetic-energy gradients [180], they employ KRR to directly approximate it:

$$\rho^{\mathrm{ML}}[v](x) = \sum_{i=1}^{M} \beta_i(x)\, k(v, v_i), \tag{35}$$

where $k(v, v_i)$ is a Gaussian kernel measuring the similarity between external potentials, and the coefficients $\beta_i(x)$ weight the contribution of each training potential at spatial grid point x. This non-parametric form works efficiently for one-dimensional systems, but cannot be used in three dimensions because the number of coefficients scales cubically with grid size. To extend the approach to three-dimensional molecules, the authors introduced a basis-set representation of the density:

$$\rho^{\mathrm{ML}}[v](x) = \sum_{l=1}^{L} u^{(l)}[v]\, \phi_l(x), \tag{36}$$

where the machine-learning model is tasked with predicting the coefficients $u^{(l)}[v]$ corresponding to a chosen set of orthogonal basis functions $\phi_l(x)$. Each coefficient is learned via an independent KRR over the training potentials, made possible by the orthogonality of the basis, which decouples the regression problem. For molecular systems, the authors employ an artificial Gaussian-type external potential and use a Fourier (plane-wave) basis representation. Trained on densities and energies obtained with the Perdew–Burke–Ernzerhof (PBE) functional, they reproduce PBE-level results for test systems such as benzene, where a second KRR mapping $\rho \mapsto E$ yields energies. The learned functional can also drive structural relaxations of a water molecule and generate a stable molecular-dynamics trajectory for malonaldehyde, correctly describing its intramolecular proton transfer. Although the approach remains dependent on the XC functional used for training, it established that machine-learned mappings can reconstruct physically accurate densities and energies without cubic-scaling SCF iterations, validating the concept of direct ML surrogates for the HK map. Since this work demonstrated the potential, but also the limitation of supervised learning of the HK map, this inspired a Δ–learning approach from Bogojeski *et al* [181]. Here, the authors obtain a $\Delta E_{\mathrm{ML}}^{\mathrm{CC-DFT}}$ energy which is the difference between the predicted DFT energy and CC energy for a specific system. This is learned via KRR as a function of a PBE-DFT density and is thus added to the PBE-DFT energy to obtain a CC-level energy prediction, demonstrating the ability of ML to learn maps between density and CC energy, enabling CC-level MD simulations at cheap DFT cost.

Moreover, the above method was used by Bai *et al* [182] to learn the HK map to excited state densities, and then a further map to energies. They use linear-response-TDDFT (PBE0) to train the model, enabling good agreement between predicted and ab-initio excited state energies along MD trajectories for malonaldehyde, as well as a convergence to an error <0.2 kcal mol^{-1} for unseen configurations of the molecule (after 5000 training samples). Relatedly, an earlier method, Tsubaki and Mizoguchi's quantum deep field (QDF) [183], uses the HK map but instead of learning $v(\mathbf{r}) \to \rho(\mathbf{r})$ as above, it learns an approximate $\rho(\mathbf{r}) \to v(\mathbf{r})$ map to enforce HK consistency while treating ρ as an unsupervised latent. Given only structure and atomization energy, QDF trains a learned linear combination of atomic orbital density generator together with an energy functional, penalizing mismatch between the predicted and true external potentials and energies. On QM9, it extrapolates to larger, out-of-distribution molecules with ~ 3 kcal mol^{-1} MAE, showing that energy-only supervision plus HK consistency can regularize density learning without explicit ρ labels using the HK map.

6.3. Atom-centered basis expansions

However, to efficiently and reliably capture tensorial targets such as dipoles and polarizabilities, Grisafi *et al* [184, 185] developed the Symmetry-adapted gaussian process regression (SA-GPR) framework, which introduces a tensorial kernel between local atomic environments: the λ-SOAP [50]. This tensorial kernel enforces rotational covariance of the expansion coefficients and measures local geometric similarity. Using this framework, the group represents the electron density through an atom-centered basis expansion, whose additivity and spatial locality allow the model to be transferable and capable of extrapolating to larger systems:

$$\begin{aligned}\rho(\mathbf{r}) &= \sum_i \sum_k c_k^i \, \phi_k(\mathbf{r} - \mathbf{R}_i) \\ &= \sum_i \sum_{nlm} c_{nlm}^i \, R_{nl}(|\mathbf{r} - \mathbf{R}_i|) \, Y_{lm}\left(\widehat{\mathbf{r} - \mathbf{R}_i}\right),\end{aligned} \tag{37}$$

where the coefficients c_k^i are optimized by the model using SA-GPR so that they transform covariantly under rotations of the underlying atomic environments. Each basis function ϕ_k is constructed from contracted Gaussian-type orbitals (GTOs), decomposed into radial R_{nl} and spherical-harmonic Y_{lm} components.

Since these basis functions are not orthogonal, all coefficients become coupled through their overlap matrix, thus the model must simultaneously learn the optimal basis decomposition along with the regression for the coefficients. As a result, when trained on DFT-level densities for butadiene and butane, the model successfully extrapolated to octane with a mean absolute density error of only $\approx$ 1.4%. However, the basis-set representation error remains a limiting factor: when the predicted densities were used for to calculate energies via XC functional evaluation, the results deviated by several kcal mol^{-1} from DFT. To improve on these results, Fabrizio *et al* [186] ensured that the model only has to optimize the expansion coefficients by making the method compatible with density-fitting or resolution of the identity (RI) auxiliary basis sets. Therefore, its accuracy and its ability to generalize improved, allowing the model to accurately predict the density and electrostatic potential of dimers in the BioFragment database [23]. However, capturing long-distance interactions remained a challenge. Therefore, Lewis *et al* expanded this method to periodic systems using symmetry-adapted learning of three-dimensional electron densities (SALTED) [187]. With a linear approximation of the density expansion coefficients and the inclusion of cell translation vectors, Al, Si and ice density predictions were obtained with a root mean squared error (RMSE) less than 2% with fewer than 100 DFT training densities. Another notable result was the model's extrapolation from 4 molecule ice to a 512 supercell within 5 meV per atom of the reference DFT XC and electrostatic energies. In fact, predicting the density and determining these energy values was shown to be more accurate than just predicting the energy, highlighting the ability of density to represent underlying information for ML models to learn. Nevertheless, SALTED becomes inefficient for large or many-atom systems because its use of a non-orthogonal atom-centered basis causes strong coupling between all atomic density components, making the kernel regression scale unfavorably. As such, Grisafi *et al* [188] discretize the density and update the minimization procedure of the loss function to obtain a much more efficient model, capable of simulating bulk liquid water and achieving chemical accuracy in 65% of the test cases from the QM9 dataset. Although the previous SA-GPR models included a cutoff term in their λ-SOAP kernel expression, long-range interactions play a large role in accurately modeling metal surfaces. Therefore, incorporating these into the atomic representation along with explicit equivariance for a finite-field model allowed Grisafi *et al* [189] to accurately simulate electronic responses of metal electrodes up to 60 times faster than DFT.

An alternative methodology also using expansions of atom-centered functions is the Anisotropic Analytical Model of Density (A2MD) [190]. As the name suggests, Cuevas-Zuviría and Pacios use a sum of anisotropic and isotropic functions to analytically calculate each atom's contribution to the total density. Although the fixed analytical form constrains the range of atomic species that can be described, the accompanying NN (A2MDnet) efficiently predicts the coefficients of this expansion, achieving approximately linear scaling of computational cost with system size, albeit with lower accuracy. The approach was later extended in A3MD [191], which incorporates MPNNs. This enabled density predictions for small proteins when restricting the message-passing to intra-monomer interactions, leading to a good fit between the predictions and the DFT calculated density shape/grid. However, the authors emphasize a key problem that plagues density learning in general: molecular properties are very sensitive to errors in densities since they lie 'downstream'. Thus, even a 1% error in density can lead to much larger errors in energy and other properties. Moreover, atom-centered models struggle to expand to more training examples, particularly if they have many elements.

6.4. Grid based approaches

As such, since electronic density can also be seen as the sum of the density at many grid points in 3D, Schmidt *et al* [192] used a grid-based representation of the density. They predicted DFT electron densities on voxelized spatial grids using Bayesian linear regression for crystalline structures such as the Al-Ni alloy. Likewise, Alred *et al* [193] take each atom's charge, their distance to a grid point, and the external potential at the grid point to train a random forest [194] model to predict energy and charge density for sulfur cross-linked nanotubes. Although highly specific, good accuracy is achieved: 0.25% error from DFT.

In parallel, Chandrasekaran *et al* [195] then define and combine rotationally invariant scalar, vector, and tensor fingerprints for grid points. These Gaussian fingerprints serve as the inputs for a 3 layer NN (each with 300 neurons) that learn the corresponding charge density, but also the LDOS spectrum for each grid point with an extra bidirectional recurrent NN layer. A key benefit of this representation is that not many MD snapshots are needed to provide enough training data since the properties are predicted for each grid point, and each snapshot contains many grid points. This method focused on Al and polyethylene (PE) and showed ability to generalize to unseen configurations of both systems with a density RMSE of below 10^{-3} e Å^{-3}. Furthermore, by summing the per-grid-point LDOS prediction, an accurate DOS was obtained. This representation and method was then extended to predict the charge densities of hydrocarbons with an 8 layer model [196]. Charge densities with a 0.91%–1.58% error and energies with a 0.026–1.52 eV RMSE were obtained, but the model struggled with generalizing to unseen structures. By combining this grid point representation of systems and the SOAP representation, Achar *et al* [197] developed DeepCDP. This model takes in atomic positions and uses the SOAP fingerprints to construct power spectrum vectors which are then passed through a model similar to that in Chandrasekaran *et al* [195]. Innovatively, they also enforce a constant number of electrons, which allows them to predict these properties for charged molecules since the model can thus infer the presence of excess charge. Pushing the grid-point paradigm further, Fiedler and Cangi *et al* [176, 177] developed the materials learning algorithms (MALA) framework, which uses spectral neighbor analysis potential bispectrum descriptors computed on a real-space grid as inputs to a feed-forward NN that predicts the LDOS at each grid point. From the predicted LDOS, the electron density and total free energy are recovered via numerical integration and post-processing, effectively making MALA a density prediction method whose intermediate representation is energy-resolved rather than spatial. Trained on simulation cells containing only 256 beryllium atoms, the model maintained constant accuracy across increasing system sizes (256 to 2048 atoms), with energy errors well within chemical accuracy and even below the 10 meV atom^{-1} threshold considered the gold standard for MLIPs. The framework was even accurate when applied to a 131 072-atom beryllium slab containing a stacking fault, achieving up to three orders of magnitude speedup over conventional DFT.

Building on the same grid-point philosophy, Pathrudkar *et al* [198] extended the grid-point framework to multi-million atom metallic and alloy systems (Al, Si, SiGe) using Bayesian NNs. Beyond the scale, their work introduced two capabilities absent from other density learning approaches: calibrated uncertainty quantification, decomposed into epistemic and aleatoric contributions, and transfer learning across system sizes, which drastically reduces the DFT training data required. The model's uncertainty estimates remained stable from small training cells to a 4 million atom density prediction of Al, providing a route to assessing reliability at scales where DFT verification is infeasible.

More recently, this framework was extended to predict electron densities across the composition space of concentrated alloys [164], tackling a combinatorial challenge since the number of required training compositions grows rapidly with the number of alloying species. To combat this, the authors employ Bayesian active learning, in which the model's own uncertainty estimates are used to select the most informative compositions for the next round of training, which reduced the required training data by a factor of 2.5 for ternary (SiGeSn) and 1.7 for quaternary (CrFeCoNi) systems compared to systematic sampling. The resulting models generalized accurately across the full composition space, even including systems with vacancy defects and species segregation not seen during training.

A more physically grounded approach to grid-based density prediction was introduced by Zepeda-Núñez *et al* [199], who proposed deep density, a symmetry-preserving NN. In contrast to models relying on their descriptors for symmetry requirements, deep density analytically embeds permutation, translation, and rotation invariance into the functional form of the network itself. This explicit encoding can often allow for relatively higher efficiency and accuracy of the model. Here, the symmetry-preserving construction allows the model to recover near-DFT accuracy ($\approx$ 1% error) for H_2O and Al densities and to generalize efficiently from 32 atoms to larger configurations, albeit with error $\approx$ 5% for Al with 256 atoms.

While deep density demonstrated that invariance can be embedded directly within the network's functional form, later work sought to reintroduce symmetry preservation at the level of the input representation itself. Indeed, Lv *et al* [200] introduced Deep Charge, a deep-learning model that predicts the electron density $\rho(\mathbf{r})$ from features of a single 'one-shot' DFT calculation for crystalline solids. This is enabled by deep potential-type local atomic embeddings [201] that preserve permutation, translation, and rotation invariance, coupled with a feed-forward network that outputs the real-space density on a grid while enforcing charge conservation and spatial smoothness in the loss function. Trained on diverse datasets including bulk Al and Si, Al–Mg alloys, amorphous Si, and liquid water, Deep Charge achieves mean absolute errors of 5×10^{-4}–8×10^{-4} e Å^{-3} (corresponding to relative errors of approximately 0.05%–0.4%) compared to self-consistent DFT densities.

Although these grid-based approaches demonstrate that spatially resolved learning of the charge density is possible, they often rely on large neural architectures or millions of grid-point evaluations, which make both training and deployment computationally demanding. To overcome this, Focassio *et al* [202] proposed a compact, physics-informed alternative: a linear model based on a Jacobi–Legendre many-body polynomial expansion of local atomic environments (JLCDM) [203]. In this framework, the electron density at each grid point $\mathbf{r}_g$ is expressed as a sum of explicitly defined one-, $\rho_i^{(1)}$, two-, $\rho_{ij}^{(2)}$, and three-body, $\rho_{ijk}^{(3)}$, contributions:

$$\rho\left(\mathbf{r}_g\right) = \sum_i \rho_i^{(1)}\left(\mathbf{r}_g\right) + \sum_{i \neq j} \rho_{ij}^{(2)}\left(\mathbf{r}_g\right) + \sum_{i \neq j,\, i \neq k,\, j \neq k} \rho_{ijk}^{(3)}\left(\mathbf{r}_g\right) + \ldots. \tag{38}$$

Each term is expanded in a set of Jacobi–Legendre polynomials of interatomic distances and angles. Thus, because the model is linear in its parameters, it can be trained by simple linear regression, which offers a speedup compared to large feed-forward, or convolutional NNs (which are the focus of the next section). When evaluated across the full grid, similar accuracy to that obtained with DNNs is yielded for diverse systems such as Al, Mo, and 2D MoS_2, while remaining orders of magnitude cheaper than deep neural grid surrogates or SCF cycles.

Finally, del Rio *et al* [204] extend grid-based density learning into a full deep-learning framework that emulates DFT itself. Their model first predicts the electronic charge density directly from the atomic configuration using GTO descriptors, for which the optimal basis is learned from data. The resulting density descriptors, combined with atomic fingerprints, then serve as inputs to predict key observables, including the total energy, forces, stress, and DOS (figure 11). This architecture closely mirrors the logic of DFT, where the charge density determines all properties, and demonstrates improved accuracy and transferability across molecular and polymeric systems compared to single-step property predictors. This exemplifies the motivation behind learning the underlying electronic structure rather than target properties alone. For materials containing C, H, O, and N, the model achieves a *valence* charge-density error below 2% even for out-of-distribution cases, an impressive result that surpasses many atom-centered methods, and would correspond to an even smaller percentage error when compared to the *total* charge density. Furthermore, like many of the other methods, it offers linear scaling as shown in figure 12, which exemplifies the practicality of such methods.

6.5. Use of CNNs, MPNNs, and equivariance

The advent and development of CNNs [205, 206] enabled the potential for more efficient learning of $x \times y \times z$ 3D voxelized grids by exploiting spatial locality through shared convolutional filters. Indeed, in an early work, Sinitskiy and Pande [207] employ a U-net [208] architecture to simultaneously predict DFT-level density and energy corrections to HF/cc-VDZ calculations for the QM9 dataset. This leads to a MAE compared to DFT of 1.07 kcal mol^{-1}. Lee and Kim [209] outline a more recent approach targeting residual density using a U-Net-type CNN. Their DeepSCF model not only takes a base density as input, but it is also provided with atomistic fingerprints. These modifications, along with data enhancements to account for the lack of rotational equivariance in CNNs, led to an MAE of 0.01 ev/atom for the total energy and a 1.67% deviation from the reference density for molecules in the TABS dataset [210]. Moreover, with this representation, the model could generalize to periodic systems, and a 1644 atom system, for which the density error was 2.10%. This highlighted the power of CNNs when coupled with grid representations, albeit, as mentioned, that these ML architectures do not include the equivariance that is characteristic of chemical systems.

Another ML model apt for the representations of molecular systems are Graph based NNs [54, 55] such as A3MD mentioned above. This was first shown by Gong *et al* [211] using a crystal graph convolution neural network [212] at each grid point to predict the DFT density for that particular point. Although this also scales linearly once trained, it is inefficient as it generates a graph representation

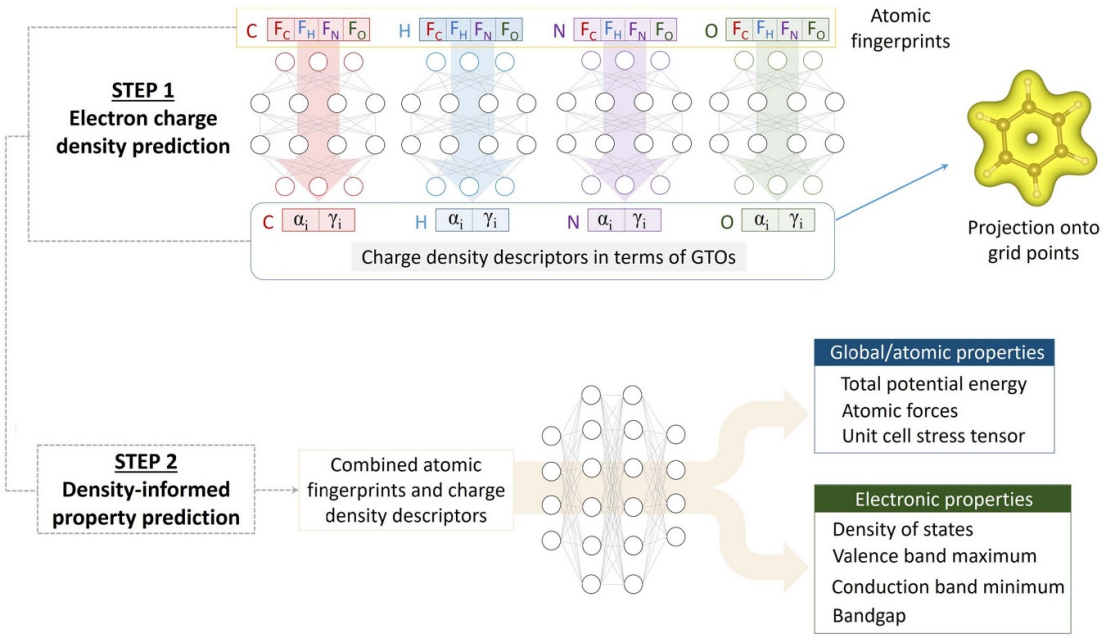

Figure 11. Architecture of the NN, and the workflow used to predict key observables using the density descriptors, analogous to DFT. Reproduced from [204]. CC BY 4.0.

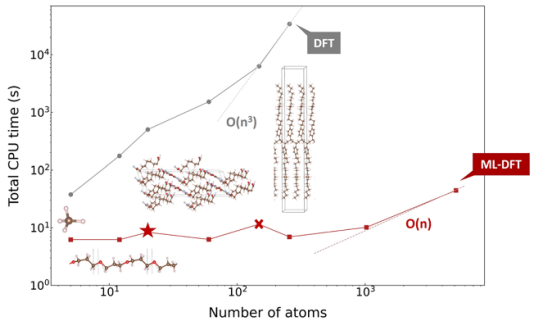

Figure 12. CPU time in seconds for DFT, which scales cubically, and the ML-DFT, which scales linearly with size, like most other methods in this section. Reproduced from [204]. CC BY 4.0.

and a local model for each grid point. Nonetheless, the model is trained on zeolites and polymers and achieves an average RMSE of 1 e Å^{-3}, showing that graphical representations of matter are viable for density prediction.

In order to allow a single graph to learn a representation of the molecule, Jørgensen and Bhowmik [213] employ an invariant MPNN (DeepDFT), trained on QM9 molecules, oxide Li-ion cathodes, and liquid ethylene (PBE) data. By constructing per-atom features from local environments, the model attains an average density MAE $< 0.55\%$ across all systems, which is found to be smaller than the variation introduced by changing the XC functional, illustrating the robustness of message passing. However, these invariant descriptors limit the efficiency and accuracy of the models [57, 214] as they ignore how tensorial quantities should transform. As such, Jørgensen and Bhowmik [215] upgrade DeepDFT to an equivariant MPNN, which markedly reduces errors on the same molecular/solid/liquid benchmarks; using the predicted densities in single-point non-SCF energy evaluations yields MAEs below 5×10^{-4} eV atom^{-1}. Together, these results show that enforcing the correct transformation laws improves accuracy and transfer for learned electron densities. This framework was later extended to charged systems [216] by incorporating the total system charge as an input parameter, enabling accurate density predictions for charged defective perovskites, $LiCoO_2$ with lithium vacancies, and charged diamond-based structures.

Similarly, building on recent advances in equivariant learning, Qiao *et al* [217] (OrbNet-Equi) replace point-cloud features with AO-space inputs from a cheap mean-field calculation (GFN-xTB) and design

a unitary-equivariant network in the AO basis. Rather than predicting voxel values, OrbNet-Equi outputs coefficients of an atom-centered RI (density-fitting) basis for $\rho(\mathbf{r})$ (similarly to (37)), achieving $\approx$0.191 % mean density error on BFDb-SSI [23]) dimers and $\approx$0.206% on QM9, improving over SA-GPR/DeepDFT baselines. Likewise, Lee *et al* [218] use a Euclidean Graph CNN [62] and an RI basis (37) to predict densities for DNA structures. Indeed, being trained on base-pair steps, its atom-centered decomposition of density and repeating motifs in DNA allowed the model to generalize to 5 base-pairs with an average error of 1.06%, from which key electrostatic potentials can be derived. Moreover, due to the almost linear scaling of the method, a prediction for a 108 654 electron DNA structure was possible. This exemplifies the advantageous inclusion of equivariance and scalability of these models. Furthermore, using this RI basis and a E3NN-inspired GNN, Rackers *et al* [214], achieved low (0.4%) density error for 30-molecule water clusters and accurately modeled long-range forces between molecules. Furthermore, they trained the model on CCSD densities and observed the expected jump in accuracy, showing that these density models can (only) learn up to the theory which they are trained on, which offers a path to higher accuracy, but is also a limitation of supervised ML as we stated above. This also allowed them to model the density for 6400 water atoms at near CCSD level (which, if using classic CCSD, would take longer than the age of the Universe).

Pushing equivariant density prediction to much larger and more chemically diverse datasets, Koker *et al* [219] introduced ChargE3Net, an E(3)-equivariant GNN that incorporates higher-order tensor representations beyond the vector features used in PaiNN-based models. Trained on over 100 000 materials from the Materials Project database, ChargE3Net exceeds the performance of prior methods on both QM9 molecules and crystalline materials. Crucially, the authors demonstrated a practical application as a DFT accelerator: using ChargE3Net's predicted densities to initialize SCF calculations on unseen materials reduced the number of self-consistent iterations by 26.7%, and non-self-consistent DFT calculations from these densities yielded near-DFT accuracy for electronic and thermodynamic properties at a fraction of the cost. Analysis attributed the accuracy gains specifically to the higher-order equivariant features, which better capture systems with large angular variations in their charge density.

While probe-based methods like ChargE3Net achieve high accuracy, they require neural message passing over millions of grid points, limiting scalability. Fu *et al* [220] addressed this accuracy–efficiency dilemma from the orbital-based side, identifying three key ingredients for scalable density prediction. First, they augmented atom-centered orbital basis functions with virtual orbitals placed at bond midpoints, directly tackling the above–mentioned limitation that atom-centered bases struggle to capture interatomic density features. Second, they constructed an even-tempered Gaussian basis with learnable exponents, allowing the basis set expressivity to be tuned during training rather than fixed *a priori*. Third, they employed a high-capacity equivariant backbone (eSCN) to predict the orbital coefficients. On the QM9 benchmark, this approach achieved state-of-the-art accuracy while being approximately 30$\times$ faster than prior probe-based methods such as ChargE3Net and InfGCN. Moreover, by adjusting the model and basis set sizes, accuracy–efficiency trade-offs of up to 171$\times$ speedup with only slight degradation in accuracy were demonstrated. This result illustrates that the atom-centered basis set representation error limitation discussed earlier can be substantially mitigated through learnable bases and virtual orbital augmentation, challenging the need for more expensive grid-based representations.

Interestingly, while the progression from invariant to equivariant architectures described above has been a major driver of accuracy improvements, Li *et al* [221] recently demonstrated that the choice of input representation can be equally important. Drawing on the image super-resolution literature from computer vision, they treat the electron density as a 3D grayscale image and use a convolutional residual network to refine a crude superposition of atomic densities into an accurate ground-state density. Because the input is itself a real-space density field, the predictions are implicitly equivariant to rotations and translations without requiring explicit equivariant construction. On QM9, this approach outperforms prior density prediction methods, including equivariant models. The model also generalizes to unseen molecular conformations and, through fine-tuning on limited data ($\approx$8% of the target dataset), achieves 0.33% density error for systems containing chemical elements and charge states absent from training. Together with the training-level dependence observed by Rackers *et al*, this result underscores that accuracy in density learning is governed not only by architectural sophistication but also by the quality of the training theory and the choice of how the problem is represented.

6.6. Neural operators

Neural operators [222] extend regression from finite-dimensional vectors to mappings between function spaces. Instead of learning a conventional parametric model $f_\theta : \mathbb{R}^n \to \mathbb{R}^m$ which maps fixed-length input

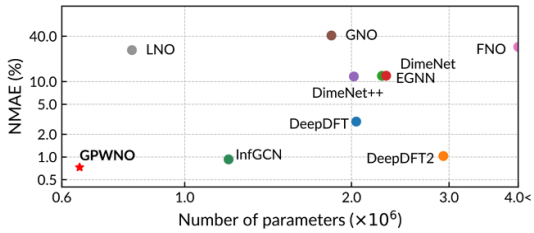

Figure 13. Normalized mean absolute error (NMAE) of the density for various methods mentioned in this section. Illustrates the efficiency of this method (GPWNO); it achieves the most accurate performance with the lowest amount of parameters. Reproduced from [227]. CC BY 4.0.

vectors to output vectors, one learns an operator

$$\mathcal{G}_\theta : \mathcal{A} \to \mathcal{U}, \qquad u = \mathcal{G}_\theta(a),$$

where $a(\mathbf{x})$ and $u(\mathbf{x})$ are continuous fields defined over a spatial domain $\Omega \subset \mathbb{R}^d$. In the context of electronic-structure learning, such models directly approximate the functional mapping $v(\mathbf{r}) \mapsto \rho(\mathbf{r})$, rather than predicting scalar properties derived from it.

Cheng and Peng [223] first applied this concept to electron densities, framing the problem as SE(3)-equivariant operator learning on $\mathbb{R}^3$. They represent the density via a multicentric expansion in atom-centered spherical harmonics with Gaussian radial functions (much like SA-GPR approaches),

$$\hat{\rho}(\mathbf{r}) = \sum_{a \in \text{atoms}} \sum_{n\ell m} c_{a,n\ell m}\, \phi_{n\ell m}(\mathbf{r} - \mathbf{R}_a),$$

where the coefficients $c_{a,n\ell m}$ are generated by an SE(3)-equivariant message-passing network. Moreover, to correct residual basis-set errors, a coordinate-dependent residual operator $r(\mathbf{x})$ refines the prediction pointwise,

$$\rho(\mathbf{x}) = \hat{\rho}(\mathbf{x}) + r(\mathbf{x}).$$

However, unlike SA-GPR, which learns the expansion coefficients via GPR with a fixed kernel between atomic environments, Cheng and Peng's model learns a parametric, system-independent transformation rule that generalizes across geometries and discretizations. The resulting architecture, interpreted as a graphon convolutional operator (InfGCN), provides a continuous analog of graph convolution over infinitely many nodes. This model's combination of function mapping, atom-centered expansions, and equivariant MPNNs that have made up the earlier parts of this section, leads to higher accuracy than many of the previous methods cited on QM9, cubic inorganic materials, and small molecules across MD trajectories. It also outperforms many of the previous neural operators that the authors specifically adapted for this mapping [224–226].

Expanding this framework to periodic systems, a PW contribution is added to the GTO decomposition. Indeed, Kim and Ahn [227] propose a neural-operator architecture that represents the electron density by decomposing it into complementary basis channels: a PW component that captures low-frequency, long-range structure under periodic boundary conditions, and a GTO component that captures high-frequency, near-nuclear features:

$$\rho(\mathbf{r}) = \rho_{\text{PW}}(\mathbf{r}) + \rho_{\text{GTO}}(\mathbf{r})$$

where $\rho_{\text{GTO}}(\mathbf{r})$ is calculated like in Cheng and Peng's method with PBCs and $\rho_{\text{PW}(\mathbf{r})}$ is based on Li *et al*'s Fourier neural operator (FNO) [225], but uses a PW basis. From this, equivariance is maintained and periodicity is achieved. This approach leads to state-of-the-art accuracy and a minimal number of parameters for the QM9 dataset (figure 13) as well as for small molecules across MD trajectories and, as expected, for periodic materials too.

Lastly, these neural operators have also been shown to be able to propagate the density in time. Indeed, Shah and Cangi [228] develop a model based on the FNO that learns the time-propagation operator in TDDFT for 1D diatomics. Once trained on reference TDDFT trajectories, the learned propagator reproduces the evolution of TD density and thus dipole moment after processing for a 1D diatomic under the influence of an oscillating laser pulse with orders-of-magnitude lower cost than explicit numerical integration.

6.7. Use of NQS

Finally, for completeness, we also note that there is literature on the overlap between ML the wavefunction and the electronic density. Indeed, Cheng *et al* [229] combine the Psiformer ansatz with a neural electron real-space density (NERD) head. This allows it to generate a very accurate wavefunction for the system (as shown in the part on NQSs), and use this wavefunction to generate training data for the NERD head for N_e electrons according to:

$$\rho(\mathbf{x}) = \frac{N}{\langle \Psi | \Psi \rangle} \int d\mathbf{x}_2 \dots d\mathbf{x}_{N_e} |\Psi(\mathbf{x}, \mathbf{x}_2, \dots, \mathbf{x}_{N_e})|^2. \tag{39}$$

NERD is an energy based model that then learns to predict this quantity. Due to the extreme accuracy of the wavefunction, the training data provided (by the model itself) allows it to outperform FCI/aug-cc-pVQZ for Be and Li atoms, and capture MR character of H_4 that CCSD could not. Moreover, with such accurate densities, calculated properties such as effective potentials and Hellmann-Feynman forces for small diatomic dissociation curves are also accurate. Nevertheless, this method is subject to the same limitations as those in the NQS section, thus its applicability to larger systems is constrained by the efficacity of the NQS approaches.

6.8. Summary

In summary, learning the electron density $\rho(\mathbf{r})$ provides a principled route to many observables-total energies via explicit functionals, electrostatic potentials and multipole moments, Hellmann–Feynman forces and stress, dipoles and polarizabilities (10). However, the fidelity of these properties is ultimately limited by the accuracy of the density itself, or the functionals used to map to certain properties (i.e. XC functional). Small systematic errors such as near-nuclear cusps, long-range decay, or symmetry violations can be amplified when evaluating derived quantities, so physically consistent constraints (normalization to N_e, positivity, SE(3) symmetry or periodicity, and smoothness) to maximize accuracy are important. The field has progressed from kernel methods and atom-centered expansions to grid models, equivariant GNNs, neural operators, and NQS-based formulations each trading off expressivity, cost, and built-in physics.

We anticipate further development of the neural operator approach, perhaps toward hybrid approaches that couple neural operators with explicit physics priors (charge conservation, asymptotics, response constraints). Furthermore, training on more accurate QC methods, as illustrated by Cheng *et al* [229] and Rackers *et al* [214], has a big impact on predictive accuracy, and thus future models might need to incorporate this to achieve better accuracy. Extensions to real-time electron dynamics via learned propagators, and to finite-temperature and excited-state densities, suggest a broader role for density learning as a fast, transferable interface between electronic structure and data-driven simulation.

7. Conclusion

Through the exploration of the recent advances in multiple areas of ML for electronic structure prediction, we have shown the rapid growth in popularity and the ensuing rise in accuracy across all areas. Indeed, having defined our 'ML Jacob's ladder', we have presented methods which learn the entire wavefunction down to those that predict the charge density, with tradeoffs in accuracy and scaling being prevalent as we compare methods. However, it is important to note that for supervised learning the data the models are trained on limits in potential accuracy. Therefore, generating more high accuracy datasets (such as CCSD(T)-level) with increasing diversity remains an essential part to obtaining chemically accurate models. The unsupervised NQSs have shown excellent accuracy, and the ability to be directly used as chemically accurate training data for sub-models exemplifies an interesting approach to this issue. Nevertheless, NQSs, even with the recent architectural upgrades, are limited to smaller systems. Despite this, the emergence of generalizable models across rungs in the ladder promises to make these models more applicable in the *ab initio* simulation community. We believe a key factor in this progress has been the inspiration from the ever-accelerating field of ML, and we envision further development in ML for QC to come from similar sources.

As such, we expect the field of ML for electronic structure to keep growing quickly, as it has been doing over the past years. The influence of these models will continue to grow as they aim to make a generalizable and accurate model that can allow researchers to limit their dependency and resources allocated to the original QC methods.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Tim S Hindges 0000-0002-0647-7163
Ju Li 0000-0002-7841-8058

References

[1] Becke A D 2014 *J. Chem. Phys.* **140** 18A301
[2] Merz K M 2014 *Acc. Chem. Res.* **47** 2804–11
[3] Jain A, Shin Y and Persson K A 2016 *Nat. Rev. Mater.* **1** 15004
[4] Heryanto H, Ardiansyah A, Rahmat R and Tahir D 2024 *JOM J. Miner. Met. Mater. Soc.* **76** 4629–42
[5] Teale A M *et al* 2022 *Phys. Chem. Chem. Phys.* **24** 28700–81
[6] Swiss National Supercomputing Centre 2024 CSCS annual report 2024: shaping global collaborations in high-performance computing (available at: https://report2024.cscs.ch/) (Accessed 11 Dec 2025)
[7] Heinen S, Schwilk M, Von Rudorff G F and Von Lilienfeld O A 2020 *Mach. Learn.: Sci. Technol.* **1** 025002
[8] Vogiatzis K D, Ma D, Olsen J, Gagliardi L and De Jong W A 2017 *J. Chem. Phys.* **147** 184111
[9] Hansch C, Maloney P P, Fujita T and Muir R M 1962 *Nature* **194** 178–80
[10] Wei J, Chu X, Sun X-Y, Xu K, Deng H-X, Chen J, Wei Z and Lei M 2019 *InfoMat* **1** 338–58
[11] Kulik H J 2020 *WIREs Comput. Mol. Sci.* **10** e1439
[12] Walters W P and Barzilay R 2021 *Acc. Chem. Res.* **54** 263–70
[13] Kulik H J *et al* 2022 *Electron. Struct.* **4** 023004
[14] Shilpa S, Kashyap G and Sunoj R B 2023 *J. Phys. Chem. A* **127** 8253–71
[15] Griesemer S D, Xia Y and Wolverton C 2023 *Nat. Comput. Sci.* **3** 934–45
[16] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A and Müller K R 2021 *Chem. Rev.* **121** 10142–86
[17] Wan K, He J and Shi X 2024 *Adv. Mater.* **36** 2305758
[18] Jacobs R *et al* 2025 *Curr. Opin. Solid State Mater. Sci.* **35** 101214
[19] Unke O T, Chmiela S, Gastegger M, Schütt K T, Sauceda H E and Müller K R 2021 *Nat. Commun.* **12** 7273
[20] Levine D S *et al* 2025 arXiv:2505.08762
[21] Ramakrishnan R, Dral P O, Rupp M and Von Lilienfeld O A 2014 *Sci. Data* **1** 140022
[22] Scherbela M, Gerard L and Grohs P 2024 *Nat. Commun.* **15** 120
[23] Burns L A, Faver J C, Zheng Z, Marshall M S, Smith D G A, Vanommeslaeghe K, MacKerell A D, Merz K M and Sherrill C D 2017 *J. Chem. Phys.* **147** 161727
[24] Rasmussen C E and Williams C K 2006 *Gaussian Processes for Machine Learning* (MIT Press)
[25] Saunders C, Gammerman A and Vovk V 1999 *ICML-1998 Proc. 15th Int. Conf. on Machine Learning*
[26] Hornik K, Stinchcombe M and White H 1989 *Neural Netw.* **2** 359–66
[27] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2015 *J. Chem. Theory Comput.* **11** 2087–96
[28] Zaspel P, Huang B, Harbrecht H and Von Lilienfeld O A 2018 *J. Chem. Theory Comput.* **15** 1546–59
[29] Dral P O, Owens A, Dral A and Csányi G 2020 *J. Chem. Phys.* **152** 204110
[30] Vinod V, Kleinekathöfer U and Zaspel P 2024 *Mach. Learn.: Sci. Technol.* **5** 015054
[31] Atz K, Isert C, Böcker M N, Jiménez-Luna J and Schneider G 2022 *Phys. Chem. Chem. Phys.* **24** 10775–83
[32] Tang H, Xiao B, He W, Subasic P, Harutyunyan A R, Wang Y, Liu F, Xu H and Li J 2025 *Nat. Comput. Sci.* **5** 144–54
[33] Perdew J P 2001 Jacob's ladder of density functional approximations for the exchange-correlation energy *AIP Conf. Proc.* **577** 1–20
[34] Kalita B, Li L, McCarty R J and Burke K 2021 *Acc. Chem. Res.* **54** 818–26
[35] Fiedler L, Shah K, Bussmann M and Cangi A 2022 *Phys. Rev. Mater.* **6** 040301
[36] Pederson R, Kalita B and Burke K 2022 *Nat. Rev. Phys.* **4** 357–8
[37] Hermann J, Spencer J, Choo K, Mezzacapo A, Foulkes W M C, Pfau D, Carleo G and Noé F 2023 *Nat. Rev. Chem.* **7** 692–709
[38] Battaglia S 2023 Machine learning wavefunction *Quantum Chemistry in the Age of Machine Learning* (Elsevier) pp 577–616
[39] Bartlett R J and Musiał M 2007 *Rev. Mod. Phys.* **79** 291–352
[40] Echenique P and Alonso J L 2007 *Mol. Phys.* **105** 3057–98
[41] Park J W, Al-Saadon R, MacLeod M K, Shiozaki T and Vlaisavljevich B 2020 *Chem. Rev.* **120** 5878–909
[42] Bishop C M 2006 *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer)
[43] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436–44
[44] Goodfellow I, Bengio Y, Courville A and Bengio Y 2016 *Deep Learning* vol 1 (MIT Press)
[45] Strout D L and Scuseria G E 1995 *J. Chem. Phys.* **102** 8448–52
[46] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864–71
[47] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133–8
[48] Mohr S, Ratcliff L E, Genovese L, Caliste D, Boulanger P, Goedecker S and Deutsch T 2015 *Phys. Chem. Chem. Phys.* **17** 31360–70
[49] Runge E and Gross E K U 1984 *Phys. Rev. Lett.* **52** 997–1000
[50] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev. B* **87** 184115
[51] Rumelhart D E, Hinton G E and Williams R J 1986 *Nature* **323** 533–6
[52] Behler J and Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
[53] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C and Sun M 2020 *AI Open* **1** 57–81

[54] Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural message passing for quantum chemistry *Int. Conf. on Machine Learning* (PMLR) pp 1263–72
[55] Schütt K, Kindermans P J, Sauceda Felix H E, Chmiela S, Tkatchenko A and Müller K R 2017 *Advances in Neural Information Processing Systems* vol 30
[56] Gasteiger J, Giri S, Margraf J T and Günnemann S 2022 Fast and uncertainty-aware directional message passing for non-equilibrium molecules *Published at the Machine Learning for Molecules Workshop at NeurIPS 2020. Author Name Changed From Johannes Klicpera to Johannes Gasteiger*
[57] Schütt K, Unke O and Gastegger M 2021 Equivariant message passing for the prediction of tensorial properties and molecular spectra *Int. Conf. on Machine Learning* (PMLR) pp 9377–88
[58] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 *Advances in Neural Information Processing Systems* vol 30
[59] Maziarka Ł, Danel T, Mucha S, Rataj K, Tabor J and Jastrzebski S 2020 arXiv:2002.08264
[60] Liao Y L, Wood B, Das A and Smidt T 2023 arXiv:2306.12059
[61] Von Lilienfeld O A, Ramakrishnan R, Rupp M and Knoll A 2015 *Int. J. Quantum Chem.* **115** 1084–93
[62] Geiger M and Smidt T 2022 arXiv:2207.09453
[63] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E and Kozinsky B 2022 *Nat. Commun.* **13** 2453
[64] Bochkarev A, Lysogorskiy Y, Ortner C, Csányi G and Drautz R 2022 *Phys. Rev. Res.* **4** L042019
[65] Foulkes W M C 2020 Variational wave functions for molecules and solids *Topology, Entanglement and Strong Correlations* ed E Pavarini and EKoch, vol 10 (*Forschungszentrum JüLich: Modeling and Simulation*)
[66] Jastrow R 1955 *Phys. Rev.* **98** 1479–84
[67] Ceperley D 1978 *Phys. Rev. B* **18** 3126–38
[68] Pfau D, Spencer J S, Matthews A G D G and Foulkes W M C 2020 *Phys. Rev. Res.* **2** 033429
[69] Carleo G and Troyer M 2017 *Science* **355** 602–6
[70] Rumelhart D E and McClelland J L 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (The MIT Press)
[71] Feynman R P and Cohen M 1956 *Phys. Rev.* **102** 1189–204
[72] Holzmann M and Moroni S 2019 *Phys. Rev. B* **99** 085121
[73] Luo D and Clark B K 2019 *Phys. Rev. Lett.* **122** 226401
[74] Hermann J, Schätzle Z and Noé F 2020 *Nat. Chem.* **12** 891–7
[75] Kingma D P 2014 arXiv:1412.6980
[76] Martens J and Grosse R 2015 Optimizing neural networks with Kronecker-factored approximate curvature *Int. Conf. on Machine Learning* (PMLR) pp 2408–17
[77] Spencer J S, Pfau D, Botev A and Foulkes W M C 2020 arXiv:2011.07125
[78] Ren W, Fu W, Wu X and Chen J 2023 *Nat. Commun.* **14** 1860
[79] Li X, Fan C, Ren W and Chen J 2022 *Phys. Rev. Res.* **4** 013021
[80] Gerard L, Scherbela M, Marquetand P and Grohs P 2022 *Advances in Neural Information Processing Systems* vol 35 pp 10282–94
[81] Yoshioka N, Mizukami W and Nori F 2021 *Commun. Phys.* **4** 106
[82] Pescia G, Han J, Lovato A, Lu J and Carleo G 2022 *Phys. Rev. Res.* **4** 023138
[83] Cassella G, Sutterud H, Azadi S, Drummond N D, Pfau D, Spencer J S and Foulkes W M C 2023 *Phys. Rev. Lett.* **130** 036401
[84] Ewald P P 1921 *Ann. Phys., Lpz.* **369** 253–87
[85] Wilson M, Moroni S, Holzmann M, Gao N, Wudarski F, Vegge T and Bhowmik A 2023 *Phys. Rev. B* **107** 235139
[86] Pescia G, Nys J, Kim J, Lovato A and Carleo G 2024 *Phys. Rev. B* **110** 035108
[87] Li X, Li Z and Chen J 2022 *Nat. Commun.* **13** 7895
[88] Vicentini F *et al* 2022 *SciPost Phys. Codebases* 7
[89] Schätzle Z, Szabó P B, Mezera M, Hermann J and Noé F 2023 *J. Chem. Phys.* **159** 094108
[90] Renaud N 2025 arXiv:2506.09743
[91] Sharir O, Levine Y, Wies N, Carleo G and Shashua A 2020 *Phys. Rev. Lett.* **124** 020503
[92] Barrett T D, Malyshev A and Lvovsky A I 2022 *Nat. Mach. Intell.* **4** 351–8
[93] Li R *et al* 2023 arXiv:2307.08214
[94] Scherbela M, Gao N, Grohs P and Günnemann S 2025 arXiv:2504.06087
[95] von Glehn I, Spencer J S and Pfau D 2022 arXiv:2211.13672
[96] Viteritti L L, Rende R and Becca F 2023 *Phys. Rev. Lett.* **130** 236401
[97] Gu Y, *et al* 2025 arXiv:2507.02644
[98] Geier M, Nazaryan K, Zaklama T and Fu L 2025 *Phys. Rev. B* **112** 045119
[99] Teng Y, Dai D D and Fu L 2025 *Phys. Rev. B* **111** 205117
[100] Foster A, Schätzle Z, Szabó P B, Cheng L, Köhler J, Cassella G, Gao N, Li J, Noé F and Hermann J 2025 arXiv:2506.19960
[101] Schmitt M and Heyl M 2020 *Phys. Rev. Lett.* **125** 100503
[102] Gutiérrez I L and Mendl C B 2022 *Quantum* **6** 627
[103] Sinibaldi A, Giuliani C, Carleo G and Vicentini F 2023 *Quantum* **7** 1131
[104] Gartner M, Mazzanti F and Zillich R 2022 *SciPost Phys.* **13** 025
[105] Medvidović M and Sels D 2023 *PRX Quantum* **4** 040302
[106] Nys J, Pescia G, Sinibaldi A and Carleo G 2024 *Nat. Commun.* **15** 9404
[107] Carleo G, Becca F, Schiró M and Fabrizio M 2012 *Sci. Rep.* **2** 243
[108] Choo K, Carleo G, Regnault N and Neupert T 2018 *Phys. Rev. Lett.* **121** 167204
[109] Entwistle M T, Schätzle Z, Erdman P A, Hermann J and Noé F 2023 *Nat. Commun.* **14** 274
[110] Pathak S, Busemeyer B, Rodrigues J N B and Wagner L K 2021 *J. Chem. Phys.* **154** 034101
[111] Lu Z and Fu W 2023 arXiv:2311.17595
[112] Pfau D, Axelrod S, Sutterud H, Von Glehn I and Spencer J S 2024 *Science* **385** eadn0137
[113] Szabó P B, Schätzle Z, Entwistle M T and Noé F 2024 *J. Chem. Theory Comput.* **20** 7922–35
[114] Li Z, Lu Z, Li R, Wen X, Li X, Wang L, Chen J and Ren W 2024 *Nat. Comput. Sci.* **4** 910–9
[115] Schätzle Z, Szabó P B, Cuzzocrea A and Noé F 2025 arXiv:2503.19847
[116] Scherbela M, Reisenhofer R, Gerard L, Marquetand P and Grohs P 2022 *Nat. Comput. Sci.* **2** 331–41

[117] Gao N and Günnemann S 2021 arXiv:2110.05064
[118] Gao N and Günnemann S 2023 Generalizing neural wave functions *Int. Conf. on Machine Learning* (PMLR) pp 10708–26
[119] Brown T *et al* 2020 *Advances in Neural Information Processing Systems* vol 33 pp 1877–901
[120] Radford A *et al* 2021 Learning transferable visual models from natural language supervision *Int. Conf. on Machine Learning* (PMLR) pp 8748–63
[121] Zhang Y-H and Di Ventra M 2023 *Phys. Rev. B* **107** 075147
[122] Scherbela M, Gerard L and Grohs P 2023 *Advances in Neural Information Processing Systems* vol 36 pp 23902–20
[123] Welling M and Teh Y W 2011 Bayesian learning via stochastic gradient Langevin dynamics *Proc. 28th International Conf. on Machine Learning* (*ICML-11*) pp 681–8
[124] Rende R, Viteritti L L, Becca F, Scardicchio A, Laio A and Carleo G 2025 *Nat. Commun.* **16** 7213
[125] Aldegunde M, Kermode J R and Zabaras N 2016 *J. Comput. Phys.* **311** 173–95
[126] Ryabov A, Akhatov I and Zhilyaev P 2020 *Sci. Rep.* **10** 8000
[127] Lei X and Medford A J 2019 *Phys. Rev. Mater.* **3** 063801
[128] Nagai R, Akashi R and Sugino O 2022 *Phys. Rev. Res.* **4** 013106
[129] Bystrom K and Kozinsky B 2022 *J. Chem. Theory Comput.* **18** 2180–92
[130] Dick S and Fernandez-Serra M 2020 *Nat. Commun.* **11** 3509
[131] Kirkpatrick J *et al* 2021 *Science* **374** 1385–9
[132] Luise G, *et al* 2025 arXiv:2506.14665
[133] Gao N, Eberhard E and Günnemann S 2024 arXiv:2410.07972
[134] Bystrom K, Falletta S and Kozinsky B 2024 *Phys. Rev. B* **110** 075130
[135] Snyder J C, Rupp M, Hansen K, Muller K-R and Burke K 2012 *Phys. Rev. Lett.* **108** 253002
[136] Snyder J C, Rupp M, Hansen K, Blooston L, Muller K-R and Burke K 2013 *J. Chem. Phys.* **139** 224104
[137] Mitchell M A J, Del Aguila Ferrandis T and Sanvito S 2024 arXiv:2412.08143
[138] Manzhos S, Lüder J, Golub P and Ihara M 2025 arXiv:2502.05411
[139] Meyer R, Weichselbaum M and Hauser A W 2020 *J. Chem. Theory Comput.* **16** 5685–94
[140] Manzhos S, Lüder J and Ihara M 2023 *J. Chem. Phys.* **159** 234115
[141] Golub P and Manzhos S 2019 *Phys. Chem. Chem. Phys.* **21** 378–95
[142] Sun L and Chen M 2024 *Phys. Rev. B* **109** 115135
[143] Sun L and Chen M 2024 *Electron. Struct.* **6** 035006
[144] Remme R, Kaczun T, Scheurer M, Dreuw A and Hamprecht F A 2023 *J. Chem. Phys.* **159** 144113
[145] Kasim M F and Vinko S M 2021 *Phys. Rev. Lett.* **127** 126403
[146] Li H, Wang Z, Zou N, Ye M, Xu R, Gong X, Duan W and Xu Y 2022 *Nat. Comput. Sci.* **2** 367–77
[147] Zhouyin Z, Gan Z, Liu M, Pandey S K, Zhang L and Gu Q 2024 arXiv:2407.06053
[148] Gong X, Li H, Zou N, Xu R, Duan W and Xu Y 2023 *Nat. Commun.* **14** 2848
[149] Li H, Tang Z, Fu J, Dong W-H, Zou N, Gong X, Duan W and Xu Y 2024 *Phys. Rev. Lett.* **132** 096401
[150] Zhong Y, Liu S, Zhang B, Tao Z, Sun Y, Chu W, Gong X-G, Yang J-H and Xiang H 2024 *Nat. Comput. Sci.* **4** 615–25
[151] Zhang C, Zhong Y, Tao Z-G, Qin X, Shang H, Lan Z, Prezhdo O V, Gong X-G, Chu W and Xiang H 2025 *Nat. Commun.* **16** 2033
[152] Tully J C 1990 *J. Chem. Phys.* **93** 1061–71
[153] Tully J C 2012 *J. Chem. Phys.* **137** 22A301
[154] Barbatti M 2011 *WIREs Comput. Mol. Sci.* **1** 620–33
[155] Crespo-Otero R and Barbatti M 2018 *Chem. Rev.* **118** 7164–207
[156] Barbatti M, Ruckenbauer M, Plasser F, Pittner J, Granucci G, Persico M and Lischka H 2014 *WIREs Comput. Mol. Sci.* **4** 26–33
[157] Dral P O, Barbatti M and Thiel W 2018 *J. Phys. Chem. Lett.* **9** 5660–3
[158] Dral P O and Barbatti M 2021 *Nat. Rev. Chem.* **5** 388–405
[159] Westermayr J, Gastegger M, Menger M F S J, Mai S, González L and Marquetand P 2019 *Chem. Sci.* **10** 8100–7
[160] Westermayr J and Marquetand P 2021 *Chem. Rev.* **121** 9873–926
[161] Yarkony D R 1996 *Rev. Mod. Phys.* **68** 985–1013
[162] W Domcke, D R Yarkony and H Köppel ed 2004 *Conical Intersections: Electronic Structure, Dynamics and Spectroscopy* (World Scientific)
[163] Domcke W and Yarkony D R 2012 *Annu. Rev. Phys. Chem.* **63** 325–52
[164] Pathrudkar S, Taylor S, Keripale A, Gangan A S, Thiagarajan P, Agarwal S, Marian J, Ghosh S and Banerjee A S 2025 *npj Comput. Mater.* **11** 378
[165] Li H, Tang Z, Gong X, Zou N, Duan W and Xu Y 2023 *Nat. Comput. Sci.* **3** 321–7
[166] Wang Y *et al* 2024 *Sci. Bull.* **69** 2514–21
[167] Zhong Y, Yu H, Yang J, Guo X, Xiang H and Gong X 2024 *Chin. Phys. Lett.* **41** 077103
[168] Kaniselvan M, Miller B K, Gao M, Nam J and Levine D S 2025 arXiv:2510.00224
[169] Gong X, Louie S G, Duan W and Xu Y 2024 *Nat. Comput. Sci.* **4** 752–60
[170] Wang Y, Li H, Tang Z, Tao H, Wang Y, Yuan Z, Chen Z, Duan W and Xu Y 2024 arXiv:2401.17015
[171] Li Y, Tang Z, Chen Z, Sun M, Zhao B, Li H, Tao H, Yuan Z, Duan W and Xu Y 2024 *Phys. Rev. Lett.* **133** 076401
[172] Umeno Y and Kubo A 2019 *Comput. Mater. Sci.* **168** 164–71
[173] Ben Mahmoud C, Anelli A, Csányi G and Ceriotti M 2020 *Phys. Rev. B* **102** 235130
[174] Kong S, Ricci F, Guevarra D, Neaton J B, Gomes C P and Gregoire J M 2022 *Nat. Commun.* **13** 949
[175] Ellis J A, Fiedler L, Popoola G A, Modine N A, Stephens J A, Thompson A P, Cangi A and Rajamanickam S 2021 *Phys. Rev. B* **104** 035120
[176] Fiedler L, Modine N A, Schmerler S, Vogel D J, Popoola G A, Thompson A P, Rajamanickam S and Cangi A 2023 *npj Comput. Mater.* **9** 115
[177] Cangi A *et al* 2025 *Comput. Phys. Commun.* **314** 109654
[178] Brockherde F, Vogt L, Li L, Tuckerman M E, Burke K and Müller K R 2017 *Nat. Commun.* **8** 872
[179] Bogojeski M, Brockherde F, Vogt-Maranto L, Li L, Tuckerman M E, Burke K and Müller K R 2018 arXiv:1811.06255
[180] Ribeiro R F, Lee D, Cangi A, Elliott P and Burke K 2015 *Phys. Rev. Lett.* **114** 050401
[181] Bogojeski M, Vogt-Maranto L, Tuckerman M E, Müller K-R and Burke K 2020 *Nat. Commun.* **11** 5223
[182] Bai Y, Vogt-Maranto L, Tuckerman M E and Glover W J 2022 *Nat. Commun.* **13** 7044
[183] Tsubaki M and Mizoguchi T 2020 *Phys. Rev. Lett.* **125** 206401

[184] Grisafi A, Wilkins D M, Csányi G and Ceriotti M 2018 *Phys. Rev. Lett.* **120** 036002
[185] Grisafi A, Fabrizio A, Meyer B, Wilkins D M, Corminboeuf C and Ceriotti M 2019 *ACS Cent. Sci.* **5** 57–64
[186] Fabrizio A, Grisafi A, Meyer B, Ceriotti M and Corminboeuf C 2019 *Chem. Sci.* **10** 9424–32
[187] Lewis A M, Grisafi A, Ceriotti M and Rossi M 2021 *J. Chem. Theory Comput.* **17** 7203–14
[188] Grisafi A, Lewis A M, Rossi M and Ceriotti M 2023 *J. Chem. Theory Comput.* **19** 4451–60
[189] Grisafi A, Bussy A, Salanne M and Vuilleumier R 2023 *Phys. Rev. Mater.* **7** 125403
[190] Cuevas-Zuviría B and Pacios L F 2020 *J. Chem. Inf. Modeling* **60** 3831–42
[191] Cuevas-Zuviría B and Pacios L F 2021 *J. Chem. Inf. Modeling* **61** 2658–66
[192] Schmidt E, Fowler A T, Elliott J A and Bristowe P D 2018 *Comput. Mater. Sci.* **149** 250–8
[193] Alred J M, Bets K V, Xie Y and Yakobson B I 2018 *Compos. Sci. Technol.* **166** 3–9
[194] Breiman L 2001 *Mach. Learn.* **45** 5–32
[195] Chandrasekaran A, Kamal D, Batra R, Kim C, Chen L and Ramprasad R 2019 *npj Comput. Mater.* **5** 22
[196] Kamal D, Chandrasekaran A, Batra R and Ramprasad R 2020 *Mach. Learn.: Sci. Technol.* **1** 025003
[197] Achar S K, Bernasconi L and Johnson J K 2023 *Nanomaterials* **13** 1853
[198] Pathrudkar S, Thiagarajan P, Agarwal S, Banerjee A S and Ghosh S 2024 *npj Comput. Mater.* **10** 175
[199] Zepeda-Núñez L, Chen Y, Zhang J, Jia W, Zhang L and Lin L 2021 *J. Comput. Phys.* **443** 110523
[200] Lv T, Zhong Z, Liang Y, Li F, Huang J and Zheng R 2023 *Phys. Rev. B* **108** 235159
[201] Zeng J *et al* 2023 *J. Chem. Phys.* **159** 054801
[202] Focassio B, Domina M, Patil U, Fazzio A and Sanvito S 2023 *npj Comput. Mater.* **9** 87
[203] Domina M, Patil U, Cobelli M and Sanvito S 2023 *Phys. Rev. B* **108** 094102
[204] Del Rio B G, Phan B and Ramprasad R 2023 *npj Comput. Mater.* **9** 158
[205] Lecun Y, Bottou L, Bengio Y and Haffner P 1998 *Proc. IEEE* **86** 2278–324
[206] Krizhevsky A, Sutskever I and Hinton G E 2012 *Advances in Neural Information Processing Systems* vol 25
[207] Sinitskiy A V and Pande V S 2018 arXiv:1809.02723
[208] Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ed N Navab, J Hornegger, W M Wells and A F Frangi (Springer) pp 234–41
[209] Lee R-G and Kim Y-H 2024 *npj Comput. Mater.* **10** 248
[210] Blair S A and Thakkar A J 2014 *Comput. Theor. Chem.* **1043** 13–16
[211] Gong S, Xie T, Zhu T, Wang S, Fadel E R, Li Y and Grossman J C 2019 *Phys. Rev. B* **100** 184103
[212] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[213] Jørgensen P B and Bhowmik A 2020 arXiv:2011.03346
[214] Rackers J A, Tecot L, Geiger M and Smidt T E 2023 *Mach. Learn.: Sci. Technol.* **4** 015027
[215] Jørgensen P B and Bhowmik A 2022 *npj Comput. Mater.* **8** 183
[216] Tawfik S A, Gupta S and Venkatesh S 2025 *J. Phys. Chem. A* **129** 2117–22
[217] Qiao Z, Christensen A S, Welborn M, Manby F R, Anandkumar A and Miller T F 2022 *Proc. Natl Acad. Sci.* **119** e2205221119
[218] Lee A J, Rackers J A and Bricker W P 2022 *Biophys. J.* **121** 3883–95
[219] Koker T, Quigley K, Taw E, Tibbetts K and Li L 2024 *npj Comput. Mater.* **10** 161
[220] Fu X, Rosen A, Bystrom K, Wang R, Musaelian A, Kozinsky B, Smidt T and Jaakkola T 2024 *Advances in Neural Information Processing Systems* vol 37 pp 9727–52
[221] Li C, Sharir O, Yuan S and Chan G K L 2025 *Nat. Commun.* **16** 4811
[222] Azizzadenesheli K, Kovachki N, Li Z, Liu-Schiaffini M, Kossaifi J and Anandkumar A 2024 *Nat. Rev. Phys.* **6** 320–8
[223] Cheng C and Peng J 2023 *Advances in Neural Information Processing Systems* vol 36 pp 61960–84
[224] Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A and Anandkumar A 2020 arXiv:2003.03485
[225] Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A and Anandkumar A 2020 arXiv:2010.08895
[226] Lu L, Jin P, Pang G, Zhang Z and Karniadakis G E 2021 *Nat. Mach. Intell.* **3** 218–29
[227] Kim S and Ahn S 2024 arXiv:2402.04278
[228] Shah K and Cangi A 2024 arXiv:2407.09628
[229] Cheng L *et al* 2025 *J. Chem. Phys.* **162** 034120