Large-scale materials modeling at quantum accuracy: *Ab initio* simulations of quasicrystals and interacting extended defects in metallic alloys

Sambit Das*
University of Michigan
Department of Mechanical Engineering
Ann Arbor, MI, USA

Bikash Kanungo*
University of Michigan
Department of Mechanical Engineering
Ann Arbor, MI, USA

Vishal Subramanian*
University of Michigan
Department of Materials Science & Engineering
Ann Arbor, MI, USA

Gourab Panigrahi
Indian Institute of Science
Department of Computational and Data Sciences
Bangalore, Karnataka, India

Phani Motamarri
Indian Institute of Science
Department of Computational and Data Sciences
Bangalore, Karnataka, India

David Rogers
Oak Ridge National Laboratory
Scientific Computing Group
Oak Ridge, TN, USA

Paul M. Zimmerman
University of Michigan
Department of Chemistry
Ann Arbor, MI, USA

Vikram Gavini†
University of Michigan
Department of Mechanical Engineering
Ann Arbor, MI, USA

ABSTRACT

Ab initio electronic-structure has remained dichotomous between achievable accuracy and length-scale. Quantum many-body (QMB) methods realize quantum accuracy but fail to scale. Density functional theory (DFT) scales favorably but remains far from quantum accuracy. We present a framework that breaks this dichotomy by use of three interconnected modules: (i) `invDFT`: a methodological advance in *inverse* DFT linking QMB methods to DFT; (ii) `MLXC`: a machine-learned density functional trained with `invDFT` data, commensurate with quantum accuracy; (iii) `DFT-FE-MLXC`: an adaptive higher-order spectral finite-element (FE) based DFT implementation that integrates `MLXC` with efficient solver strategies and HPC innovations in FE-specific dense linear algebra, mixed-precision algorithms, and asynchronous compute-communication. We demonstrate a paradigm shift in DFT that not only provides an accuracy commensurate with QMB methods in ground-state energies, but also attains an unprecedented performance of 659.7 PFLOPS (43.1% peak FP64 performance) on 619,124 electrons using 8,000 GPU nodes of Frontier supercomputer.

*Sambit Das, Bikash Kanungo, Vishal Subramanian contributed equally to this work
†Also with University of Michigan, Department of Materials Science & Engineering.

SC '23, November 12–17, 2023, Denver, CO, USA

ACM ISBN 979-8-4007-0109-2/23/11…$15.00
https://doi.org/10.1145/3581784.3627037

CCS CONCEPTS

• **Computing methodologies** → **Quantum mechanic simulation**; **Massively parallel and high-performance simulations**.

KEYWORDS

Quantum simulation, Inverse problems, Density functional theory, Machine learning, Finite elements, Exascale computing, Scalability, Heterogeneous architectures, Mixed precision, Quasicrystals, Lightweight alloys

ACM Reference Format:
Sambit Das, Bikash Kanungo, Vishal Subramanian, Gourab Panigrahi, Phani Motamarri, David Rogers, Paul M. Zimmerman, and Vikram Gavini. 2023. Large-scale materials modeling at quantum accuracy: *Ab initio* simulations of quasicrystals and interacting extended defects in metallic alloys . In *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23), November 12–17, 2023, Denver, CO, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581784.3627037

1 JUSTIFICATION FOR ACM GORDON BELL PRIZE

Largest materials simulation involving 619,124 electrons at an accuracy commensurate with quantum many-body methods, which is 100× larger in system-size, >100× improvement in time-to-solution (3.3×10^{-2}sec/GS/electron), compared to state-of-the-art quantum-accurate methods. Unprecedented sustained performance of 659.7 PFLOPS (43.1% FP64-peak) for any *ab initio* ground-state (GS) calculation[1].

[1]High watermark for sustained performance is 64 PFLOPS on New Sunway (5% FP64-peak) [37].

2 PERFORMANCE ATTRIBUTES

Category of achievement	peak performance, scalability, time-to-solution
Type of method used	N/A
Results reported on the basis of	whole application including I/O
Precision reported	mixed precision
System scale	results measured on full-scale system
Measurement mechanism	timers and FLOP count

3 OVERVIEW: BREAKING THROUGH THE BARRIER ON ACCURACY/LENGTH-SCALE TRADE-OFF IN *AB INITIO* METHODS

The design and discovery of new materials, be it for energy storage, lightweight structures or quantum devices, relies on quantum mechanical *ab initio* methods to predict wide ranging materials properties. The efficacy of an *ab initio* method for materials design depends on simultaneously attaining quantum accuracy (1-5 mHa in energy/atom) and accessing large length-scales demanded by materials applications. However, *ab initio* methods have been restricted by a Pareto-like frontier, where any gain in accuracy is met with a loss in accessible length-scale (see Fig. 1). On one extreme (Level 4 and beyond) lie the Quantum many-body (QMB) methods—configuration interaction (CI), coupled cluster (CC), and quantum Monte Carlo (QMC)—which offer a direct solution to the many-electron Schrödinger equation, and, hence, attain quantum accuracy. However, they suffer from high computational complexity of $O(e^N) - O(N^6)$ (N: number of electrons), limiting their reach to $O(10^3)$ electrons. Density functional theory (DFT) (Level 1-3), on the other hand, provides a formally exact reduction of the many-electron Schrödinger equation to an equivalent system of non-interacting single electrons [1], offering an $O(N^3)$ *ab initio* method that can scale. The cornerstone of this simplification is the exchange-correlation (XC) functional, which encapsulates the quantum many-electron interactions. The XC functional, expressed as an energy (E_{xc}) or a potential (v_{xc}), is known to be a universal functional of the electron density ($\rho(\mathbf{r})$), making DFT, in principle, an exact theory. However, in practice, the unavailability of this universal XC functional necessitates the use of approximations. Existing approximations, in increasing accuracy and complexity, can be categorized into three levels [2]: local density approximation (LDA) at Level 1, generalized gradient approximation (GGA) at Level 2, and hybrid functionals at Level 3. However, as evident from Fig. 1, all levels of XC approximation remain far from quantum accuracy. Further, conventional approaches to model the XC functional, based on either idealized model systems or semi-empirical fitting, make systematic improvements difficult. Thus, with the high computational complexity of QMB methods on one hand and the inaccuracies of the XC approximations in DFT on the other, breaking through the accuracy/length-scale trade-off has remained elusive.

In this work, we present an advance that breaks through the barrier on accuracy/length-scale trade-off to realize large-scale *ab initio* simulations approaching quantum-accuracy (see Fig. 1). The main

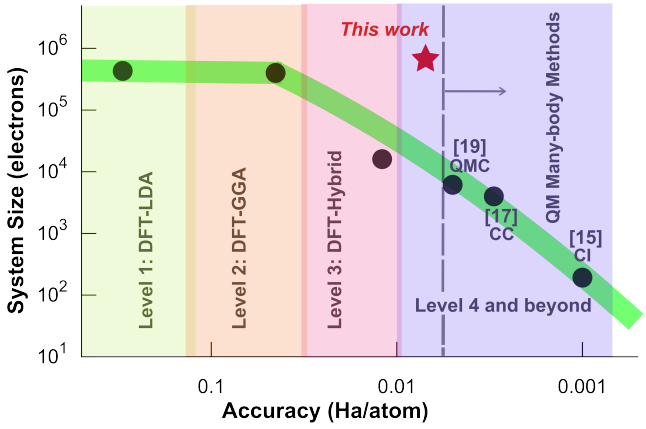

Figure 1: The barrier of accessible system size at various levels of increasingly accurate theories in *ab initio* electronic structure calculations. Accuracy of various levels of theory is based on ground-state energies from benchmark data sets.

idea is to integrate machine-learned modeling of the XC functional, informed by data from QMB methods, with various numerical and HPC innovations for large-scale DFT. This is realized by interfacing three modules:

(i) `invDFT`: A methodology advance [3, 4] providing an accurate and scalable solution to the numerically challenging *inverse* DFT problem of finding the *exact* XC potentials corresponding to electron densities from QMB methods. This forms our interface between QMB methods and DFT.

(ii) `MLXC`: A physics-informed deep neural network (DNN) based XC functional trained using the exact XC potentials obtained from `invDFT`.

(iii) `DFT-FE-MLXC`: A highly scalable hybrid CPU/GPU DFT code based on spatially adaptive and systematically convergent higher-order spectral finite-element (FE) basis that interfaces with `MLXC` to enable large-scale DFT calculations at quantum accuracy. Here, we build upon our `DFT-FE` code [5] (Finalist, 2019 Gordon Bell prize [6]) with a host of HPC innovations—high arithmetic intensity FE-cell-level linear algebra, strategic use of mixed precision, asynchronous compute and communication schemes—to enhance the scalability and performance on exascale architectures. These HPC innovations are also used in our scalable implementation of `invDFT`.

The combination of the above modules paves a path for quantum-accurate *ab initio* calculations on systems with $O(10^5)$ electrons—a staggering improvement over the $O(10^3)$ electrons accessible to existing QMB methods (see 'Level 4 & beyond' in Fig. 1). Notably, we attain a sustained performance of 659.7 PFLOPS on 8,000 Frontier nodes at a throughput efficiency of 43.1% of FP64 peak, which is unprecedented for any *ab initio* calculation of ground-state properties, let alone calculations approaching quantum accuracy. This advance unlocks wide-ranging scientific applications from structural materials, energy materials, chemical sciences to catalysis, hitherto constrained by the accuracy/length-scale trade-off in *ab initio* methods. We demonstrate the capabilities of this framework by focusing on two such important scientific problems that are computationally challenging, owing to the inherent long-ranged electronic and

atomic interactions, and the requirement of higher quantum accuracy that has been elusive at such scales. The first application problem relates to quasicrystals—a novel class of materials with long-ranged order but aperiodic structure [7]—which demonstrate exotic physical, magnetic, and electronic properties compared to conventional crystals [8–10]. One long-standing question in quasicrystals has been the direct quantification of their thermodynamic stability relative to crystalline phases with the same composition. The developed framework has been used to address this challenge for the first time, where the competition between bulk energies and surface energies of quasicrystals compared to the crystalline phase is revealed. The YbCd quasicrystal nanoparticles investigated in this study required accurate ground-state calculations, with structural relaxation, on system sizes of ∼2,000 atoms (∼40,000 electrons (e^-)). The second example pertains to the study of magnesium (Mg) pyramidal (<c+a>) dislocation system, which is critical to the design and realization of lightweight structural alloys [11]. The main technological hurdle in the industrial use of Mg alloys is its low ductility, making it challenging to form complex parts. The ductility of Mg alloys can be improved by activating the pyramidal dislocation system [12], which in turn is dependent on: (i) the energetics of pyramidal I, II dislocations; (ii) the interaction of pyramidal dislocations with other defects—substitutional solutes, and extended defects such as twin and grain boundaries. In [6], we computed the cell-size converged energy difference of pyramidal I and II dislocations (relaxed configurations), ΔE^{I-II}, in Mg—an outstanding question in the field due to the need for highly accurate large-scale DFT calculations. We obtained ΔE^{I-II}=16 meV per nanometer dislocation length, a result requiring simulation of ∼10,000 atoms (∼100,000 e^-) to attain cell-size convergence. Herein, we focus on the interaction of dislocations with other defects—transition metal solutes and twin boundary—and demonstrate the ability to accurately compute the energetics of interacting extended defects using simulations with $O(10^5)$ e^-. Further, in both science applications, the electronic structure may not be well described by existing XC functionals due to the presence of transition metals. Thus, the capability of `DFT-FE-MLXC` to conduct fast and quantum-accurate large-scale materials simulations is instrumental in addressing these problems.

4 CURRENT STATE OF THE ART

The development of increasingly accurate and efficient *ab initio* methods is a subject of unending quest. Below, is a review of the state-of-the-art in the two paradigms of *ab initio* methods—QMB methods and DFT. In Table 1 we provide the attributes of a subset of the works discussed below.

Quantum many-body (QMB) methods: Among the QMB methods, configuration interaction (CI), coupled cluster (CC) and quantum Monte-Carlo (QMC) constitute the state-of-the-art, offering the quantum accuracy of 1-5 mHa in ground-state energies (Level 4 & beyond). The CI [13] method offers the most accurate description by expanding the many-electron wavefunction, $\Psi_{\text{QMB}}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \in \mathbb{R}^{3N}$, as a linear combination of all Slater determinants possible in a given basis. However, due to combinatorial increase in the number of determinants, the problem is intractable beyond $N \sim O(10)$ [14]. Incremental full-CI has pushed the limits of CI by providing a polynomial scaling approximation to full-CI, while retaining similar accuracy. However, its steep $O(N^8)$ scaling restricts $N \sim O(10^2)$ [15]. The CC [16] methods are similar in approach to CI, albeit with approximated linear coefficients for the Slater determinants, and, hence, are less accurate than CI. While typical CC methods scale as $O(N^6)$, recently, linear-scaling methods have been proposed [17]. However, linear-scaling CC methods have a high computational prefactor and have not yet been demonstrated on multi-node, resulting in walltime of 18 days for $N \sim 4000$. QMC, in its most popular variant of diffusion Monte Carlo [18], offers an $O(N^3)$ stochastic Green's function based approach that, although less accurate than CI and CC, is still quantum accurate with ∼ 5mHa accuracy in ground-state energies. However, QMC suffers from a high computational prefactor that limits its present reach to $N \sim O(10^3)$ [19].

Density functional theory (DFT): DFT, by virtue of reducing the many-electron problem to an effective single-electron problem, can handle larger system sizes than QMB methods. Nonetheless, DFT is faced with the dual challenges of accuracy and efficiency that impede its use for large-scale materials modeling. The accuracy challenge, more acute of the two, stems from the use of approximations to the XC functional that encapsulates the quantum many-body interactions as a mean-field of the electron density. As noted earlier, all the three levels of XC approximation—LDA (Level 1), GGA (Level 2), and hybrid (Level 3)—are considerably away from quantum accuracy. This accuracy limitation plagues all past efforts at large-scale DFT, including those listed in Table 1 and discussed below. Besides accuracy, DFT confronts an efficiency challenge from its $O(N^3)$ scaling, entailed in solving the non-linear Kohn-Sham (KS) eigenvalue problem (see Sec. 5). Additionally, meaningful prediction of materials properties places stringent demands on discretization schemes, requiring discretization errors below 10^{-4} Ha/atom and 10^{-4} Ha/Bohr in energy and forces. As a result, routine DFT calculations are limited to $O(10^3)$ electrons. Past attempts to improve efficiency progressed on two complementary paths. The first aimed at efficient spatial discretization schemes that reduce the cost of KS eigenvalue problem. The planewave (PW) basis—for its systematic convergence and efficiency enabled by its spectral convergence—remains the most popular basis and has been adopted in many DFT codes, such as `VASP` [21], `Quantum Espresso` [22], and `QBox` [23] (2006 Gordon Bell prize). Despite the popularity, the extended nature of the PW basis greatly impedes its parallel scalability. The other popular basis in quantum chemistry codes is the Gaussian basis [24, 25], or those based on atomic orbitals [27]. The `CP2K` code [25] was a key component in the work of [26] (2019 Gordon Bell). However, the Gaussian basis lacks systematic convergence, making it difficult to attain the desired accuracy, especially for metallic systems. These limitations have spurred the development of systematically convergent real-space methods based on finite-difference (FD) [28–30] discretization ([31] 2011 Gordon Bell prize) and the finite-element (FE) [5, 32, 33] basis ([6] Finalist, 2019 Gordon Bell prize), both providing good parallel scalability. Another direction for improving efficiency of DFT calculations entails the development of reduced-order scaling approaches that exploit the locality of the density matrix, resulting in $O(N)$-$O(N^2)$ scaling [34, 35]. In [36] (Finalist, 2016 Gordon Bell prize) large-scale *ab initio* molecular dynamics was showcased by employing an $O(N)$

Table 1: State-of-the-art for various levels of theory in electronic structure calculations (see Fig. 1) and their various attributes. For DFT calculations (Levels 1-3), only works using a complete basis and methods that are generically applicable to any materials system are provided, which is the case for DFT-FE and DFT-FE-MLXC. Section 4 provides a broader discussion, including other approaches. AE and PSP denote all-electron and pseudopotential calculations, respectively. Wall times, as reported, are provided per self consistent field (SCF) iteration or per full ground-state (GS) calculation. Typically, a ground-state calculation comprises of 30-50 SCF iterations.

	Work	Basis	Versatility AE/PSP	Benchmark System	Machine (CPU cores/GPUs)	Wall time (mins)	PFLOPS (% of peak)
Level 1	RSDFT [31] (2011)	FD	PSP	Si nanowire 107K atoms, 430K e^-	K (450K cores)	73.6 / SCF	7.1 (43.6%)
	QBox [23] (2008)	PW	PSP	Mo (1K atoms, 12K e^-)×8 k-pts¶	BlueGene/L (125K cores)	8.8 / SCF	0.2 (56.5%)
Level 2	DFT-FE [6] (2019)	FE	AE/PSP	Mg dislocation 10K atoms, 100K e^-	Summit (22,800 GPUs)	2.4 / SCF	46 (27.8%)
	PARSEC [30] (2023)	FD	PSP	Si nanocluster 100K atoms, 400K e^-	Frontera (115K cores)	2,808 / GS	–
Level 3	Hybrid DFT, ACE [38] (2017)	PW	PSP	Si bulk 4,096 atoms, 16K e^-	NERSC Cori-KNL (8K cores)	30 / SCF	–
Level 4 & beyond	QMCPACK [20] (2018)	PW	PSP	NiO supercell 128 atoms, 1,536 e^-	Titan (18000 GPUs)	294.7 / GS	–
	QMCPACK [19] (2020)	PW	PSP	NiO supercell 512 atoms, 6,144 e^-	–	–	–
	LNO-CCSD(T) [17] (2019)	Gaussian	AE	Lipid transfer protein 1,023 atoms, 3,980 e^-	Intel Xeon PC (6 cores)	26,064 / GS	–
	iFCI, QChem [15] (2021)	Gaussian	AE	Transition metal complex 47 atoms, 192 e^-	–	–	–
	MCSCF, NWChem [14] (2017)	Gaussian	AE	Cr trimer 3 atoms, 72 e^-	Cori Haswell (2048 cores)	57.8 / SCF	–
	This Work DFT-FE-MLXC	FE	AE/PSP	Extended defects in Mg-Y alloy I. (36K atoms, 76K e^-)×4 k-pts¶ II. (74K atoms, 155K e^-)×4 k-pts¶	Frontier (19,200 GCDs) (64,000 GCDs)	 3.7 / SCF 8.6 / SCF	 226.3 (49.3%) 659.7 (43.1%)

¶ In simulations using k-point sampling (QBox, DFT-FE-MLXC), the total number of electrons in the supercell are obtained by multiplying with the number of k-pts: number of electrons in the supercell for QBox are 96K e^-, DFT-FE-MLXC system I are 302K e^- and system II are 619K e^-.

method. In [37] (Finalist, 2022 Gordon Bell prize), the quasi-2D geometry was exploited to conduct large-scale DFT calculations with $O(N^{1.5})$ scaling using a discontinuous Galerkin formulation with PEXSI solver [35]. The work of [37] is the high watermark for sustained performance in ground-state calculations—64 PFLOPS on New Sunway at 5% efficiency. While reduced-order scaling methods provide access to larger systems sizes, they are not generically applicable to all materials systems. Notably, on the accuracy front, all the aforementioned works used either LDA or GGA approximations for the XC functional (Level 1, 2 in Fig. 1, Table 1). Hybrid DFT calculations (Level 3) were demonstrated using PW basis on 4,000 atoms of bulk Silicon in [38].

Overall, the current state-of-the-art in electronic structure methods is either limited by the accuracy of DFT calculations due to approximations in the XC functional, or limited by the accessible system sizes of $O(10^3)$ electrons using QMB methods. The framework developed in this work integrates the accuracy of QMB methods with the computational efficiency afforded by DFT to enable large-scale materials simulations at quantum accuracy. Notably, as will be demonstrated, using XC functionals constructed from *exact* XC potentials corresponding to QMB densities, DFT-FE-MLXC is able to achieve systematically convergent materials simulations on $O(10^5)$ electrons at an accuracy commensurate with QMB methods.

5 INNOVATIONS REALIZED

This work combines the accuracy provided by QMB methods with the efficiency of DFT to access larger length scales at quantum accuracy, which are otherwise beyond the reach of conventional approaches. This ability to perform large-scale, fast, and systematically convergent materials simulations (on generic materials systems) at an accuracy commensurate with QMB methods is a result of advances in the methods and algorithms as well as HPC

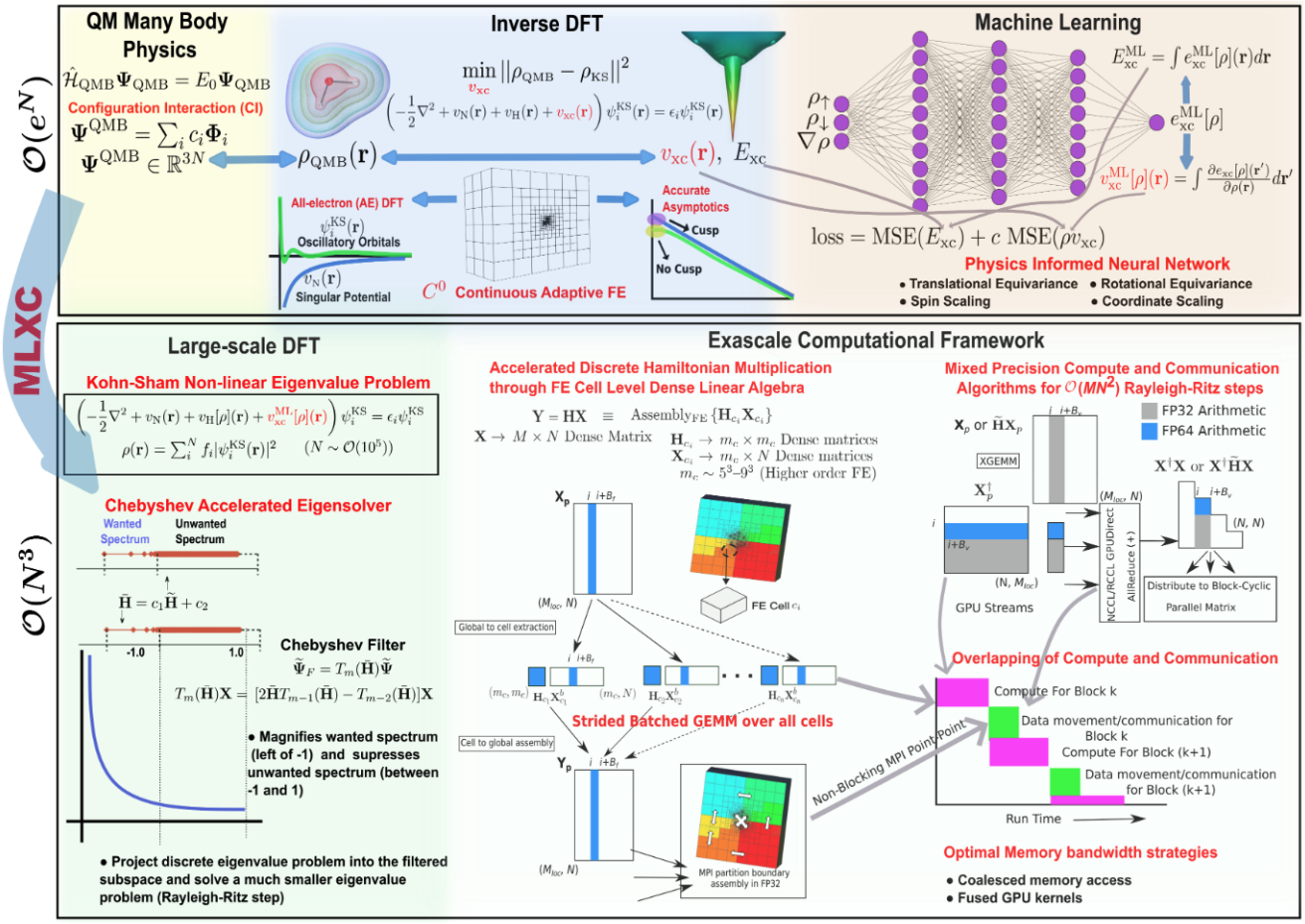

Figure 2: Overview of our approach, enabling large-scale materials simulations at quantum accuracy

innovations. Figure 2 provides an overview of the various aspects of this work. Below, we discuss the details that are central to the developed capabilities.

As noted, DFT provides a formally exact reduction of the many-electron Schrödinger equation to an equivalent system of non-interacting electrons in an effective mean field governed by the electron density ($\rho(\mathbf{r})$). Computing the ground-state in DFT entails a self consistent field (SCF) iteration solving the non-linear Kohn-Sham (KS) eigenvalue problem:

$$\left(-\frac{1}{2}\nabla^2 + v_{\mathrm{N}}(\mathbf{r}) + v_{\mathrm{H}}[\rho](\mathbf{r}) + v_{\mathrm{xc}}[\rho](\mathbf{r})\right)\psi_i^{\mathrm{KS}}(\mathbf{r}) = \epsilon_i \psi_i^{\mathrm{KS}}(\mathbf{r}), \quad (1)$$

where $\rho(\mathbf{r}) = \sum_i f_i |\psi_i^{\mathrm{KS}}(\mathbf{r})|^2$ is electron density; $\{\epsilon_i, \psi_i^{\mathrm{KS}}\}$ are the eigenpairs of the KS Hamiltonian; v_{N} and v_{H} are electrostatic potentials corresponding to the nuclei and electron density; v_{xc} is XC potential; f_i denotes the occupancy of the ith state, evaluated using the Fermi-Dirac distribution. The accuracy afforded by DFT is solely governed by the XC functional, $E_{\mathrm{xc}}[\rho]$, with $v_{\mathrm{xc}}[\rho] = \frac{\delta E_{\mathrm{xc}}}{\delta \rho}$. The existing approximations for XC in DFT remain far from quantum accuracy (see Fig. 1, Fig. 3), and this remains a major hurdle.

5.1 Inverse DFT: Solving an open problem

Inverse DFT [39] provides a powerful link between QMB methods and DFT by finding the *exact* XC potential ($v_{\mathrm{xc}}(\mathbf{r})$) corresponding to an electron density ($\rho_{\mathrm{QMB}}(\mathbf{r})$) obtained from a QMB calculation. Subsequently, the $\{\rho_{\mathrm{QMB}}, v_{\mathrm{xc}}^{\mathrm{exact}}\}$ pairs obtained from inverse DFT allows us to learn v_{xc} as a functional of ρ (i.e., model $v_{\mathrm{xc}}[\rho]$), and hence, model the XC functional in DFT. Though seemingly simple, an accurate solution to the inverse DFT problem has remained an open problem for past 30 years, owing to its numerical challenges. Previous attempts suffered from ill-conditioning, resulting in spurious oscillations and/or non-unique solutions in v_{xc}. This ill-conditioning stems from: (i) incompleteness of the Gaussian basis employed [40]; (ii) incorrect asymptotics in the Gaussian densities obtained from QMB methods [41], i.e., lack of cusp at nuclei and incorrect far-field decay. We recently proposed an accurate and robust solution to inverse DFT, using FE basis [3, 4]. We use systematically convergent (complete) adaptive higher-order FE to: (i) render the discrete inverse problem well-conditioned; (ii) efficiently handle the oscillatory electronic fields, characteristic of all-electron DFT. Importantly, we mitigate the spurious artifacts arising from the Gaussian densities by: (i) adding a cusp-correction to $\rho_{\mathrm{QMB}}(\mathbf{r})$

near the nuclei; (ii) applying the physical $-1/r$ boundary condition on $v_{\text{xc}}(\mathbf{r})$ in the far-field. These asymptotic corrections rely on two key attributes of FE: C^0 continuity (which admits a cusp) and the seamless handling of general boundary conditions. Formally, inverse DFT finds the $v_{\text{xc}}(\mathbf{r})$ that minimizes the difference between target ($\rho_{\text{QMB}}(\mathbf{r})$) and KS ($\rho_{\text{KS}}(\mathbf{r})$) densities, subject to the condition that $\rho_{\text{KS}}(\mathbf{r})$ is obtained from the KS eigenvalue problem (cf. 'Inverse DFT' in Fig. 2). We solve this as a PDE-constrained optimization that results in two key equations. The first is the KS eigenvalue problem (Eq. 1), albeit with v_{xc} being unknown (i.e., updated iteratively). The second is the adjoint equation, given as

$$\left(-\frac{1}{2}\nabla^2 + v_{\text{N}}(\mathbf{r}) + v_{\text{H}}(\mathbf{r}) + v_{\text{xc}}(\mathbf{r}) - \epsilon_i\right) p_i(\mathbf{r}) = g_i(\mathbf{r})\,, \qquad (2)$$

where $p_i(\mathbf{r})$ is the adjoint function enforcing the KS eigenvalue problem for $\psi_i^{\text{KS}}(\mathbf{r})$, and $g_i(\mathbf{r})$ is defined in terms of $\rho_{\text{QMB}}(\mathbf{r})$, $\rho_{\text{KS}}(\mathbf{r})$ and $\psi_i^{\text{KS}}(\mathbf{r})$. Subsequently, $u(\mathbf{r}) = \sum_i p_i(\mathbf{r})\psi_i^{\text{KS}}(\mathbf{r})$ provides the iterative update to $v_{\text{xc}}(\mathbf{r})$. Our `invDFT` module encapsulates this approach.

5.2 Learning density functionals from exact XC potentials

We use the $\{\rho_{\text{QMB}}, v_{\text{xc}}^{\text{exact}}\}$ pairs, obtained from the `invDFT` module, to learn a deep neural-network (DNN) based XC functional ($E_{\text{xc}}^{\text{ML}}[\rho]$). We note that learning $E_{\text{xc}}^{\text{ML}}[\rho]$ entails a map from a field (ρ) to a scalar ($E_{\text{xc}}^{\text{ML}}$), which can result in a brittle model. Thus, we transform the learning into a field-to-field map by recasting it in terms of the XC energy density ($e_{\text{xc}}^{\text{ML}}[\rho]$), as $E_{\text{xc}}^{\text{ML}}[\rho] = \int e_{\text{xc}}^{\text{ML}}[\rho](\mathbf{r})\, d\mathbf{r}$. We model $e_{\text{xc}}^{\text{ML}}[\rho]$ as

$$e_{\text{xc}}^{\text{ML}}[\rho](\mathbf{r}) = \rho^{4/3}(\mathbf{r})\phi(\mathbf{r})F^{\text{DNN}}(\rho, \xi, s)\,. \qquad (3)$$

In the above, F^{DNN} is modeled as a DNN with ρ, ξ, s as input descriptors, where ξ is the relative spin density $\xi(\mathbf{r}) = (\rho_\uparrow(\mathbf{r}) - \rho_\downarrow(\mathbf{r}))/\rho(\mathbf{r})$; $\phi(\mathbf{r}) = \frac{1}{2}\left((1+\xi(\mathbf{r}))^{4/3} + (1-\xi(\mathbf{r}))^{4/3}\right)$; and $s(\mathbf{r}) = (3\pi^2)^{1/3}|\nabla\rho(\mathbf{r})|/(2\rho^{4/3}(\mathbf{r}))$. This form of $e_{\text{xc}}^{\text{ML}}$, by design, satisfies translational and rotational equivariance. The explicit inclusion of spin information via ξ is crucial in describing spin-polarized systems. The choice of the $\rho^{4/3}$ and ϕ as prefactors enforces known coordinate- and spin-scaling relations. Given that the exact $e_{\text{xc}}(\mathbf{r})$ is unknown, we use a composite loss function comprising of mean squared errors (MSE) in XC energy (E_{xc}) and density-weighted XC potential ($\rho_{\text{QMB}} v_{\text{xc}}$), with $v_{\text{xc}}^{\text{ML}}(\mathbf{r}) = \frac{\delta e_{\text{xc}}^{\text{ML}}[\rho](\mathbf{r})}{\delta\rho(\mathbf{r})}$ inexpensively obtained via back-propagation. Our DNN has 5 layers, 80 neurons/layer, with ELU activation. We train the DNN using a small training set of $\{\rho_{\text{QMB}}, v_{\text{xc}}^{\text{exact}}\}$ for two molecules (H_2 and LiH) and three atoms (Li, N, and Ne). The resultant model, termed `MLXC`, is tested against a standard thermochemistry dataset [42] of main group molecules, routinely used to evaluate XC functionals. Figure 3 compares accuracy of `MLXC` against widely used XC approximations (Level 1-3), in terms of energy/atom, with `MLXC` remarkably better than existing XC approximations. Notably, `MLXC` attains an accuracy of 7 mHa/atom—very close to accuracy of QMB methods. The promising accuracy of `MLXC`, trained with limited data, underscores the importance of using exact XC potentials and physics-informed modeling to improve the XC functional.

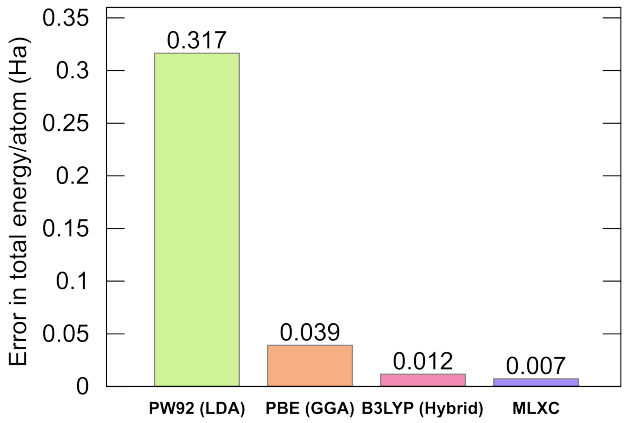

Figure 3: Comparison of `MLXC` with XC approximations in DFT.

5.3 Efficient and scalable solver strategies

The solution to KS eigenvalue problem (Eq. 1) constitutes most of the computational cost in DFT. In *Inverse DFT*, each iteration of PDE-constrained optimization is dominated by: (i) KS eigenvalue problem (same as DFT); (ii) adjoint problem (Eq. 2). We provide efficient numerical strategies to solve the eigenvalue and adjoint problems, discretized in an adaptive higher-order spectral (Löwdin orthonormalized) FE basis [43].

5.3.1 Adjoint Solve. We solve $(\widetilde{\mathbf{H}} - \epsilon_i\mathbf{I})\widetilde{\mathbf{p}}_i = \widetilde{\mathbf{g}}_i$ using a preconditioned block-MINRES solver. The key idea is to construct Krylov subspaces, within the MINRES solver, for blocks of $\widetilde{\mathbf{p}}_i$'s, i.e., apply $\widetilde{\mathbf{H}}$ to a block $\widetilde{\mathbf{P}}$ of $\widetilde{\mathbf{p}}_i$'s. This allows us to exploit high arithmetic intensity linear algebra tailored for FE (see Sec. 5.4.1). We also precondition the MINRES solver with the inverse diagonal of the discrete Laplacian, an inexpensive yet effective preconditioner. Notably, this provides a $\sim 5\times$ reduction in the number of MINRES iterations.

5.3.2 Chebyshev filtered eigensolver (ChFES). We solve the discrete KS eigenvalue problem, $\widetilde{\mathbf{H}}\widetilde{\boldsymbol{\psi}}_i = \epsilon_i\widetilde{\boldsymbol{\psi}}_i$, by employing a fast and efficient eigensolver for $N \sim O(10^5)$ (see 'Large-scale DFT' in Fig. 2). We use the Chebyshev filtering procedure [43, 44] that exploits the fact that the eigenstates of interest are the occupied states at lower end of eigenspectrum (wanted eigenspectrum). A scaled-and-shifted Hamiltonian ($\bar{\mathbf{H}} = c_1\widetilde{\mathbf{H}} + c_2$) is constructed such that the 'wanted eigenspectrum' of discrete KS Hamiltonian ($\widetilde{\mathbf{H}}$) is mapped to (-∞, -1) and the remainder to (-1,1). Chebyshev polynomials grow fast in (-∞, -1) but take small values in [-1,1]. Thus, upon applying a Chebyshev polynomial filter of degree m ($T_m(\bar{\mathbf{H}})$) on an input subspace $\widetilde{\boldsymbol{\Psi}}_{\text{in}}$, we obtain a resultant subspace of Chebyshev filtered vectors, $\widetilde{\boldsymbol{\Psi}}_f$, that is a good approximation to the 'wanted eigenspace'. The approximation error decreases systematically with m. We efficiently cast the Chebyshev filtering as recursive evaluation of $\bar{\mathbf{H}}\widetilde{\boldsymbol{\Psi}}_{\text{in}}$ (see CF in Algorithm 1). Further, the Chebyshev filter is applied to a block of wavefunction vectors simultaneously which lends to efficient implementation (Sec. 5.4.1). Subsequently, the desired eigenstates are computed by projecting the discrete KS eigenvalue problem onto the subspace spanned by $\widetilde{\boldsymbol{\Psi}}_f$ (see RR in Algorithm 1).

Algorithm 1 ChFES procedure: $[\widetilde{\Psi}, \mathbf{D}] = \text{ChFES}\left(\widetilde{\Psi}_{\text{in}}, \bar{\mathbf{H}}\right)$

$\widetilde{\Psi}$: $M \times N$ matrix M: # FE basis (DoF) N: # electrons.
$\widetilde{\Psi}_{\text{in}} = \widetilde{\Psi}$ from previous SCF step.

1: [CF] Chebyshev polynomial filtering: $\widetilde{\Psi}_f = T_m(\bar{\mathbf{H}})\widetilde{\Psi}_{\text{in}}$, $T_{m+1}(\bar{\mathbf{H}}) = 2\bar{\mathbf{H}}\,T_m(\bar{\mathbf{H}}) - T_{m-1}(\bar{\mathbf{H}})$. $(O(MN))$
2: [CholGS]: Cholesky Gram-Schmidt Orthonormalization
  a: [CholGS-S] Overlap matrix, $\mathbf{S} = \widetilde{\Psi}_f^{\dagger}\widetilde{\Psi}_f$. $(O(MN^2))$
  b: [CholGS-CI] Cholesky Inverse. Compute $\mathbf{L}^{-1}$, s.t. $\mathbf{S} = \mathbf{L}\mathbf{L}^{\dagger}$. $(O(N^3))$
  c: [CholGS-O] Orthonormalization: $\widetilde{\Psi}_o = \widetilde{\Psi}_f \mathbf{L}^{-1^{\dagger}}$. $(O(MN^2))$
3: [RR] Perform Rayleigh-Ritz step:
  a: [RR-P] Projected Hamiltonian: $\hat{\mathbf{H}} = \widetilde{\Psi}_o^{\dagger}\bar{\mathbf{H}}\widetilde{\Psi}_o$. $(O(MN^2))$
  b: [RR-D] Diagonalization: $\hat{\mathbf{H}}\mathbf{Q} = \mathbf{Q}\mathbf{D}$. $(O(N^3))$
  c: [RR-SR] Subspace rotation: $\widetilde{\Psi} = \widetilde{\Psi}_o\mathbf{Q}$. $(O(MN^2))$

5.4 HPC Innovations

The overall trend in pre-exascale and exascale architectures has been a significant increase in single-node peak compute performance relative to modest increases in inter-node and intra-node data movement bandwidths. Thus, high throughput performance at large node counts requires strategies that increase arithmetic intensity but reduce data movement. To that end, we use a combination of four broad strategies to boost the GPU acceleration and extreme parallel scaling of the key computational kernels in `DFT-FE-MLXC` (Algorithm 1) and `invDFT` (adjoint solve). Below, we discuss these strategies and provide relevant performance benchmarks for them.

5.4.1 FE cell level dense linear algebra. $\mathbf{Y}^b = \mathbf{H}\mathbf{X}^b$ is the main computational kernel in CF and adjoint solve, involving the FE discretized sparse matrix, $\mathbf{H}$, and a dense matrix, $\mathbf{X}^b$, representing a column block of $\widetilde{\Psi}_{\text{in}}$ in Algorithm 1 (or $\widetilde{\mathbf{P}}$ in adjoint solve) with block size B_f. This operation incurs high memory access costs if performed using conventional approaches at the global level. We significantly reduce the access costs by recasting $\mathbf{Y}^b = \mathbf{H}\mathbf{X}^b$ as $\mathbf{Y}^b = \text{Assembly}_{\text{FE}}\left\{\mathbf{H}_{c_i}\mathbf{X}^b_{c_i}\right\}$ (cf. 'Exascale Computational Framework' Fig. 2), where $\text{Assembly}_{\text{FE}}$ denotes the assembly operation of contributions from all FE cells into $\mathbf{Y}^b$. Using this form, which exploits the FE cell structure, we efficiently perform small dense-dense matrix multiplications by using massive fine-grained parallelism available on GPUs through `xGEMMStridedBatched` BLAS calls. Importantly, we employ a recent reformulation [33] of the problem that decouples the FE mesh nodes from the positions of nuclei, allowing the use of higher-order FE of polynomial degree $p = 6-8$, as compared to $p = 4-5$ used previously. This provides faster convergence in the discretization error ($O(h^{2p})$ in energy, h being the size of the FE cell) leading to significant reduction in degrees of freedom (DoF) required to achieve the desired discretization errors of 10^{-4} Ha/atom and 10^{-4} Ha/Bohr in energy and ionic forces. Coincidentally, a large p also results in relatively larger FE cell matrix sizes—$9^3 \times 9^3$ for $\mathbf{H}_{c_i}$ for $p = 8$—that provides higher throughput efficiency on GPUs.

These implementation innovations provided high throughput performance for CF step on hybrid CPU-GPU architectures, see Fig. 4. Performance is measured on a dislocation system in Mg-Y alloy with (6,016 atoms, 12,041 e^-)×2k-points, using $p = 8$. The

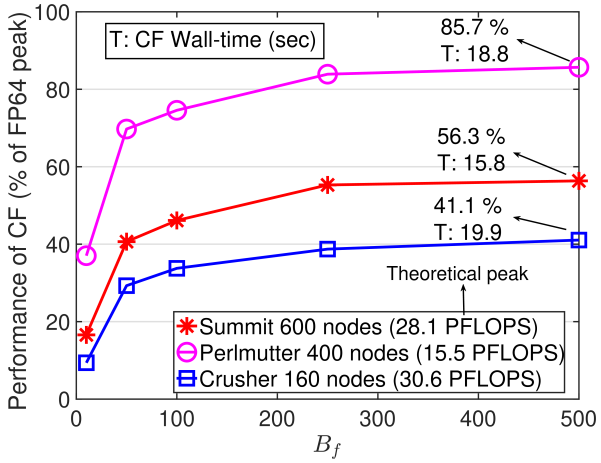

Figure 4: Chebyshev filtering (CF) performance for various wavefunction block sizes B_f, using DislocMgY (see Sec 6.2) with (6,016 atoms, 12,041 e^-)×2k-pts. FE DoF: $\sim 96 \times 10^6$**; FE poly. degree: 8.**

performance, measured as percentage of the theoretical FP64 peak FLOPS, increases with block size B_f due to improved arithmetic intensity of `xGEMMStridedBatched` operations, improved utilization of memory and interconnect bandwidths in level 1 BLAS and FE partition boundary communication operations in CF step (cf. Fig. 2). Comparing the performance achieved for $B_f = 500$ on various architectures, we obtain high throughput efficiencies of 56.3% and 41.1% on Summit NVIDIA V100 and Crusher AMD M1250X GPU nodes. The 1.4× reduction of throughput efficiency on Crusher relative to Summit correlates well with the 1.7× increase in the peak FP64 FLOPS to HBM memory bandwidth ratio of a Crusher node relative to a Summit node. On Perlmutter A100 GPU nodes, we achieve a very high efficiency of 85.7%. This is attributed to the use of FP64 tensor cores that provide 2× higher peak FLOPS compared to FP64 cores, which we verified via NVIDIA's Nsight profiler[2].

5.4.2 Mixed precision algorithms. We employed mixed precision strategies to reduce communication and computational costs in the key computational steps (Algorithm 1). First, in CF step, the $\text{Assembly}_{\text{FE}}$ operation in $\mathbf{HX}$ kernel (Sec. 5.4.1) involves point-to-point MPI communication across FE domain decomposition partition boundaries. The FE nodes on partition boundaries are far fewer than the total number of FE nodes. In view of this, we use FP32 arithmetic for boundary communication, and this has been observed to retain FP64 accuracy, while reducing the communication cost by $\sim 2\times$. Further, as the Chebyshev filtered wavefunctions ($\widetilde{\Psi}_{\text{f}}$) iteratively approach the eigenvectors corresponding to the N lowest eigenstates, the off-diagonal entries of $\mathbf{S} = \widetilde{\Psi}_f^{\dagger}\widetilde{\Psi}_f$, computed in the CholGS step, converge to zero as the SCF approaches convergence. Thus, as shown in Fig. 2, we compute the diagonal block entries of $\mathbf{S}$ in FP64 arithmetic, while the off-diagonal entries are computed in FP32 arithmetic. We also use similar mixed-precision strategies in the RR step, significantly reducing the computational

[2]We are unable to verify usage of AMD MI250X's FP64 matrix cores on Crusher due to technical issues of using the ROCm profiler in the GNU compiler environment, which is the only environment that presently worked for compiling the full DFT-FE-MLXC application.

prefactor of the $O(MN^2)$ steps. The accuracy and robustness of these mixed precision strategies have recently been demonstrated in [5, 33], with the error in energy and forces being well within the target discretization accuracy.

5.4.3 Asynchronous GPU compute and data-movement. The blocked approach employed in ChFES (cf. Fig. 2) allowed us to devise strategies to overlap GPU computations with data movement—MPI communication and host-device data transfers. In the CF step, $\mathbf{HX}$ is computed using column blocks of $\mathbf{X}$. If $\mathbf{X}^k$ denotes the kth block of $\mathbf{X}$, the GPU compute of $\mathbf{HX}^k$ is executed concurrently with partition boundary communication calls in computation of $\mathbf{HX}^{k-1}$ (previous block). Considering CholGS-S step, $\mathbf{S} = \mathbf{X}^\dagger\mathbf{X}$, the GPU computations of $\mathbf{S}_p^k = \mathbf{X}_p^\dagger\mathbf{X}_p^k$ for block-k (in every task p) are executed concurrently with GPU aware `allreduce` collective operation in forming $\mathbf{S}^{k-1} = \sum_p \mathbf{S}_p^{k-1}$, followed by data movement of copying $\mathbf{S}^{k-1}$ to a parallel (ScaLAPACK) matrix. Our implementation makes use of two GPU streams, one for compute operations, the other handling data movement. We pass the data movement tagged GPU stream *id* to GPU direct optimized collective communication libraries (NCCL/RCCL), as will be discussed below. Along similar lines, we also implemented asynchronous compute and data movement for the RR-P and RR-SR steps.

5.4.4 Efficient use of GPU aware MPI libraries. We further boost the strong scaling of CF, CholGS and RR steps by efficient use of hardware aware MPI communication libraries that exploit fast interconnects between the GPUs inside a node. For FE boundary communication in CF, we perform non-blocking point-point MPI communication using GPU aware MPI library such as the Cray MPICH on Frontier/Crusher and Perlmutter, which provides around 1.5× speedup in wall-times for the CF step. In the CholGS and RR steps, we use GPU aware NCCL and RCCL libraries to perform `allreduce` calls. Our internal benchmarks on RCCL with AWS-OFI-RCCL plugin demonstrate an order of magnitude improvement in `allreduce` bus bandwidth achieved on Frontier compared to Cray MPICH. However, we restrict RCCL usage in `DFT-FE-MLXC` to under 1000 Frontier nodes, as beyond that RCCL has stability issues[3].

We now present the improvements in strong scaling performance realized by our mixed precision and asynchronous compute/data movement. We consider YbCd quasicrystal nanoparticle comprising of 1,943 atoms (40,040 e^-). Figure 5 shows the wall-time per SCF iteration step for 240 to 1,920 Summit nodes. As evident, the mixed precision and asynchronous compute/data movement strategies provide a substantial improvement of 1.8× in the minimum wall-time over the baseline. The strong parallel scaling efficiency at 1920 nodes improved to 54% from the baseline of 36% at the same 1920 nodes.

6 HOW PERFORMANCE WAS MEASURED

6.1 Systems and Environment

All simulations are executed on OLCF Frontier/Crusher[4], Summit and NERSC Perlmutter supercomputers. Frontier is currently the fastest supercomputer, with ~1.8 exaFLOPS FP64 theoretical peak.

[3]This is a known issue. When fixed, we anticipate up to a 1.5× reduction in computational times for the $O(MN^2)$ steps.

[4]Crusher is a test system with same architecture as Frontier

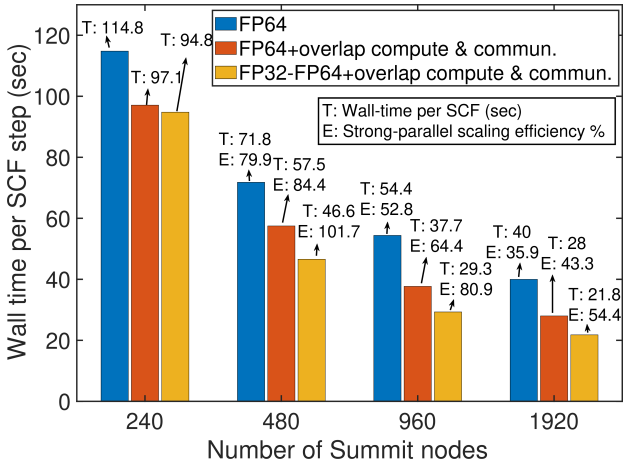

Figure 5: Strong scaling of `DFT-FE-MLXC` on Summit. Case study: YbCd quasicrystal nanoparticle ($Yb_{295}Cd_{1648}$) with 1,943 atoms, 40,040 e^-. FE DoF: 75,069,290.

Each Frontier node consists of 4 AMD MI250X GPUs with two Graphic Compute Dies (GCDs) in each GPU and 64-core AMD "Optimized 3rd Gen EPYC" CPU. The theoretical peak FP64 performance[5] per GPU in the above machines are 47.8 TFLOPS, 7.8 TFLOPS and 9.7 TFLOPS for Frontier, Summit and Perlmutter, respectively, which are used in our throughput efficiency analysis. On Frontier, we have compiled `DFT-FE-MLXC` using ROCm/5.4.0, GNU/11.2.0 and Cray-MPICH/8.1.26.

6.2 Applications used to Measure Performance

We consider two challenging application problems due to the presence of transition metal elements, where existing XC approximations are deficient, and the requirement of large length scales. Our first application problem is computing the bulk and surface energy of a Tsai-type icosahedral quasicrystal [10]–$YbCd_{5.7}$. This is aimed at understanding size-dependent stability of the aperiodic, long-range ordered quasicrystal relative to their crystal counterparts. Accordingly, we choose a large $Yb_{295}Cd_{1648}$ nanoparticle with 1,943 atoms, 40,040 e^- (Fig. 6) as our benchmark system.

The second application problem is the magnesium (Mg) pyramidal (<c+a>) dislocation system. Accurate *ab initio* computed energetics of <c+a> dislocations and their interactions with other defects are crucial for aiding the design of lightweight structural alloys [11]. However, these calculations require well-converged simulations on systems with many tens of thousands of atoms that have been out of reach. We demonstrate the capability of computing interactions between extended defects in Mg-1 at.% Y alloy, Y being a transition metal element. Accordingly, we create four benchmark systems at relevant length-scales: (i) "*DislocMgY*" is a pyramidal II <c+a> screw dislocation interacting with an Y solute in the dislocation core. This system contains (6,016 atoms, 12,041 e^-)×2 k-points, with 2 k-points used for the BZ zone sampling along the periodic dislocation line direction, for a total of 24,082 e^- in the supercell; (ii) "*TwinDislocMgY(A)*" comprises of a pyramidal II <c+a> screw dislocation interacting with reflection twin boundary (pyramidal I

[5]Considering only the vector registers, not the FP64 Tensor/Matrix cores.

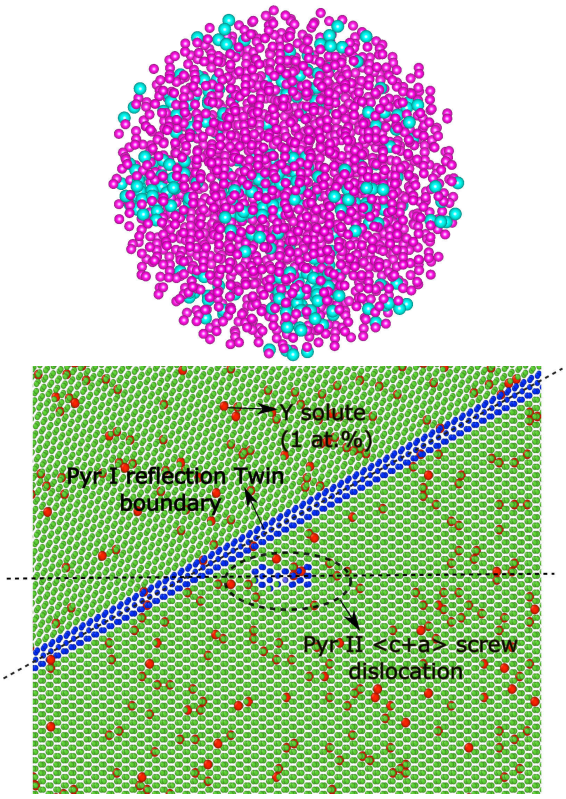

Figure 6: Benchmark systems. Top: YbCd quasicrystal ($Yb_{295}Cd_{1648}$) nanoparticle: 1,943 atoms, 40,040 e^-. Width of nanoparticle: ~3 nm. FE DoF: $\sim 75 \times 10^6$. Bottom: TwinDislocMgY(C) system: reflection twin boundary (extended defect) in pyramidal I plane interacting with a <c+a> pyramidal II screw dislocation (line defect) in Mg-1%Y alloy with (74,164 atoms, 154,781 e^-)×4 k-points, for a total of 619,124 e^- in the supercell. FE DoF: $\sim 1.7 \times 10^9$.

plane) in a random environment of Y solutes at 1 at.% concentration. This system comprises of (36,344 atoms, 75,667 e^-)×4 k-points, for a total of 302,668 e^-; (iii) "*TwinDislocMgY(B)*" is a larger version of TwinDislocMgY(A) comprising of (74,164 atoms, 154,781 e^-)×3 k-points, for a total of 464,343 e^- in the supercell; and (iv) "*TwinDislocMgY(C)*" (cf. Fig. 6) which is the largest system comprising of (74,164 atoms, 154,781 e^-)×4 k-points, for a total of 619,124 e^- in the supercell. FE mesh parameters are chosen to provide discretization errors of $\sim 10^{-4}$ Ha/atom and $\sim 10^{-4}$ Ha/Bohr in energy and ionic forces, respectively.

6.3 Measurement Methodology

Time measurements for the various computational steps and the total run-times in `invDFT` and `DFT-FE-MLXC` were obtained using a combination of `MPI_Barrier`, `MPI_Wtime` and `cudaDeviceSynchronize` / `hipDeviceSynchronize`. FLOP counts on GPUs were measured for the key computational steps: CF, CholGS-S, CholGS-O, RR-P and RR-SR. In the case of the CF step, we measured the FLOP count for the DislocMgY system using `nvprof` on Summit. We measured the FLOP counts at two different MPI tasks, and used the average FLOP count per MPI task multiplied by the total number of MPI tasks to obtain the total FLOP count. The total FLOP count thus obtained is expected to be very close to explicitly measuring and adding FLOP counts for all MPI tasks, as the load-balanced partitioning in `DFT-FE-MLXC` results in an almost equal number of FE DoF in each MPI task[6]. Subsequently, we compute the CF FLOP count for the largest systems TwinDislocMgY(A),(B),(C) using the linear scaling relation of CF's FLOP count with respect to number of FE cells and wavefunctions, with the computational prefactor being the same for DislocMgY and TwinDislocMgY(A),(B),(C) systems as they have the same FE mesh parameters and Chebyshev polynomial degree. Next, in the case of $O(MN^2)$ CholGS-S, CholGS-O, RR-P and RR-SR steps involving relatively large `GEMM` operations of $M \times N$ and $N \times N$ sized matrices, we manually compute a lower bound [7] for the FLOP count as $\alpha * 4 * N * M * N$. The factor 4 results from complex datatype usage for the aforementioned k-point sampling in DislocMgY and TwinDislocMgY(A),(B),(C) systems, and α is either 1 or 2 dependent upon whether matrix Hermiticity is exploited.

7 PERFORMANCE RESULTS

We demonstrate parallel scaling performance, time-to-solution and sustained performance of interlinked `invDFT` and `DFT-FE-MLXC` framework on pre-exascale (Perlmutter, Summit) and exascale (Frontier) machines. First, we demonstrate the performance of `invDFT` on molecular systems involving accurate all-electron inverse DFT calculations. Subsequently, using the `MLXC` functional trained on exact XC potentials obtained from `invDFT`, we demonstrate the performance of `DFT-FE-MLXC` on large-scale quasicrystal nanoparticles and extended defect interactions in metallic alloys. We use ONCV pseudopotentials for all `DFT-FE-MLXC` simulations.

7.1 Scalability & Time-to-Solution

7.1.1 `invDFT`. We demonstrate the performance of `invDFT` module using ortho-benzyne (C_6H_4), a strongly correlated system—a paradigmatic case where existing XC approximations perform poorly. Employing the various GPU acceleration strategies discussed in Sec. 5.4, we attain a 17.7× CPU-GPU speedup (in node-hours) on 4 nodes of Perlmutter. We present the strong scaling of the GPU-accelerated `invDFT` on Perlmutter in Fig. 7. We attain a 5.2× speedup, reducing the wall time per iteration from 104 sec on 4 nodes to 20 sec on 32 nodes. Given the typical 500−600 iterations in inverse DFT calculations, the advances in `invDFT`—preconditioned block-MINRES, the Chebyshev filtered eigensolver, FE cell level dense linear algebra—now make possible the evaluation of *exact XC potentials*, computed only once and stored for each system, in merely ~ 3 hours of wall time. This is a 50× improvement over a wall time of ~7 days needed in our previous implementation [3]. These advances in `invDFT` not only solve a hitherto open problem of accurate solution to the inverse DFT problem, but also enable rapid generation of exact XC potentials—an aspect that we expect will further spur the development of `MLXC` as training data becomes

[6] We verified this on a small system with 250K FE DoF and using 6 MPI tasks on Summit, where we find the difference in the FLOP count between the two approaches for the CF step is ~ 3%.

[7] Due to the blocked approach utilized in the CholGS and RR steps, the actual FLOP count is around 5% higher than estimated using the manual approach based on internal checks on medium scale system sizes.

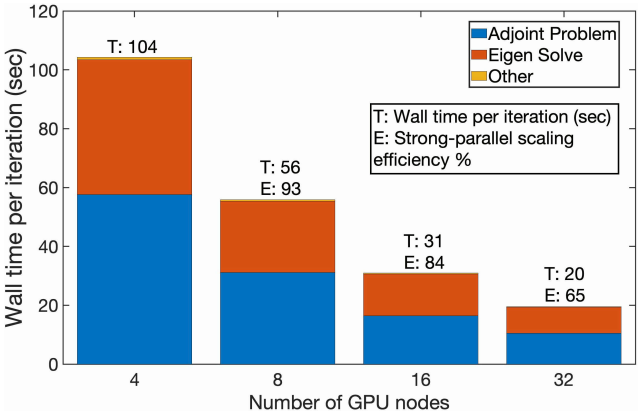

Figure 7: Strong scaling of `invDFT`. Case study: Ortho-benzyne, C_6H_4 (strongly correlated system).

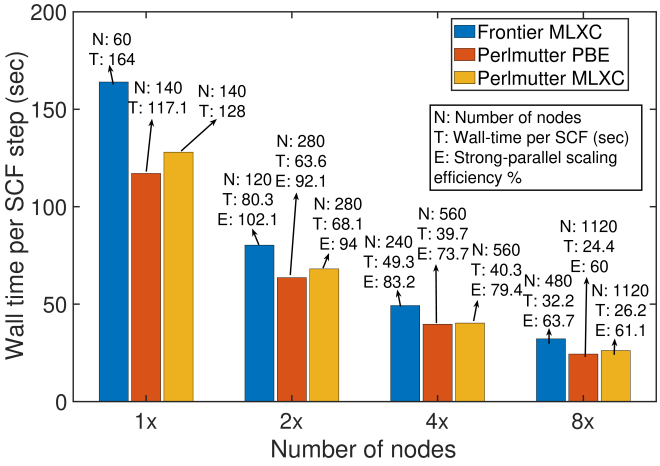

Figure 8: Strong scaling of `DFT-FE-MLXC`. Case study: YbCd quasicrystal nanoparticle ($Yb_{295}Cd_{1648}$) with 1,943 atoms, 40,040 e^-. FE DoF: 75,069,290.

more readily available to improve `MLXC` in sophistication and target accuracy.

7.1.2 `DFT-FE-MLXC`. We examine the strong scaling of `DFT-FE-MLXC` on YbCd quasicrystal (cf. Fig. 6) consisting of 1,943 atoms (40,040 e^-). The study conducted on Frontier and Perlmutter GPU nodes is shown in Fig. 8. We obtain a remarkable ~80% strong scaling efficiency at 240 Frontier nodes (39.1K DoF/GCD) and 560 Perlmutter nodes (33.5K DoF/GPU). Even at 1,120 Perlmutter nodes, with just 16.8K DoF/GPU, we attain a scaling efficiency of ~60%, reducing the walltime/SCF to ~25 sec from ~125 sec on 140 nodes—a relative speedup of 5×. Also, from Fig. 8, we note that the Level 4+ `MLXC` functional incurs only a small overhead over Level 2 `PBE` functional, with similar wall-times on Perlmutter.

To underscore the implications of this strong-scaling performance, we conduct a full ground-state calculation on YbCd quasicrystal nanoparticle with `MLXC` functional on Perlmutter using 1,120 nodes. The timings are reported in Table 2. It is notable that we are able to complete the full ground-state of a 40,000 e^- system at Level 4+ quantum-accuracy in ~30 mins.

Table 2: Time-to-solution (in sec) of YbCd quasicrystal nanoparticle (40,040 e^-) using 1,120 Perlmutter nodes.

Initialization	Total SCF	Total run
69	2023 (34 SCF steps[8])	2092

7.2 Sustained Performance: Extended defects in Mg-Y alloy

We demonstrate the performance of `DFT-FE-MLXC` on interacting extended defects in Mg-Y alloy—TwinDislocMgY(A),(B),(C) with 302,668 e^-, 464,343 e^-, 619,124 e^- in the supercells, respectively. Table 3 reports the sustained performance measured for these calculations. On the TwinDislocMgY(C) system, the largest system in this work with M=1.7 billion (FE DoF) and N=356,000 eigenstates (~605 trillion wavefunction values), we obtain a sustained performance of 659.7 PFLOPS on 8,000 nodes of Frontier (43.1% throughput efficiency), which is *unprecedented* for electronic structure ground-state calculations. On the TwinDislocMgY(A) system, we obtain a sustained performance of 226.3 PFLOPS on 2,400 Frontier nodes with 49.3% throughout efficiency. Further, on the TwinDislocMgY(B) system, we obtain a sustained performance of 508.9 PFLOPS on 6,000 Frontier nodes (44.4% efficiency), demonstrating consistently high performance on various system sizes and node counts. We remark that in spite of using a sufficiently large block size of 250 in the Chebyshev filtering (CF) step for the TwinDislocMgY(A),(B),(C) systems, we observe a drop in efficiency for the CF step to ~30% in comparison to ~40% efficiency obtained on the smaller DislocMgY system using 160 nodes (cf. Fig. 4). This is attributed to the present instability in the Frontier machine beyond ~1,000 nodes that prevented us from running the larger simulations with optimal GPU aware MPI for the FE partition boundary communication, while the DislocMgY system simulation was able to use optimal GPU aware routing settings[9]. We anticipate further improvement in efficiency for TwinDislocMgY(A),(B),(C) upon being able to use optimal hardware aware communication patterns for large node counts on Frontier.

Notably, this sustained performance using ML-XC, providing an accuracy commensurate with QMB methods, is a staggering 10× improvement over the previous high watermark[10] of 64 PFLOPS (5% efficiency on new Sunway) [37] obtained for a Level 2 XC functional. The overall high throughput efficiency obtained for these calculations is a consequence of the high efficiency realized for all main kernels (cf. Table 3). A wall-time of ~4–8 mins per SCF iteration for ~300,000-600,000 e^- using `MLXC` marks an improvement of 100× in system-size and > 100× in time-to-solution over state-of-the-art QMB methods[11]. It underscores that large-scale materials simulations at an accuracy commensurate with quantum

[8] includes multiple passes of Chebyshev filtering in the initial SCF step
[9] RCCL with OFI plugin is also not used in these large-scale runs due to aforementioned instability issues (cf. Sec 5.4.4).
[10] The high watermark for ground-state DFT calculations using a complete basis, and a method that is generically applicable for any materials system, is 46 PFLOPS (28% efficiency on Summit) [6].
[11] Comparing with QMC [20]—the most efficient of QMB methods. Since both QMC and `DFT-FE-MLXC` are $O(N^3)$ scaling, their time-to-solution is expected to scale similarly. Comparing `DFT-FE-MLXC` (Tables 2 & 3) with QMC based on data in [20], this amounts to 220–350× speedup in time-to-solution in terms of sec/GS/electron.

Table 3: Wall-time and sustained performance for a single SCF iteration of TwinDislocMgY(A) (302,668 e^- in supercell), TwinDislocMgY(B) (464,343 e^- in supercell) and TwinDislocMgY(C) (619,124 e^- in supercell) systems, with a breakdown for the key steps shown for TwinDislocMgY(A),(C). Simulations performed on Frontier using 2,400 nodes (FP64 peak: 458.9 PFLOPS), 6,000 nodes (FP64 peak: 1147.2 PFLOPS) and 8,000 nodes (FP64 peak: 1529.6 PFLOPS) for TwinDislocMgY(A),(B),(C) systems, respectively. FLOP count for operations—CholGS-CI, RR-D, discrete Hamiltonian construction (DH), electrostatic potential solve (EP) and Others—that constitute a minor portion of the total FLOP count are not measured, though their wall-times are included in the total time.

System	Wall-time (sec)	FLOP count (PFLOP)	PFLOPS (% of FP64 peak)
TwinDislocMgY(A)	223	50,456.7	**226.3** (**49.3**%)
TwinDislocMgY(B)	499.4	254,147.5	**508.9** (**44.4**%)
TwinDislocMgY(C)	513.7	338,863.4	**659.7** (**43.1**%)
Breakdown for TwinDislocMgY(A)			
Step	Wall-time (sec)	FLOP count (PFLOP)	PFLOPS (% of FP64 peak)
CF	102.3	14,854.2	145.2 (31.6%)
CholGS-S	24.8	6,917.3	278.9 (60.8%)
CholGS-CI	3.8	-	-
CholGS-O	12.1	6,917.3	571.7 (124.6%)
RR-P	22.7	7,341.7	323.4 (70.5%)
RR-D	9.7	-	-
RR-SR	23.5	13,834.6	588.7 (128.3%)
DC	3.3	591.6	179.3 (39.1%)
DH+EP+Others	20.8	-	-
Breakdown for TwinDislocMgY(C)			
Step	Wall-time (sec)	FLOP count (PFLOP)	PFLOPS (% of FP64 peak)
CF	135.4	57,809.5	427 (27.9%)
CholGS-S	79.3	54,428.9	686.4 (44.9%)
CholGS-CI	8.8	-	-
CholGS-O	49.6	54,428.9	1097.4 (71.7%)
RR-P	66.7	61,035.7	915.1 (59.8%)
RR-D	22.3	-	-
RR-SR	93.5	108,857.9	1164.3 (76.1%)
DC	4.3	2,302.5	535.5 (35%)
DH+EP+Others	53.8	-	-

many-body methods are now possible on systems with $O(10^5)$ electrons.

8 IMPLICATIONS

Conducting large length-scale *ab initio* calculations at quantum accuracy is a cherished, yet, elusive goal in materials modeling. *Ab initio* methods suffer from a longstanding accuracy and length-scale dichotomy—QMB methods provide quantum accuracy but scale poorly; DFT can scale but is far from quantum accuracy. `DFT-FE-MLXC` breaks through this dichotomy. We realize, for the first time, calculations on $O(10^5)$ electrons while being commensurate with quantum accuracy. The combination of `invDFT` and `MLXC` that lends `DFT-FE-MLXC` an exceptional 7 mHa/atom accuracy, marks only the beginning of a new and systematic means to model accurate XC functionals in DFT. It now paves the way to the coveted 1 mHa/atom accuracy by use of more expressive and sophisticated forms for `MLXC`. This would, invariably, demand more training data to model the `MLXC`'s. Therein, the low time-to-solution attained in `invDFT` will enable rapid generation of *exact* XC potentials. These developments can unlock the door to numerous consequential scientific applications, heretofore hindered by the unavailability of a large-scale quantum accurate method. `DFT-FE-MLXC` is capable of fast, systematically converged, ground-state calculations at close to quantum accuracy, with wall-time per SCF of ~8 mins on large-scale systems with ~600,000 electrons—hitherto infeasible by any state-of-the-art method—as showcased on quasicrystals and *realistic* metallic alloys with defects. These two case studies on the thermodynamics of quasicrystals and defect interaction in Mg alloys has direct bearing on the design of novel magnetic, and lightweight structural materials, respectively. As `DFT-FE-MLXC` is generic and material-agnostic, it can aid in tackling a diverse set of key scientific and technological problems, including, designing new catalytic materials for clean fuel production, devising materials and mechanisms for CO_2 sequestration, pharmaceutical drug development, discovering novel qubit materials for quantum computers, to name a few.

ACKNOWLEDGEMENTS

We acknowledge our collaboration and discussions with Wenhao Sun and Woohyean Baek on the science of quasicrystals. V.G. and P.M.Z. acknowledge the support from DOE-BES (DE-SC0022241) under the auspices of which the computational framework connecting QMB methods and DFT was developed. V.G. and S.D. acknowledge DOE-BES (DE-SC0008637) for supporting the development of DFT-FE and the study of the energetics of extended defects in Mg alloys. B.K. acknowledges support from Toyota Research Institute that funded initial development and implementation of inverse DFT. P.M. and G.P. acknowledge the support from the Department of Science and Technology India (Startup research grant SRG/2020/002194) and the Ministry of Education India (Prime Minister's Research Fellowship) for the development of GPU matrix-free frameworks employed in the electrostatics treatment of DFT calculations. V.G. also acknowledges AFOSR (FA9550-21-1-0302) that supported mathematical analysis of inverse degenerate eigenvalue problems. This work used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work also used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

[1] W. Kohn, L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140(4A) (1965) A1133.

[2] Becke, A. D. (2014). Perspective: Fifty years of density-functional theory in chemical physics. J. Chem. Phys., 140(18), 18A301.

[3] B. Kanungo, P.M. Zimmerman, V. Gavini, Exact exchange-correlation potentials from ground-state electron densities. Nature communications 10(1) (2019), 4497.

[4] B. Kanungo, J. Hatch, P. M. Zimmerman, & V. Gavini, Exact and model exchange-correlation potentials for open-shell systems. arXiv preprint arXiv:2305.15620 (2023).

[5] P. Motamarri et al., DFT-FE–A massively parallel adaptive finite-element code for large-scale density functional theory calculations. Comp. Phys. Commun. 246 (2020), 106853.

[6] S. Das et al., Fast, scalable and accurate finite-element based ab initio calculations using mixed precision computing: 46 PFLOPS simulation of a metallic dislocation system. In Proceedings of the 2019 International Conference for High Performance Computing, Networking, Storage and Analysis, Art. No. 2.

[7] D. Shechtman et al., Metallic phase with long-range orientational order and no translational symmetry. Phys. Rev. Lett. 53(20) (1984), 1951.

[8] K. Deguchi et al., Quantum critical state in a magnetic quasicrystal. Nature materials 11(12) (2012), 1013.

[9] K. Kamiya et al., Discovery of superconductivity in quasicrystal. Nature communications 9(1) (2018), 154.

[10] H. Takakura et al., Atomic structure of the binary icosahedral Yb–Cd quasicrystal. Nature materials 6(1) (2007), 58.

[11] T. M. Pollock, Weight loss with magnesium alloys, Science 328(5981) (2010) 986.

[12] Z. Wu et al., Mechanistic origin and prediction of enhanced ductility in magnesium alloys. Science 359(6374) (2018) 447.

[13] Sherrill, C. D., & Schaefer III, H. F. (1999). The configuration interaction method: Advances in highly correlated approaches. In Advances in quantum chemistry (Vol. 34, pp. 143-269). Academic Press.

[14] K.D. Vogiatzis et al., Pushing configuration-interaction to the limit: Towards massively parallel MCSCF calculations. J. Chem. Phys. 147(18) (2017), 184111.

[15] A.E. Rask, P.M. Zimmerman, Toward full configuration interaction for transition-metal complexes. J. Phys. Chem. A 125(7) (2021), 1598.

[16] Bartlett, R. J., & Musiał, M. (2007). Coupled-cluster theory in quantum chemistry. Reviews of Modern Physics, 79(1), 291.

[17] P.R. Nagy, Approaching the basis set limit of CCSD (T) energies for large molecules with local natural orbital coupled-cluster methods. J. Chem. Theory Comput. 15(10) (2019), 5275.

[18] W.M.C. Foulkes et al., Quantum Monte Carlo simulations of solids. Rev. Mod. Phys. 73(1) (2001), 33.

[19] Towards QMCPACK Performance Portability. URL: https://ecpannualmeeting.com/assets/overview/posters/QMCPACK_Kent_ECP_Houston_2020.pdf

[20] https://asc.llnl.gov/sites/asc/files/2020-09/coral2_qmcpack_vfeb13.pdf

[21] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54(16) (1996), 11169.

[22] P. Giannozzi et al., Advanced capabilities for materials modelling with QUANTUM ESPRESSO, J. Phys. Condens. Matter. 29(46) (2017), 465901.

[23] F. Gygi et al., Large-scale electronic structure calculations of high-Z metals on the BlueGene/L platform. In Proceedings of the 2006 ACM/IEEE conference on Supercomputing.

[24] M. Valiev et al., NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations, Comput. Phys. Commun. 181(9) (2010), 1477.

[25] J. Hutter et al., CP2K: atomistic simulations of condensed matter systems, Wiley Interdiscip. Rev.: Comput. Mol. Sci. 4(1) (2014) 15.

[26] A. N. Ziogas et al., A data-centric approach to extreme-scale ab initio dissipative quantum transport simulations. In Proceedings of the 2019 International Conference for High Performance Computing, Networking, Storage and Analysis, Art. No. 1.

[27] V. Blum et al., Ab initio molecular simulations with numeric atom-centered orbitals. Comp. Phys. Commun. 180 (2009), 2175.

[28] J. R. Chelikowsky, N. Troullier, Y. Saad, Finite-difference-pseudopotential method: Electronic structure calculations without a basis, Phys. Rev. Lett. 72 (1994), 1240.

[29] Q. Xu et al., SPARC: Simulation Package for Ab-initio Real-space Calculations. Software X 15 (2021), 100709.

[30] M. Dogan, K.H. Liou, J.R. Chelikowsky, Solving the electronic structure problem for over 100,000 atoms in real-space, Physical Review Materials, 7(6) (2023), L063001.

[31] Y. Hasegawa et al., First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer, In Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis.

[32] J.E. Pask, P.A. Sterne, Finite element methods in ab initio electronic structure calculations, Modell. Simul. Mater. Sci. Eng. 13 (2005), R71.

[33] S. Das, et al., DFT-FE 1.0: A massively parallel hybrid CPU-GPU density functional theory code using finite-element discretization. Comp. Phys. Commun. 280 (2022), 108473.

[34] D.R. Bowler, T. Miyazaki, Calculations for millions of atoms with density functional theory: linear scaling shows its potential, J Phys. Condens. Matter 22(7) (2010), 074207.

[35] L. Lin, et al. Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion, J Phys. Condens. Matter 25 (2013), 295501.

[36] J.-L. Fattebert et al., Modeling dilute solutions using first-principles molecular dynamics: computing more than a million atoms with over a million cores. In Proceedings of 2016 International Conference for High Performance Computing, Networking, Storage and Analysis.

[37] W. Hu et al., 2.5 Million-Atom Ab Initio Electronic-Structure Simulation of Complex Metallic Heterostructures with DGDFT. In Proceedings of 2022 International Conference for High Performance Computing, Networking, Storage and Analysis.

[38] W. Hu, et al. Adaptively compressed exchange operator for large-scale hybrid density functional calculations with applications to the adsorption of water on silicene. J. Chem. Theory Comput. 13(3) (2017), 1188.

[39] Y. Shi, A. Wasserman, Inverse Kohn–Sham density functional theory: progress and challenges. J. Phys. Chem. Lett. 12(22) (2021), 5308.

[40] T. Heaton-Burgess, F.A. Bulat, W. Yang, Optimized effective potentials in finite basis sets. Phys. Rev. Lett. 98(25) (2007), 256401.

[41] A.P. Gaiduk, I.G. Ryabinkin, V.N. Staroverov, Removal of basis-set artifacts in Kohn–Sham potentials recovered from electron densities. J. Chem. Theory Comput. 9(9) (2013), 3959.

[42] L.A. Curtiss et al. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. J. Chem. Phys. 106(3) (1997), 1063.

[43] P. Motamarri et al., Higher-order adaptive finite-element methods for Kohn-Sham density functional theory, J. Comput. Phys. 253 (2013), 308.

[44] Y. Zhou et al., Self-consistent-field calculations using Chebyshev-filtered subspace iteration, J. Comput. Phys. 219(1) (2006), 172.