LEARN THE FIELD, CONSTRUCT THE MATRIX: RETHINKING TARGETS FOR KOHN–SHAM DFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning is increasingly coupled with Kohn–Sham density functional theory (DFT), yet it remains unclear which electronic representation provides the most effective interface to the solver. We compare four candidates—charge density ρ, effective potential V_{eff}, Hamiltonian H, and density matrix D—on a unified 43k-crystal benchmark, evaluating native prediction accuracy, closed-loop self-consistent-field (SCF) acceleration, and downstream property recovery. Although H and D lie closest to the eigensolution, their learned counterparts realize little of the large acceleration available to exact matrix states. In contrast, learned ρ and V_{eff} accelerate SCF on 90.3% and 84.3% of test crystals, with median loop speedups of $1.18\times$ and $1.17\times$. More strikingly, matrices reconstructed from learned real-space fields outperform dedicated matrix-input routes on 97.6–99.9% of test crystals after one solver step, while residual V_{eff} further improves downstream recovery under matched electronic work. These results show that proximity to the eigensolution is a poor criterion for choosing a learning target. Instead, effective ML–DFT interfaces should expose quantities from which the solver can analytically reconstruct constrained electronic states, rather than requiring neural networks to learn those constraints directly.

1 INTRODUCTION

Machine learning is increasingly used either to initialize the self-consistent-field (SCF) loop of Kohn–Sham density functional theory (DFT) (Kohn & Sham, 1965), preserving the DFT fixed point, or to bypass part of the solver for direct electronic-property prediction. These two roles have evolved around different learning targets: the charge density ρ is commonly used for SCF initialization (Koker et al., 2024; Chen et al., 2025; Li et al., 2025a), whereas the Hamiltonian H and density matrix D are increasingly learned for electronic-structure and property prediction (Li et al., 2022; Yu et al., 2024; Shao et al., 2023; Hazra et al., 2024). It remains unclear whether these choices reflect an intrinsic advantage of the representations themselves or simply the tasks for which they have historically been developed.

These targets occupy different positions in the Kohn–Sham computational graph (Figure 1). The density ρ generates the effective potential V_{eff}, which enters the Hamiltonian H; solving the generalized eigenproblem yields the spectrum and occupied subspace, from which the density matrix D and a new density are constructed. This ordering suggests a natural but largely untested hypothesis: representations closer to the eigensolution should provide more effective targets for property prediction, while ρ should remain preferable for SCF initialization. Existing studies cannot directly test this hypothesis because the targets are typically evaluated with different datasets, error metrics, and decoding procedures.

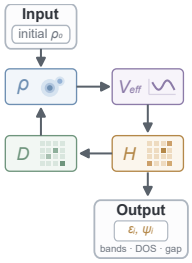


Figure 1: **Representations in the SCF cycle.**

One field in this cycle has been left out. In the Kohn–Sham equations the density reaches the spectrum only through the local effective potential $V_{\text{eff}}(\mathbf{r})$: V_{eff} is the field the Hamiltonian actually contains, whereas a predicted ρ must still pass through the long-range Hartree kernel and the nonlinear exchange–correlation map before it can move an eigenvalue. Yet V_{eff} is conventionally treated as a quantity derived from ρ rather than as a learning target in its own right, even though the quantum-chemistry literature has long found a superposition of atomic potentials to be at least as good an SCF initial guess as the superposition of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

atomic densities (SAD) used as the standard baseline (Lehtola, 2019). Like $\rho(\mathbf{r})$, V_{eff} is a periodic real-space scalar field whose representation is uniform across materials, but it enters the Kohn–Sham operator directly before eigensolution; predicting it introduces its own requirements, including a consistent potential gauge and subsequent assembly with kinetic, nonlocal, and overlap terms.

Following the ρHD-43K data-generation protocol (Chen et al., 2026), we generate converged V_{eff} fields for the same frozen periodic split and evaluate learned ρ, V_{eff}, H, and D under a unified framework. We assess each representation along three axes: native prediction accuracy, closed-loop SCF acceleration, and downstream property recovery. Each target is evaluated through its natural solver deployment path, complemented by a matched eigensolve that gives every representation identical electronic work and so isolates representation effects from subsequent electronic updates.

Our contributions are:

- We introduce a 43k-crystal benchmark for the gauge-fixed local effective potential, together with a gauge-aware equivariant field model and a solver-ready interface, enabling V_{eff} to be evaluated alongside ρ, H, and D for both SCF initialization and property recovery.
- We show that the pre-registered learned V_{eff} and ρ pipelines both accelerate SCF: they reduce iteration counts for 84.3% and 90.3% of test crystals and achieve median loop speedups of $1.17\times$ and $1.18\times$, respectively; a non-pre-registered ChargE3Net-ΔV_{eff} residual reaches $1.26\times$. In contrast, the evaluated learned matrix initializers realize little of the acceleration exposed by their exact-oracle counterparts.
- We find that real-space fields form a more solver-compatible interface than the evaluated matrix pipelines: after one solver step, field-decoded matrices outperform dedicated matrix-input routes for 97.6–99.9% of test crystals. Within the field family, SCF speedups are similar; the non-pre-registered ChargE3Net-ΔV_{eff} residual has lower matched-eigensolve errors than ρ on all six reported properties.

2 Background and Related Work

2.1 Electronic Representations in the Kohn–Sham SCF Map

A Kohn–Sham self-consistent-field (SCF) iteration propagates the electronic solution through several coupled representations,

$$\rho^{(n)} \to V_{\text{eff}}^{(n)} \to H^{(n)} \to \{\epsilon_i, C_i\}^{(n)} \to D^{(n)} \to \rho_{\text{out}}^{(n)},$$

before the output density is mixed to form the next iterate. For a local or semilocal functional,

$$V_{\text{eff}}[\rho](\mathbf{r}) = v_{\text{loc}}(\mathbf{r}) + v_{\text{H}}[\rho](\mathbf{r}) + v_{\text{xc}}[\rho](\mathbf{r}), \tag{1}$$

and, in a localized atomic-orbital basis,

$$H_{\mu\nu} = T_{\mu\nu} + V_{\mu\nu}^{\text{nl}} + \langle\varphi_\mu|V_{\text{eff}}|\varphi_\nu\rangle. \tag{2}$$

The eigensolution of H determines the occupied subspace and hence the density matrix D, from which ρ is reconstructed.

These quantities provide different but coupled views of the same self-consistent electronic solution. The density ρ and effective potential V_{eff} are real-space scalar fields with a representation that is uniform across chemical species, whereas H and D are orbital-space matrices whose block structure and transformation laws depend on the atomic basis and composition. The density additionally satisfies electron-number normalization, while V_{eff} is defined only up to an additive constant. Thus, moving from ρ toward H brings the representation closer to the eigensolution but also increases its structural complexity. In this sequence, V_{eff} is a natural intermediate target: it retains the uniform real-space representation of ρ while directly entering Hamiltonian construction.

2.2 Learning Electronic Fields and Matrices

Machine-learning approaches to electronic structure have largely developed along two parallel directions. Real-space density models, from early symmetry-adapted approaches (Grisafi et al.,

Table 1: **Representative machine-learning approaches for electronic structure, 2022–2026.** Rows are grouped by learned target and sorted by year within each group; no prior work listed here learns V_{eff}. SCF column: ✓, the learned state is tested as an initializer inside a closed self-consistent loop; Bypass, properties are read from it without iterating to self-consistency; –, only the learned target itself is scored. Unified: multiple electronic representations are aligned on the same structures and compared under one protocol. Mol./Per.: molecules/periodic crystals; MP: Materials Project.

Work	Domain	Dataset / chemical scope	Target	Primary evaluation	SCF	Unified
Jørgensen & Bhowmik (2022)	Both	QM9, liquid electrolyte, NMC	ρ	Density accuracy	–	–
ChargE3Net (2024)	Both	QM9, NMC, >100k MP crystals	ρ	Density and SCF acceleration	✓	–
ECD (2025)	Per.	140k PBE crystals; 7k HSE	ρ	Density; SCF; PBE→HSE	✓	–
Li et al. (2025a)	Mol.	Unseen geometries and elements	ρ	Density accuracy, SCF init.	✓	–
ELECTRA (2025)	Mol.	QM9 densities	ρ	Density accuracy, SCF init.	✓	–
DeepH (2022)	Per.	Crystals, low-dim. families	H	Matrix error; bands, DOS	Bypass	–
DeepH-E3 (2023)	Per.	Graphene, MoS_2, twisted, SOC	H	Matrix error; bands, topology	Bypass	–
QHNet (2023)	Mol.	MD17 molecules	H	H accuracy, SCF steps	✓	–
HamGNN (2023)	Both	Molecules, C/Si, SiO_2, Bi–Se	H	Matrix error; bands, wavefns.	Bypass	–
QH9 (2024)	Mol.	130k QM9 molecules, dynamics	H	H benchmark, SCF init.	✓	–
Zhong et al. (2024)	Per.	Materials Project, broad chemistry	H	Spectra and screening	Bypass	–
Gong et al. (2024)	Per.	Plane-wave labels in an AO basis	H	Reconstruction and bands	Bypass	–
QHFlow (2025)	Mol.	MD17 and QH9	H	H accuracy, SCF init.	✓	–
NextHAM (2026)	Per.	∼17k structures, >60 elements	H	H accuracy, bands	Bypass	–
UniHam (2026)	Per.	>100k crystals, 72 elements, SOC	H	H MAE, bands, DOS	Bypass	–
QMLearn (2023)	Mol.	Water through benzene, propanol	D	Energy, forces, gaps, AIMD	Bypass	–
Hazra et al. (2024)	Mol.	Selected molecules, configs.	D	SCF conv., forces, AIMD	✓	–
ρHD-43K (2026)	Per.	43k crystals, aligned labels	ρ, H, D	Native error, closed-loop SCF	✓	✓
This work (2026)	**Per.**	**43k crystals, one frozen split**	$\boldsymbol{\rho}$, $\boldsymbol{V_{\text{eff}}}$, $\boldsymbol{H}$, $\boldsymbol{D}$	**Native, SCF, and properties**	✓	✓

2018; Fabrizio et al., 2019; Lewis et al., 2021) to equivariant molecular and periodic architectures (Jørgensen & Bhowmik, 2022; Koker et al., 2024; Elsborg et al., 2025; Li et al., 2025b; Elsborg et al., 2026), have demonstrated accurate prediction of electronic densities and related observables. In parallel, molecular benchmarks such as QH9 (Yu et al., 2024) and orbital-equivariant models such as QHNet (Yu et al., 2023), DeepH (Li et al., 2022), DeepH-E3 (Gong et al., 2023), HamGNN (Zhong et al., 2023), and NextHAM (Yin et al., 2026) have shown that Hamiltonians can also be learned accurately, from relatively restricted molecular or material chemistries to increasingly diverse periodic systems (Zhong et al., 2024; Huang et al., 2026). These studies establish that both fields and matrices are learnable, but they are typically evaluated with representation-specific metrics and downstream tasks, leaving unclear whether the representation closest to the eigensolution is also the most effective learning target. Our work addresses this question by comparing real-space fields and orbital-space matrices within one periodic-material benchmark.

This distinction is especially relevant for chemically diverse crystals. Real-space fields retain the same scalar representation across elements, whereas orbital-space matrices must accommodate species-dependent orbital channels, basis-dependent block structures, and their associated equivariance. Conversely, matrices are directly connected to the eigensolution and are therefore natural targets for spectral prediction. Existing work has largely optimized models for a chosen target; here we instead treat the choice of electronic representation itself as the object of study.

2.3 SCF Initialization and Unified Evaluation

Learned densities (Koker et al., 2024; Chen et al., 2025; Li et al., 2025a; Elsborg et al., 2025), density matrices (Hazra et al., 2024), and Hamiltonians (Yu et al., 2023; 2024; Kim et al., 2025) have all been used to initialize or accelerate SCF calculations. Chen et al. (2026) provided a unified periodic benchmark for ρ, H, and D by evaluating all three representations on the same frozen dataset and within the same DFT solver. It showed that target-space accuracy does not directly translate into closed-loop SCF acceleration, but it focused primarily on initialization and did not include V_{eff} or a systematic comparison of downstream property recovery. The present work extends this setting from an SCF-initialization benchmark to a broader representation benchmark by adding V_{eff} and evaluating all four targets under the same framework.

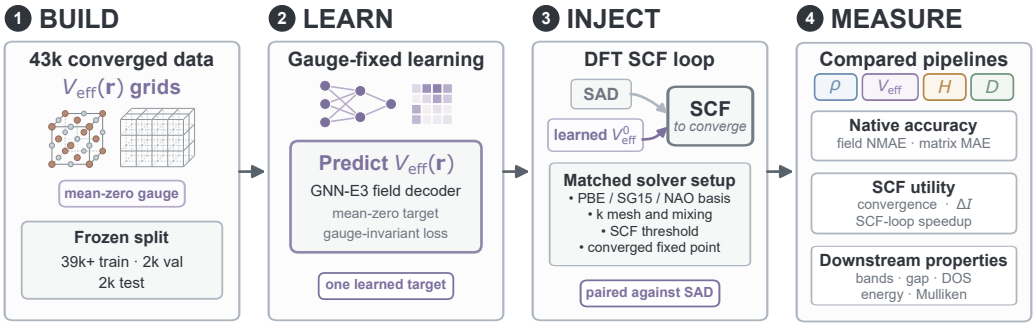


Figure 2: **V_{eff}-centered benchmark and evaluation workflow.** The four stages, covered in turn in Sections 3.1–3.5, are construction of the V_{eff} dataset, learning of the gauge-fixed field, injection into the SCF solver, and evaluation at the native-error, closed-loop, and downstream-property levels. The ρ, D, and H pipelines are controlled comparators.

Machine-learned exchange–correlation functionals and potentials (Snyder et al., 2012; Nagai et al., 2020; Li et al., 2021; Kirkpatrick et al., 2021) address a related but distinct problem: they modify the density-to-potential map and are evaluated repeatedly inside the SCF loop. Here, V_{eff} is instead learned as a representation of the converged electronic state and used without changing the underlying DFT functional.

Table 1 places these choices side by side. Only the periodic ρHD comparison and this work align more than one representation on one frozen split and score both closed-loop SCF and, here, downstream recovery.

3 Benchmark and Evaluation Protocol

Figure 2 lays out the benchmark: one protocol carries all four targets from dataset construction through solver injection to evaluation, so the ρ, D, and H pipelines are controlled comparators rather than separately tuned systems. The three evaluation levels are not interchangeable. Native error is expressed in each representation's own support and units and is therefore comparable only within a representation, whereas the closed-loop and downstream levels are defined on shared outcomes and carry the cross-target comparisons.

3.1 Frozen Dataset and Coverage

Following the ρHD-43K protocol (Chen et al., 2026), we generate converged V_{eff} grids for the same Alexandria-derived inorganic crystals (Schmidt et al., 2021), using the same calculation settings and frozen 39,851/2,000/2,000 train/validation/test partition. After excluding a small number of failed calculations, the benchmark contains 43k crystals with aligned electronic-structure labels. All calculations use PBE (Perdew et al., 1996), SG15 pseudopotentials (Schlipf & Gygi, 2015), and a double-ζ plus polarization numerical atomic-orbital basis in ABACUS v3.11.0-beta.4 (Zhou et al., 2025); Appendix Table 8 gives the raw file structure of each target. The dataset spans broad structural and chemical diversity: most structures contain 5–32 atoms per cell, concentrated around 9–16 atoms, and most crystals contain three or four chemical species, while the element-occurrence map shows broad but nonuniform coverage of the periodic table (Figure 3). The frozen test split closely follows the full dataset in both structural size and chemical composition, which matters because grid size and chemical complexity directly affect both field inference and SCF cost. Appendix C.1 lists the failed calculations per split and the sample-alignment and leakage audit.

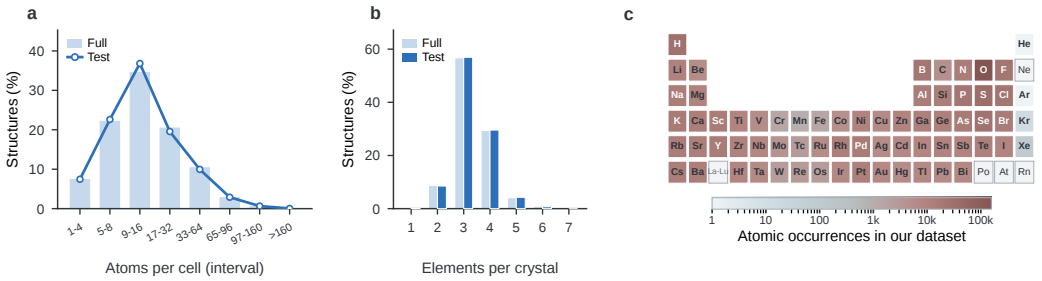


Figure 3: **Coverage of our dataset and its frozen test set.** (a) Atoms per cell and (b) chemical species per crystal, for the full dataset and the frozen test split; (c) elemental occurrences in the full dataset. The test split follows the full-dataset distributions in (a) and (b).

3.2 Gauge-Fixed Target and Equivariant Decoder

For crystal x with a local potential $V_x(\mathbf{r})$ in Rydberg, the normalized training target is

$$y_x(\mathbf{r}) = \frac{V_x(\mathbf{r}) - \langle V_x \rangle_{\Omega_x}}{s_{\text{train}}}, \qquad s_{\text{train}} = 3.036109\ \text{Ry}, \tag{3}$$

where Ω_x is the cell and s_{train} is the training-set median of per-structure 95th-percentile absolute deviations, recomputable from the frozen training split and released at full precision with the configuration. This convention removes the physically irrelevant cellwise constant and is fixed for training, evaluation, and export. Appendix Table 9 lists the physical constraints each target carries and which of them an equivariant regressor satisfies by construction.

The model retains the periodic E(3)-equivariant scalar decoder of ChargE3Net (Koker et al., 2024) and evaluates arbitrary grid probes, avoiding padding between material-specific FFT grids. Three interaction layers use a 4 Å cutoff, multiplicity 256, and channels through $\ell_{\max} = 3$. A mean-centered, tail-weighted SmoothL1 objective ($\beta = 0.05$) replaces the density loss; the point weight $1 + 0.25 \min(|y|, 4)$ mildly emphasizes sharp potential regions without imposing a Poisson equation or electron-count constraint. Multiplicity 256 was selected on a frozen validation panel: it improves NMAE by 14.5% over 128, while 500 adds no measurable gain. As residual controls, ChargE3Net and EdenGNN (Li et al., 2025b) are also trained on the correction $V - V_0$ to the SAD-derived potential. The cutoff, complete architecture, optimization settings, and design pilot are reported in Appendices C.8–C.9 and Table 17.

3.3 Solver Injection

We de-normalize the predictions, remove the numerical cell mean, and write one checksummed ABACUS cube per test crystal. The customized path reads the local potential before the first diagonalization. SAD, the learned initializer, and the exact-V_{eff} oracle share the 10^{-7} threshold, 100-iteration limit, and all other settings. The oracle is not deployable; it separates prediction error from interface headroom. Appendix C.4 gives the timing and retry audit.

3.4 Native Error and Closed-Loop Utility

Appendix Table 6 summarizes the three ordered evaluation levels, each of which supports a different claim. For Levels II and III we pre-register one primary pipeline per target: direct ChargE3Net (Koker et al., 2024) for ρ, residual HamGNN-Matrix (Zhong et al., 2023) for D, residual NextHAM (Yin et al., 2026) for H, and direct ChargE3Net for V_{eff}. Because the four targets differ in output type and in the constraints they must satisfy (Section 2.1), one shared width, loss, or parameter budget would favor whichever target best matched that arbitrary design; we therefore keep architectures and objectives target appropriate and freeze one primary row per representation before inspecting the properties, and read the result as a ranking of pipelines rather than of representations. Model selection and capacity controls are in Appendix C.8.

For crystal i in the valid test set $\mathcal{T}$, let $\widetilde{V} = V - \langle V \rangle_{\Omega_i}$ denote the gauge-aligned grid. The reported potential error averages a per-crystal normalized absolute error,

$$E_V = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \frac{\sum_{\mathbf{r} \in \Omega_i} |\widetilde{\widehat{V}}_i(\mathbf{r}) - \widetilde{V}_i^*(\mathbf{r})|}{\sum_{\mathbf{r} \in \Omega_i} |\widetilde{V}_i^*(\mathbf{r})|}. \tag{4}$$

The other representations retain their native support and units rather than being forced into a common numerical scale.

For initialized crystal i, positive ΔI_i means that the initializer saves SCF steps. We report

$$\Delta I_i = I_i^{\mathrm{SAD}} - I_i^{\mathrm{init}}, \qquad R_{\mathrm{fewer}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[i \in \mathcal{C}, \Delta I_i > 0], \qquad S_{\mathrm{SCF}} = \underset{i \in \mathcal{C}}{\mathrm{median}} \frac{T_i^{\mathrm{SAD}}}{T_i^{\mathrm{init}}}, \tag{5}$$

where $N = 2{,}000$, $\mathcal{C}$ contains strict convergence pairs, and T is summed SCF-iteration time. Missing predictions, failures, and timeouts count as unsuccessful in R_{fewer} and are excluded from the conditional median S_{SCF}, which is therefore always read together with the convergence coverage $|\mathcal{C}|/N$.

3.5 Downstream Property Recovery

For property q, the primary statistic is the median paired discrepancy

$$E_q = \underset{i \in \mathcal{V}_q}{\mathrm{median}} d_q(\widehat{q}_i, q_i^*), \tag{6}$$

where d_q is the aligned eigenvalue MAE, normalized-L_1 DOS error, or the absolute gap, energy-per-atom, or Mulliken-population error, and $\mathcal{V}_q$ is the set of test crystals scored by every pipeline compared.

No Level III readout iterates to self-consistency: in the terms of Table 1 all are bypass readouts, and they differ only in how much electronic work follows the prediction. Table 4 arranges them in three blocks that share structures, basis, pseudopotentials, k points, and output definitions. The *reference* block starts ABACUS from its atomic-density (SAD) guess and takes one electronic step, fixing what a non-learned start yields. In the *solver-native one-step* block, ρ, D, and H take one ABACUS step, while ChargE3Net-ΔV_{eff} is diagonalized at fixed potential; only ρ receives an extra density update. This supports deployment claims. In the *matched eigensolve* block, EdenGNN-ΔV_{eff} is converted outside ABACUS to a Kohn–Sham Hamiltonian (D via $D \to \rho \to V_{\mathrm{eff}} \to H$) and decoded by one generalized eigensolve with no density update, so all routes receive identical electronic work; this block supports cross-representation interpretation. Energies in both learned blocks are diagnostics rather than variational ground-state energies. Throughout, we remove the cS Hamiltonian gauge, align spectra to the Fermi level, and use 0.10 eV DOS broadening.

We call a paired difference *resolved* when a 10,000-resample paired bootstrap interval over materials excludes zero. Every rate keeps the full 2,000-crystal test denominator with missing or failed cases counted as failures; medians are taken over the crystals each comparison can score. Exact denominators and decoder definitions are in Appendix C.5.

4 Results and Discussion

4.1 Level I: Native Prediction Accuracy

The converged local field is learnable to 1.00% of its own scale with a residual model. On the test split, direct ChargE3Net-V_{eff} reduces the gauge-aligned NMAE from 6.38% for the SAD-derived potential to 1.50% ($4.26\times$; Table 2), with per-crystal mean/median/p95 MAE of 0.0140/0.0134/0.0219 Ry and mean RMSE 0.0198 Ry, evaluated on all 2.36 billion FFT voxels of the test grids rather than on training probes. Normalizing each block by its own SAD-derived reference gives the one cross-representation reading the table supports: the learned fields improve on their atomic references by $10.4\times$ (ρ) and $6.38\times$ (V_{eff}), the best learned H by only $1.40\times$, and no learned D beats its reference. Direct regression leads for ρ, while the updated ChargE3Net residual leads for V_{eff}; for matrices only residuals on the atomic reference come close to or beat it. The matrix

Table 2: **Level I: native prediction accuracy** on the test split; lower is better. Errors compare only within a block; "× vs SAD" (reference error / model error) compares across blocks. Shaded: SAD-derived reference; bold: best learned model; Δ: residual on the reference. ρ and V_{eff}: normalized full-grid L_1 error (Eq. 4); D, H: entrywise MAE, H on the converged support. ρ, D, and H rows, including ELECTRAFI (Elsborg et al., 2025; 2026) and EdenGNN-$\Delta\rho$ (Li et al., 2025b), are from Chen et al. (2026); test counts in Appendix C.1.

Representation	Model / reference	Predicted state	Native metric	Test error	× vs SAD
ρ	SAD-derived reference	ρ_0	full-grid ϵ_ρ (%)	4.51	1.00
ρ	ELECTRAFI-ρ	$\hat{\rho}_*$	full-grid ϵ_ρ (%)	0.951	4.74
ρ	ChargE3Net-ρ	$\hat{\rho}_*$	full-grid ϵ_ρ (%)	**0.435**	**10.4**
ρ	ChargE3Net-$\Delta\rho$	$\rho_0 + \widehat{\Delta\rho}$	full-grid ϵ_ρ (%)	1.55	2.92
ρ	EdenGNN-$\Delta\rho$	$\rho_0 + \widehat{\Delta\rho}$	full-grid ϵ_ρ (%)	1.67	2.70
V_{eff}	SAD-derived reference	$V_{\text{eff},0}$	gauge-aligned NMAE (%)	6.38	1.00
V_{eff}	ChargE3Net-V_{eff}	$\hat{V}_{\text{eff},*}$	gauge-aligned NMAE (%)	1.50	4.26
V_{eff}	ChargE3Net-ΔV_{eff}	$V_{\text{eff},0} + \widehat{\Delta V}_{\text{eff}}$	gauge-aligned NMAE (%)	**1.00**	**6.38**
V_{eff}	EdenGNN-ΔV_{eff}	$V_{\text{eff},0} + \widehat{\Delta V}_{\text{eff}}$	gauge-aligned NMAE (%)	1.77	3.61
D	SAD-derived reference	D_0	entrywise MAE (10^{-4})	0.583	1.00
D	HamGNN-Matrix-D	$\hat{D}_*$	entrywise MAE (10^{-4})	5.39	0.11
D	HamGNN-Matrix-ΔD	$D_0 + \widehat{\Delta D}$	entrywise MAE (10^{-4})	**1.63**	**0.36**
H	SAD-derived reference	H_0	support MAE (10^{-4} Ry)	2.13	1.00
H	HamGNN-Matrix-H	$\hat{H}_*$	support MAE (10^{-4} Ry)	24.5	0.087
H	DeepH-E3-H	$\hat{H}_*$	support MAE (10^{-4} Ry)	26.5	0.080
H	NextHAM-ΔH	$H_0 + \widehat{\Delta H}$	support MAE (10^{-4} Ry)	**1.52**	**1.40**

references are strong for a structural reason: H_0 is assembled by the solver from the SAD density, so its kinetic, nonlocal, and local-pseudopotential terms, sparsity, and overlap consistency are exact and only the Hartree–exchange–correlation part carries error, leaving a self-consistent correction $H^* - H_0$ with a median norm of 5.3% of H^*; D_0 is the eigensolution of H_0 and inherits its electron count and idempotency. A direct regressor must reproduce the whole matrix and errs by about as much as a typical supported entry, almost entirely in inter-atomic blocks where H_0 errs $14\times$ less (Appendix C.3); a residual model learns only the correction. Among the matrices only NextHAM-ΔH beats its reference; both ΔV_{eff} models do too ($6.38\times$ for ChargE3Net, $3.61\times$ for EdenGNN), but only the former beats the direct model ($4.26\times$), which remains pre-registered. These errors are inputs to Levels II and III, not evidence of solver or property utility.

4.2 Level II: Closed-Loop SCF Acceleration

The pre-registered direct ChargE3Net-V_{eff} initializer uses fewer iterations than SAD on 84.3% of the test crystals, with a median loop speedup of $1.172\times$ (95% CI $[1.166, 1.179]$; $1.111\times$ including the 13.2 s neural inference, Appendix C.4), whereas the exact potential converges in one step ($15.90\times$). The deployable gain is robust but captures only a fraction of the interface headroom (Table 3).

Pointwise field accuracy alone does not explain the saving. For the primary model, full-grid NMAE correlates with iterations saved only weakly and with the counterintuitive sign (Spearman $r_s = +0.30$), because larger and chemically heavier cells are both harder to predict and slower to converge, whereas the first-step density deviation relative to SAD predicts it ($r_s = -0.58$; Appendix Figure 5). Across models, the updated ChargE3Net-ΔV_{eff} leads both native error (1.00%) and SCF gain ($1.26\times$, 92.6% fewer-step). Yet native error is not a sufficient ranking: EdenGNN-ΔV_{eff} is less accurate than the direct model (1.77% versus 1.50%) but accelerates more ($1.22\times$/90.4% versus $1.17\times$/84.3%). A density residual must still pass through the long-range Hartree kernel and the nonlinear exchange–correlation map before entering H, making SCF performance sensitive to density-error modes that may be weakly reflected by pointwise metrics. A potential residual instead enters H directly, so the first diagonalization is already performed with the corrected operator and returns its corresponding density. Acceleration depends on where the initial state lands in the solver's basin, not on how small its error is. The residual comparisons were not pre-registered; direct ChargE3Net-V_{eff} remains the pre-registered comparison.

Table 3: **Level II: closed-loop SCF acceleration** against SAD: $\Delta I = I^{\mathrm{SAD}} - I^{\mathrm{init}}$, "fewer than SAD" counts $\Delta I > 0$, and loop speedup is the median $T^{\mathrm{SAD}}/T^{\mathrm{init}}$. Rates use the 2,000-crystal test split, counting failed runs as failures; iteration and speedup medians use strict convergence pairs. Oracle rows inject the exact converged state. Bold: best learned result per block, not necessarily better than SAD. Accounting in Appendices C.4 and C.8.

Representation	Model / initializer	Type	Conv.	Med. I	Med. ΔI [IQR]	Fewer than SAD, n (%)	Loop speedup vs SAD
all	ABACUS SAD default	baseline	99.65%	16	0 [0, 0]	–	1.00×
ρ	True charge density	oracle	99.60%	1	+15 [+14, +17]	1992 (99.6)	15.91×
ρ	ELECTRAFI-ρ	direct	99.20%	15	+1 [0, +2]	1162 (58.1)	1.06×
ρ	ChargE3Net-ρ	direct	99.35%	**14**	**+3** [+**2**, +**3**]	**1806 (90.3)**	**1.18×**
ρ	ChargE3Net-$\Delta\rho$	residual	99.15%	18	−1 [−3, 0]	367 (18.4)	0.93×
ρ	EdenGNN-$\Delta\rho$	residual	99.25%	18	−1 [−3, 0]	367 (18.4)	0.93×
V_{eff}	True effective potential	oracle	98.85%	1	+15 [+14, +17]	1992 (99.6)	15.90×
V_{eff}	ChargE3Net-V_{eff}	direct	99.55%	14	+2 [+1, +3]	1686 (84.3)	1.17×
V_{eff}	ChargE3Net-ΔV_{eff}	residual	99.60%	**13**	**+3** [+**2**, +**4**]	**1852 (92.6)**	**1.26×**
V_{eff}	EdenGNN-ΔV_{eff}	residual	99.45%	13	+3 [+2, +4]	1807 (90.4)	1.22×
D	True density matrix	oracle	99.65%	1	+15 [+13, +17]	1993 (99.7)	14.56×
D	HamGNN-Matrix-D	direct	99.65%	18	−2 [−3, −1]	90 (4.5)	0.90×
D	HamGNN-Matrix-ΔD	residual	99.60%	**16**	**0** [−**1**, +**1**]	**845 (42.3)**	**1.00×**
H	True Hamiltonian	oracle	99.65%	1	+15 [+13, +17]	1993 (99.7)	15.19×
H	HamGNN-Matrix-H	direct	99.55%	22	−5 [−8, −3]	0 (0.0)	0.75×
H	DeepH-E3-H	direct	99.60%	23	−6 [−10, −4]	1 (0.05)	0.73×
H	NextHAM-ΔH	residual	99.50%	**17**	**−1** [−**3**, +**1**]	**710 (35.5)**	**0.97×**

Table 4: **Level III: property recovery without self-consistency.** Median absolute errors; gap on gapped crystals only. Spec. ± 2: eigenvalue MAE near Fermi. One-step cost excludes inference, property output, and injection overhead (CPU for reference/one-step, GPU for matched); cross-run GPU times are not controlled. Bold: better field route (paired 95% interval excludes zero); †extra density update. Both V_{eff} rows are non-pre-registered residuals; the pre-registered direct control is Appendix Table 11. Protocol in Appendix C.5.

Block	Input	Model / decode route	Cost (s)	Downstream observables					
				All bands (eV)	Spec. ± 2 (eV)	DOS L_1	Gap (eV)	Energy (meV/atom)	Mulliken (e)
reference	SAD	ABACUS atomic density	27.7	1.1087	0.3658	0.740	0.897	133	0.393
solver-native one-step	ρ	ChargE3Net	30.7	0.0840	0.0350	0.109	0.049	**0.822**†	0.032
	V_{eff}	ChargE3Net-ΔV_{eff}		**0.0418**	**0.0200**	**0.059**	**0.037**	2.91	**0.017**
	H	NextHAM-ΔH		0.5509	0.2424	0.467	0.202	36.7	0.262
	D	HamGNN-Matrix-ΔD		0.9044	0.3135	0.603	0.552	86.6	0.397
matched eigensolve	ρ	PBE reconstruction + eigensolve	22.7	0.0739	0.0324	0.088	0.049	8.88	0.032
	V_{eff}	ChargE3Net-ΔV_{eff} + eigensolve	49.4	**0.0451**	**0.0224**	**0.058**	**0.044**	**4.23**	**0.020**
	H	NextHAM-ΔH + eigensolve	3.9	0.2527	0.1921	0.250	0.319	2745	0.569
	D	$D \to \rho \to V_{\mathrm{eff}} \to H$ + eigensolve	52.4	0.7046	0.3396	0.529	0.602	1193	0.381

Across targets, position in the SCF cycle does not order the outcome. The pre-registered ChargE3Net-ρ reduces iterations on more crystals than the direct V_{eff} model (90.3% versus 84.3%), but their loop-time gains are unresolved (paired median difference $+0.0013$, 95% CI $[-0.0027, +0.0042]$). The matrices are the sharper control: exact D and H expose $14.56\times$ and $15.19\times$ of headroom, comparable to the exact density ($15.91\times$), yet the best learned initializers deliver $1.00\times$ and $0.97\times$. Residual NextHAM improves on H_0 by $1.40\times$ natively without any median acceleration, since H_0 already sits close to H^* (Section 4.1). Proximity to the eigensolve therefore does not predict whether approximation error lands inside the solver's basin of rapid convergence.

4.3 Level III: Downstream Property Recovery

In the solver-native block of Table 4, non-pre-registered ChargE3Net-ΔV_{eff} improves on ρ for spectra, DOS, gaps, and Mulliken populations (all paired intervals exclude zero). ρ retains the lower energy error (0.822 versus 2.91 meV/atom), aided by its extra density update. The pre-registered direct model is Appendix Table 11.

Table 5: **Matrices decoded from each learned input after one ABACUS step,** with the cS gauge removed, hence different from Table 2. Medians over the five-route common test set; lower is better; bold: best. "vs H/D": fraction of crystals on which the decoded matrix beats the corresponding NextHAM/HamGNN input route after its one-step decode; dashes: a route against itself.

			Decoded H		Decoded D	
Input	Model	ρ NMAE (%)	MAE (meV)	vs H	MAE (10^{-4})	vs D
ρ	ChargE3Net-ρ	0.330	0.290	97.6%	0.289	99.8%
V_{eff}	ChargE3Net-ΔV_{eff}	**0.255**	**0.247**	**98.0%**	**0.243**	**99.9%**
H	NextHAM-ΔH	2.60	1.870	—	2.02	74.3%
D	HamGNN-Matrix-ΔD	3.79	2.933	32.4%	3.15	—

With matched electronic work, non-pre-registered ChargE3Net-ΔV_{eff} lowers all six errors relative to ρ (all bands 0.0451 versus 0.0739 eV; energy 4.23 versus 8.88 meV/atom; Mulliken 0.020 versus 0.032 e). Paired intervals exclude zero, although the gap advantage is small (0.044 versus 0.049 eV). H decodes fastest, but its energy and population errors remain high (2,745 meV/atom, 0.569 e). D requires the longer $D \to \rho \to V_{\text{eff}} \to H$ decode and retains large spectral errors.

On the common one-step set, decoded Hamiltonians from learned ρ and ChargE3Net-ΔV_{eff} err by 0.290 and 0.247 meV, against 1.870 and 2.933 meV from the dedicated H and D inputs (Table 5). Field routes beat the corresponding matrix-input routes on 97.6%–99.9% of crystals, favoring matrix construction over direct matrix input. Forces remain a limit of non-stationary one-step states: the original direct ChargE3Net-V_{eff} control has a 5.73 eV/Å median force error, against 0.57 eV/Å for ρ (Appendix Table 16); neither residual potential model was force-audited.

4.4 Interpretation and Task Boundaries

Two mechanisms displace graph position as the ordering principle. The first is where the output constraints of Section 2.1 are discharged: projecting a field through Eq. 2 imposes the operator structure and the subsequent eigensolve imposes the solver's filling rule, so a route that constructs its matrix inherits both analytically while a route that regresses one must recover them from data. The occupation audit measures the residue: the archived reference $D(R)$ is 3.3% out of range from real-space truncation, the learned matrix 15.0%, and the latter exceeds the former on every evaluated test crystal, so a small entrywise loss does not buy an admissible occupied subspace. Greater capacity may close that gap empirically; it cannot remove the constraint from the output space. Appendix Table 9 makes the asymmetry explicit: a field has a composition-independent layout and at most one global constraint, whereas a matrix has species-dependent blocks and global constraints that entrywise regression does not enforce.

The second is conditioning, which decouples spectral from state accuracy. First-order eigenvalue shifts average the field error against $|\psi_i|^2$ and are forgiving, whereas eigenvector corrections carry inverse level spacings and are amplified near degeneracies, so a route can be right in the spectrum and wrong in the occupied state, as direct H is. V_{eff} escapes the long-wavelength amplification of the $\rho \to V_{\text{eff}}$ map, which is what buys its spectra, but it supplies only the local term of Eq. 2 and never a self-consistent density. Neither mechanism is read off the number of arrows in a computational graph (Appendix Figure 4).

5 Conclusion

We scored ρ, V_{eff}, H, and D on 43,704 crystals under one three-level protocol. Proximity to the eigensolution did not predict utility: H and D expose as much oracle headroom as the fields and realize the least of it, while a learned potential accelerates the solver and rebuilds those matrices better than the matrix pipelines. What separates these pipelines is where the constraints of the Kohn–Sham map are discharged—constructing a matrix inherits the operator structure and filling rule, regressing one does not. This study suggests a target the solver can reconstruct, so that constraints are inherited rather than learned, and score it in the loop—native error did not survive deployment here. Capacity-matched comparison, magnetic and non-PBE systems, and cross-code transfer remain open (Appendix A).

AI USE STATEMENT

Generative AI tools were used only for manuscript revision and polishing and for code debugging. The authors reviewed the assisted text and code and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Every number in this paper is produced by a recorded pipeline rather than a one-off run. The frozen split, the intersection accounting, and the exact per-metric denominators are given in Appendix C.5 and Table 7; the complete model and optimization configuration, including the seed, the arithmetic precision, and the SHA-256 of the reported checkpoint, is in Table 17. The target normalization is defined in closed form in Eq. 3, with its train-only scale released at full precision, so the learning target can be reconstructed exactly. Capacity selection precedes any test-set inference and is reported with paired intervals in Appendix C.8. Closed-loop runs are audited per material with input-cube checksums, solver initialization-branch verification, and immutable terminal records; nonconvergence is retained rather than dropped. The anonymized repository at `https://anonymous.4open.science/r/Veff_bench-0D1E` releases the model and preprocessing code, the reported checkpoint and checksum, the frozen full-dataset ID split, six model-ready examples with an ABACUS cube template, portable training, inference, export, and SCF-gate scripts, environment specifications, and compact machine-readable validation, SCF, and downstream aggregates. The full 43k-grid corpus, per-material solver directories, site-specific executables, pseudopotentials, and orbital files are too large or not licensed for redistribution; the released scripts and provenance records document their reconstruction. Reproducing the closed-loop audit additionally requires an ABACUS build supporting V_{eff} initialization and the specified SG15 pseudopotential and double-ζ plus polarization orbital files.

ETHICS STATEMENT

This work computes and learns electronic-structure quantities for inorganic crystals drawn from a public database, and involves no human subjects, personal data, or human annotation. The intended use is to reduce the cost of first-principles screening, which lowers the energy footprint of a workload that is currently a large consumer of supercomputing time; the corresponding risk is that a learned initializer is silently trusted where it has not been validated. We mitigate this in the design of the method rather than only in its description: injection changes the initial condition of an unmodified PBE calculation, so a converged run still reaches the same DFT fixed point and the predictions do not alter the reported physics. The non-self-consistent read-out path does not carry that guarantee, and we therefore report its failure modes explicitly, including one formal SCF nonconvergence, the six reproducible charge-path decode failures, the median total-energy error of 4.23 meV/atom in that read-out, and the one-step force errors that led us to withhold any force claim.

REFERENCES

Pin Chen, Zexin Xu, Qing Mo, Hongjin Zhong, Fengyang Xu, and Yutong Lu. ECD: A machine learning benchmark for predicting enhanced-precision electronic charge density in crystalline inorganic materials. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=SBCMNc3Mq3`.

Pin Chen, Jiang Li, and Yutong Lu. Evaluating predicted densities, hamiltonians, and density matrices as periodic SCF initializers. *arXiv preprint arXiv:2609.15151*, 2026. URL `https://arxiv.org/abs/2609.15151`.

Jonas Elsborg, Luca Thiede, Alán Aspuru-Guzik, Tejs Vegge, and Arghya Bhowmik. Electra: A cartesian network for 3d charge density prediction with floating orbitals. In *Advances in Neural Information Processing Systems*, 2025.

Jonas Elsborg, Felix Ærtebjerg, Luca Thiede, Alán Aspuru-Guzik, Tejs Vegge, and Arghya Bhowmik. Global plane waves from local gaussians: Periodic charge densities in a blink. *arXiv preprint arXiv:2601.19966*, 2026.

Alberto Fabrizio, Andrea Grisafi, Benjamin Meyer, Michele Ceriotti, and Clemence Corminboeuf. Electron density learning of non-covalent systems. *Chemical science*, 10(41):9424–9432, 2019.

Xiaoxun Gong, He Li, Nianlong Zou, Runzhang Xu, Wenhui Duan, and Yong Xu. General framework for E(3)-equivariant neural network representation of density functional theory hamiltonian. *Nature Communications*, 14(1):2848, 2023. doi: 10.1038/s41467-023-38468-8.

Xiaoxun Gong, Steven G. Louie, Wenhui Duan, and Yong Xu. Generalizing deep learning electronic structure calculation to the plane-wave basis. *Nature Computational Science*, 4, 2024. doi: 10.1038/s43588-024-00701-9.

Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M Wilkins, Clemence Corminboeuf, and Michele Ceriotti. Transferable machine-learning model of the electron density. *ACS central science*, 5(1):57–64, 2018.

S. Hazra, U. Patil, and S. Sanvito. Predicting the one-particle density matrix with machine learning. *Journal of Chemical Theory and Computation*, 20(11):4569–4578, 2024. doi: 10.1021/acs.jctc.4c00042.

Yuewen Huang, Pin Chen, and Yutong Lu. UniHam: A large-scale SOC-complete dataset and benchmark for hamiltonian learning in materials. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9486–9497. ACM, 2026. doi: 10.1145/3770855.3817531. URL https://doi.org/10.1145/3770855.3817531.

Peter Bjørn Jørgensen and Arghya Bhowmik. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Computational Materials*, 8(1):183, 2022.

Seongsu Kim, Nayoung Kim, Dongwoo Kim, and Sungsoo Ahn. High-order equivariant flow matching for density functional theory hamiltonian prediction. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.18817.

James Kirkpatrick, Brendan McMorrow, David H P Turban, Alexander L Gaunt, James S Spencer, Alexander G D G Matthews, Annette Obika, Louis Thiry, Meire Fortunato, David Pfau, Lara Roman Castellanos, Stig Petersen, Alexander W R Nelson, Pushmeet Kohli, Paula Mori-Sánchez, Demis Hassabis, and Aron J Cohen. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science*, 374(6573):1385–1389, 2021.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

Teddy Koker, Keegan Quigley, Eric Taw, Kevin Tibbetts, and Lin Li. Higher-order equivariant neural networks for charge density prediction in materials. *npj Computational Materials*, 10(1):161, 2024.

Susi Lehtola. Assessment of initial guesses for self-consistent field calculations. superposition of atomic potentials: Simple yet efficient. *Journal of chemical theory and computation*, 15(3):1593–1604, 2019.

Alan M Lewis, Andrea Grisafi, Michele Ceriotti, and Mariana Rossi. Learning electron densities in the condensed phase. *Journal of chemical theory and computation*, 17(11):7203–7214, 2021.

Chenghan Li, Or Sharir, Shunyue Yuan, and Garnet Kin-Lic Chan. Image super-resolution inspired electron density prediction. *Nature Communications*, 16:4811, 2025a. doi: 10.1038/s41467-025-60095-8.

He Li, Zun Wang, Nianlong Zou, Meng Ye, Runzhang Xu, Xiaoxun Gong, Wenhui Duan, and Yong Xu. Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nature Computational Science*, 2(6):367–377, 2022. doi: 10.1038/s43588-022-00265-6.

Li Li, Stephan Hoyer, Ryan Pederson, Ruoxi Sun, Ekin D Cubuk, Patrick Riley, and Kieron Burke. Kohn-sham equations as regularizer: Building prior knowledge into machine-learned physics. *Physical Review Letters*, 126(3):036401, 2021.

Xiwen Li, Zaizhou Xin, Hongyu Yu, Yang Zhong, Xingao Gong, and Hongjun Xiang. Efficient E(3)-equivariant framework for universal charge density prediction. *arXiv preprint arXiv:2510.00788*, 2025b.

Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Materials*, 6(1):43, 2020.

John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

Martin Schlipf and François Gygi. Optimization algorithm for the generation of ONCV pseudopotentials. *Computer Physics Communications*, 196:36–44, 2015. doi: 10.1016/j.cpc.2015.05.011.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel A. L. Marques. Crystal graph attention networks for the prediction of stable materials. *Science Advances*, 7(49): eabi7948, 2021. doi: 10.1126/sciadv.abi7948.

Xuecheng Shao, Lukas Paetow, Mark E. Tuckerman, and Michele Pavanello. Machine learning electronic structure methods based on the one-electron reduced density matrix. *Nature Communications*, 14:6281, 2023. doi: 10.1038/s41467-023-41953-9.

John C Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. Finding density functionals with machine learning. *Physical Review Letters*, 108(25):253002, 2012.

Shi Yin, Zujian Dai, Xinyang Pan, and Lixin He. Advancing universal deep learning for electronic-structure hamiltonian prediction of materials. In *International Conference on Learning Representations*, 2026. arXiv:2509.19877.

Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Efficient and equivariant graph networks for predicting quantum Hamiltonian. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL `https://proceedings.mlr.press/v202/yu23i.html`.

Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Qh9: A quantum hamiltonian prediction benchmark for qm9 molecules. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang Zhong, Hongyu Yu, Mao Su, Xingao Gong, and Hongjun Xiang. Transferable equivariant graph neural networks for the hamiltonians of molecules and solids. *npj Computational Materials*, 9(1):182, 2023.

Yang Zhong, Hongyu Yu, Jihui Yang, Xingyu Guo, Hongjun Xiang, and Xingao Gong. Universal machine learning Kohn–Sham Hamiltonian for materials. *Chinese Physics Letters*, 41(7):077103, 2024. doi: 10.1088/0256-307X/41/7/077103.

Weiqing Zhou, Daye Zheng, Qianrui Liu, et al. ABACUS: An electronic structure analysis package for the AI era. *The Journal of Chemical Physics*, 163(19):192501, 2025. doi: 10.1063/5.0297563.

A Limitations and Scope

The study covers non-magnetic Alexandria crystals under PBE, SG15 pseudopotentials, a double-ζ plus polarization NAO basis, and one ABACUS mixing protocol; it does not test surfaces, defects, spin polarization, strongly correlated functionals, or alternative mixing schemes. The 147 in-split structures without successful potential exports may introduce selection bias.

The four models are target appropriate rather than capacity matched. Accuracy comparisons therefore rank the evaluated pipelines, not representations in the abstract; a capacity-matched study of all four would be needed for the latter claim. The observed occupation-bound violations describe the evaluated direct D output, while an eigensolve enforces occupations according to the solver's filling rule. Reported loop speedup excludes inference and cube I/O and includes one 100-iteration nonconvergence. Finally, the gauge and independent decoders are tied to the frozen basis, overlap, operators, pseudopotentials, and sampled k points, so they establish recoverability in this interface rather than transfer across codes or electronic-structure settings.

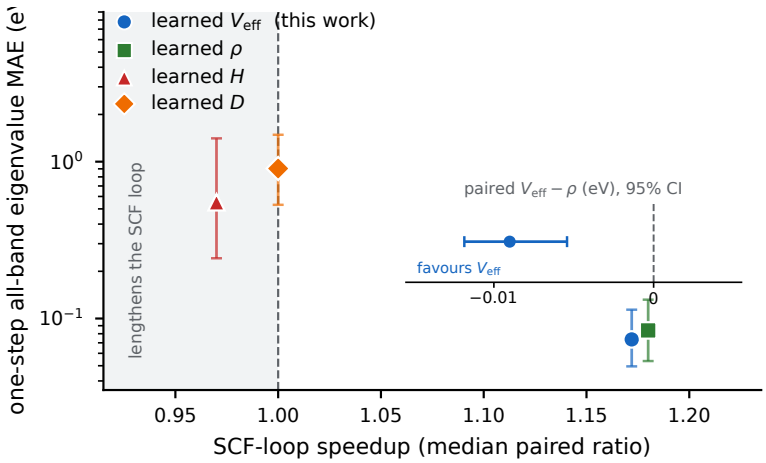


Figure 4: **SCF utility versus spectral accuracy for the original direct-model control.** Median loop speedup is compared with one-step all-band eigenvalue MAE; bars show interquartile ranges, and the inset gives the paired $V_{\text{eff}} - \rho$ speedup difference.

B Additional Cross-Level View

C Reproducibility Details

Table 6 summarizes the three evaluation levels of Section 3.4, which Sections 4.1–4.3 report in turn.

Table 6: **Three-level evaluation protocol.** Each level tests a distinct claim, so native errors are not interpreted as substitutes for solver or property performance.

Level	Task and protocol	Primary measures	Question answered
I: Native prediction accuracy	Reconstruct each converged target in its native support	Gauge-aligned field NMAE; target-specific matrix or density error	Is the learned electronic object accurate on its own representation?
II: Closed-loop SCF acceleration	Inject the state into ABACUS and pair it with SAD and an exact-state oracle	Convergence, fewer-step rate, ΔI, and paired SCF-loop speedup	Does the initializer reach the same DFT fixed point more efficiently?
III: Downstream property recovery	Read properties without self-consistency, through a solver-native one-step decode and a matched eigensolve	Bands, DOS, gap, energy, and Mulliken-population errors	What information is recoverable, and does the ranking survive controlled electronic work?

C.1 Data Accounting

Table 7: Intersection of the frozen ρHD-43K split with successful converged effective-potential exports. Missing records are excluded in place.

Split	Frozen source	Veff usable	Missing Veff
Training	39,851	39,719	132
Validation	2,000	1,992	8
Test	2,000	1,993	7

Attrition on the test split. Failure-inclusive fewer-step rates and V_{eff} convergence use the complete 2,000-structure frozen test split as denominator. For the primary direct V_{eff} pipeline, the split loses cases at four points, of two distinct kinds. Seven structures have no converged V_{eff} reference at all, because the reference DFT export itself did not succeed; these are label-availability failures, and no error can be defined for them. One further structure, `agm005068910`, has a converged reference but no successful full-grid prediction, and one, `agm006233407`, has both but reaches the

100-iteration limit when its predicted potential is injected; these two are method failures. Six more are lost to reproducible charge-path decode failures in the five-way downstream intersection. The ladder therefore forks after the 1,992 structures that have both a converged reference and a full-grid prediction, because the two downstream levels do not impose the same requirement: $2{,}000 \to 1{,}993$ (converged reference available) $\to 1{,}992$ (prediction also available), and from there $\to 1{,}991$ for the strict closed-loop SCF pair, which additionally requires the injected run to converge, and independently $\to 1{,}986$ for the five-way downstream intersection, which does not require SCF convergence but does require every representation to decode. The 100-iteration case therefore leaves the closed-loop set but remains in the downstream set, and $1{,}986 = 1{,}992 - 6$ rather than $1{,}991 - 6$. Every case lost at any of these steps is counted as a failure in the primary pipeline's failure-inclusive rates; medians, which require a per-material error to exist, are necessarily formed over the surviving set, whose size each table caption states. The one structure with a prediction but no reference, `agm003157136`, enters no evaluation. The EdenGNN residual potential uses the same 1,992-ID panel and has 1,989 strict SCF pairs: two runs reach the iteration limit and one fails during MPI initialization. The learned ρ and all D and H convergence percentages in Table 3 follow the runnable populations of 1,999, 1,997, and 1,997 in the cited source; the new ρ oracle is 1,992/2,000 from verified charge-cube restarts. The paired ρ–V_{eff} loop-speedup comparison uses the 1,977 crystals on which both converge. The blocks of Table 2 are scored on 1,999 (ρ), 1,992 (V_{eff}), and 1,997 (D and H) test crystals; the ρ, D, and H counts are those of Chen et al. (2026).

The Veff intersection contains 43,704 unique material IDs and has zero split leakage. Sixty-two successful potential calculations that are absent from the frozen source split remain excluded. The closed-loop set is the exact 1,992-ID intersection between full-grid predictions and remote converged-potential references; its two one-sided set differences—one local prediction without a remote reference and one remote reference without a local prediction—are recorded in the workflow manifest. The exported cubes hold 2,358,826,488 scalar values in total.

The original direct-model property-completion manifest contains 7,968 rows (four representations for each of 1,992 crystals). Scientific analysis succeeds for 7,962 rows: V_{eff}, D, and H complete 1,992/1,992 cases each, while the charge path completes 1,986/1,992. Three retries on different nodes reproduce the six charge failures, so they remain in the success denominator. The separate immutable SAD one-step audit completes 1,992/1,992 cases; only its derived per-material metrics are joined to the four-representation manifest. The machine-readable sources for Table 4 and Appendix Tables 13 and 5 record medians, quartiles, paired intervals, and exact per-metric denominators; raw retry subsets are retained rather than overwritten. The independent decoders additionally contain 1,992/1,992 learned-density spectra and 1,992/1,992 learned-potential spectra, each with 1,952 finite ± 2 eV comparisons and 936 systems in the reference-gap-above-0.01-eV subset. Their separate 24-system pre-registered oracle controls pass before the full expansions; the potential path also passes its constant-field overlap control. The independent Energy audit is complete for all three representations on 1,992/1,992 systems; all fixed-N occupation and eigenspectrum-recompute checks pass to numerical precision. The separate EdenGNN-ΔV_{eff} independent audit also covers 1,992/1,992 systems for spectra, DOS, Energy, and Mulliken; its matched paired comparisons are recorded with the frozen manifest. A separate solver-native audit for each residual model also succeeds on all 1,992 systems; the updated ChargE3Net residual supplies the one-step row of Table 4. Strict-common summaries, paired intervals, and checkpoint hashes are retained for both residual models.

C.2 Data Format and Target Constraints

Table 8 gives the structure of the released files, and Table 9 the physical constraints each target carries.

Table 8: **Raw file structure of the four targets**, with counts for one test crystal (6 atoms, $45 \times 45 \times 108$ grid). Cube files follow the Gaussian cube layout; matrix files are NumPy archives holding one dense block per atom pair and lattice vector $\mathbf{R}$. Learned H and D enter the solver as ABACUS CSR text (per-$\mathbf{R}$ values, column indices, and row pointers).

Target	File	Header / metadata	Data layout	Values	Units
ρ	`chg.cube`; binary `*.restart`	Comment line; spin count and Fermi energy; atom count and origin; grid sizes N_1, N_2, N_3 with voxel vectors; one line per atom (Z, valence, position)	$N_1 N_2 N_3$ values, z fastest, then y and x (218,700 here); the restart stores the same density as full-precision plane-wave coefficients within the 480 Ry cutoff	Float, strictly positive (2×10^{-3} to 1.09 here); integrates to the valence electron count (64 here)	e/Bohr^3
V_{eff}	`pot.cube`	As for ρ, without the Fermi energy	Same grid and ordering; defined up to a constant, whose cell mean is removed for training (Eq. 3)	Float, signed (-19.6 to $+0.88$ here, 39% of voxels negative); deep wells at the nuclei	Ry
H	`output_HR0.npz`	`lattice_vectors` (3×3, Bohr); `atom_info` (species, Z, fractional position per atom); `orbital_info_`t (zeta, l, m per orbital of species t)	One dense block `mat_`i`_`j`_`R_x`_`R_y`_`R_z of shape $n_i \times n_j$ per atom pair and $\mathbf{R}$; $n = 13$, 18, and 27 for the three species here (1,440 blocks)	Float64, signed (-2.18 to 3.21 here); 79% of stored entries below 10^{-4} in magnitude	Ry
D	`output_DM0.npz`	As for H	Same block layout as H	Float64, signed; on-site diagonal in $[0, 2]$ (spin-unpolarized occupations)	–
S	`output_SR.npz` (auxiliary)	As for H	Same block layout (674 blocks here); needed to decode H and D	Float64; on-site diagonal exactly 1	–

Table 9: **Physical constraints carried by each target.** The middle column lists what an E(3)-equivariant regressor satisfies by construction; the right column must be learned or enforced separately. A field has a composition-independent scalar layout; a matrix has species-dependent block shapes and global constraints that entrywise losses do not see.

Target	Output shape	Satisfied by construction	Learned or enforced separately
ρ	One scalar per grid point	Periodicity; rotation invariance	$\rho \geq 0$; $\int_\Omega \rho = N_e$
V_{eff}	One scalar per grid point	Periodicity; rotation invariance; additive constant removed by the mean-zero target	–
H	$n_i \times n_j$ block per atom pair and $\mathbf{R}$, n set by species	Blockwise equivariance ($l_i \otimes l_j$)	Hermiticity $H_{ij}(\mathbf{R}) = H_{ji}(-\mathbf{R})^\top$ unless outputs are symmetrized; gauge $H \to H + cS$; sparsity pattern
D	As for H	Blockwise equivariance ($l_i \otimes l_j$)	Hermiticity; electron count $\sum_{\mathbf{R}} \mathrm{Tr}[D(\mathbf{R})S(-\mathbf{R})] = N_e$; idempotency $DSD = 2D$ at zero smearing; occupations in $[0, 2]$

C.3 Size of the Self-Consistent Hamiltonian Correction

For all 1,999 test crystals with a converged Hamiltonian H^*, we compare H^*, the SAD-started H_0, and each learned H on the nonzero support of H^*, the support used by Table 2, in the Ry units of the ABACUS export without rescaling. The audit reproduces the table: crystal-mean MAEs are 2.27×10^{-4} (H_0), 2.46×10^{-3} (HamGNN-Matrix-H), 2.62×10^{-3} (DeepH-E3-H), and 1.54×10^{-4} Ry (NextHAM-ΔH). Medians over crystals follow. The self-consistent correction is small: $\|H^* - H_0\|_F / \|H^*\|_F = 5.3\%$ (5th–95th percentile 2.3–10.7%). Supported entries average 2.9×10^{-3} Ry in magnitude, so the direct models, at 2.4 and 2.6×10^{-3} Ry, err by about one typical entry, whereas H_0 errs by 2.0×10^{-4} Ry. On-site blocks (same atom, $\mathbf{R} = 0$) hold 0.4% of the supported entries, and 97% of the direct models' absolute error lies in the inter-atomic blocks, where their MAE of 2.3–2.5×10^{-3} Ry is $14\times$ that of H_0 (1.7×10^{-4} Ry); on-site the gap is about $3\times$ (1.6×10^{-2} against 5.7×10^{-3} Ry). The kinetic, nonlocal, and local-pseudopotential parts of the inter-atomic blocks do not depend on the density and are therefore exact in H_0, which is where the direct regressors lose most. NextHAM-ΔH lowers both parts, to 1.3×10^{-4} Ry inter-atomic and 2.4×10^{-3} Ry on-site.

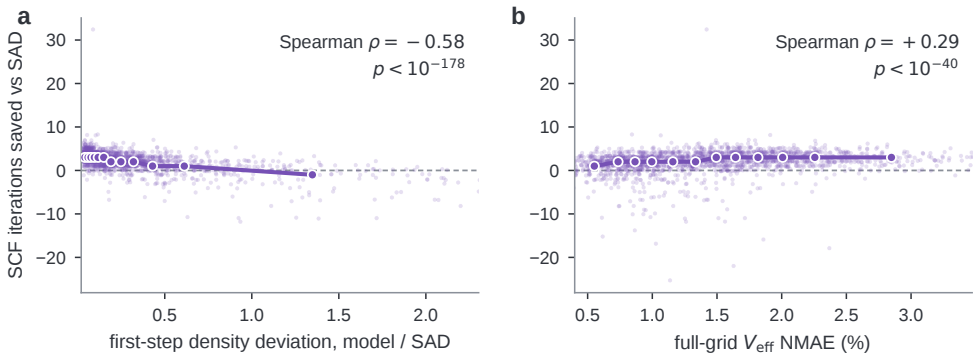


Figure 5: **What predicts closed-loop acceleration.** Iterations saved against SAD for the 1,991 strict pairs, jittered vertically, with binned medians. (a) The first-step density deviation of the injected field relative to SAD predicts the saving. (b) Full-grid V_{eff} NMAE does not: its weak positive association reflects material difficulty, since larger and chemically heavier cells are both harder to predict and slower to converge from SAD.

C.4 Closed-Loop Timing Audit

The primary model yields 1,991 strict pairs in Table 3. Of the 2,000 test crystals, 1,742 (87.1%) are faster by SCF-loop time, and the median loop speedup of $1.172\times$—formed over the strict pairs, the only cases in which a paired ratio exists—has a paired-bootstrap 95% interval of $[1.166, 1.179]$. Median neural inference is 13.23 s per crystal on this allocation. Including inference but excluding cube serialization lowers the median speedup to $1.111\times$ ($[1.104, 1.117]$) with 1,547 (77.4%) time-faster crystals; the ratio of summed reference time to summed inference-plus-SCF time is $1.047\times$. Against the strict-pair denominator of 1,991 the same three rates read 87.5%, 77.7%, and 84.7%, so no conclusion turns on the choice. The sole learned failure, `agm006233407`, reaches the 100-iteration limit and remains in the denominator; `agm005068910` has a converged reference but no successful prediction and is likewise counted as a failure.

Table 10 gives the closed-loop effect of capacity on the identical 1,991 pairs, with counts also expressed on the common denominator. The superseded multiplicity-128 checkpoint reaches a $1.111\times$ loop speedup, against $1.172\times$ for the primary model on the same pairs; paired by material over the 1,990 crystals on which both converge, the primary model uses fewer iterations on 848, ties on 860, and uses more on 282, a mean saving of 0.465 iteration. Figure 5 separates this from pointwise accuracy across materials.

The four-layer variant reproduces the primary result at $1.172\times$ ($[1.165, 1.178]$) under a second receptive-field setting; because its interval overlaps that of the pre-registered row, Table 3 retains the pre-registered model rather than replacing the primary checkpoint across all three levels after seeing test-set results.

Table 10: Closed-loop effect of V_{eff} model capacity, evaluated on the identical 1,991 strict pairs. Fewer-step counts are expressed as a fraction of the common 2,000-structure test split. The multiplicity-256 row is the pre-registered primary model reported in Table 3. Loop speedup excludes model inference; the last column includes it.

V_{eff} variant	Fewer steps, n (%)	SCF loop spd.	Infer.-incl. spd.
Multiplicity 128 (superseded)	1592 (79.6)	$1.111\times$	—
Multiplicity 256 (primary)	**1686 (84.3)**	$\mathbf{1.172}\times$	$1.111\times$
Multiplicity 256, 4 layers	1683 (84.2)	$1.172\times$	$1.094\times$

Before the full run we tested three frozen IDs end to end. The ABACUS logs confirm both the `VEFF` initialization branch and the exact cube path. The learned potentials converge in 12, 13, and 15 electronic iterations. These cases validate field units, ordering, serialization, and solver entry.

They are a smoke gate, not an estimate of test-set convergence or acceleration, and are excluded from every aggregate statistic in the paper.

C.5 Level III Protocol, Denominators, and Cost Accounting

Protocol definitions. Each solver-native route runs one electronic decode: the SAD control starts from ABACUS atomic charge with no external learned initialization, and predicted ρ, D, and H use `scf_nmax=1`, while predicted ChargE3Net-ΔV_{eff} uses a fixed-potential non-self-consistent diagonalization. The matched eigensolve of Table 4, called the independent decoder below, ends in one generalized Kohn–Sham eigensolve and never enters ABACUS for any reported quantity, populations included. Fully converged calculations are omitted from Table 4: their property errors vanish by construction, and their convergence and timing belong to Level II (Table 3).

Denominators. The strict-common learned solver-native comparison uses $n = 1{,}986$ for all-band, ± 5 eV, DOS, Energy, and Mulliken errors, $n = 1{,}946$ for the ± 2 eV spectrum, and $n = 934$ for gaps above 0.01 eV. The independent audit uses $n = 1{,}992$, 1,952, 1,992, and 936 for all-band, ± 2 eV, ± 5 eV, and gap errors, respectively; Energy and Mulliken use $n = 1{,}992$ for each matched field route. The six charge-path failures affect only the solver-native block. The fully converged learned-initialization errors are measured on 1,991 converged pairs (1,951 for ± 2 eV); Mulliken output was not requested in those production runs.

Table 11: **Original direct ChargE3Net-V_{eff} controls.** Median absolute errors for the pre-registered potential model. The native one-step row uses $n = 1{,}986$ (1,946 for ± 2 eV; 934 gapped); the matched eigensolve uses $n = 1{,}992$ (1,952 for ± 2 eV; 936 gapped).

Decode	Cost (s)	All bands (eV)	Spec. ± 2 (eV)	DOS L_1	Gap (eV)	Energy (meV/atom)	Mulliken (e)
Solver-native one-step	30.7	0.0736	0.0324	0.101	0.066	8.27	0.028
Matched eigensolve	19.6	0.0641	0.0319	0.082	0.066	8.27	0.028

Provenance of the independent Energy column. The ρ and H values (8.878 and 2,745 meV/atom) come from the solver-free energy audit, whose predictions for those two representations are fixed inputs reproduced from Chen et al. (2026) and are therefore unaffected by any change to the V_{eff} model. The EdenGNN-ΔV_{eff} matched-eigensolve value in Table 4 is 4.229 meV/atom (checkpoint `6608bb77`); the ChargE3Net residual one-step value is 2.906 meV/atom (`864eebec`). The earlier ChargE3Net-V_{eff} value (8.270 meV/atom) in Appendix Table 11 is computed from the pre-registered primary checkpoint of Table 17, SHA-256 prefix `49898478`, through the `capacity_20260903` projection. That same audit tree also retains an earlier V_{eff} energy of 15.027 meV/atom obtained from a superseded checkpoint (`a090b84a`, the pre-capacity-study model) under the `learned_npy` projection; it is kept for the audit trail and is *not* the value reported here. Both projections record their checkpoint and prediction root in the manifest summaries we release, so the two can be told apart without reference to this note. The added D value (1,193 meV/atom) comes from the complete 1,992-case HamGNN-Matrix-ΔD independent-decoder audit, not either field run.

Independent reconstruction. The density path independently reconstructs V_{eff} with SG15, FFT Hartree, and PBE components before the same AO projection and eigensolve; it consumes no reference-potential values. The D path contracts the corrected $D_0 + \widehat{\Delta D}$ to ρ, reconstructs PBE V_{eff}, projects an AO Hamiltonian, and then diagonalizes it; its unoccupied bands come from this reconstructed H. The independent V_{eff} Mulliken population is reconstructed as $\sum_R \mathrm{diag}[D(R)S(-R)]$ and grouped by atom. For the independent Energy audit, fixed-electron occupations define D, which is contracted to ρ before a common PBE candidate-state functional is evaluated; the converged reference is passed through the same reconstruction. This is an independent functional audit, not a band-energy error or a strict variational bound.

Independent-decoder results in detail. EdenGNN-ΔV_{eff} has median all-band, ± 2 eV, ± 5 eV, and sampled-gap errors of 0.0451, 0.0224, 0.0235, and 0.0441 eV, respectively; its DOS error is 0.0584, Energy error 4.229 meV/atom, and Mulliken error 0.0201 e. Paired against ρ, all six

Table 12: **Original direct ChargE3Net-V_{eff} one-step state completion.** Median errors on the same 1,986-crystal common set as Table 5. "vs H/D" compares decoded matrices with the dedicated matrix-input routes under the one-step protocol.

Model	ρ NMAE (%)	H MAE (meV)	vs H	D MAE (10^{-4})	vs D
ChargE3Net-V_{eff}	0.358	0.338	96.6%	0.341	99.7%

Table 13: **Paired uncertainty for the solver-native ChargE3Net-ΔV_{eff} versus ChargE3Net-ρ routes.** Δ is the median paired error difference $V_{\text{eff}} - \rho$; negative values favor V_{eff}. Intervals are 95% paired-bootstrap intervals over materials (10,000 resamples). The density route receives an extra density update, so these are deployment outcomes rather than intrinsic representation rankings.

Metric	n	Median Δ	95% interval	Resolved winner
All bands (eV)	1,986	-0.03711	$[-0.04062, -0.03443]$	V_{eff}
Spectrum ± 2 eV (eV)	1,946	-0.00989	$[-0.01140, -0.00877]$	V_{eff}
DOS L_1	1,986	-0.03510	$[-0.03896, -0.03017]$	V_{eff}
Gap (eV)	934	-0.00684	$[-0.00981, -0.00459]$	V_{eff}
Energy (meV/atom)	1,986	$+1.293$	$[+1.196, +1.436]$	ρ
Mulliken (e)	1,986	-0.01349	$[-0.01461, -0.01236]$	V_{eff}

matched-eigensolve property differences favor EdenGNN with 10,000-resample 95% intervals excluding zero. The gap advantage is marginal: the paired median improvement is 0.00339 eV, interval $[0.00015, 0.00769]$ eV, and EdenGNN wins on 499/936 gapped crystals.

For the original direct ChargE3Net-V_{eff} control of Appendix Table 11, median all-band, ± 2 eV, ± 5 eV, and sampled-gap errors are 0.0739, 0.0323, 0.0355, and 0.0494 eV for ρ; 0.0641, 0.0319, 0.0341, and 0.0664 eV for ChargE3Net-V_{eff}; and 0.253, 0.192, 0.189, and 0.319 eV for direct learned H. The corresponding DOS errors are 0.0880, 0.0817, and 0.2502. The reconstructed D route yields 0.705, 0.340, 0.364, and 0.602 eV, respectively, with DOS error 0.5294. The population comparison is more separated in the tail than at the median. Independent-decoder Mulliken errors have quartiles 0.0169/0.0282/0.0480 e for direct ChargE3Net-V_{eff} against 0.0177/0.0319/0.0595 e for ρ, but the 95th percentiles are 0.0945 against 0.1458 e and the worst cases 0.292 against 1.041 e: the charge path misassigns a full electron on its hardest crystal while the potential path never exceeds 0.3 e. A workflow that must trust a learned state without checking it is governed by that tail rather than by the median. The independently reconstructed ChargE3Net-V_{eff}, ρ, and H Mulliken errors are 0.0282, 0.0319, and 0.5689 e, all three formed from the candidate's own fixed-electron density matrix without entering ABACUS. On the paired comparison the median V_{eff}-minus-ρ difference is -0.00318 e with a 95% interval of $[-0.00425, -0.00201]$ and V_{eff} lower for 56.3% of the 1,992 materials, so the direct-model population advantage is resolved here and in Appendix Table 14; the independent ρ median (0.03185 e) differs from its solver-native one-step value (0.03189 e) by 4×10^{-5} e, confirming that Mulliken populations are insensitive to which of the two decoding protocols produced them. The median absolute Energy errors are 8.878, 8.270, and 2,745 meV/atom for ρ, direct ChargE3Net-V_{eff}, and H, so the low spectral error of direct H does not imply an accurate occupied density or total-energy functional. The V_{eff} value is obtained without entering ABACUS and on a different sample ($n = 1{,}992$ against $n = 1{,}986$), yet agrees with the solver-native one-step median of 8.269 meV/atom in Appendix Table 11 to within 0.001 meV/atom, which is an independent check on the offline functional evaluation rather than a coincidence. In paired spectral tests against direct H, V_{eff} is better for 91.0%, 92.7%, 95.2%, and 85.0% of the respective valid sets, and the median direct-H minus V_{eff} difference is positive for all four spectral metrics with bootstrap 95% intervals excluding zero. Both independent decoders pass pre-registered 24-system oracle controls before the full audits: maximum all-band errors are 0.1625 meV for the density decoder and 0.122 meV for the potential projector, against a 1 meV gate.

Paired versus marginal ratios. For the original direct ChargE3Net field comparison, the medians of paired per-material V_{eff}-to-ρ ratios are 1.01 for the ± 2 eV spectrum and 0.946 for DOS, and only the DOS interval excludes zero. The two summaries agree in direction; the paired form is the stricter one and is what we use when calling a difference resolved. For the original direct ChargE3Net-V_{eff} control, decoded D and H beat their dedicated one-step matrix-input routes on 99.7% and

Table 14: **Paired uncertainty for the original direct ChargE3Net solver-native field pipelines.** Δ is the median paired error difference $V_{\text{eff}} - \rho$; negative values favor V_{eff}. Intervals are 95% paired-bootstrap intervals over materials (10,000 resamples), so they quantify test-set sampling uncertainty. Because the two paths perform different electronic updates, these intervals compare deployment outcomes rather than intrinsic representation quality.

Metric	n	Median Δ	95% interval	Resolved winner
All bands (eV)	1,986	-0.00901	$[-0.01185, -0.00542]$	V_{eff}
Spectrum ± 2 eV (eV)	1,946	$+0.00045$	$[-0.00083, +0.00172]$	Neither
Spectrum ± 5 eV (eV)	1,986	-0.00117	$[-0.00291, +0.00050]$	Neither
DOS L_1	1,986	-0.00439	$[-0.00813, -0.00136]$	V_{eff}
Gap (eV)	934	$+0.01025$	$[+0.00493, +0.01492]$	ρ
Energy (meV/atom)	1,986	$+6.153$	$[+5.775, +6.511]$	ρ
Mulliken (e)	1,986	-0.00326	$[-0.00431, -0.00207]$	V_{eff}

96.6% of paired systems (Appendix Table 12); its decoded ρ beats the charge route on 44.7%. The corresponding EdenGNN-ΔV_{eff} rates are 99.9%, 98.0%, and 66.4% on the same set (Table 5).

Cost accounting. The cost column reports the median time per crystal to turn a learned state into a solved Hamiltonian. Neural inference is excluded; it adds a median 13.2 s for the direct ChargE3Net-V_{eff} model (Appendix C.4). Property output is also excluded, since it is the same scripted post-processing for every route. The reference and solver-native blocks are ABACUS CPU time. Every route in them performs one electronic step, one Hamiltonian build and one diagonalization sweep over the k points. The SAD and D runs execute exactly that, in 37.2 and 41.6 s at the median, of which 8.0 and 9.8 s is the ABACUS output stage; the table reports the remaining 27.7 and 30.7 s, the latter for every learned route. The recorded V_{eff} and H runs repeat the diagonalization sweep once and read their inputs from text files, for 66.1 and 67.3 s in total; these overheads belong to the injection interface rather than to the representation. The ρ runs write complete outputs but exit before printing the solver's timing report. The separate ChargE3Net residual fixed-potential run records a 70 s median duration including interface and output overhead, not a new normalized cost estimate. The matched block is GPU decoding time, with the same eigensolve batch and projection chunk sizes. EdenGNN-ΔV_{eff} was audited in a separate GPU run; its observed 49.4 s median is not a controlled timing comparison. All figures below concern the original direct ChargE3Net field runs and are stage-wise medians. Reconstructing V_{eff} from ρ takes 2.3 s, AO projection 15.3 s, and the generalized eigensolve 3.9 s; the last two are shared by the field routes, with paired V_{eff}-to-ρ ratios of 1.00. The D route adds 30.1 s for DM-to-ρ contraction, and direct H needs only the eigensolve. The direct ChargE3Net potential route was cheaper than the density route, at $0.89\times$ by paired per-material ratio (95% CI $[0.875, 0.905]$), because it skips only the reconstruction step. The independent path returns an electronic state at the reported decoder cost; its offline Energy audit adds occupied-D construction, $D \to \rho$ contraction, and functional evaluation, and this property-specific audit time is excluded from the common decoder-cost column and recorded in the machine-readable summary.

C.6 Decoded-State and Force-Protocol Controls

For the independent Energy audit, a generalized eigensolve is followed by fixed-electron Gaussian occupations ($\sigma = 0.001$ Ry), inverse Fourier transformation of the occupied density matrix, and contraction in the same numerical atomic-orbital basis. We then evaluate the symmetric difference

$$\Delta E = \Delta\left\{ \mathrm{Tr}[DH_0] + \int \rho(\mathbf{r}) v_{\text{loc}}(\mathbf{r})\, d\mathbf{r} + E_{\text{H}}[\rho] + E_{\text{xc}}^{\text{PBE}}[\rho] - TS \right\}, \tag{7}$$

where H_0 contains the frozen kinetic and nonlocal operators. The ionic Ewald term cancels, and the mean-zero local-potential gauge cancels at fixed electron count. Passing the reference D through the same real-space decoder avoids assigning its discretization error only to the candidate. Across all 1,992 systems and the three originally audited ρ, V_{eff}, and H representations, the maximum electron-count and eigenvalue-recompute errors are 2.01×10^{-11} electrons and 2.25×10^{-10} eV, respectively; both maxima are set by the H path, the V_{eff} path staying below 1.82×10^{-12} electrons and 4.48×10^{-12} eV. The median absolute closure error against the original ABACUS density is 1.58 meV/atom. Consequently, Table 4 reports absolute candidate–reference functional differences

as an audit, not as a strict variational upper bound. Table 15 lists the component-level and oracle-spectrum controls that the density decoder passes before any learned input is evaluated.

Table 15: **Independent density-decoder controls.** The one-system G1 checks compare independently reconstructed potential components with synchronous ABACUS outputs. G2 evaluates exact-density spectra on the pre-registered 24-system orbital-coverage set. Learned-density errors are reported as outcomes and are not pass/fail criteria.

Role	Metric	Gate	Result
G1 component	$v_{\mathrm{loc}} + v_H$ max error (Ry)	$< 10^{-4}$	1.72×10^{-5}
G1 total	exact-ρ V_{eff} max error (Ry)	$< 10^{-4}$	1.77×10^{-5}
G2 oracle ($n = 24$)	all-band MAE, median / max (meV)	max < 1	$0.0416/0.1625$
Learned ρ ($n = 24$)	all-band / ± 2 / ± 5 MAE (eV)	—	$0.09195/0.02807/0.04288$
Learned ρ ($n = 24$)	sampled-gap absolute error (eV)	—	0.00524

Table 16: One-system force-protocol audit on `agm001003669`. Values are force-component MAE in eV/Å against a converged reference. Exact-input controls verify each decoder, whereas learned inputs expose the non-stationary one-step force. This feasibility control is not a dataset-level benchmark and is not included in Table 4.

Input	Exact decoder floor	Learned input	Run status
ρ	0.000003	0.5654	post-output crash
D	0.000017	5.6666	completed
H	0.000016	3.2995	completed
V_{eff}	0.000090	5.7266	completed

Table 16 is intentionally a feasibility gate. A defensible force benchmark would instead report force error as a function of electronic convergence—for example after $k = 1, 2, 4, \ldots$ SCF steps—paired with the added solver cost. No full-scale force claim is made from the present single-system control.

C.7 Model and Training Configuration

Table 17: Pre-registered primary V_{eff} model and optimization settings. This is the checkpoint used for the primary direct V_{eff} rows of Tables 2, 3, and 4; the superseded mul-128 configuration is reported in Table 18.

Setting	Value
Interaction layers / neighbors	3 / 48
Multiplicity / $\ell_{\max}$	256 / 3
Cutoff / radial basis	4 Å / 24 Gaussian
Optimizer / learning rate	Adam / 5×10^{-4}
Random seed / arithmetic	20260831 / FP32
Training probes / validation probes	512 / 1,024 per structure
GPU processes / batch per process	4 / 4
Early stopping	patience 8, minimum 10^{-5} Ry
Target scale	train-median per-structure p95 $\lvert V - \langle V \rangle \rvert$
Best validation MAE	0.0142273 Ry (step 198,640)
Final step / status	200,000 / step budget reached
Checkpoint SHA-256 (first 8)	`49898478`

Training records every optimizer step and validation epoch, with atomic best and latest checkpoints. Full-grid evaluation uses the recorded best checkpoint and stores its SHA-256 hash. The Veff-specific objective applies gauge alignment before SmoothL1 and evaluates validation MAE/RMSE/NMAE after restoring Rydberg units.

C.8 Validation-Set Capacity Selection

Because the four representations are not given a shared parameter budget, the V_{eff} capacity must be shown to be selected without reference to test data. We compare full-grid variants on a frozen

256-structure panel drawn from the validation partition; the panel was sampled without inspecting labels and every comparison below precedes any test-set inference.

Table 18: Capacity selection on the frozen 256-structure validation panel. Errors are gauge aligned. Deltas are paired mean differences against the previous best checkpoint with 95% bootstrap intervals.

Variant	NMAE (%)	MAE (Ry)	Paired ΔNMAE vs. previous best
Multiplicity 128 (superseded)	1.8356	0.017030	—
Multiplicity 256 (primary)	**1.5458**	**0.014200**	-14.45%, $[-0.297, -0.227]$, 236/256 wins
Multiplicity 500	1.5456	0.014225	-0.02%, $[-0.056, +0.041]$, 144/256 wins

Doubling the multiplicity from 128 to 256 reduces validation NMAE by 14.45% with a paired interval that excludes zero. Doubling it again to 500 produces a mean change of -0.02% with an interval that crosses zero and a win rate indistinguishable from chance, and it is worse on the low-$|G|$ error diagnostics. The primary model is therefore at its capacity ceiling under this architecture and objective: the reported V_{eff} results cannot be improved by further widening, and they are not an artifact of an unbounded parameter search. Table 10 shows the corresponding closed-loop effect, where the multiplicity-128 and multiplicity-256 checkpoints give $1.11\times$ and $1.17\times$ median SCF-loop speedups on the identical 1,991 strict pairs, and fewer-step rates of 79.6% and 84.3% on the common denominator. Improving the checkpoint therefore improves the closed loop, but that between-checkpoint trend does not carry over to the between-material comparison of Figure 5, where pointwise NMAE does not predict which crystals are accelerated.

C.9 Controlled Pilot Ablation

The design pilot uses one hash-selected 256/64/64 split, seed 20260819, 4,000 updates, 512 training probes, and 1,024 validation/test probes per structure. The ordinary-L1 variant changes gauge alignment, robust loss, and tail weighting together and therefore tests the complete objective package rather than any one term. The cutoff comparison keeps the gauge-aware objective fixed. Table 19 reports both comparisons. These sampled-grid results guided configuration selection but are not used as full-dataset performance claims.

Table 19: Controlled pilot results on 64 held-out structures. The reported MAE/RMSE are gauge aligned and use 1,024 probes per structure.

Objective	Cutoff (Å)	MAE (Ry)	RMSE (Ry)	NMAE (%)
Gauge SmoothL1 + tail	6	0.7564	1.4399	74.90
Ordinary L1	6	0.7933	1.5437	77.70
Gauge SmoothL1 + tail	4	**0.4299**	**0.8458**	**43.01**

C.10 Closed-Loop Audit Protocol

Each material-specific Slurm allocation runs the exact-potential control and learned-potential condition sequentially with the same node and MPI resources. The runner verifies the input cube checksum, the `VEFF` initialization branch, the pre-SCF potential diagonalization marker, the exact path reported by ABACUS, and the final convergence marker. Completed successful conditions are immutable and skipped on retry. Formal nonconvergence is not retried; pre-execution infrastructure failures may be resubmitted and remain visible in the audit trail. The completed audit contains 3,984 terminal condition records and no pending calculations. All oracle runs converge; 1,991 of 1,992 learned runs converge, with `agm006233407` retained as formal 100-iteration nonconvergence.