



Future Roadmap

What remains to be implemented,
what the method cannot currently do,
and what is not yet adequately tested

Leandro Seixas

Version 26.8.19

Purpose

This document is deliberately unflattering. The user and technical guides describe what Poraquê does; this one records what it does *not* do, how far off it is where it falls short, and what would have to be built to close each gap.

Every number quoted here was measured on the shipped dataset during the 26.8.19 cycle, and the measurement is named so it can be repeated. Where a limitation is fundamental to the method rather than a defect in the code, that is stated — the distinction decides whether the fix is an afternoon or a research programme.

Note

The single most important fact in this document: **the electrostatic machinery is correct and the difficulty lies elsewhere**. Poraquê's $E_{\alpha Z}$, Ewald and Hartree terms reproduce VASP's PSCENC, TEWEN and DENC to better than 0.01 eV in a difference. The accuracy problems below come from what the pseudo-valence fields *omit*, not from how they are integrated.

Contents

Purpose	1
1 Where the method stands	3
1.1 What is verified	3
1.2 What is not accurate	3
2 What needs to be implemented	5
2.1 Priority 1: the PAW one-centre and non-local terms	5
2.2 Priority 2: charge normalisation as a training constraint	5
2.3 Priority 3: chemical transferability	6
2.4 Priority 4: spin polarisation, validated	6
2.5 Priority 5: the stress tensor	6
3 Known limitations	7
3.1 Fundamental to the method	7
3.2 Implementation limitations	7
4 Areas needing more rigorous testing	8
4.1 Untested because the data does not exist	8
4.2 Tested too narrowly	8
4.3 Not exercised in the ordinary test run	8
4.4 Suggested order of work	9

CHAPTER 1

Where the method stands

1.1 What is verified

The following are checked against exact references and hold to the tolerance quoted. They are not expected to be sources of future error.

Quantity	Reference	Agreement
Ewald energy	sc/bcc/fcc Madelung constants	3×10^{-6} rel.
Ewald forces	finite differences of the energy	4×10^{-9} eV/Å
Electron-ion force	finite differences of $\int \rho V_{\text{ext}}$	10^{-7} eV/Å
$E_{\alpha Z}$	VASP PSCENC	2×10^{-4} eV
Ewald energy	VASP TEWEN (Δ)	0.004 eV
Hartree energy	VASP DENC (Δ)	0.001 eV
PBE exchange	Dirac limit on a uniform gas	machine precision
PW92 correlation	published values	machine precision

1.2 What is not accurate

1.2.1 Energy differences

With VASP's *own* ρ and τ — that is, with the operators removed from the problem entirely — energy differences between gold structures of the same composition carry a mean absolute error of 0.18–0.24 eV per atom. The true differences span 0.125 eV per atom. **The error is larger than the signal.**

Running the full pipeline with the shipped operators, the error rises to 0.87–1.93 eV per atom.

Source	MAE on ΔE	Relative to signal
Exact fields (method ceiling)	0.18–0.24 eV/atom	1.5–2×
Predicted fields (as shipped)	0.87–1.93 eV/atom	7–15×
VASP spread being resolved	0.125 eV/atom	—

The ceiling is set by the neglected PAW one-centre and non-local pseudopotential energies. Those are often described as a per-atom constant, and to a first approximation they are: the offset between Poraquê and VASP is constant to about a part in 10^4 across the dataset. But the constant is ≈ 1154 eV per atom, so a part in 10^4 of it is still 0.1 eV per atom — comparable to everything being predicted.

Caveat

The total energy is a cancellation of terms of order 10^4 eV down to a result of order 10^0 eV. Getting ΔE right to 10 meV/atom therefore demands a relative accuracy of about 10^{-6} in the fields. The operators currently deliver 10^{-2} .

1.2.2 Forces

Forces were implemented this cycle (`poraquê.physics.forces`) and the implementation is exact: each analytic force agrees with a finite difference of the term it differentiates to 10^{-7} eV/Å. The *physics* is incomplete.

Against VASP's TOTAL-FORCE, using VASP's own density, the mean absolute error is 0.83 eV/Å on forces whose mean magnitude is 1.66 eV/Å. The correlation with the reference is +0.39 and -0.01 on the two structures that carry an OUTCAR. The magnitude is right; the direction is not.

Two reasons, in order of size:

1. **The Hellmann–Feynman force is complete only for a local pseudopotential.** A PAW dataset contributes a projector force and a one-centre force, and neither is recoverable from ρ , τ and V_{loc} sampled on a grid. Gold's 5d projectors are strong.
2. **The cancellation amplifies density error.** The electron–ion and ion–ion terms are each ≈ 100 eV/Å and cancel to ≈ 0.5 eV/Å. A relative error ϵ in ρ therefore arrives in the force amplified roughly 200-fold.

1.2.3 Charge normalisation

The shipped `ext2chg` operator predicts densities whose electron count drifts by up to 1.7% — nearly five electrons out of 297. Since every electrostatic term is at least linear in ρ and the Hartree energy is quadratic, this alone moved total energies by tens of eV, and by a different amount for each structure, so it did not cancel in a difference.

This is now repaired at inference: `ChargeDensity.normalized` rescales to the count the pseudopotentials fix, which is exactly known. It halved the end-to-end ΔE error. It is a *repair*, not a cure — a density whose integral is wrong by 1.7% has the wrong shape too, and rescaling does not fix the shape.

CHAPTER 2

What needs to be implemented

2.1 Priority 1: the PAW one-centre and non-local terms

This is the single change that would move Poraquê from qualitative to quantitative, and it gates both energies and forces.

What is missing The one-centre (augmentation-sphere) kinetic, Hartree and exchange–correlation energies, and the non-local projector energy $\sum_{ij} \rho_{ij} D_{ij}$. Together these are ≈ 1154 eV per atom for gold, of which the environment-dependent part is the few eV that decides every energy difference.

Why it is hard They are not functionals of the plane-wave density. They need the PAW occupancies ρ_{ij} , which live inside the augmentation spheres and are written to `CHGCAR` as the augmentation records that `collect_augmentation` currently copies verbatim without interpreting.

The tractable route The augmentation records *are* in the data, and the `POTCAR` parser already reads the radial tables. A third learned map — $\rho \mapsto \{\rho_{ij}\}$, an on-site quantity with a small output dimension per atom — is a far better-posed learning problem than the field maps, because the target is a handful of numbers per atom rather than a field. Combined with an implementation of the one-centre energy expressions, this would close most of the gap.

Effort Substantial. Weeks, and it requires reading the PAW formalism carefully rather than pattern-matching VASP output.

2.2 Priority 2: charge normalisation as a training constraint

Normalising at inference repairs the symptom. The operator should be *unable* to produce a density with the wrong electron count in the first place, by construction rather than by correction.

Implementation

The machinery is half-present: `poraque.ml.physics.electron_count_loss` exists and `training.physics.electron_count_weight` is wired through the config, but it defaults to zero and the shipped model was trained without it. A hard normalisation *head* — rescaling the network’s output to the target count inside the forward pass, so the constraint holds exactly and the gradient is aware of it — is the stronger option and does not exist yet. Compare the two against the same split before choosing.

2.3 Priority 3: chemical transferability

The dataset is 17 gold structures. Nothing in any reported number speaks to whether the operators transfer across chemistry, and the reports say so in their caveat box. Before Poraquê can be claimed to predict densities, it needs a dataset spanning at least a period and a few bonding types.

This also gates the validation that `tests/test_energy_differences.py` is written for but cannot currently run: N_2 against isolated nitrogen, CO , and the cohesive energy of diamond. The tests exist and skip; they need the reference calculations *and* operators that have seen those elements.

2.4 Priority 4: spin polarisation, validated

The `ISPIN=2` code path was implemented this cycle and is exercised only by synthetic fixtures:

- `SpinDensity` reads and writes VASP’s two-block `CHGCAR` and round-trips to full precision;
- `FieldPairDataset` detects two channels from the file rather than from a flag, and refuses a flag that contradicts the data;
- the two channels are normalised separately (`transforms.Channelwise`), because ρ integrates to hundreds of electrons while m integrates to a few μ_B and one shared scale normalises neither;
- the FNO takes and returns two channels, and the counts survive a checkpoint round trip.

What is missing is any spin-polarised reference data at all. Until an `ISPIN=2` dataset exists, none of this is known to be *correct*, only self-consistent.

Note

One physics gap is already known within the implemented path: the exchange–correlation energy in `poraquê.physics.energy` is spin-unpolarised. Feeding it a spin-polarised system through `SpinDensity.as_charge_density` silently evaluates the LDA/PBE functional on the total density, which is wrong for a magnetic system. Implementing the spin-scaling relation for exchange and the PW92 polarised branch for correlation is well-defined, textbook work and should accompany the first real `ISPIN=2` dataset.

2.5 Priority 5: the stress tensor

`Poraquê.get_stress` still raises. The stress needs the energy’s response to a strain, which deforms the cell *and* the grid the fields live on — so it is not simply another Hellmann–Feynman term. Given that forces are not yet accurate enough for relaxation, the stress is correctly sequenced after Priority 1.

Known limitations

3.1 Fundamental to the method

Absolute energies are not DFT energies. The fields are pseudo-valence quantities and VASP's atomic reference `EATOM` is absent. The offset is ≈ 1154 eV per atom for gold. This will never be fixed and does not need to be — only differences are claimed.

The energy is a difference of large numbers. Discussed above. It sets a ceiling on what any field accuracy can buy and should inform how much effort goes into the operators versus into the missing terms.

An FNO is resolution-flexible, not resolution-indifferent. Evaluating far from the training grid density extrapolates in a way no metric in the package detects. The inference script warns; nothing enforces.

Symbolic distillation searches semi-local functionals only. The features are evaluated at a point, so whatever the operator learned that is non-local cannot appear in the distilled expression and shows up as residual. The reports state this beside every fit.

3.2 Implementation limitations

No forces good enough for dynamics. Geometry optimisation and molecular dynamics remain out of reach. `compute_forces` exists and is documented as indicative.

Gaussian pseudo-ion fallback is out of distribution. Without a `POTCAR` the external potential falls back to a Gaussian model that differs from the tabulated one by a relative L^2 of ≈ 0.13 , far outside anything the operators saw. The calculator warns once and proceeds; forces refuse outright, since there is no form factor to differentiate.

Spin-polarised exchange–correlation is not implemented. See the note above.

τ is single-channel under `ISPIN=2`. VASP writes one `TAUCAR` block regardless, so `chg2tau` is two channels in and one out. Whether the spin-resolved kinetic energy density is needed for a spin-polarised kinetic functional is an open question, not a settled one.

CHAPTER 4

Areas needing more rigorous testing

The suite is 602 tests and covers the numerics well. The gaps below are not about coverage percentages — they are the specific claims that are currently *unverified*, ordered by how badly a silent failure would hurt.

4.1 Untested because the data does not exist

Molecular and cohesive energetics. `tests/test_energy_differences.py` contains the N_2 , CO and diamond cases the project needs, parameterised against a documented layout under `PORAQUE_REFERENCE_DATA`. They skip. These are differences between systems of *different* composition, where the per-atom offset does not cancel — the hardest thing that will ever be asked of the calculator, and currently the least tested.

Forces on a perturbed molecule. `tests/test_forces.py` carries stretched- N_2 and stretched- CO cases which skip for the same reason. A stretched diatomic is the cleanest possible force test: one bond, one non-zero component per atom, equal and opposite.

Anything spin-polarised. All 25 spin tests use synthetic fixtures.

4.2 Tested too narrowly

Forces against VASP. Two structures, both gold, both from the same generation run — they are the only ones in the dataset carrying an `OUTCAR`. The force comparison is written as a characterisation with a regression bound, not as a physics claim, precisely because two structures of one element cannot support one.

Energy differences. Seventeen structures, one element, two distinct atom counts. The per-atom-constant assumption underlying every difference is verified only within gold.

Resolution convergence. Energies were compared at 32^3 and at the native $\approx 128^3$ and agree to 0.2 eV in a difference, which is reassuring but is two points on one system. There is no systematic convergence study of any reported quantity against grid density.

4.3 Not exercised in the ordinary test run

Symbolic distillation end to end. PySR is an optional dependency carrying a Julia toolchain, so the tests cover the feature construction, the asymptotic-limit checker and the report

rendering, but never run a search. The load-order workaround for the `juliacall/torch` segfault is therefore untested by CI — and it is the kind of failure that appears only after an hour of training.

Fine-tuning. Configuration validation is tested; the fine-tuning path itself is not run against a real base checkpoint.

Committee disagreement. The ranking machinery is tested on synthetic ensembles. Its correlation with true error — the only property that makes it useful for active learning — is measurable with `-against` but is not part of any test.

4.4 Suggested order of work

1. Generate reference calculations for N_2 , isolated N, CO , isolated C and O, and diamond. This unblocks eight skipped tests and is a day of compute.
2. Generate one `ISPIN=2` dataset — bcc iron is the obvious choice — and validate the spin path against it.
3. Train with `electron_count_weight` enabled and measure whether it beats inference-time normalisation on the same split.
4. Begin the PAW one-centre work, which everything quantitative depends on.

Note

Items 1–3 are each about a day and each convert an unverified claim into a tested one. Item 4 is the research programme. The ordering is deliberate: it would be a mistake to start item 4 while the cheap verification is still outstanding, because item 4’s success can only be measured against the tests items 1–2 create.