Physics-informed neural-network based extrapolations of nonlinear force-free magnetic fields in the Solar corona

Student Name: Andrew Low
Supervisor Name: Anthony R. Yeates
Submitted as part of the degree of M.Sc. MISCADA, Academic Year 2025-26 to the Board of Examiners in the Department of Computer Sciences, Durham University

Abstract — Modelling the coronal magnetic field is crucial to understanding and forecasting space weather. Routine magnetic field observations are largely confined to the solar photosphere as the corona is diffuse and optically thin making direct measurements of the magnetic field there inherently challenging; consequently the coronal field must be inferred by extrapolation from photospheric measurements. Recently, physics-informed neural networks (PINN) have shown promise for this problem. This dissertation evaluates the effectiveness of PINNs for coronal extrapolations and compares them against a classical Grad–Rubin method. A semi-analytical field is used as a controlled benchmark across several grid resolutions. On matched hardware the PINN method achieves errors comparable to those of the classical method while requiring - for a uniform N=400 grid - approximately one-sixth of the wall time and one-tenth of the sampled GPU-board energy. Both methods are also applied to observations of several solar active regions and evaluated using a suite of quality metrics. Separately, we report and patch unnecessary dynamic memory allocation in FastQSL - a field-line tracing CUDA kernel - accelerating isolated tracing-kernel performance by 33-89 times across grid sizes on the analytical test case.

Keywords: solar corona / solar magnetic fields / nonlinear force-free field extrapolation / physics-informed neural networks / Grad–Rubin methods / solar active regions / magnetic helicity / GPU acceleration

1 INTRODUCTION

The solar corona is an extremely hot, tenuous plasma whose macroscopic structure is organised mostly by magnetic fields. Electric currents in active regions store magnetic energy that can be released through flares and coronal mass ejections (CMEs). Eruptive events can launch disturbances that drive geomagnetic storms when they reach Earth (Fogg et al., 2026). Routine observations do not measure the three-dimensional coronal field directly. Instead, the Zeeman effect allows the vector field at the photosphere, the Sun's visible surface layer, to be inferred from the polarisation of spectral absorption lines (Borrero et al., 2010). This leaves the inverse problem of extrapolating a plausible, physically consistent field $\mathbf{B}(\mathbf{x})$ throughout the coronal volume from noisy and incomplete boundary data.

For active regions evolving slowly compared with their Alfvén crossing time, the standard working model is a nonlinear force-free field (NLFFF),

$$(\nabla \times \mathbf{B}) \times \mathbf{B} = \mathbf{0}, \qquad \nabla \cdot \mathbf{B} = 0. \tag{1}$$

The first equation requires electric current to follow magnetic field lines; the second excludes magnetic monopoles. The model greatly reduces the inference problem, but the measured photosphere is not itself force-free. Comparison studies have shown that extrapolations remain sensitive to the processing of the boundary data and the choice of numerical method (Schrijver et al., 2006; Metcalf et al., 2008; DeRosa et al., 2009, 2015).

This dissertation compares the physics-informed neural network (PINN) method, using the NF2 codes from the seminal work of Jarolim et al. (2023), with current-field iteration (CFIT; Wheatland, 2006). NF2 represents the field as a continuous function of position and trains a PINN against the boundary data and residuals of the governing equations (Raissi et al., 2019). CFIT is a classical Grad–Rubin method that alternates between transporting the force-free parameter along field lines and updating the magnetic field. Their accuracy and computational cost are first measured against a semi-analytical NLFFF solution (Low & Lou, 1990), before extrapolating coronal fields from observational data.

Research questions. This dissertation asks three questions.

1. How accurately and consistently do the PINN

and CFIT methods recover a known nonlinear force-free field from identical lower-boundary data?

2. How does extrapolation quality vary with computational cost?

3. Which conclusions remain supported when no volumetric ground truth exists?

Contributions. To answer these questions, we provide a matched-hardware comparison of NF2 and CFIT on an analytical benchmark, evaluate both methods on three observed active regions using a common metric suite, as well as profile and patch an allocation bottleneck in a FastQSL (Zhang et al., 2022), a CUDA field-line tracer used by our CFIT implementation.

2 SCIENTIFIC BACKGROUND AND RELATED WORK

2.1 Magnetograms and the force-free model

NASA's Solar Dynamics Observatory (SDO) provides near-continuous observations of the Sun from geosynchronous orbit (Pesnell et al., 2011). Its Helioseismic and Magnetic Imager (HMI) infers the photospheric magnetic field from the Zeeman effect (Scherrer et al., 2011) by measuring the Stokes profiles I, Q, U and V across the Fe I 6173 Å line, and a Milne–Eddington inversion estimates the field strength, inclination γ and image-plane azimuth ϕ. The azimuthal dependence of the linear-polarisation terms is

$$Q \propto \sin^2\gamma \cos 2\phi, \tag{2}$$
$$U \propto \sin^2\gamma \sin 2\phi. \tag{3}$$

Rotating the transverse field by 180° gives

$$\cos[2(\phi+\pi)] = \cos 2\phi, \tag{4}$$
$$\sin[2(\phi+\pi)] = \sin 2\phi. \tag{5}$$

The two azimuths therefore produce the same Q and U: the observation determines an axis but not which way along it the transverse field points. This ambiguity must be resolved computationally. The Space-weather HMI Active Region Patch (SHARP) Cylindrical Equal-Area (CEA) products used here have already undergone the HMI pipeline's minimum-energy disambiguation and heliographic remapping (Hoeksema et al., 2014; Bobra et al., 2014). They are uncertain measurements of the lower boundary, rather than observations of the coronal volume. The transverse components are generally less certain than the line-of-sight component, and the spatial derivatives used to infer vertical current further amplify noise. This is particularly important for Grad–Rubin methods, which construct α (and current density) from the measured field.

2.1.1 From MHD to NLFFF

The force-free equations arise from the magnetohydrodynamic (MHD) equations under assumptions appropriate to the quasi-static, magnetically dominated corona. For a conducting fluid, momentum conservation is

$$\rho\left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\right] = -\nabla p + \mathbf{J}\times\mathbf{B} + \rho\mathbf{g} + \mu\nabla^2\mathbf{v}, \tag{6}$$

where ρ, p, $\mathbf{v}$ and μ are the mass density, gas pressure, bulk velocity and dynamic viscosity, respectively. An NLFFF seeks a quasi-static magnetic equilibrium. This is a reasonable approximation when the Alfvén crossing time is much shorter than the evolution time of the photospheric boundary,

$$\tau_A = \frac{L}{v_A} \ll \tau_{\mathrm{ev}}. \tag{7}$$

where L is the characteristic length scale. For active regions these timescales are typically minutes and hours to days, respectively (Venkatakrishnan & Ravindra, 2003). This separation permits the coronal field to be treated as an approximate equilibrium. Neglecting bulk acceleration and viscous stress reduces equation (6) to magnetostatic balance,

$$\mathbf{0} = -\nabla p + \mathbf{J}\times\mathbf{B} + \rho\mathbf{g}. \tag{8}$$

The remaining approximation is magnetic dominance. Plasma beta compares gas and magnetic pressure,

$$\beta = \frac{p_{\mathrm{gas}}}{p_{\mathrm{mag}}} = \frac{2\mu_0 p_{\mathrm{gas}}}{B^2}. \tag{9}$$

In the low-β active-region corona, pressure and gravitational forces are small relative to the characteristic magnetic force (Gary, 2001). Equation (8) therefore becomes

$$\mathbf{J}\times\mathbf{B} = \mathbf{0}, \tag{10}$$

This is the force-free condition. Electric current is permitted, but it must be locally parallel or antiparallel to the magnetic field.

To write this condition using $\mathbf{B}$ alone, Ampère–Maxwell gives

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}. \tag{11}$$

The displacement-current term is smaller by order $(V/c)^2$ for a characteristic evolution speed V. Coronal evolution is non-relativistic, so $\mathbf{J} \simeq \mu_0^{-1} \nabla \times \mathbf{B}$. Substitution into equation (10), together with the absence of magnetic monopoles, gives the NLFFF system

$$\boxed{(\nabla \times \mathbf{B}) \times \mathbf{B} = \mathbf{0}, \qquad \nabla \cdot \mathbf{B} = 0.} \tag{12}$$

Because the curl must be parallel to the field, a scalar $\alpha(\mathbf{x})$ exists such that

$$\nabla \times \mathbf{B} = \alpha \mathbf{B}. \tag{13}$$

Taking its divergence and using $\nabla \cdot \mathbf{B} = 0$ gives

$$0 = \nabla \cdot (\alpha \mathbf{B}) = \mathbf{B} \cdot \nabla \alpha + \alpha \nabla \cdot \mathbf{B} \quad \Longrightarrow \quad \mathbf{B} \cdot \nabla \alpha = 0. \tag{14}$$

Thus α has dimensions of inverse length and is constant along each field line, although it may differ between lines. This is why Grad–Rubin methods can transport boundary values of α along field lines.

The force-free approximation applies most closely in the corona. In the photosphere, gas pressure, gravity, bulk flows and non-force-free stresses remain significant; measurements indicate that the field becomes approximately force-free only some 400 km above the surface (Metcalf et al., 1995). This is a fundamental difficulty because the boundary data are measured in the photosphere. To combat this, preprocessing of the data is common which balance agreement with the magnetogram against force-free consistency and spatial smoothness (Wiegelmann et al., 2006). NF2 instead retains the magnetogram as a soft boundary constraint alongside the coronal physics losses.

2.2 Model hierarchy and benchmark field

The behaviour of α defines a natural hierarchy. A field with no electric current, known as a potential field, has $\alpha = 0$ and is the unique minimum-energy field for a specified normal boundary flux (Sakurai, 1982; Alissandrakis, 1981). It is inexpensive and reproducible but since it contains neither current nor free magnetic energy, cannnot provide a realistic physical representation of an active region field. Nevertheless it is a useful baseline. A linear force-free field has a constant non-zero α, allowing global twist but not spatially localised currents. A nonlinear force-free field allows α to vary between field lines and accommodates the shearing, twist and excess energy relevant to active regions, while remaining far cheaper than computing the full time-dependent MHD (Wiegelmann & Sakurai, 2021). For a volume V, the magnetic energy is $E \propto \int_V \|\mathbf{B}\|^2 \, \mathrm{d}V$ and the free energy is the excess above the corresponding potential field. This free energy powers flares and coronal mass ejections, but its extrapolation is particularly sensitive to errors in the transverse field component (Sun et al., 2012).

Magnetic helicity measures the linkage, twist and writhe of magnetic flux and is topologically invariant in the ideal MHD regime. The ordinary volume integral definition, $H = \int_V \mathbf{A} \cdot \mathbf{B} \, dV$ (where $\mathbf{A}$ is the magnetic vector potential), depends on gauge in an magnetically open volume, which is commonplace in astrophysical contexts, so it is often replaced by "relative" helicity - helicity with respect to a potential field sharing the normal boundary flux (Berger & Field, 1984; Finn & Antonsen, 1985). It is a useful metric of non-potential topology, provided numerical divergence errors are also reported alongside it.

A class of nonlinear force-free equilibria derived by Low & Lou (1990), which is specified semi-analytically (it requires the numerical solution of an ordinary differential equation), produces an axisymmetric field; by translating its source below the $z = 0$ plane and tilting its axis of symmetry, we obtain a magnetic field that varies in all three Cartesian coordinates. Because the exact field is known throughout the computation volume it is useful as a test case for extrapolation methods and allows us to distinguish between agreement with the reference solution from satisfaction of the force-free equations. The configuration used in this work is specified in Section 4.1.

2.3 Classical extrapolation methods

Grad–Rubin methods use equation (13) directly. Given an initial field, they trace field lines to transport the lower boundary values of α through the volume; given the resulting current, they update the field and repeat. Prescribing α on one magnetic polarity gives a mathematically well-posed boundary problem, but the two polarities inferred

from real magnetograms are generally inconsistent. Polarity choice and treatment of field lines leaving the box can materially change the extrapolation (Grad & Rubin, 1958; Wheatland, 2006; Amari et al., 2006).

Optimisation methods instead adjust a gridded field to reduce force-free and divergence residuals (Wheatland et al., 2000; Wiegelmann, 2004). We flag this family of methods in particular as they are the closest classical analogue of the PINN method: both minimise violations of the same physical equations, but use different field representations and optimisers.

Magnetofrictional methods approach the problem from the time-dependent force-free induction partial differential equation $\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{v} \times \mathbf{B})$, and "relax" an initial field by setting the velocity proportional to the Lorentz force as an artificial "friction". Their result may depend on the initial state, driving and numerical diffusion (Toriumi et al., 2020). Boundary integral approaches can be used directly for potential fields but become implicit when nonlinear currents must also be determined.

No family of methods removes the incompatibility and uncertainty in the photospheric observations.

Blind analytical tests found the best extrapolations in strong-field, current-carrying regions and substantial sensitivity to the imposed boundaries (Schrijver et al., 2006). Simulated filament arcade tests likewise showed that extrapolation quality depended on how closely the lower boundary approached the force-free regime (Metcalf et al., 2008). Observed region comparisons then identified limited field of view, boundary uncertainty and the non-force-free photosphere as major sources of disagreement (DeRosa et al., 2009). Increasing spatial resolution improved boundary self-consistency but did not remove the spread in energy, helicity or boundary changes (DeRosa et al., 2015).

These studies predate NF2; although Jarolim et al. (2023) compare NF2 with a classical grid-based optimisation method, to our knowledge, no previous study has performed a controlled comparison between a PINN-based NLFFF method and an implementation of the Grad–Rubin method.

2.4 Physics-informed neural networks

A PINN represents an unknown field with a neural network (Raissi et al., 2019). For a differential equation $\mathcal{F}(u) = 0$ in Ω and boundary condition $\mathcal{B}(u) = 0$, substitution of u_θ defines local residuals $r_\mathcal{F} = \mathcal{F}(u_\theta)$ and $r_\mathcal{B} = \mathcal{B}(u_\theta)$. A residual measures violation at one coordinate; the loss function aggregates those violations into a scalar value optimised during training. In general, the loss function can be written as

$$\mathcal{L}(\theta) = \frac{1}{N_c}\sum_{i=1}^{N_c} \|r_\mathcal{F}(\mathbf{x}_i;\theta)\|^2 + \frac{1}{N_b}\sum_{j=1}^{N_b} \|r_\mathcal{B}(\mathbf{x}_j;\theta)\|^2 . \tag{15}$$

Here θ denotes all trainable network parameters; $\mathbf{x}_i$ and $\mathbf{x}_j$ are interior collocation and boundary coordinates, respectively; and N_c and N_b are their corresponding sample counts. Interior collocation points are coordinates where the field is sampled and evaluated and are not observation data. Automatic differentiation supplies any required spatial derivatives. For NLFFF extrapolation, u_θ is $\mathbf{B}_\theta$ and the physics residuals are the two terms in equation (12); Section 3.1 gives the precise normalisation and weighting of NF2's loss function.

An activation is the nonlinear function applied inside each hidden layer of a network; the choices examined in this work include ReLU, tanh and sine. Smooth activations are important because the loss differentiates the network with respect to its inputs. ReLU, $f(x) = \max(0, x)$, is piecewise linear and has a zero second derivative almost everywhere, so it cannot learn non-zero curvature from a second-order residual. Even with smooth activations, standard fully connected networks tend to learn low spatial frequencies first. Sinusoidal representation networks (SIRENs) reduce this spectral bias and are the architecture used by NF2 (Sitzmann et al., 2020; Jarolim et al., 2023).

We apply PINNs to several a range of toy problems from simple differential equations through to a linear force-free field in order to illustrate these architectural claims in Appendix A.

3 METHODOLOGY

The PINN and CFIT methods solve the same nonlinear force-free field equations but represent the unknown field in different ways. With PINNs, we adjust the parameters of a continuous neural field to reduce its boundary and differential residuals. CFIT stores the field representation on a discrete grid and alternates between transporting electric current along field lines and updating the field produced by that current. Increasing grid resolution changes CFIT's numerical state during

solving, whereas sampling a trained PINN model more densely during evaluation changes only its output discretisation. Neither method's extrapolated field can be treated as correct; both are assessed independently in Section 4.

Software implementation. We use the published implementation of NF2 Jarolim et al. (2023) [1]. We implement a CFIT solver following Wheatland (2006), with a "self-consistency" procedure following Wheatland & Régnier (2009) for active region extrapolations; UFiT (Aslanyan et al., 2024) [2] and FastQSL (Zhang et al., 2022) [3] perform field-line tracing on the CPU and GPU respectively. Relative helicity is evaluated with FLHtools (Yeates & Page, 2018) [4]. The codes developed for this project are available online.[5]

3.1 PINN extrapolations

NF2 uses a sinusoidal representation network (SIREN), a fully connected neural network with sine activations (Sitzmann et al., 2020; Jarolim et al., 2023), with a 3D coordinate vector as input and a 3D $\mathbf{B}$ field vector as output. For normalised coordinates $\mathbf{x}$, its hidden layers have the form

$$\mathbf{h}_{\ell+1} = \sin[\omega_\ell(W_\ell \mathbf{h}_\ell + \mathbf{b}_\ell)], \qquad \mathbf{h}_0 = \mathbf{x}, \quad (16)$$

where W_ℓ and $\mathbf{b}_\ell$ are trainable parameters and ω_ℓ sets the spatial-frequency scale of the layer. The complete training workflow is summarised in Algorithm B.1 in Appendix B.

Physical coordinates are normalised using

$$\widetilde{\mathbf{x}} = \mathbf{x}_{\mathrm{Mm}}/L_0, \qquad \Delta\widetilde{x} = n_{\mathrm{bin}}\Delta_{\mathrm{native}}/L_0, \quad (17)$$

where L_0 is a scaling factor, megametres per dimensionless spatial unit. The SHARP magnetograms have $\Delta_{\mathrm{native}} = 0.36\,\mathrm{Mm}$, so factor-two binning (which we use in this work) gives 0.72 Mm pixels. This scale also enters spatial derivatives in the loss through $\nabla_{\mathrm{Mm}} = L_0^{-1}\nabla_{\widetilde{\mathbf{x}}}$; so it is not just plot metadata. For Low–Lou extrapolations, dimensionless coordinates with $L_0 = 1$ are used.

Two NF2 parameterisations are used in this project. For Low–Lou, the network predicts a vector potential $\mathbf{A}_\theta$ and sets $\mathbf{B}_\theta = \nabla \times \mathbf{A}_\theta$. Solenoidality then holds by construction, although the force-free term requires second derivatives of $\mathbf{A}_\theta$ which increases training time. For the observational models, we predict $\mathbf{B}_\theta$ directly and penalise its divergence. At interior collocation points, the losses and total objective are

$$\mathcal{L}_{\mathrm{NF2}} = \lambda_b\mathcal{L}_b + \lambda_p\mathcal{L}_p + \lambda_{\mathrm{ff}}\mathcal{L}_{\mathrm{ff}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}, \quad (18)$$

$$\mathcal{L}_{\mathrm{ff}} = \frac{1}{N_c}\sum_i \frac{\|(\nabla \times \mathbf{B}_\theta)_i \times \mathbf{B}_{\theta,i}\|^2}{\|\mathbf{B}_{\theta,i}\|^2 + \varepsilon}, \quad (19)$$

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{N_c}\sum_i |\nabla \cdot \mathbf{B}_{\theta,i}|^2. \quad (20)$$

$\mathcal{L}_b$ measures disagreement with the supplied B_z field, while $\mathcal{L}_p$ applies the potential-field condition on the lateral and top boundaries. For the vector-potential model, $\lambda_{\mathrm{div}} = 0$ because solenoidality is built into the representation. These are soft constraints as the optimiser may reduce one term while another remains comparatively large. The NF2 training loss diagnoses its own optimisation but cannot be used to rank PINNs against CFIT.

NF2 draws new interior collocation points at every optimiser step. Each loss term is averaged over its own batch before its λ weight is applied, so batch size controls spatial coverage and sampling variance rather than the nominal relative weight of the loss terms.

With uniformly sampled collocation points, the first mean in equation (15) estimates an equal-volume average of the squared physics residual. NF2 can instead bias its sampling towards the lower corona by drawing $U \sim \mathcal{U}(0,1)$ and setting

$$z = z_{\min} + (z_{\max} - z_{\min})U^p, \quad (21)$$

where p is the height-sampling exponent. The choice $p = 1$ is uniform in height, whereas $p > 1$ preferentially samples lower heights; NF2 now uses $p = 2$ by default although at release used uniform sampling. The horizontal coordinates remain uniformly sampled. Consequently, the sampling distribution provides an effective spatial weighting even though each loss is batch-averaged, since more frequently sampled regions contribute to more gradient updates.

NF2 also provides an optional height based field-strength scaling, distinct from the sampling distribution. For each sampled height z_j, it calculates

$$S_j = \left\langle \left(\|\mathbf{B}(x,y,z_j)\| + \varepsilon\right)^2 \right\rangle_{x,y} \quad (22)$$

where $\varepsilon = 10^{-8}$ is a small constant added for numerical stability, $\langle\cdot\rangle_{x,y}$ is the mean over the sampled horizontal positions at height z_j, and divides

[1] https://github.com/RobertJaro/NF2
[2] https://github.com/Valentin-Aslanyan/UFiT
[3] https://github.com/peijin94/FastQSL
[4] https://github.com/antyeates1983/flhtools
[5] https://github.com/ackl/solar-corona-pinns

the selected pointwise physics losses by S_j. The scaling factor is excluded from gradient calculation. This prevents the stronger field near the lower boundary from dominating the loss solely through its magnitude and gives weak-field heights greater relative influence. Section 4 states the sampling exponent and whether this scaling is used for each study.

3.2 CFIT extrapolations

We implement the current-field iteration method following Wheatland (2006) which uses the Grad–Rubin formulation (Grad & Rubin, 1958); it is not the only possible implementation but we chose it as it has best known time complexity $O(N^4)$ documented in the literatue. This method starts from the potential field matching the observed B_z and supplies α_0 on either the positive or negative lower boundary polarity. Each iteration solves

$$\mathbf{B}^{(k)} \cdot \nabla \alpha^{(k)} = 0, \qquad \alpha^{(k)}|_{S_\pm} = \alpha_0, \tag{23}$$

$$\nabla \times \mathbf{B}^{(k+1)} = \alpha^{(k)} \mathbf{B}^{(k)}, \quad \nabla \cdot \mathbf{B}^{(k+1)} = 0. \tag{24}$$

The first equation states that α is constant along each field line of the current iterate. The classic Runge–Kutta method is then used to trace both directions from every point of the grid and bilinear interpolation is used to read α_0 at the lower footpoint. A line receives current only if it reaches a valid footpoint on the chosen polarity; a line leaving through a side or the top is assigned $\alpha = 0$.

For Low–Lou, exact lower B_z and α_0 are supplied. For SHARP data, $\alpha_0 = (\nabla \times \mathbf{B})_z / B_z$ is calculated from the observed horizontal field at every pixel on the chosen polarity. The source $\alpha^{(k)}\mathbf{B}^{(k)}$ determines the next current-carrying field through a Fourier transform. Writing the update as $\mathbf{B}^{(k+1)} = \mathbf{B}_0 + \mathbf{B}_J^{(k)}$, the potential field satisfies $B_{0,z}(x, y, 0) = B_{\mathrm{obs,z}}$ and the non-potential current carrying component is constructed to satisfy $B_{J,z}(x, y, 0) = 0$. Thus CFIT uses the lower normal field directly as a hard boundary condition, rather than fitting it through a loss as NF2 does. The Fourier solve requires mirroring the current cube and padding it with zeroes (to the nearest 2^n); the padding is to mitigate the influence of the periodicity from the FFT. We shall see that in practice this limits the problem size - as with *all* classical NLFFF methods it requires $O(N^3)$ memory (grid of N^3 points), but the extended current-density array with mirroring and padding needed for the FFT field update introduces a substantial constant factor memory overhead (at least 2^3 times), presenting a trade-off between this $O(N^4)$ algorithm and alternative CFIT implementations with higher time complexity.

The lower current layer is duplicated, placing the reflection plane half a grid cell below $z = 0$. Wheatland (2006) found that this better represented the symmetry of the mirrored current paths and improved test-case accuracy, at the expense of a small error in the normal-field boundary condition.

The choice of polarity affects the solution because the measured photosphere generally supplies inconsistent values of α at the two footpoints of a field line. We follow the convergence criterion of Wheatland (2006) - when the current-weighted force-free angle improves by less than one per cent without increasing, or after 30 iterations.

Field-line tracing dominates the cost of each CFIT iteration, so two tracing backends are compared. Universal Fieldline Tracer (UFiT) (Aslanyan et al., 2024) performs the integration with Fortran and OpenMP on CPUs. FastQSL (Zhang et al., 2022) performs this operation with CUDA on NVIDIA GPUs. Both backends process the grid points in batches of at most 200,000 tracing seeds; the UFiT wrapper copies the full field into its Fortran array for each batch, whereas FastQSL retains one device copy across batches. Algorithm B.2 in Appendix B gives the corresponding high-level CFIT workflow.

3.3 Computational cost and profiling

Computations for both extrapolation methods were carried out on various nodes (using exclusive access) on HPC clusters at Durham University. Here we will describe the general methods used to profile them.

Where available we utilise Slurm's energy accounting plugin which reports `ConsumedEnergyRaw` in joules from the node's IPMI instrumentation. This is denoted in our results by E_{node} and represents aggregate entire node energy over the accounting interval. It does not provide granular energy attribution to one process or component.

We spawn a sibling process with each extrapolation to sample GPU power at approximately 1 Hz using `nvidia-smi` and integrated using the

measured timestamps,

$$E_{\text{GPU}} \approx \sum_k \frac{P_k + P_{k+1}}{2}(t_{k+1} - t_k). \tag{25}$$

This gives us an estimate of GPU-board energy: it excludes power usage from everything else e.g. CPU, main memory, storage and cooling. Since the power sensor may update more slowly than it is queried short transients can be missed (Burtscher et al., 2014; Yang et al., 2024).

For profiling of FastQSL to measure the effect of our dynamic allocation patch, we use Nsight Compute on a local GTX 1050 as GPU hardware counters were inaccessible on the HPC nodes.

As noted in the previous subsection, all classical NLFFF methods required $O(N^3)$ memory, with our adopted CFIT implementation requiring $O(I \cdot N^4)$ work - grid-based optimization and relaxation methods require $O(I \cdot N^3)$, where I is the number of iterations performed. PINN time complexity is instead $O(I \cdot C \cdot P)$ governed by C the number of collocation evaluations and P total network parameters, and cannot generally be expressed solely in terms of the (output) grid resolution N.

4 EVALUATION DESIGN

4.1 Study cases

Low–Lou field. The controlled case is the nonlinear axisymmetric solution of Low & Lou (1990). We use the parameters: $n = m = 1$, source depth 0.3, inclination $\pi/4$, and domain $[-1, 1] \times [-1, 1] \times [0, 2]$. A grid of width N has shape $(N, N, N, 3)$, origin $(-1, -1, 0)$ and isotropic spacing $2/(N-1)$.

The PINNs receive the exact lower B_z and corresponding potential field on the outer faces. CFIT instead receives exact lower B_z and α on one polarity. The NF2 loss weights are $(\lambda_b, \lambda_p, \lambda_{\text{ff}}, \lambda_{\text{div}}) = (1, 10, 0.1, 0)$; the default lower height biased sampling is used, with no height scaling applied.

Since the PINN method with NF2 has no built-in convergence criteria unlike CFIT, only a maximum number of iterations, we test a range of network architectures along with a set of 26 training time budgets - intervals of 10s between 50s and 300s - with the model evaluated at each interval while fixing the grid size to $N = 200$ to investigate the models' performance. The upper wall-time limit was chosen based on representative convergence times from early tests of the CFIT implementation. The tested architectures (5 network sizes) have total parameter counts between 512 and 2048. The upper limit was chosen as Jarolim et al. (2023) found no substantial benefit with larger networks beyond this size. Extrapolations were done with 3 random seeds for each network size. The resulting 390 output fields are compared with both polarity branches from CFIT. These runs measure the stochastic spread inherent to neural-network optimisation and highlight a capacity-cost trade-off. Each NF2 run uses $1,000$ optimiser steps per epoch with a ceiling of 20 epochs; the wall-time budget determines the number actually completed.

4.1.1 Active regions

The observational cases include the published benchmark from NF2's debut and two active regions associated with the major geomagnetic storms of May and October 2024. The two 2024 regions are useful as contrasting sources; the compound May event produced the strongest geomagnetic storm since 2003, while the October event reached a similar scale despite a simpler solar driver.

Both regions had $\beta\gamma\delta$ Hale classifications, but the region associated with May produced multiple CMEs that interacted during propagation, whereas the other produced a single Earth directed halo CME. The compound May event encountered a comparatively quiet magnetosphere; the October CME arrived after preceding activity had already primed the magnetosphere.

Both events produced prominent aurorae and vast visibility beyond the usual high-latitude terrestial zones, including many urban locations.

Solar flares are classified according to their peak soft X-ray flux, as measured by the Geostationary Operational Environmental Satellites (GOES), using the logarithmic A, B, C, M and X classes. Each class represents a tenfold increase in peak flux, while the numerical multiplier specifies the position within that class

(Fogg et al., 2026).

AR 11158. Following Jarolim et al. (2023) this case uses the HMI Active Region Patch (HARP) 377 observation of National Oceanic and Atmospheric Administration Active Region (NOAA AR) 11158 at 2011-02-15 00:00 TAI. NF2 and the self-consistency conditioned CFIT branches of both polarities (see Section 5.3.1 for motivation details) are exported to a common $296 \times 184 \times 160$ grid with isotropic 0.72 Mm spacing. For NF2 we follow the hyperparamters used in

Jarolim et al. (2023); λ_b decays exponentially from 1000 to 1 over 50,000 steps, with $(\lambda_p, \lambda_{\rm ff}, \lambda_{\rm div}) = (1, 0.1, 0.1)$. Training uses five epochs of 20,000 optimiser steps. Interior collocation is uniform in height ($p = 1$), without height scaling.

AR 13664. The AR 13664 case uses HARP 11149 observations at 07:48 and 10:36 TAI on 2024-05-09, bracketing an GOES X2.2 flare which occured at 09:13 TAI, plus seven further frames sampled the denser 12 minute cadence from 08:36 to 09:48 TAI. Three NF2 seeds and the conditioned CFIT branches are evaluated on a common $511 \times 239 \times 111$ grid with 0.72 Mm spacing and a height of 79.2 Mm. The NF2 loss weights are $(\lambda_b, \lambda_p, \lambda_{\rm ff}, \lambda_{\rm div}) = (1, 10, 10^{-3}, 10^{-2})$, with height-wise field-strength scaling applied to the physics losses. Default lower height biased sampling exponent $p = 2$ is used, *with* height scaling applied. Each model is trained for 15 epochs of 10,000 optimiser steps.

AR 13848 time series. With the AR 13848 case we test a longer time series using HARP 12001 observations from 2024-10-04 to 2024-10-09 at a 12 minute cadence, in an attempt to invesitage NF2's ability to achieve quasi-realtime extrapolations. Only the initial frame is trained from scratch, for one epoch of 100,000 optimiser steps. Each subsequent time step is initialised from the preceding model checkpoint and trained for 2,000 optimiser steps. Nine independent PINNs are also trained from scratch for 100,000 steps as controls for 12:00 TAI each day in the series. The evaluation grid again has 0.72 Mm spacing and extends to 160 Mm. For the initial NF2 model, λ_b decays exponentially from 1000 to 1 over 50,000 steps and is fixed at 1 during continuation; $(\lambda_p, \lambda_{\rm ff}, \lambda_{\rm div}) = (10, 0.1, 0.1)$. We again use the default lower height biased scaling and no height scaling.

Case	Field grid	CFIT FFT grid
Low–Lou $N = 800$	800^3	$1024\times1024\times2048$
AR 11158	$296\times184\times160$	$512\times256\times512$
AR 13664	$511\times239\times111$	$512\times256\times256$
AR 13848	$477\times328\times222$	512^3

Table 1: Evaluation grid sizes and the corresponding CFIT current-field FFT grid sizes after mirroring and padding. AR 13848 was evaluated with NF2 only; its CFIT size is hypothetical.

4.2 Hardware context

As the comparisons span several CPU and GPU systems, we note that wall timings should be interpreted only within the context of matched hardware stated in the results. The full hardware allocation inventory is given in Table B.1 in Appendix B.3.

4.3 Evaluation metrics

Since no single scalar score can establish if an extrapolation is correct, we report a suite of metrics across genres which are standard in the literature. These measure agreement with a reference field, consistency with the NLFFF equations, magnetic energy and boundary agreement. By default, any summations are carried out over Ω - the set of all points in the volume; when we use the subscript "int", sums are restricts to $\Omega_{\rm int}$, with all six boundary faces removed (where compact finite difference stencil is used). Here we introduce the most pertinent metrics - detailed definitions of all metrics are collected in Appendix B.2.

For the analytic Low–Lou case, the primary metric is a relative field error

$$\epsilon_B = \frac{\|\mathbf{B} - \mathbf{B}^\star\|_2}{\|\mathbf{B}^\star\|_2}, \tag{26}$$

where $\mathbf{B}^\star$ is the exact field and $\|\cdot\|_2$ denotes the L_2 norm. Physical consistency with the force-free equations is reported by the current-weighted force-free angle

$$\theta_{J,\rm int} = \arcsin\left(\frac{\sum_{i\in\Omega_{\rm int}} \|\mathbf{J}_i \times \mathbf{B}_i\|/\|\mathbf{B}_i\|}{\sum_{i\in\Omega_{\rm int}} \|\mathbf{J}_i\|}\right) \tag{27}$$

and the mean fractional flux imbalance

$$\langle|f_i|\rangle_{\rm int} = \frac{1}{M_{\rm int}} \sum_i \left| \frac{\oint_{\partial V_i} \mathbf{B}\cdot d\mathbf{S}}{\oint_{\partial V_i} \|\mathbf{B}\|\, dS} \right|. \tag{28}$$

Both vanish in the ideal limit (Wheatland et al., 2000; Wheatland, 2006; Schrijver et al., 2006).

For voxel volume ΔV, magnetic energy is $E = \Delta V \sum_i \|\mathbf{B}_i\|^2/(2\mu_0)$. Low–Lou cases use $q_E = E/E^\star$; observational cases for both methods use a common potential field constructed from the observed lower B_z to report $E_{\rm free} = E - E_{\rm pot}$ and $\eta_{\rm free} = E_{\rm free}/E_{\rm pot}$.

Relative magnetic helicity is evaluated with the gauge-invariant volume integral

$$H_R = \int_V (\mathbf{A} + \mathbf{A}_p)\cdot(\mathbf{B} - \mathbf{B}_p)\, dV, \tag{29}$$

where $\mathbf{B}_p$ is a potential reference sharing the flux-balanced normal field on every face, and $\mathbf{A}$ and $\mathbf{A}_p$ are the corresponding vector potentials (Berger & Field, 1984; Finn & Antonsen, 1985).

Non-solenoidal "contamination" due to numerical methods can be quantified using the energy ratio E_{div}/E (Valori et al., 2013, 2016). We report an E_{div}/E-like metric calculated using the FLHtools vector potential extrapolation. It measures energy associated with the failure of the sampled field to equal $\nabla \times \mathbf{A}$. We calibrate this derived metric against the DCT-based Valori-style reference described in Appendix B.2.1. The resulting empirical screen is

$$\left(\frac{E_{\mathrm{div}}}{E}\right)_{\mathrm{FLHTOOLS}} \leq 0.6. \tag{30}$$

This is a screening limit for our implementation, not the published Valori threshold.

At the bottom boundary, the discrepancy from the observed vector field is

$$\begin{aligned} B_{\mathrm{diff},i} &= \|\max(|\mathbf{B}_i - \mathbf{B}_{\mathrm{obs},i}| - \mathbf{s}_i, \mathbf{0})\|, \\ \langle B_{\mathrm{diff}} \rangle_{\mathrm{obs}} &= \frac{1}{M_{\mathrm{obs}}} \sum_{i \in \Omega_{\mathrm{obs}}} B_{\mathrm{diff},i}. \end{aligned} \tag{31}$$

where $M_{\mathrm{obs}} = |\Omega_{\mathrm{obs}}|$ is the number of observed lower-boundary pixels and $\mathbf{s}_i$ contains component uncertainties provided in the SHARP data products.

Timing and energy comparisons use the measurement boundaries defined in Section 3.3.

5 RESULTS

5.1 CFIT validation and performance

Our CFIT solver closely reproduces the convergence reported by Wheatland (2006): the force-free angle metrics agree broadly with their reported values with lower iteration counts. (Table 2).

Although grid refinement strongly improves the force-free and solenoidal residuals, it does not materially reduce the error against the exact field. This is consistent with CFIT assigning zero current to open field lines rather than receiving the exact field on every face.

UFiT scales well across physical CPU cores, reducing both runtime (and energy usage), but gains little from simultaneous multithreading. Although the field-line tracing portion remains dominant, the serial remainder of the computation limits parallel efficiency which is expected as per Amdahl's law (full measurements are given in Appendix C.1.1).

5.1.1 Moving tracing to the GPU

As field-line tracing dominates each CFIT iteration, we replaced UFiT with the CUDA-based FastQSL tracer. This moves the dominant computation onto a GPU, matching the accelerator used by NF2 more closely. Based on profiling of UFiT with a fixed problem size, we see that it lies on the compute-bound side of the ridge point (Figure C.1), so CPU memory bandwidth is not the governing roofline ceiling. This provided additional motivation for evaluating a massively parallel CUDA tracer. While profiling FastQSL we identified unnecessary per-thread dynamic memory allocations within the CUDA kernel; at the tested problem size and concurrency on an NVIDIA A100 GPU, these allocations exhausted the default CUDA heap size of 8MB which surfaced as an `CUDA_ERROR_ILLEGAL_ADDRESS` error as the kernel tried to dereference a null pointer. Simply replacing the per-thread dynamic allocations with stack local variables to mitigate this also accelerates the isolated `TraceAllBline` kernel by factors of 33-89 across the tested grid resolutions (this does not accelerate the remaining CFIT stages); machine-specific roofline metrics are given in Appendix C.1.2. From this section onwards all extrapolations with the CFIT method use the FastQSL tracing codes with our patch.

5.2 Low–Lou analytical benchmark

We find that NF2 and CFIT attain comparable global errors relative to the analytical solution but, as could be expected, disagree materially with each other on their outputs and exhibit different cost scaling. Visualisations of extrapolations using matched hardware are shown in Figure 1.

Across the same evaluation grids, CFIT becomes steadily more force free with refinement, whereas neither NF2's relative field error nor its force-free consistency varies monotonically (Figure 3). Because this model predicts a vector potential and uses no divergence-loss term, its small E_{div}/E is primarily a property of the representation and its sampled curl, rather than evidence that optimisation learned solenoidality.

Examining the ensemble evaluated on a $N = 200$ grid shows a quality-cost trade-off rather than a monotonic ordering by network size or training time. Most error reduction occurs at the start, after which the curves for each architecture continue to cross each other while energy usage rises approximately linearly with training time (which

N	Published iter.	Iter.	ϵ_B	$\epsilon_{J,\mathrm{int}}$	$\theta_{J,\mathrm{int}}$	$\langle\lvert f_i\rvert\rangle_{\mathrm{int}}$	$\langle B_{\mathrm{diff}}\rangle$ (G)	$\langle B_{\mathrm{diff},z}\rangle$ (G)	q_E	IPMI node (kJ)
25	8	6	0.2280	0.3479	12.765°	4.05×10^{-3}	4.894	1.383	1.0974	24.47
50	10	8	0.2205	0.2698	3.840°	5.18×10^{-4}	4.608	0.662	0.9881	31.81
100	11	10	0.2238	0.2404	1.295°	7.64×10^{-5}	4.504	0.325	0.9404	50.14
200	13	12	0.2273	0.2299	0.474°	1.19×10^{-5}	4.472	0.162	0.9188	340.82

Table 2: Positive-polarity Low–Lou results with CFIT method using the UFiT tracer. The published iteration counts are from Wheatland (2006); the remaining columns report the metrics defined in Section 4.3.

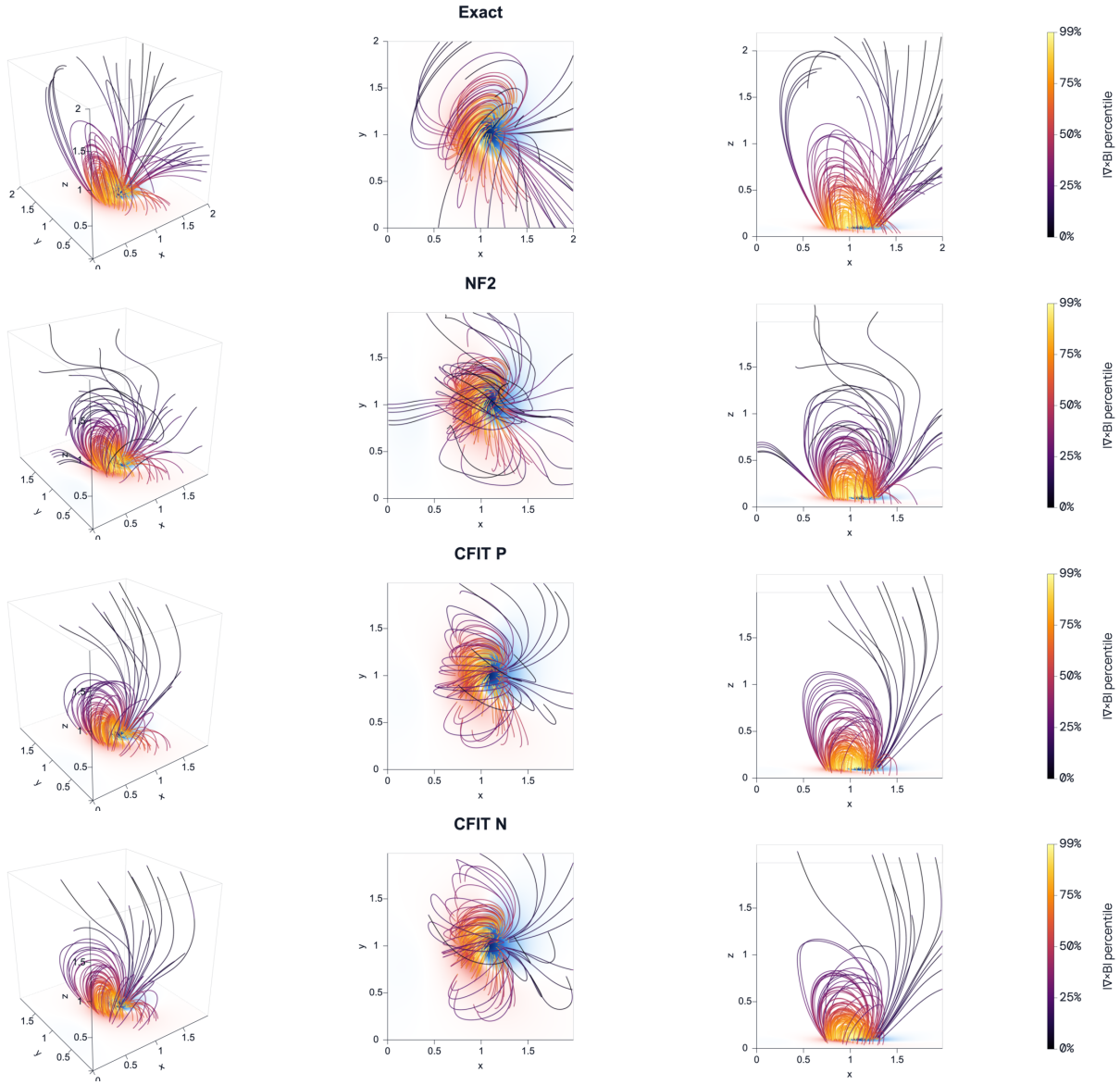


Figure 1: Field lines for the exact field, NF2 and the two CFIT polarity branches at $N = 400$, rendered using VTK.js. The field lines are coloured based on current density. Coordinates are dimensionless.

is to be expected with a constant workload). (Figure 2). At a fixed wall-time budget, larger networks also complete fewer optimiser steps. The crossings therefore compare quality at fixed cost, not network capacity after an equal number of updates.

Table 3 compares the exact solution, representative NF2 results and CFIT. End-to-end timings for the NF2 runs are included as evaluating the learnt field onto the discrete grid is not free. Hardware placement and the stochastic nature of the Adam optimiser used for the PINN method both contribute to the spread.

Short NF2 runs reach CFIT-like field errors at lower apparent cost, but this comparison is not hardware matched.

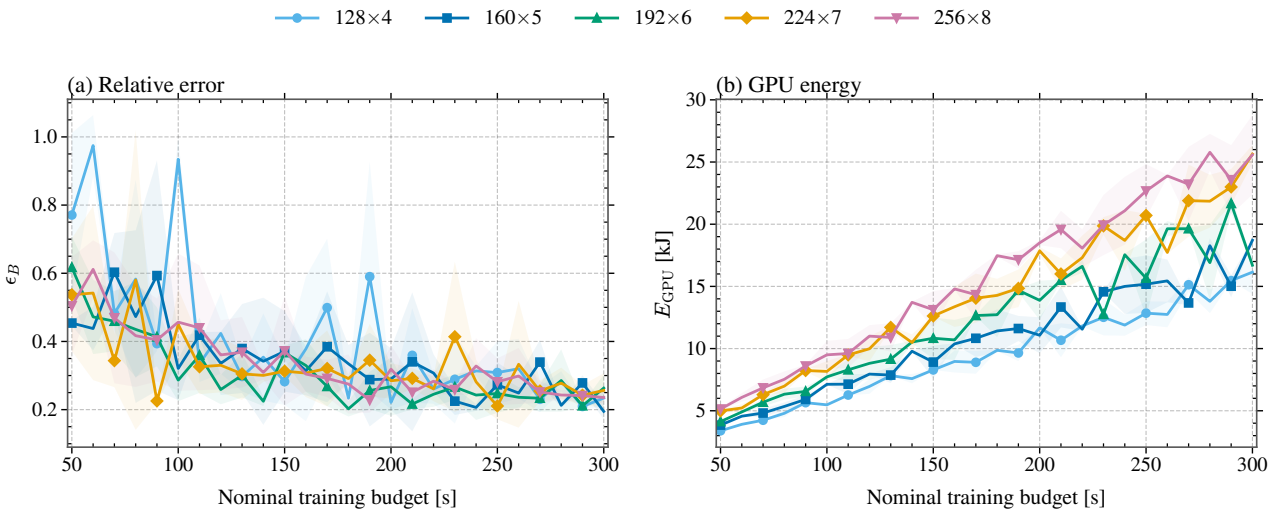


Figure 2: Relative field error and utilised GPU-board energy as functions of the training time for five NF2 network architectures. Curves show means from the three random seeds and shaded regions show standard deviations.

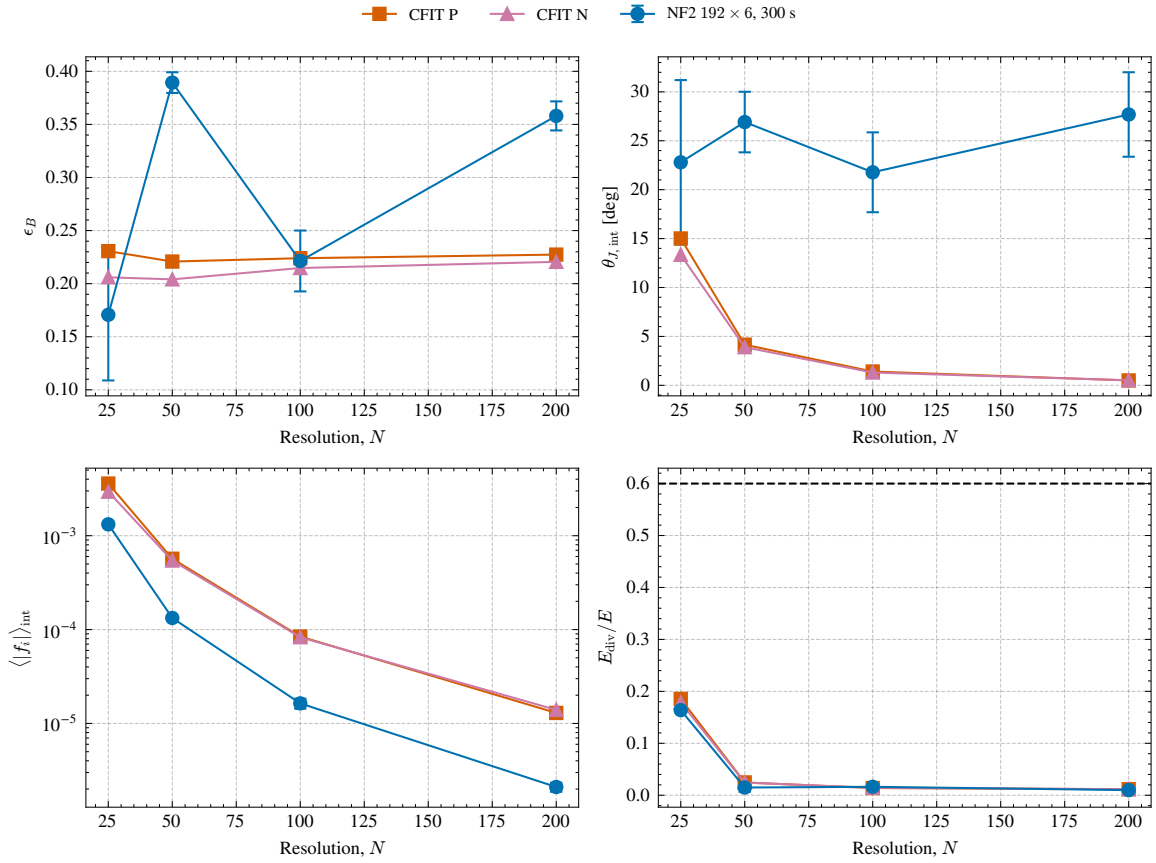


Figure 3: Low–Lou metrics evaluated on the same $N = 25$, 50, 100 and 200 grids. The NF2 curves show means and standard deviations from three random seeds for a 192×6 PINN trained for 300 s. The CFIT curves show the positive- and negative-polarity FastQSL branches. The dashed line marks our E_{div}/E quality limit of 0.6; the fractional-flux metric is plotted logarithmically.

We remove the main hardware confound by matching hardware with a comparison at the $N = 400$ grid. (Table 4). At the sampled operating point, NF2 reaches the CFIT field-error regime at lower cost but is less force free. The disagreement between the CFIT polarity branches is smaller than either inter-method disagreement, so polarity choice within CFIT does not account for the dif-

Metric	Exact	NF2 224 × 7	NF2 192 × 6	NF2 160 × 5	FastQSL P	FastQSL N
ϵ_B	0	0.2254 ± 0.0074	0.2163 ± 0.0152	0.1948 ± 0.0233	0.2273	0.2205
$\epsilon_{J,\mathrm{int}}$	0	0.2962 ± 0.0222	0.2196 ± 0.0109	0.1773 ± 0.0257	0.2297	0.2021
C_{vec}	1	0.9754 ± 0.0004	0.9817 ± 0.0007	0.9836 ± 0.0002	0.9740	0.9755
C_{CS}	1	0.5345 ± 0.0238	0.4302 ± 0.0269	0.5635 ± 0.0760	0.4673	0.4362
E'_n	1	0.6186 ± 0.0109	0.5905 ± 0.0130	0.6470 ± 0.0118	0.5768	0.5764
E'_m	1	0.2161 ± 0.0268	0.1329 ± 0.0215	0.2621 ± 0.0539	0.1886	0.1636
$\theta_{J,\mathrm{int}}$ (deg)	0.101	19.826 ± 0.276	18.529 ± 1.014	16.569 ± 3.077	0.497	0.518
$\langle \lvert f_i \rvert \rangle_{\mathrm{int}}$	1.06×10^{-6}	$(1.885 \pm 0.054) \times 10^{-6}$	$(2.205 \pm 0.119) \times 10^{-6}$	$(2.240 \pm 0.001) \times 10^{-6}$	1.29×10^{-5}	1.39×10^{-5}
q_E	1	0.9483 ± 0.0570	0.8029 ± 0.0445	0.9803 ± 0.0541	0.9188	0.9223
E_{div}/E	0.01197	0.0103 ± 0.0014	0.01128 ± 0.00017	0.01219 ± 0.00034	0.01164	0.01164
H_R (Low–Lou units)	127.759	119.49 ± 8.56	90.89 ± 5.28	109.35 ± 1.85	84.26	86.49
$\langle B_{\mathrm{diff}} \rangle$ (G)	0	5.078 ± 0.545	3.773 ± 0.475	2.667 ± 0.180	4.689	4.368
$\langle B_{\mathrm{diff},z} \rangle$ (G)	0	2.496 ± 0.261	2.185 ± 0.158	1.306 ± 0.023	0.163	0.164
Training budget (s)	—	90	210	300	—	—
E2E (s)	—	171.7	291.3	380.1	214.9	214.9
GPU (kJ)	—	8.24	15.50	18.71	10.04	9.55
Node (kJ)	—	131.11	219.07	288.22	159.35	158.86

Table 3: Selected $N = 200$ Low–Lou quality–cost results. The exact column is the analytical field evaluated on the same grid - the exact solution does not have perfect force-free or solenoidal metrics due to the discrete nature of the grid. NF2 values are means of three seeds and their standard deviations. Relative helicity is expressed in the common dimensionless Low–Lou normalisation of one code unit per gauss; a physical value in Mx^2 requires physical field and length scales.

ference from NF2. The solenoidal metric $\langle |f_i| \rangle_{\mathrm{int}}$, however, changes little between methods at fixed resolution and decreases for the analytical field as the grid is refined.

CFIT's mirrored, padded FFT volume for the 800^3 field exceeded available allocated memory (Table 1), so the matched comparison was not extended. This is an implementation ceiling of the present solver, not an inherent limitation of Grad–Rubin methods.

5.3 AR 11158

Our PINN extrapolation agrees broadly with the published metrics, although it does not reproduce them exactly as show in Table 5. Although we followed the same neural network configuration and hyperparameters where published, the NF2 codes have since undergone many updates which may explain this discrepancy. The quoted angle and divergence use the definitions of Jarolim et al. (2023), not our interior-domain definitions. We construct a common potential field from the observed boundary B_z for metrics to facilitate a comparison baseline between methods.

5.3.1 NF2–CFIT comparison

Without the self-consistency procedure of Wheatland & Régnier (2009), CFIT reaches its force-free stopping criterion but produces grossly unphysical fields. After conditioning, the positive branch converges while the negative branch reaches the iteration limit. The boundary metrics then expose the methods' different constraints: NF2 follows the full observed vector more closely, whereas CFIT preserves its prescribed normal component (Table 6), an expected consequence of its hard normal field boundary condition.

Unlike in the Low–Lou comparison, E_{div}/E separates the methods strongly here. Its ordering follows neither full-vector nor normal component boundary agreement. For NF2 the boundary loss weighting coefficients are ten times larger than either physics loss coefficient. NF2's closer full-vector boundary fit but poorer E_{div}/E is qualitatively consistent with this hyperparameter configuration.

Like in the Low–Lou comparison, CFIT polarity branches' metrics generally remain closer to one another than either is to NF2 (Table 7 and Figure 4).

5.4 AR 13664 flare pair

Across the extrapolations of the AR 13664 X2.2 flare, the boundary comparison retains the trade-off we saw for AR 11158: NF2 better matches the full vector field, while CFIT is more nearly force free and reproduces the prescribed B_z much more closely. The E_{div}/E metric, however, reverses its ordering between the two active regions and does not deteriorate with the large CFIT energy and helicity changes.

Our NF2 extrapolations for AR 13664 gives the divergence loss ten times the force-free weight and

Metric	Exact	NF2 192 × 6, seed 0	FastQSL P	FastQSL N
ϵ_B	0	0.2266	0.2296	0.2234
$\epsilon_{J,\mathrm{int}}$	0	0.2489	0.2251	0.1984
C_{vec}	1	0.9764	0.9735	0.9750
C_{CS}	1	0.4614	0.4674	0.4363
E'_n	1	0.5878	0.5751	0.5740
E'_m	1	0.1814	0.1887	0.1639
$\theta_{J,\mathrm{int}}$ (deg)	0.0256	16.098	0.210	0.220
$\langle\lvert f_i\rvert\rangle_{\mathrm{int}}$	1.32×10^{-7}	4.84×10^{-7}	2.32×10^{-6}	2.55×10^{-6}
q_E	1	0.8252	0.9082	0.9106
E_{div}/E	0.00680	0.00669	0.00667	0.00665
H_R (Low–Lou units)	127.983	91.63	83.75	85.61
$\langle B_{\mathrm{diff}}\rangle$ (G)	0	4.166	4.716	4.414
$\langle B_{\mathrm{diff},z}\rangle$ (G)	0	2.356	0.081	0.081
E2E (s)	—	454.1	2532.7	3210.8
GPU (kJ)	—	17.07	182.26	234.18
Node (kJ)	—	340.66	1866.09	2349.06

Table 4: Matched-hardware Low–Lou quality–cost comparison at $N = 400$.

Metric	Jarolim et al. (2023)	Ours
θ_J (deg)	7.1115	8.7684
Normalised divergence	3.3365×10^{-4}	5.9198×10^{-4}
E (10^{33} erg)	0.9933	1.0754
$E - E_{\mathrm{pot}}$ (10^{32} erg)	1.8349	1.2735
E/E_{pot}	1.2266	1.1343

Table 5: Comparison of the AR 11158 NF2 result with Jarolim et al. (2023).

Metric	NF2	CFIT P	CFIT N
$\theta_{J,\mathrm{int}}$ (deg)	8.768	7.399	14.828
$\langle\lvert f_i\rvert\rangle_{\mathrm{int}}$	5.92×10^{-4}	1.80×10^{-4}	2.59×10^{-4}
$\langle B_{\mathrm{diff}}\rangle_{\mathrm{obs}}$ (G)	39.47	91.27	87.99
$\langle B_{\mathrm{diff},z}\rangle_{\mathrm{obs}}$ (G)	19.96	0.76	0.74
E (10^{33} erg)	1.0754	1.0959	1.0786
$E - E_{\mathrm{pot}}$ (10^{32} erg)	1.2735	1.4778	1.3056
E/E_{pot}	1.1343	1.1559	1.1377
E_{div}/E	0.23837	0.06710	0.06811
H_R (10^{42} Mx^2)	6.939	15.726	11.543
E_{job} (MJ)	4.669	9.361	

Table 6: Metrics for AR 11158 extrapolations on a common evaluation grid. E_{job} accounting scopes are listed in Table B.1.

Pair	Relative L_2	C_{vec}
NF2–CFIT P	0.4743	0.8865
NF2–CFIT N	0.4449	0.9009
CFIT P–CFIT N	0.1495	0.9889

Table 7: Pairwise field agreement on the common HARP 377 volume.

divides both physics losses by the mean $|\mathbf{B}|^2$ at each height, amplifying their contribution in weak-field layers. Its lower E_{div}/E but poorer force-free consistency than CFIT is qualitatively consistent with this balance, and the larger boundary loss weighting is consistent with the closer NF2 boundary fit. This is also the only observational NF2 case using height scaling and the only one below the E_{div}/E limit, making the scaling a plausible contributor - however we do not claim it is the only reason as the active regions, domains and training hyperparameters differ.

Across the pair of extrapolations immediately before and after the flare, NF2 gives a small decline in free energy and relative helicity whereas both CFIT branches give large increases. Across the three 150,000-step NF2 runs, seed-to-seed variation is much smaller than the NF2–CFIT separation, so random initialisation alone does not explain the incompatible changes.

Both methods place the strongest height-integrated current in the active region core traced by the 171-Å emission, but they give different surrounding topologies and different pre- to post-flare changes (Figure 5).

The total energy evolution follows the flux-driven potential baseline, while the methods give incompatible free-energy and helicity histories. Different reconciliation of inconsistent polarity constraints can produce this sensitivity (Wheatland & Régnier, 2009; Mastrano et al., 2020); we therefore cannot determine a valid flare-energy or helicity change from these extrapolations.

We compare the pre-to-post magnetic-energy changes from our extrapolations with those re-

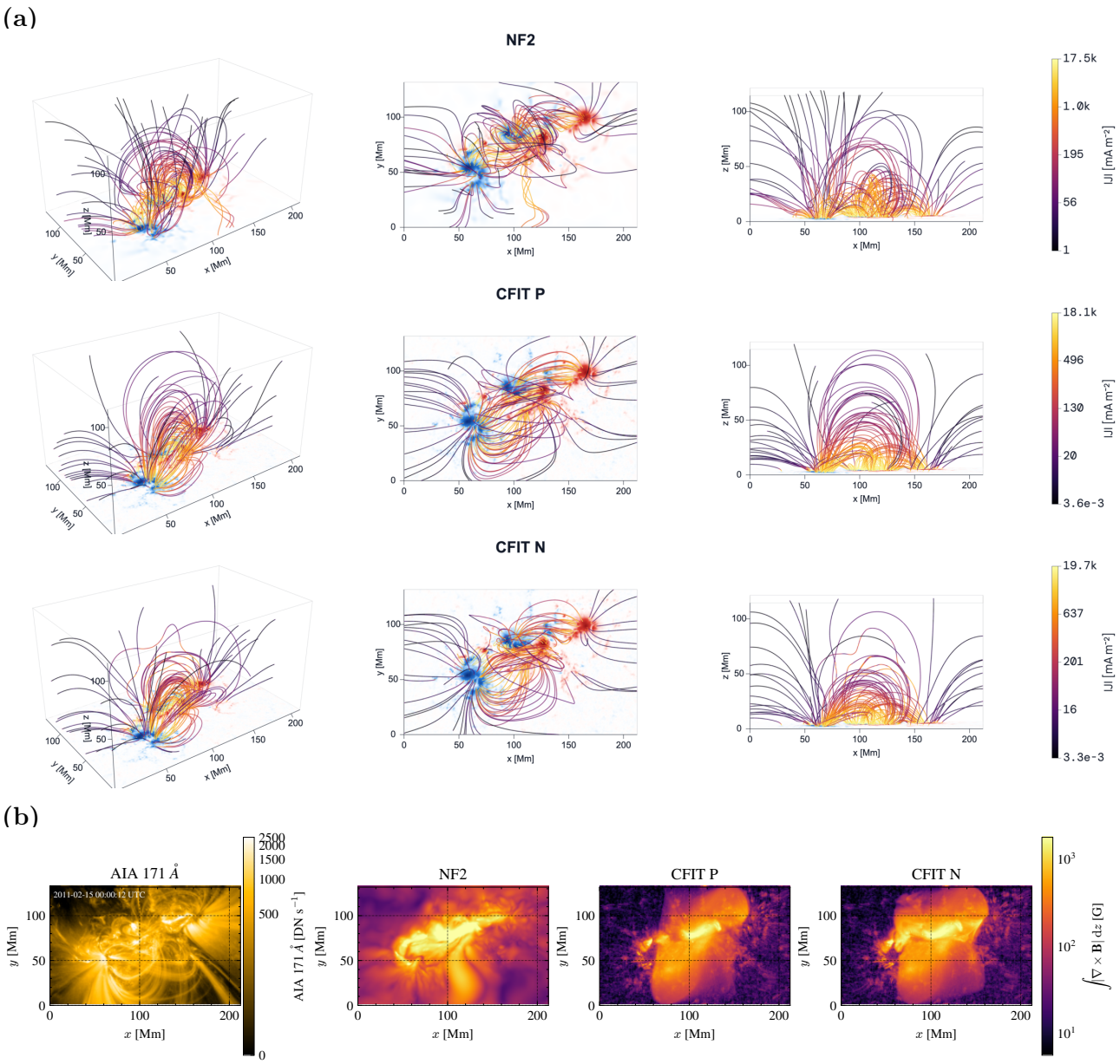


Figure 4: AR 11158 extrapolations on the common grid. **(a)** Selected field lines for NF2, CFIT P and CFIT N from top to bottom; columns show perspective, top and side views, with physical distances in Mm. Lines originate from the same lower-boundary seed locations with $|B_z| \geq 500$ G. **(b)** Atmospheric Imaging Assembly (AIA) 171-Å emission and height-integrated current density magnitude. The HMI `T_REC` is 2011-02-15 00:00:00 TAI and the AIA `DATE-OBS` is 2011-02-15 00:00:12.341 UTC. The AIA image is reprojected to the cropped, binned SHARP CEA grid and the three current maps use one logarithmic colour scale. The conditioned CFIT branches are visually more similar to each other than either is to NF2, consistent with the pairwise field scores. Field-line density is a visualisation choice, while the optically thin, The EUV emission images provides context rather than direct validation.

ported by Jarolim et al. (2024) from NF2 extrapolations of the same active region. Both comparisons use the first 12-minute records outside the GOES X2.2 interval, corresponding to 08:36 and 09:48 TAI. For our fields, we subtract the pre-flare magnetic-energy density from the post-flare value at each voxel and integrate over the common cropped volume (Table 8). Our direct-$\mathbf{B}_\theta$ NF2 extrapolations agree broadly in sign and scale with the published vector-potential NF2 results, for which $\mathbf{B}_\theta = \nabla \times \mathbf{A}_\theta$, whereas CFIT gives a substantially larger net increase. Exact agreement is not expected because the extrapolations and spatial domains are not identical. Note that $|\Delta E^-|$ is an upper estimate of depleted magnetic energy, not a measurement of flare energy.

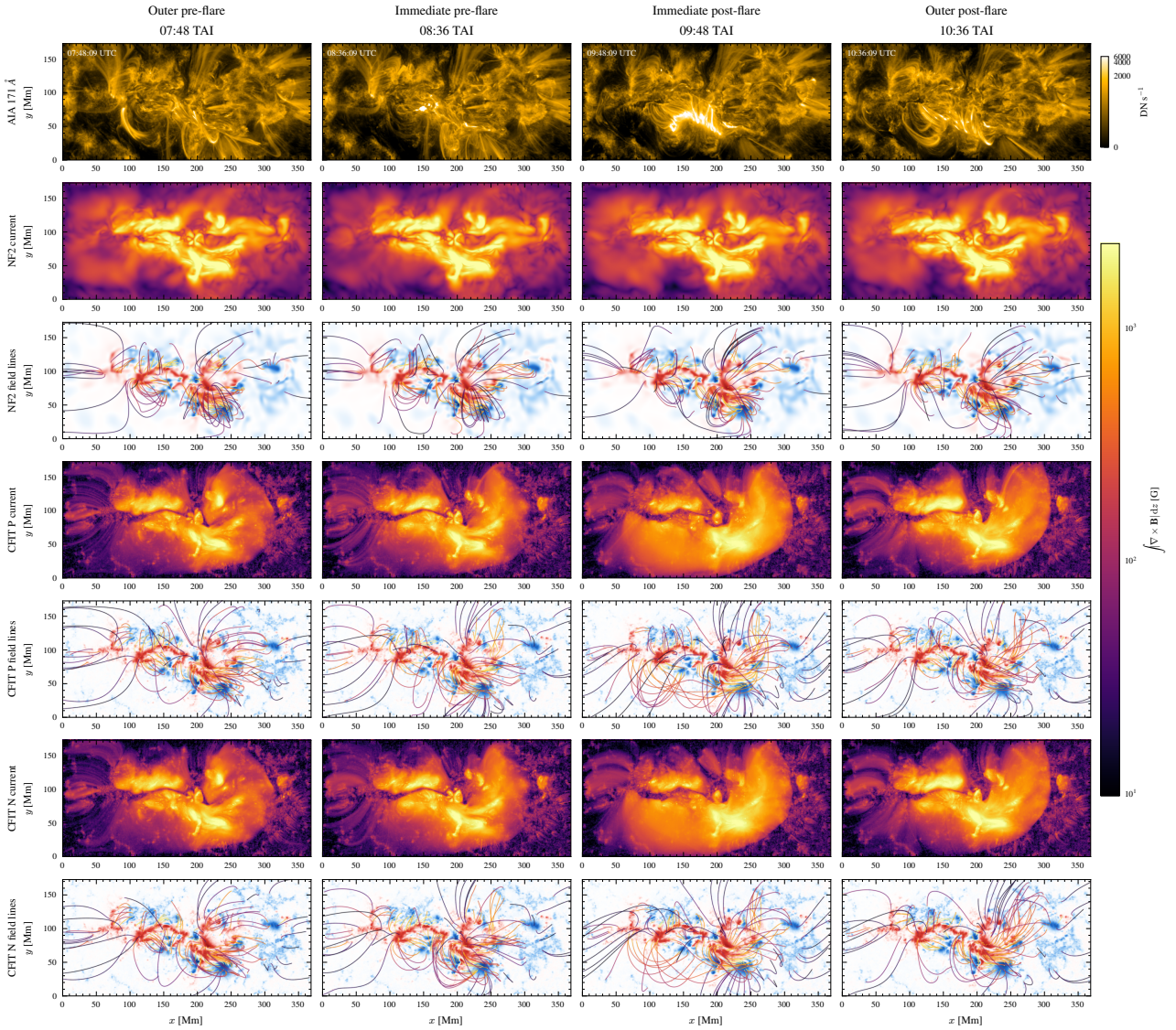


Figure 5: AR 13664 across the GOES X2.2 flare. Columns show HMI `T_REC` values 2024-05-09 07:48:00, 08:36:00, 09:48:00 and 10:36:00 TAI; the corresponding AIA `DATE-OBS` values are 07:48:09.352, 08:36:09.350, 09:48:09.349 and 10:36:09.351 UTC. The middle columns are the first 12-minute records outside the GOES flare interval, while the outer columns provide wider temporal brackets. Rows show exposure-normalised AIA 171-Å emission followed by height-integrated current and top-down field lines for NF2 and the positive- and negative-polarity conditioned CFIT solutions. The shared logarithmic colour bar applies to all twelve current maps. It does not apply to the field-line panels, whose line colours are normalised separately and are qualitative. Line density is purely a visualisation choice. The EUV emission images provides visual context and should not be taken as validation of an extrapolation.

Quantity [10^{31} erg]	Published NF2	This work: NF2	This work: CFIT
$\Delta E_{\mathrm{total}}$	4.08	7.17 [5.19, 9.45]	84.08 [76.79, 91.36]
$\lvert\Delta E^{-}\rvert$	12.16	16.22 [15.38, 17.50]	28.19 [25.83, 30.56]

Table 8: X2.2 energy-change comparison for the 08:36-09:48 TAI cadence pair. This-work entries are means with NF2 seed or CFIT polarity-branch ranges in brackets. Published values are from Jarolim et al. (2024).

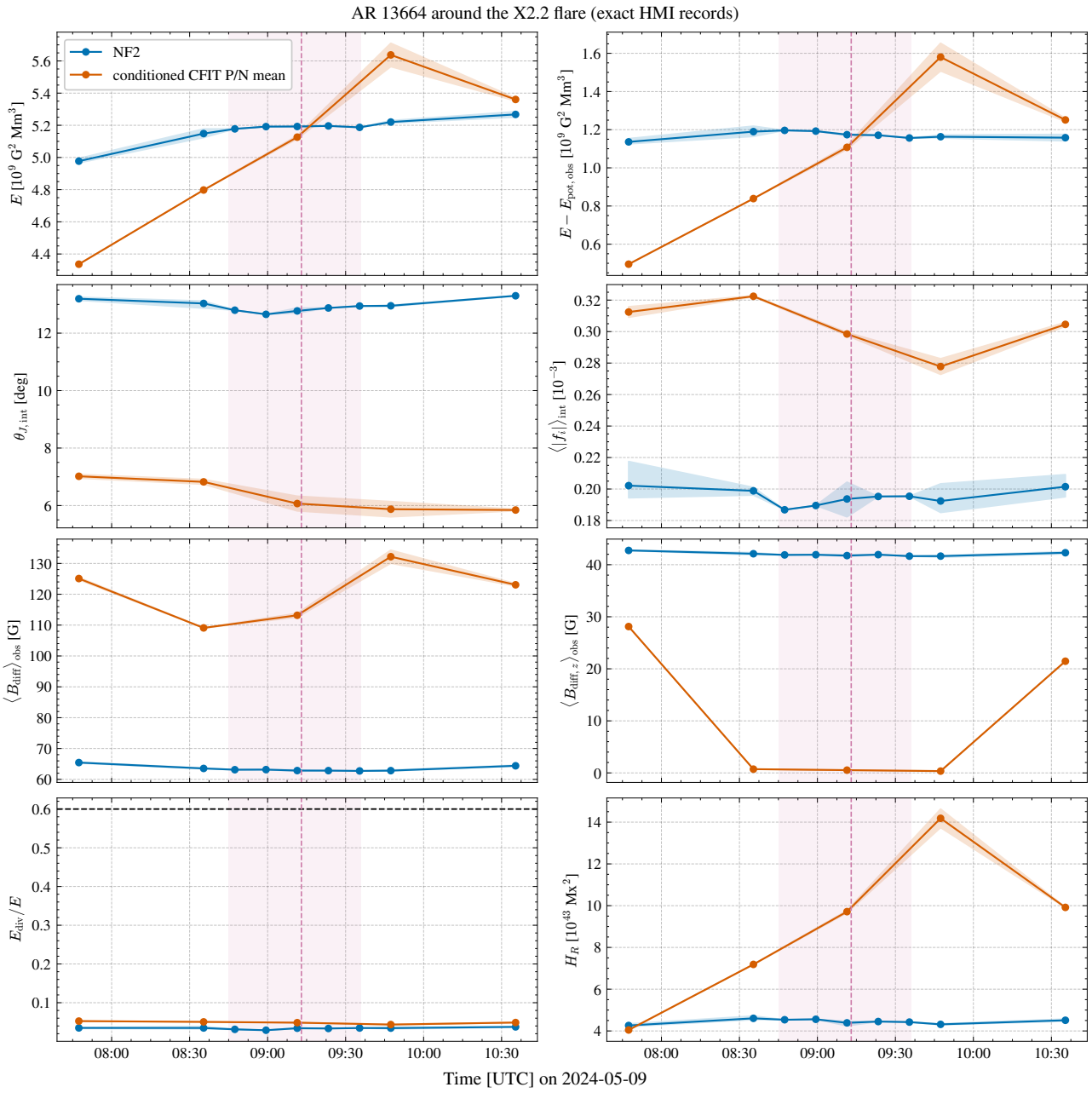


Figure 6: AR 13664 across the GOES X2.2 flare. The panels show total and free magnetic energy, force-free angle, fractional flux, full vector and normal component lower boundary discrepancies, E_{div}/E, and relative helicity. Curves show means and highlights show standard deviations spanning three NF2 seeds or the two conditioned CFIT branches. The red shaded interval marks the GOES flare and the dashed line its peak; the E_{div}/E panel marks our quality limit of 0.6.

5.5 AR 13848

The total and potential energies decline together, while E_{free} is usually negative (Figure 7). Our E_{div}/E quality limit is met by 235 of the 584 sequential fields and six of the nine independent controls, including the sequential pre-flare, peak and post-flare fields around the GOES X1.8 event. Meeting this limit does not automatically make a negative E_{free} physically interpretable. Some rejected fields nevertheless have small force-free and divergence residuals, confirming that these diagnostics measure different numerical errors. Here the force-free and divergence coefficients are equal, but during continuation the bottom and side boundary loss weight coefficients are respectively 10 and 100 times larger than either physics loss weight. Unlike the AR 13664 extrapolations, these losses are not rescaled to compensate for the field weakening with height. The force-free angle also spikes when continuation

begins, with simultaneous changes in the energy, E_{div}/E and helicity curves. Rerunning the opening frames while tracking the individual loss terms would show whether the first continued fields are under-optimised. A useful test would anneal the number of optimiser steps per frame from a larger initial value towards 2,000 as the sequence settles.

For each extrapolation, the potential-field reference E_{pot} is constructed from that extrapolation's own lower B_z. An apparently negative $E_{\mathrm{free}} = E - E_{\mathrm{pot}}$ does not therefore imply a physical negative energy reservoir but rather signals a failure of the assumptions behind Thomson's theorem, commonly driven by a large negative mixed-energy term involving the nonsolenoidal current-carrying field (Valori et al., 2013). Conversely, a positive energy difference is not sufficient evidence of reliability, because nonsolenoidal terms can cancel while remaining comparable to the inferred free energy.

The independently trained controls do not systematically reduce the E_{div}/E metric despite receiving 50 times as many optimiser steps. They do reduce $\langle B_{\mathrm{diff}} \rangle_{\mathrm{obs}}$ at every sampled frame, but generally are less force-free and more divergent. The absence of a consistent ordering in E_{div}/E between the $2,000$-step continuations and $100,000$-step controls points to the balance of loss terms or field representation rather than optimiser steps alone, although separating those effects requires further study.

6 CONCLUSION

Neither NF2 nor CFIT is uniformly the more accurate extrapolation method. On the matched $N = 400$ Low–Lou benchmark, both reach a relative field error of about 0.23, but NF2 requires 454 s and 17 kJ of GPU-board energy compared with 2533-3211 s and 182-234 kJ for CFIT. This lower cost does not make the NF2 field better in every respect: CFIT is much more force free, and fields with similar error against the analytical solution still disagree with one another. Refining the CFIT grid reduces its force-free and solenoidal residuals but not its error against the exact field. For NF2, increasing network capacity within a fixed wall-time budget reduces the available optimiser steps and does not monotonically improve the result.

The observed active regions give no basis for a general NF2–CFIT ranking. NF2 follows the full measured boundary vector more closely, whereas CFIT preserves the prescribed normal component and is generally more force free; the ordering of E_{div}/E nevertheless reverses between active regions. Across the AR 13664 flare, NF2 and CFIT infer incompatible energy and helicity changes. For AR 13848, 235 of 584 sequential fields and six of nine independent controls pass the empirical E_{div}/E screen, including the sequential pre-flare, peak and post-flare X1.8 fields. Passing this screen removes one numerical objection to their helicities; it does not resolve the path dependence or provide a method-independent flare-energy estimate.

Our encounter of the FastQSL failure had a narrower engineering cause: unnecessary dynamic allocation inside the tracing kernel exhausted the CUDA heap. Replacing it with private stack variables removed the failure and accelerated the *isolated* kernel by factors of 33-89 over the tested grids.

6.1 Limitations and next steps

Although NF2 reaches comparable global errors at the sampled matched-hardware operating point, with lower end-to-end cost, we have only one point and cannot claim that this behaviour scales. CFITs cost shifts towards field-line tracing as the grid grows: parallelisation is bound by the serial remainder, and we must pay the cost of its mirrored FFT.

The comparison also confounds field representation with outer-boundary treatment: NF2 imposes potential side and upper boundaries, whereas CFIT sets the current to zero on field lines leaving them. The matched cost point has no NF2 ensemble, and runtime and energy remain hardware- and scope-specific. With no volumetric coronal ground truth, inter-method agreement is not observational validation.

Our PINN extrapolations of active regions directly learn the **B** field: vector potential training requires slower second derivatives and was impractical for the number of extrapolations we were exploring; while direct-**B** is a fairer comparison with CFIT because it does not grant NF2 solenoidality by construction. Repeating selected cases with the vector potential form would isolate representation induced divergence.

Useful next steps would include quantitative comparison of the extrapolated magnetic field lines with coronal loops visible in AIA images aligned to the same solar coordinates, promoting these visualisations from largely qualitative illustrations to observational validation; selected repeats using the vector-potential NF2 formulation; and testing the

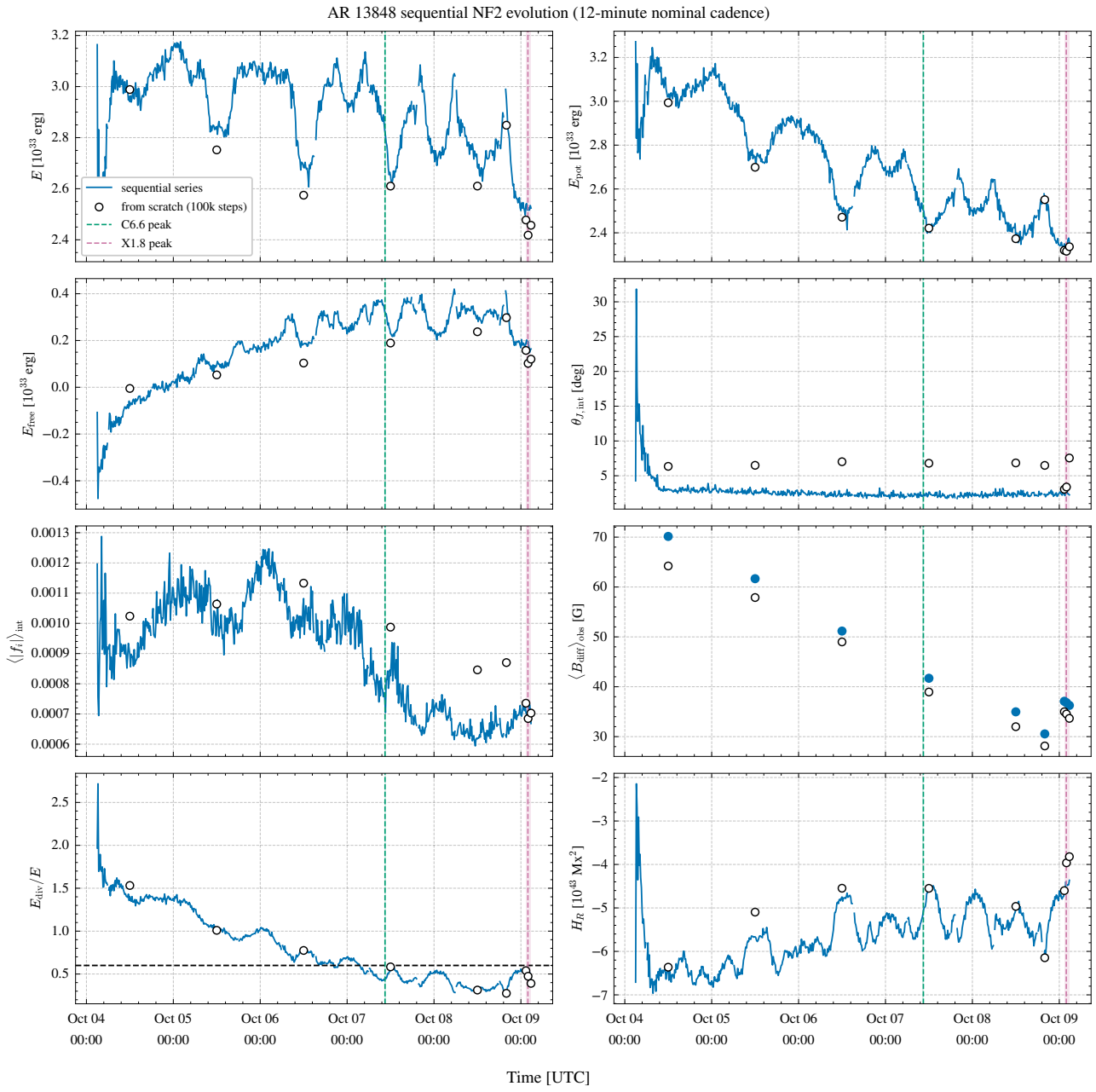


Figure 7: AR 13848 sequential NF2 series. The panels show total, potential and free energy, force-free angle, fractional flux, lower-boundary $\langle B_{\rm diff}\rangle_{\rm obs}$, $E_{\rm div}/E$, and raw FLHtools H_R. The boundary-discrepancy panel contains the nine matched samples for which the boundary residual was evaluated; the other blue curves use the full accepted cadence. The dashed lines mark the GOES C6.6 and GOES X1.8 flares. The $E_{\rm div}/E$ panel marks our quality limit of 0.6. Gaps in the plots are due to records rejected by HMI quality flags and missing observations.

empirical solenoidal-quality calibration on a larger independent field sample.

Recent physics-informed neural-operator (PINO) approaches provide another promising direction. Rather than fitting a new coordinate network for each magnetogram, these methods learn an operator mapping photospheric boundary fields directly to three-dimensional NLFFF solutions, enabling inference in under a second (Jeon et al., 2025) and incorporating spectral operator architectures with explicit force-free and solenoidal losses (Cao et al., 2025). A particularly useful extension of the present work would be to benchmark such models against both NF2 and CFIT on controlled analytical fields. Since existing neural operators are trained primarily on solutions produced by classical NLFFF codes, this should evaluate not only inference speed and pointwise accuracy, but also whether they reproduce the parent solver's systematic errors

and preserve physically important quantities such as magnetic energy, topology and relative helicity.

Finally, the method comparison would benefit from larger matched-hardware ensembles and proper profiling of each method, which was precluded here by the lack of access to GPU hardware-performance counters on the HPC nodes. Our fixed 200,000 seed tracing batch is also an untested implementation parameter. Future work should measure the runtime and memory trade-off across batch sizes and could investigate automatic selection based on problem size, detected hardware, and a short calibration run.

References

Alissandrakis, C. E. 1981, A&A, 100, 197

Amari, T., Boulmezaoud, T. Z., & Aly, J. J. 2006, A&A, 446, 691–705, doi: 10.1051/0004-6361:20054076

Aslanyan, V., Scott, R. B., Wilkins, C. P., et al. 2024, ApJ, 971, 137, doi: 10.3847/1538-4357/ad55ca

Baty, H., & Vigon, V. 2024, MNRAS, 527, 2575, doi: 10.1093/mnras/stad3320

Berger, M. A., & Field, G. B. 1984, Journal of Fluid Mechanics, 147, 133, doi: 10.1017/S0022112084002019

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549–3578, doi: 10.1007/s11207-014-0529-3

Borrero, J. M., Tomczyk, S., Kubo, M., et al. 2010, SoPh, 273, 267–293, doi: 10.1007/s11207-010-9515-6

Burtscher, M., Zecena, I., & Zong, Z. 2014, in Proceedings of Workshop on General Purpose Processing Using GPUs (ACM), 28–36, doi: 10.1145/2576779.2576783

Cao, J., Li, Q., Du, M., Wang, H., & Shen, B. 2025, Physics-informed Attention-enhanced Fourier Neural Operator for Solar Magnetic Field Extrapolations, doi: 10.48550/arXiv.2510.05351

DeRosa, M. L., Schrijver, C. J., Barnes, G., et al. 2009, ApJ, 696, 1780–1791, doi: 10.1088/0004-637x/696/2/1780

DeRosa, M. L., Wheatland, M. S., Leka, K. D., et al. 2015, ApJ, 811, 107, doi: 10.1088/0004-637x/811/2/107

Finn, J. M., & Antonsen, Jr., T. M. 1985, Comments on Plasma Physics and Controlled Fusion, 9, 111

Fogg, A. R., Lucas, A. R., Hayes, L. A., et al. 2026, Journal of Space Weather and Space Climate, 16, 2, doi: 10.1051/swsc/2025044

Gary, G. A. 2001, SoPh, 203, 71–86, doi: 10.1023/a:1012722021820

Grad, H., & Rubin, H. 1958, in Proceedings of the Second United Nations International Conference on the Peaceful Uses of Atomic Energy, Vol. 31, Geneva, 190

Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483–3530, doi: 10.1007/s11207-014-0516-8

Jarolim, R., Thalmann, J. K., Veronig, A. M., & Podladchikova, T. 2023, Nature Astronomy, 7, 1171–1179, doi: 10.1038/s41550-023-02030-9

Jarolim, R., Veronig, A. M., Purkhart, S., Zhang, P., & Rempel, M. 2024, ApJL, 976, L12, doi: 10.3847/2041-8213/ad8914

Jeon, M., Jeong, H.-J., Moon, Y.-J., Kang, J., & Kusano, K. 2025, ApJS, 277, 54, doi: 10.3847/1538-4365/adbaea

Low, B. C., & Lou, Y. Q. 1990, ApJ, 352, 343, doi: 10.1086/168541

Mastrano, A., Yang, K. E., & Wheatland, M. S. 2020, SoPh, 295, 97, doi: 10.1007/s11207-020-01663-7

Metcalf, T. R., Jiao, L., McClymont, A. N., Canfield, R. C., & Uitenbroek, H. 1995, ApJ, 439, 474, doi: 10.1086/175188

Metcalf, T. R., DeRosa, M. L., Schrijver, C. J., et al. 2008, SoPh, 247, 269–299, doi: 10.1007/s11207-007-9110-7

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2011, SoPh, 275, 3–15, doi: 10.1007/s11207-011-9841-3

Prior, C. B., & Yeates, A. R. 2014, ApJ, 787, 100, doi: 10.1088/0004-637X/787/2/100

Raissi, M., Perdikaris, P., & Karniadakis, G. 2019, Journal of Computational Physics, 378, 686–707, doi: 10.1016/j.jcp.2018.10.045

Sakurai, T. 1982, SoPh, 76, doi: 10.1007/bf00170988

Scherrer, P. H., Schou, J., Bush, R. I., et al. 2011, SoPh, 275, 207–227, doi: 10.1007/s11207-011-9834-2

Schrijver, C. J., Derosa, M. L., Metcalf, T. R., et al. 2006, SoPh, 235, 161–190, doi: 10.1007/s11207-006-0068-7

Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., & Wetzstein, G. 2020, in Advances in Neural Information Processing Systems, Vol. 33. https://arxiv.org/abs/2006.09661

Sun, X., Hoeksema, J. T., Liu, Y., et al. 2012, ApJ, 748, 77, doi: 10.1088/0004-637x/748/2/77

Thalmann, J. K., Sun, X., Moraitis, K., & Gupta, M. 2020, A&A, 643, A153, doi: 10.1051/0004-6361/202038921

Toriumi, S., Takasao, S., Cheung, M. C. M., et al. 2020, ApJ, 890, 103, doi: 10.3847/1538-4357/ab6b1f

Valori, G., Démoulin, P., Pariat, E., & Masson, S. 2013, A&A, 553, A38, doi: 10.1051/0004-6361/201220982

Valori, G., Pariat, E., Anfinogentov, S., et al. 2016, SSRv, 201, 147, doi: 10.1007/s11214-016-0299-3

Venkatakrishnan, P., & Ravindra, B. 2003, Geophysical Research Letters, 30, 2181, doi: 10.1029/2003GL018100

Wheatland, M. S. 2006, SoPh, 238, 29–39, doi: 10.1007/s11207-006-0232-0

Wheatland, M. S., & Régnier, S. 2009, ApJ, 700, L88, doi: 10.1088/0004-637X/700/2/L88

Wheatland, M. S., Sturrock, P. A., & Roumeliotis, G. 2000, ApJ, 540, 1150–1155, doi: 10.1086/309355

Wiegelmann, T. 2004, SoPh, 219, 87–108, doi: 10.1023/b:sola.0000021799.39465.36

Wiegelmann, T., Inhester, B., & Sakurai, T. 2006, SoPh, 233, 215–232, doi: 10.1007/s11207-006-2092-z

Wiegelmann, T., & Sakurai, T. 2021, Living Reviews in Solar Physics, 18, doi: 10.1007/s41116-020-00027-4

Yang, Z., Adamek, K., & Armour, W. 2024, in SC24: International Conference for High Performance Computing, Networking, Storage and Analysis (IEEE), 1–17, doi: 10.1109/SC41406.2024.00028

Yeates, A. R., & Page, M. H. 2018, Journal of Plasma Physics, 84, 775840602, doi: 10.1017/S0022377818001204

Zhang, P., Chen, J., Liu, R., & Wang, C. 2022, ApJ, 937, 26, doi: 10.3847/1538-4357/ac8d61

A TOY-PINN RESULTS

The following problems form a ladder from scalar differential equations to a force-free magnetic configuration to which we apply PINNs. Figures A.1 and A.2 show the training progression for each experiment.

The first pair deliberately uses the same tanh network and exact solution, $u(x) = \sin(\pi x)$ on $x \in [0, 6]$, but supplies different residuals:

$$u'' + \pi^2 u = 0, \tag{A.1}$$
$$u(0) = 0, \quad u'(0) = \pi,$$
$$u'' = -\pi^2 \sin(\pi x), \qquad u(0) = u(6) = 0, \tag{A.2}$$

The oscillator equation fixes the frequency but not the amplitude: its general solution is $A\sin(\pi x) + B\cos(\pi x)$. The two initial conditions select $B = 0$ and $A = 1$, but they enter the PINN as finite penalties at $x = 0$ rather than exact constraints propagated across the domain. The trained oscillation therefore loses amplitude away from the initial point. For the Poisson problem, the same trial prediction gives $r = \pi^2(1 - A)\sin(\pi x)$, so every interior point penalises the wrong amplitude and the network recovers all three periods. The different outcomes arise from the information supplied by the residual, not from the network's ability to represent the target.

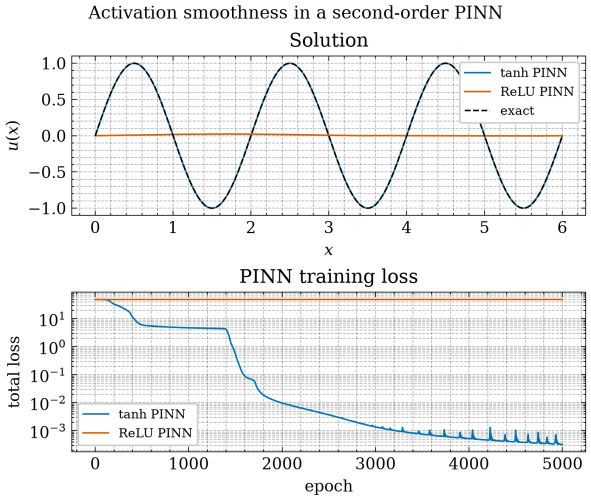
The two-dimensional Poisson problem uses $u = \sin(\pi x)\sin(\pi y)$ on $[0, 6]^2$. With width, depth, samples and optimiser fixed, the networks differ only in activation. At epoch 500 the tanh model still contains broad diagonal bands, while the NF2 SIREN has resolved the alternating cells and reaches a substantially lower final error. In this controlled comparison, the sinusoidal representation exhibits less spectral bias. A separate second-order Poisson test confirms the smoothness requirement discussed above: the tanh PINN learns the solution and reduces its loss, whereas ReLU's zero curvature almost everywhere leaves the physics loss high.

The final problem replicates the three linear force-free coronal arcades of Baty & Vigon (2024). We follow their prescribed equation, domain, test cases and numerical configuration. Because no random seed or sampled coordinates were published, both models use the same seeded realisations of the paper's fixed boundary and collocation datasets. They receive the raw coordinates (x, z) and differ only in activation and its associated initialisation: the paper's tanh network is compared with the NF2 SIREN architecture. The flux function satisfies

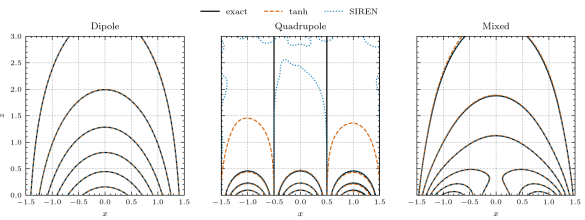
$$\nabla^2\psi + c^2\psi = 0, \qquad x \in [-L/2, L/2], \quad z \in [0, L]. \tag{A.3}$$

On an independent evaluation grid, the tanh network is more accurate for the dipole, whereas SIREN is more accurate for the quadrupole and mixed fields; the largest difference occurs for the higher-frequency quadrupole.

Together, these toy problems motivate rather than validate NF2's architecture: smooth activations are required by derivative-based losses, while sinusoidal activations can better represent higher-frequency spatial structure. Whether these advantages carry over to nonlinear force-free extrapolation is tested against the Low–Lou field in Section 5.2.



(a) Activation smoothness in a second-order PINN

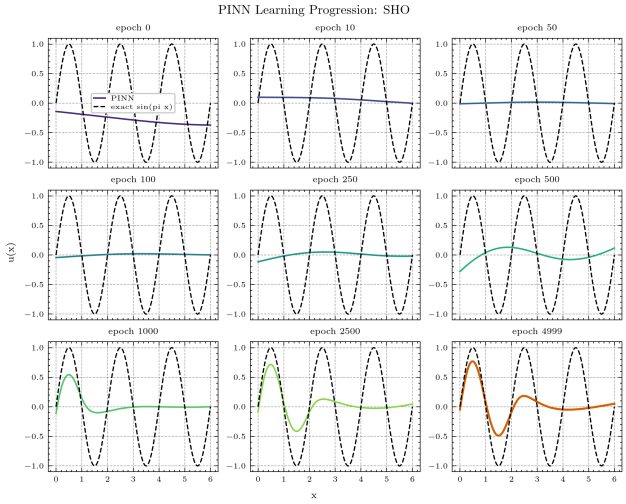


(b) Baty–Vigon coronal arcade replication

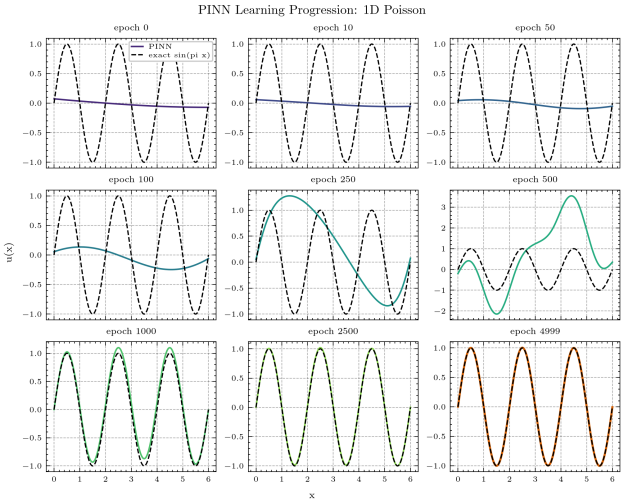
Figure A.1: Activation and coronal-arcade tests: (a) the effect of activation smoothness and (b) the arcade replication.

B METHODS AND METRICS

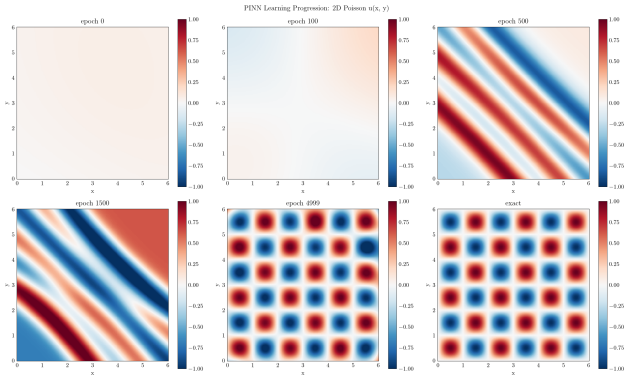
This appendix records the algorithm summaries, allocated hardware and full metric definitions used to generate the results in the main text.



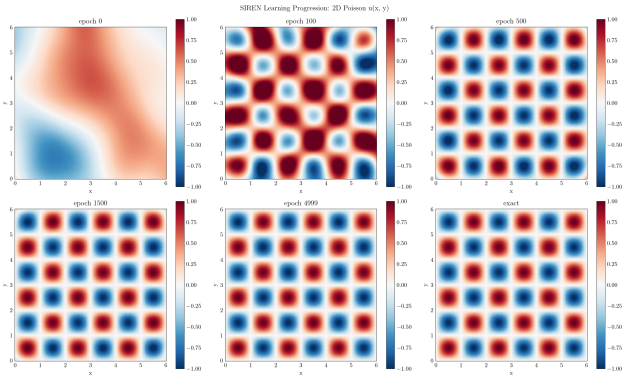
(a) Simple harmonic oscillator learning progression



(b) One-dimensional Poisson learning progression



(c) Two-dimensional Poisson learning with tanh



(d) Two-dimensional Poisson learning with SIREN

Figure A.2: Toy-PINN learning progression. (a,b) The shared one-dimensional target under two residuals; (c,d) spectral learning with tanh and SIREN.

Algorithm B.1 NF2 observational direct-**B** training

Require: Boundary field $\mathbf{B}_{\text{obs}}$, potential outer-boundary field $\mathbf{B}_{\text{pot}}$, network $\mathbf{B}_\theta(\mathbf{x})$

repeat
  Sample lower-boundary points $\mathbf{x}_b$, outer-boundary points $\mathbf{x}_p$ and interior points $\mathbf{x}_i$
  Evaluate $\mathbf{B}_b \leftarrow \mathbf{B}_\theta(\mathbf{x}_b)$
  Evaluate $\mathbf{B}_{\theta,p} \leftarrow \mathbf{B}_\theta(\mathbf{x}_p)$
  Evaluate $\mathbf{B}_i \leftarrow \mathbf{B}_\theta(\mathbf{x}_i)$
  Compute $\nabla\times\mathbf{B}_i$ and $\nabla\cdot\mathbf{B}_i$ using automatic differentiation
  $\mathcal{L}_{\text{boundary}} \leftarrow \|\mathbf{B}_b - \mathbf{B}_{\text{obs}}\|^2$
  $\mathcal{L}_{\text{potential}} \leftarrow \|\mathbf{B}_{\theta,p} - \mathbf{B}_{\text{pot}}\|^2$
  $\mathcal{L}_{\text{ff}} \leftarrow \|(\nabla \times \mathbf{B}_i) \times \mathbf{B}_i\|^2$
  $\mathcal{L}_{\text{div}} \leftarrow \|\nabla \cdot \mathbf{B}_i\|^2$
  $\mathcal{L} \leftarrow \lambda_b\mathcal{L}_{\text{boundary}} + \lambda_p\mathcal{L}_{\text{potential}} + \lambda_{\text{ff}}\mathcal{L}_{\text{ff}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$
  Update θ using Adam and $\nabla_\theta\mathcal{L}$
until the training budget is exhausted
return $\mathbf{B}_\theta$

Algorithm B.2 CFIT Grad–Rubin extrapolation

Require: Photospheric field $\mathbf{B}_{\text{obs}}$, polarity $= \pm 1$

Select the prescribed-polarity pixels from B_z
Compute $\alpha_0 = (\partial B_y/\partial x - \partial B_x/\partial y)/B_z$
Compute the potential field $\mathbf{B}_0$ from B_z
$\mathbf{B} \leftarrow \mathbf{B}_0$
repeat
  Trace field lines through $\mathbf{B}$ with numerical integration
  Map α_0 from their photospheric footpoints into the volume
  Compute the current $\mathbf{J} \leftarrow \alpha\mathbf{B}$
  Reflect $\mathbf{J}$ across the lower boundary as in Wheatland (2006), duplicating the lower layer
  Solve for $\mathbf{B}_J$ using FFTs, with $B_{J,z}(z = 0) \simeq 0$ on the discrete grid
  Update $\mathbf{B} \leftarrow \mathbf{B}_0 + \mathbf{B}_J$
until 30 iterations are completed, or the force-free angle decreases by less than 1% from the previous iteration without increasing
return $\mathbf{B}$

B.1 Extrapolation methods workflow

B.2 Full metric definitions

Let Ω contain the M points of the common evaluation grid, $\Omega_{\rm obs}$ the $M_{\rm obs}$ lower-boundary points and $\Omega_{\rm int}$ the $M_{\rm int}$ points remaining after all six boundary faces are removed. For a sampled vector field $\mathbf{X}$, define

$$\|\mathbf{X}\|_2 \equiv \left(\sum_{i\in\Omega}\|\mathbf{X}_i\|^2\right)^{1/2}, \tag{B.1}$$

$$\|\mathbf{X}\|_{2,\rm int} \equiv \left(\sum_{i\in\Omega_{\rm int}}\|\mathbf{X}_i\|^2\right)^{1/2}, \tag{B.2}$$

$$\|\mathbf{X}\|_{2,\rm obs} \equiv \left(\sum_{i\in\Omega_{\rm obs}}\|\mathbf{X}_i\|^2\right)^{1/2}. \tag{B.3}$$

Here $\|\mathbf{X}_i\|$ is the Euclidean vector magnitude; the same notation applies to a scalar component, with absolute value in place of vector magnitude. We denote a reconstructed field by $\mathbf{B}$ and the exact Low–Lou field by $\mathbf{B}^\star$.

Low–Lou recovery. The relative field and interior-current errors are

$$\epsilon_B = \frac{\|\mathbf{B}-\mathbf{B}^\star\|_2}{\|\mathbf{B}^\star\|_2}, \tag{B.4}$$

$$\epsilon_{J,\rm int} = \frac{\|\mathbf{J}-\mathbf{J}^\star\|_{2,\rm int}}{\|\mathbf{J}^\star\|_{2,\rm int}}. \tag{B.5}$$

The latter exposes derivative errors that a field-only score can miss.

The standard vector-comparison metrics of Schrijver et al. (2006) are

$$C_{\rm vec} = \frac{\sum_i \mathbf{B}_i\cdot\mathbf{B}_i^\star}{\|\mathbf{B}\|_2\|\mathbf{B}^\star\|_2}, \tag{B.6}$$

$$C_{\rm CS} = \frac{1}{M}\sum_i \frac{\mathbf{B}_i\cdot\mathbf{B}_i^\star}{\|\mathbf{B}_i\|\|\mathbf{B}_i^\star\|}, \tag{B.7}$$

$$E_n' = 1 - \frac{\sum_i\|\mathbf{B}_i-\mathbf{B}_i^\star\|}{\sum_i\|\mathbf{B}_i^\star\|}, \tag{B.8}$$

$$E_m' = 1 - \frac{1}{M}\sum_i \frac{\|\mathbf{B}_i-\mathbf{B}_i^\star\|}{\|\mathbf{B}_i^\star\|}. \tag{B.9}$$

$C_{\rm vec}$ is a global vector correlation and $C_{\rm CS}$ is the mean local directional cosine. E_n' is the complement of a globally normalised L_1 error; E_m' averages locally normalised errors and is therefore especially sensitive to weak-field regions. Unity denotes perfect agreement for all four metrics. They are not independent, but are retained for comparison with earlier extrapolation studies.

Force-free consistency. The local current–field misalignment and its current-weighted mean are

$$\sin\vartheta_i = \frac{\|\mathbf{J}_i\times\mathbf{B}_i\|}{\|\mathbf{J}_i\|\|\mathbf{B}_i\|}, \tag{B.10}$$

$$\sigma_{J,\rm int} = \frac{\sum_i\|\mathbf{J}_i\|\sin\vartheta_i}{\sum_i\|\mathbf{J}_i\|}, \tag{B.11}$$

$$\theta_{J,\rm int} = \arcsin\sigma_{J,\rm int}. \tag{B.12}$$

The sums are over $\Omega_{\rm int}$. The weighting gives greater influence to regions carrying stronger current. Zero denotes ideal alignment; the sine and angle are two forms of one metric rather than independent scores (Wheatland, 2006; Schrijver et al., 2006; Jarolim et al., 2023).

Solenoidality. The fractional-flux metric of Wheatland et al. (2000) is

$$f_i = \frac{\oint_{\partial V_i}\mathbf{B}\cdot d\mathbf{S}}{\oint_{\partial V_i}\|\mathbf{B}\|\,dS}, \qquad \langle|f_i|\rangle_{\rm int} = \frac{1}{M_{\rm int}}\sum_i|f_i|. \tag{B.13}$$

For a uniform cubic grid of cell length Δ,

$$\langle|f_i|\rangle_{\rm int} \simeq \frac{1}{M_{\rm int}}\sum_i\frac{\Delta|\nabla\cdot\mathbf{B}_i|}{6\|\mathbf{B}_i\|}. \tag{B.14}$$

This is the mean fraction of unsigned cell flux that fails to balance. It is dimensionless and vanishes for a perfectly solenoidal field; the factor 1/6 accounts for the six faces of a cubic cell (DeRosa et al., 2009; Valori et al., 2013).

Magnetic energy. For voxel volume ΔV,

$$E = \frac{\Delta V}{2\mu_0}\|\mathbf{B}\|_2^2, \tag{B.15}$$

$$E^\star = \frac{\Delta V}{2\mu_0}\|\mathbf{B}^\star\|_2^2, \qquad q_E = \frac{E}{E^\star} = \frac{\|\mathbf{B}\|_2^2}{\|\mathbf{B}^\star\|_2^2}. \tag{B.16}$$

$E^\star$ is the energy of the exact Low–Lou field, not a potential field. The ratio tests integrated field-squared energy but cannot establish correct direction or topology (Schrijver et al., 2006).

For observational fields, a common potential field constructed from $B_{{\rm obs},z}$ supplies the reference:

$$q_{\rm pot} = \frac{E}{E_{\rm pot}}, \qquad E_{\rm free} = E - E_{\rm pot}, \tag{B.17}$$

$$\eta_{\rm free} = \frac{E_{\rm free}}{E_{\rm pot}}. \tag{B.18}$$

Because $q_{\rm pot} = 1 + \eta_{\rm free}$, those ratios are not independent scores.

B.2.1 Relative magnetic helicity

Ordinary magnetic helicity is gauge dependent when magnetic flux crosses the computational boundary. We therefore use the relative helicity (Finn & Antonsen, 1985; Prior & Yeates, 2014)

$$H_{\rm R} = \int_V (\mathbf{A} + \mathbf{A}_{\rm p}) \cdot (\mathbf{B} - \mathbf{B}_{\rm p})\, {\rm d}V, \tag{B.19}$$

where $\nabla \times \mathbf{A} = \mathbf{B}$ and $\nabla \times \mathbf{A}_{\rm p} = \mathbf{B}_{\rm p}$. The potential reference $\mathbf{B}_{\rm p}$ shares the normal field of $\mathbf{B}$ on all six faces. FLHTOOLS (Yeates & Page, 2018) constructs the potential reference as $\mathbf{B}_{\rm p} = \nabla\phi$, where ϕ satisfies Laplace's equation with $\partial\phi/\partial n = \mathbf{B} \cdot \hat{\mathbf{n}}$ prescribed on all six faces. It constructs $\mathbf{A}$ and $\mathbf{A}_{\rm p}$ in the upward DeVore gauge, $A_z = A_{{\rm p},z} = 0$, with the remaining freedom fixed by imposing a two-dimensional Coulomb condition on the lower-boundary vector potentials. If the input has non-zero net boundary flux, it applies a uniform correction before the solve, so the normal-field match is to the corrected boundary data. Equation (B.19) is gauge invariant when both fields are solenoidal and their normal components agree. It is a global, signed quantity, so contributions of opposite handedness may cancel. Inputs in gauss and megametres are converted using $1\,{\rm G}^2\,{\rm Mm}^4 = 10^{32}\,{\rm Mx}^2$.

Solenoidal quality. Finite numerical divergence can contaminate relative helicity. We use the FLHTOOLS reconstruction to define

$$\mathbf{B}_{\rm s} = \nabla \times \mathbf{A}, \qquad \mathbf{B}_{\rm ns} = \mathbf{B} - \mathbf{B}_{\rm s}, \tag{B.20}$$

$$\mathbf{B}_{\rm J,s} = \mathbf{B}_{\rm s} - \mathbf{B}_{\rm p}. \tag{B.21}$$

Writing $\mathcal{E}[\mathbf{X}] = (8\pi)^{-1} \int_V |\mathbf{X}|^2\, {\rm d}V$, we take

$$E = \mathcal{E}[\mathbf{B}], \qquad E_{\rm p,s} = \mathcal{E}[\mathbf{B}_{\rm p}], \tag{B.22}$$

$$E_{\rm J,s} = \mathcal{E}[\mathbf{B}_{\rm J,s}], \qquad E_{\rm J,ns} = \mathcal{E}[\mathbf{B}_{\rm ns}], \tag{B.23}$$

$$E_{\rm p,ns} = 0. \tag{B.24}$$

Thus the proxy treats $\mathbf{B}_{\rm p}$ as solenoidal and assigns the energy of $\mathbf{B}_{\rm ns} = \mathbf{B} - \nabla \times \mathbf{A}$ to $E_{\rm J,ns}$. The component energies do not generally sum to E because the corresponding fields have non-zero cross terms. Following Valori et al. (2013) we define their difference from E as the mixed term,

$$E_{\rm mix} = E - (E_{\rm p,s} + E_{\rm J,s} + E_{\rm p,ns} + E_{\rm J,ns}), \tag{B.25}$$

$$E_{\rm div} = E_{\rm p,ns} + E_{\rm J,ns} + |E_{\rm mix}|. \tag{B.26}$$

For fields analysed with the full Valori decomposition, Thalmann et al. (2020) recommend the joint thresholds

$$\frac{E_{\rm div}}{E} < 0.08, \qquad \frac{|E_{\rm mix}|}{E_{\rm J,s}} < 0.35. \tag{B.27}$$

These limits are not applied directly to our derived metric. By default, FLHTOOLS replaces the input field with the solenoidal reconstruction $\nabla \times \mathbf{A}$. We instantiate with the parameter `clean=False` so that the helicity is evaluated for the original extrapolation. Because $E_{\rm div}$ includes $|E_{\rm mix}|$, it is not a disjoint energy fraction and may exceed E.

Calibration. This is not the exact decomposition of Valori et al. (2013), which applies a separate Helmholtz decomposition to the potential and current-carrying fields. Our metrics had already been calculated using the FLHTOOLS path when this distinction was established, and recomputing every extrapolated field with a separate Helmholtz decomposition was not feasible within the remaining project time. We therefore used the decomposition on a representative subset to calibrate, rather than replace, our metric.

This comprised of a selected subset of 17 fields: three Low–Lou fields, five AR 11158 extrapolations and nine independent AR 13848 extrapolations. Twelve satisfied the reference criterion $E_{\rm div}/E < 0.1$ and five did not. For our metric, the largest value in the accepted group was 0.583 and the smallest value in the rejected group was 0.775. The limit

$$\left(\frac{E_{\rm div}}{E}\right)_{\rm FLHTOOLS} \leq 0.6 \tag{B.28}$$

therefore reproduced all 17 reference classifications. It is a rough empirical screen for this metric and we make no claim that the reported values are equivalent to those obtained using the full Valori decomposition.

Observational boundary discrepancy. Let $\mathbf{s}_i = (s_{i,x}, s_{i,y}, s_{i,z})$ be the measured lower-boundary field component uncertainties. Following Jarolim et al. (2023),

$$B_{{\rm diff},i} = \left\| \max(|\mathbf{B}_i - \mathbf{B}_{{\rm obs},i}| - \mathbf{s}_i, \mathbf{0}) \right\|,$$
$$\langle B_{\rm diff} \rangle_{\rm obs} = \frac{1}{M_{\rm obs}} \sum_{i \in \Omega_{\rm obs}} B_{{\rm diff},i}. \tag{B.29}$$

The absolute value, subtraction and maximum act componentwise. Thus each field component is

charged only for the part of its residual exceeding its quoted uncertainty. The remaining three components form one vector magnitude per pixel, and those magnitudes are averaged over the boundary This is a mean vector error, not an RMSE or the squared norm used for the NF2 boundary training loss.

To isolate agreement in the lower-boundary normal component, we additionally report

$$\langle B_{\mathrm{diff},z}\rangle_{\mathrm{obs}} = \frac{1}{M_{\mathrm{obs}}} \sum_{i\in\Omega_{\mathrm{obs}}} \max(|B_{z,i} - B_{z,\mathrm{ref},i}| - s_{z,i}, 0)\,. \tag{B.30}$$

Here $B_{z,\mathrm{ref}} = B_{\mathrm{obs},z}$ for observations and $B_{z,\mathrm{ref}} = B_z^\star$ for Low–Lou.

For Low–Lou fields, Ω_{obs} is the complete lower face, $s_{z,i} = 0$, and the we apply normalisation setting one field code unit to one gauss.

B.3 Hardware inventory

Runtime and energy measurements depend on the machine hardware and job allocation. Table B.1 contains a lists of utilised hardware. Slurm CPU counts are schedulable logical CPUs, and memory entries are job requests rather than the nodes' installed capacity.

Peak arithmetic capability. Manufacturer specifications for the hardware in Table B.2 are available for NVIDIA V100,[6] A100,[7] H100,[8] Grace[9] and GTX 1050;[10] Intel Xeon Gold 6430;[11] and AMD EPYC 9354[12] EPYC 7702[13] and EPYC 9754.[14] An asterisk marks a derived theoretical ceiling. CPU values use either the complete host or the stated subset, together with the manufacturer base clock, vector width and maximum FMA rate so are more assumption dependent than the quoted GPU peaks. These figures are architectural ceilings; they assume that every listed core executes full-width fused multiply-adds on every cycle. Field-line tracing does not satisfy that assumption. Each Runge–Kutta step depends on the previous step, samples the field at data-dependent locations, branches at boundaries and may trace a different number of steps. Peak rates should be interpreted only as hardware context and not to normalise runtimes across algorithms.

The two tracers also use different precision. UFiT stores the field and performs the integration in FP64. The main FastQSL interpolation and Runge–Kutta path uses FP32, with FP64 retained for selected accumulators such as line length and twist. The relevant V100 peak for most FastQSL instructions is therefore about 14 TFLOP s^{-1} in FP32, rather than the 7.0 TFLOP s^{-1} FP64 value shown in Table B.2. Tensor-core rates are not relevant to either tracer.

Table B.3 show GPU-utilisation samples for selected Low–Lou and active-region runs. These are whole job averages and do not fully represent kernel occupancy; the CFIT values include CPU field updates between FastQSL tracing calls. Samples were recorded every 1 s and 10 s, respectively, and ranges span run means rather than individual samples. Comparable records were not retained for the AR 11158 or AR 13664 extrapolations.

Why peak FP64 does not predict tracer runtime. Peak FP64 throughput is a poor predictor because field-line tracing is not dense arithmetic. Each line is an ordered sequence of Runge–Kutta steps with variable length. UFiT advances each line serially on an OpenMP thread, whereas FastQSL keeps many CUDA threads in flight and can execute other warps while one waits for data. This greater available parallelism favours the CUDA implementation. Runtime also includes our implementations wrapper overhead: UFiT copies the complete field for every seed batch, whereas FastQSL transfers it to the GPU once.

[6] NVIDIA V100 datasheet.
[7] NVIDIA A100 datasheet.
[8] NVIDIA H100 specifications.
[9] NVIDIA Grace Performance Tuning Guide.
[10] NVIDIA GTX 1050 specifications.
[11] Intel Xeon Gold 6430 specifications.
[12] AMD EPYC 9354 specifications.
[13] AMD EPYC 7002 series datasheet.
[14] AMD EPYC 9004 series datasheet.

Study	HPC site	Compute hardware	Job allocation and scope
CFIT resolution validation, $N = 25$–200	COSMA5	2× AMD EPYC 9754	16 Slurm CPUs and 64 GiB host memory; both polarity branches.
UFiT strong scaling, $N = 200$	DINE2	2× Intel Xeon Gold 6430	Exclusive node: 64 Slurm CPUs and 2000 GiB; 1–64 OpenMP threads.
UFiT strong scaling, $N = 200$	NCC	2× AMD EPYC 9354	Exclusive node: 128 Slurm CPUs (64 physical cores); 1–128 OpenMP threads.
Low–Lou $N = 200$ ensemble and CFIT comparison	DINE2	2× Intel Xeon Gold 6430 + NVIDIA V100 (32 GB)	16 Slurm CPUs and 64 GiB; 390 NF2 fields. FastQSL P/N used `gc007`.
Low–Lou $N = 400$ comparison	DINE2	2× Intel Xeon Gold 6430 + NVIDIA V100 (32 GB)	64 Slurm CPUs and 1500 GiB; NF2 and FastQSL P/N on the same node.
Low–Lou $N = 800$ feasibility test	DINE2	2× Intel Xeon Gold 6430 + NVIDIA V100 (32 GB)	64 Slurm CPUs and 1500 GiB; CFIT failed when its padded FFT volume exceeded host memory.
AR 11158 / HARP 377	DINE2	2× Intel Xeon Gold 6430 + NVIDIA V100 (32 GB)	64 Slurm CPUs and 96 GiB; NF2 and conditioned CFIT P/N ran as separate jobs.
AR 13664 outer pair	COSMA8; DINE2	NF2: 72-core NVIDIA Grace Arm + H100 (GH200); CFIT: 2× Xeon Gold 6430 + V100 (32 GB)	NF2: 72 Slurm CPUs and 450 GB, three seeds per snapshot. CFIT: 64 Slurm CPUs and 256 GiB, both branches per snapshot.
AR 13664 12-min cadence	DINE2	2× Intel Xeon Gold 6430 + NVIDIA V100 (32 GB)	64 allocated Slurm CPUs; NF2 requested 64 GiB and CFIT 256 GiB.
AR 13848 sequential series and controls	COSMA8; DINE2	72-core NVIDIA Grace Arm + H100 (GH200); Xeon Gold 6430 + V100 (32 GB)	Single-node NF2 jobs: one initial fit and six serial continuation allocations on H100; seven independent controls on H100 and two on V100.
FastQSL heap-limit timing	GPU benchmark node	Host CPU not recorded; NVIDIA A100	Isolated `TraceAllBline` timings before and after the allocation patch.
FastQSL roofline counters	Local workstation	Intel Core i7-11700K + NVIDIA GTX 1050	Nsight Compute counters for $N = 25$–800; the plotted roofline point uses $N = 800$.
UFiT roofline calibration	Hamilton, `cn017`	AMD EPYC 7702	16 cores; architectural calibration only, excluded from extrapolation timings.

Table B.1: Machines and job allocations used for the scientific results and performance calibrations.

Hardware item	Peak FP64 (TFLOP/s)	Quoted value or derivation basis
CPU		
AMD EPYC 9754	9.22*	Full two-socket node: 256 cores ×2.25 GHz ×16 FP64 operations per cycle.
Intel Xeon Gold 6430	4.30*	64 cores ×2.10 GHz ×32 FP64 operations per cycle.
AMD EPYC 9354	3.33*	64 cores ×3.25 GHz ×16 FP64 operations per cycle.
AMD EPYC 7702	0.512*	16 cores ×2.00 GHz ×16 FP64 operations per cycle.
NVIDIA Grace CPU	3.55*	Half the quoted dual-Grace FP64 peak.
GPU		
NVIDIA V100 PCIe	7.0	Manufacturer-quoted FP64 peak; the FP32 peak relevant to most FastQSL instructions is about 14.
NVIDIA A100	9.7	Manufacturer-quoted PCIe/SXM scalar peaks.
NVIDIA H100 in GH200	34	H100 SXM scalar reference; GH200-specific peaks are not separately stated.
NVIDIA GTX 1050	0.058*	FP32 peak divided by the GP107 FP64 ratio of 32.

Table B.2: Peak scalar FP64 capability of the hardware used in this work.

Configuration	Runs	Mean utilisation (%)
Low–Lou $N = 200$: NF2 192×6 (V100)	3	74.8–76.8
Low–Lou $N = 200$: CFIT/FastQSL (V100)	2	18.6
Low–Lou $N = 400$: NF2 192×6 (V100)	1	55.7
Low–Lou $N = 400$: CFIT/FastQSL (V100)	2	53.4–54.4
AR 13664 cadence: NF2 (V100)	12	57.0–58.5
AR 13664 cadence: CFIT/FastQSL (V100)	2	31.3–34.4
AR 13848 series: NF2 (H100)	1	23.7
AR 13848 controls: NF2 (H100)	7	22.4–24.5
AR 13848 controls: NF2 (V100)	2	73.1–73.3

Table B.3: Available GPU utilisation for selected extrapolation runs. Values are means sampled using `nvidia-smi`.

C SUPPORTING RESULTS

C.1 CFIT performance details

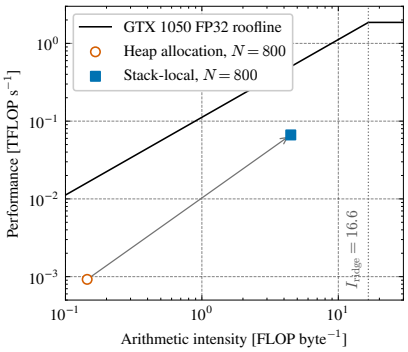
C.1.1 UFiT strong scaling

Table C.1 reports fixed-$N = 200$ UFiT solver-time scaling on the two CPU systems in Table B.1. As the one-thread runs did not complete, $p = 2$ defines $S_p = T_2/T_p$ and relative efficiency $\eta_p = S_p/(p/2)$. End-to-end timings follow the same trend.

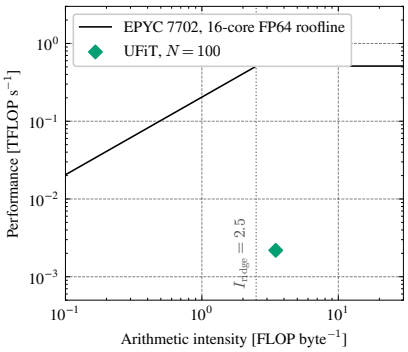
Site	p	T_{solve} (s)	S_p	η_p	Trace share (%)	IPMI node (MJ)
DINE2	2	7298	1.00	1.000	90.7	5.306
DINE2	4	3810	1.92	0.958	89.4	2.395
DINE2	8	2137	3.42	0.854	87.6	1.404
DINE2	16	1217	6.00	0.750	84.2	1.029
DINE2	32	775	9.42	0.589	79.7	0.689
DINE2	64	620	11.77	0.368	76.7	0.558
NCC	2	6870	1.00	1.000	91.0	—
NCC	4	3572	1.92	0.962	90.0	—
NCC	8	1987	3.46	0.865	88.4	—
NCC	16	1121	6.13	0.766	85.5	—
NCC	32	665	10.33	0.646	81.1	—
NCC	64	452	15.20	0.475	76.2	—
NCC	128	443	15.51	0.242	75.9	—

Table C.1: UFiT strong scaling for the positive-polarity $N = 200$ extrapolation. Speed-up and relative efficiency are normalized to the two-thread result on the same system. IPMI energy was available only on DINE2; each point is one execution, and the reconstructed fields were identical across thread counts and systems. The single executions do not quantify run-to-run variability.

C.1.2 Machine-specific roofline metrics



(a) FastQSL



(b) UFiT

Figure C.1: Machine-specific rooflines for (a) the FastQSL `TraceAllBline` kernel on a GTX 1050 and (b) UFiT on 16 EPYC 7702 cores. The FastQSL arrow shows the heap-to-stack change at $N = 800$. UFiT records 7.671×10^{11} operations in 350.959 s, or 2.19 GFLOP s^{-1} and 0.43% of the allocated-subset peak; its compute-bound position excludes memory bandwidth as the limiting ceiling but does not imply high arithmetic-unit utilisation. Each panel uses its own machine ceiling, so they are not a direct backend comparison.