Pretraining Transfer for Neural MHD Surrogates: What Generalizes, What Doesn't, and Why It Matters for Plasma Foundation Models

Stanislav DeLaurentiis*

Abstract

We study which components of neural PDE surrogate training transfer between two regimes of magnetohydrodynamic (MHD) turbulence, using the same architecture (FNO3D, 18.6 M parameters) and the same dataset family (Polymathic AI's *Well* MHD_64). With matched-architecture hyperparameter tuning on both sides and three pretraining seeds per condition, pretraining on super-Alfvénic MHD without an imposed guide field (M_A=2.0) and fine-tuning on sub-Alfvénic anisotropic MHD with a guide field (M_A=0.7) yields a 35% reduction in validation VRMSE at 1% target data. A non-MHD pretraining control on supernova hydrodynamics data underperforms the matched scratch baseline by 14%, consistent with negative transfer under wrong-domain pretraining — though this control has structural confounds (governing equations and source terms, open vs. periodic boundary conditions, channel-set and normalization mismatch, SPH-to-grid substrate origin) that prevent a clean isolation of physics-as-cause. The transfer is asymmetric across physics features: pretraining preserves the perpendicular turbulent cascade and short-horizon accuracy, but fine-tuning introduces a new late-stage rollout instability, making long-horizon (step-50) VRMSE $\approx$ 10.8 for the fine-tuned model, $\approx$ 3.5$\times$ worse than the un-fine-tuned zero-shot pretrain model ($\approx$ 3.0) and $\approx$ 5.5$\times$ worse than the from-scratch 1%-data baseline ($\approx$ 2.0) — with per-trajectory consistency across 10 held-out test trajectories. An out-of-distribution control confirms that a guide-field collapse is a property of any M_A=2.0-trained model evaluated on M_A=0.7, while a new late-stage magnetic-energy growth emerges specifically from fine-tuning. Our results highlight that MHD-matched pretraining transfers spatial statistics with high sample-efficiency on the downstream task, but reliable long-horizon emulation requires training objectives and architectural choices beyond one-step accuracy — consistent with the autoregressive-stability literature McCabe et al. [2023a], Lippe et al. [2023].

1 Introduction

Foundation models for physical simulation McCabe et al. [2025], Herde et al. [2024], McCabe et al. [2023b] promise that a single pretrained network, once trained across a diverse corpus of numerical simulations, can be fine-tuned cheaply on a new scientific task. The strongest version of this claim — that scale across diverse physics substitutes for specialization — has not been comprehensively tested at the physics-feature level for plasma. For plasma physics in particular, where potential applications range from tokamak transport emulation to astrophysical magnetohydrodynamic (MHD) surveys, two questions matter for design:

1. **Does pretraining corpus composition matter, or does any large pretrain help as a form of regularization?** If architecture-identical pretraining on non-plasma physics transfers

*Department of Astronomy, Columbia University. sod2112@columbia.edu.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

just as well as plasma pretraining, then the Polymathic bet on plasma-specific pretraining corpora is mostly a generic scale-and-compute decision. If plasma pretraining gives a specific advantage — or, stronger still, if non-plasma pretraining *hurts* — then pretraining corpora should be curated for downstream physics.

2. **What specifically transfers?** A pretrained model that matches a scratch baseline on one-step VRMSE is not automatically a good surrogate. It may fail at long-horizon rollouts, violate conservation laws, fail to propagate linear waves, or drift to a different statistical attractor during rollout. These are distinct axes and they are the axes that matter for plasma science applications.

We study both questions at controlled scale, using a single architecture (FNO3D) and a single dataset family (the Polymathic AI *Well* `MHD_64`) to avoid confounds. Our target is sub-Alfvénic MHD with an imposed background field (M_A=0.7), chosen as a proxy for guide-field plasma regimes. Our source is super-Alfvénic MHD without an imposed guide field (M_A=2.0). With matched-architecture hyperparameter tuning on both sides, we report a 35% VRMSE advantage at 1% target data from specifically-MHD pretraining, a 14% *penalty* from non-MHD pretraining (supernova hydrodynamics with radiative cooling) under the same architecture, and a detailed audit of what physics actually transfers. The non-MHD comparison carries structural confounds (governing equations and source terms, open vs. periodic boundary conditions, channel-set and normalization mismatch, SPH-to-grid substrate origin) that prevent a clean isolation of physics content; we discuss this carefully in Section 5.

2 Related Work

Neural PDE surrogates. Fourier neural operators Li et al. [2020] established spectral methods for parametric PDE emulation. Recent work scales these to multi-physics pretraining corpora: MPP McCabe et al. [2023b], Poseidon Herde et al. [2024], and Walrus McCabe et al. [2025], alongside the *Well* benchmark Ohana et al. [2024] of 15 TB of physics simulations. Walrus McCabe et al. [2025] covers 19 scenarios spanning fluids, MHD, active matter, and astrophysical turbulence in a 1.3 B-parameter transformer.

Plasma ML. Disruption prediction Kates-Harbeck et al. [2019], reactor control Degrave et al. [2022], Seo et al. [2024], and plasma surrogate modelling with FNOs Gopakumar et al. [2024] have produced domain-specific successes. Transfer between simulation regimes with controlled corpus-composition experiments of the kind we run is less studied.

Transfer learning in PDE surrogates. Prior work compares pretraining corpora on in-distribution benchmarks McCabe et al. [2023b], Herde et al. [2024] and studies scaling and OOD parameter transfer Subramanian et al. [2023], but the question of whether downstream improvements come from the pretraining task's *physics* or from pretraining acting as architecture-agnostic regularization is rarely isolated head-on. Our non-MHD control provides one such comparison, with the declared caveat that the NS and MHD datasets differ on multiple non-physics axes as well.

Autoregressive rollout stability. Long-horizon rollout instability of one-step-trained neural PDE solvers is a recognized failure mode, with proposed mitigations including pushforward training Brandstetter et al. [2022], stability-aware regularization McCabe et al. [2023a], and iterative refinement Lippe et al. [2023]. Conservation-preserving architectures Liu et al. [2024], Richter-Powell et al. [2022] target a related failure. Our finding that *fine-tuning* introduces a new late-stage rollout instability — absent in the corresponding zero-fine-tune pretrain rollout — is consistent with that literature.

3 Methods

Dataset. Polymathic AI's *Well* Ohana et al. [2024], `MHD_64`: 3D compressible isothermal MHD simulations on 64^3 periodic grids, 7 physical fields per cell (density, B_x, B_y, B_z, v_x, v_y, v_z), 100 timesteps per trajectory. The dataset covers a grid of sonic and Alfvén Mach numbers ($M_s \in \{0.5, 0.7, 1.5, 2.0, 7.0\}$, $M_A \in \{0.7, 2.0\}$). The sub-Alfvénic runs include an imposed background field $B_0 \| \hat{x}$ (verified directly in per-sample statistics: $\langle B_x \rangle \approx 1.0$, $\langle B_{y,z} \rangle \approx 0$), which is the physical source of anisotropy and the reason we treat this regime as a guide-field proxy.

Model. FNO3D (`the_well.benchmark.models.FNO`), 18.62 M parameters at `modes=12`, `hidden_channels=48`. One-step-ahead prediction. We train on adjacent-timestep pairs (Markov input: $u_t \rightarrow u_{t+1}$), which differs from the Well's standard 4-timestep-history baselines Ohana et al. [2024]; this simplifies the model but discards temporal context that may be relevant for wave-bearing MHD dynamics, and we flag it as a potential confound for rollout-stability comparisons (Limitations). Pretraining: 20 epochs on M_A=2.0 train split (3,960 windows), AdamW, `lr=1e-3`, cosine schedule, batch size 8, gradient clipping at L2 norm 1. Fine-tuning: 40 epochs on M_A=0.7 data at fractions $\{1, 10, 100\}\%$ of the 3,663 available windows, otherwise identical hyperparameters unless noted. We fine-tune the full network (no layer freezing or adapters).

Hyperparameter control. At 1% target data, we ran a coarse sweep (4 learning rates $\times$ 3 hidden widths) at seed 0 for the baseline, then refined the top 5 at 2 additional seeds. For fine-tuning we swept the learning rate at fixed `hidden=48` (pretrained-weights constraint), 3 seeds per point. *For a fair, matched-architecture comparison we report the scratch baseline at the same `hidden=48` used for fine-tuning*: best tuned matched-arch scratch baseline at 1% data: `(lr=3e-3, hidden=48)` giving 0.465 ± 0.017 (3 seeds). Best tuned fine-tune: `(lr=3e-4, hidden=48)` giving 0.301 ± 0.0003 (3 seeds). A larger-capacity `hidden=64` scratch baseline achieves 0.419 ± 0.008; we do not use it as the headline reference because fine-tuning is locked to the pretrained width.

Non-MHD control. Identical FNO3D architecture pretrained for 20 epochs on `supernova_explosion_64` (Polymathic *Well*; 3D hydrodynamics with self-gravity and radiative-cooling source terms, SPH-to-grid substrate origin, open boundary conditions, 6 physical channels) with 3,860 windows from the valid split, z-score standardization per channel (densities span orders of magnitude), and variance-masked VRMSE loss for the zero-padded channel used to match the 7-channel MHD target. Same fine-tuning protocol on M_A=0.7 1% data across 3 seeds. We emphasize that this control differs from MHD on multiple axes besides the governing equations (boundary conditions, channel-set and normalization mismatch, SPH-to-grid substrate), so it tests *wrong-domain* pretraining rather than cleanly isolated *wrong-physics* pretraining. The Well presents all data on uniform grids at constant time intervals Ohana et al. [2024], so temporal cadence is not itself a confound.

Evaluation. VRMSE Ohana et al. [2024] is our primary metric. We additionally evaluate isotropic and anisotropic power spectra $E(k)$ and $E(k_\parallel, k_\perp)$, autoregressive rollout VRMSE vs ground truth for up to 50 steps, per-step mass/magnetic-energy/kinetic-energy/$\nabla \cdot B$ conservation diagnostics, and the drift of E_B/E_K equipartition during rollout, computed on a held-out sample of 10 test trajectories per checkpoint.

4 Results

4.1 MHD pretraining gives a 35% advantage at 1% target data

Figure 1 shows validation VRMSE on M_A=0.7 as a function of the fraction of M_A=0.7 training data used, for from-scratch and pretrained-then-fine-tuned models. At 1% data, under *matched hidden width* (`hidden=48`) on both sides, MHD pretraining yields a 35% reduction in VRMSE (0.3015 ± 0.0010 vs 0.4654 ± 0.0175; 3 pretrain seeds each). The gap narrows as target data grows (7% at 10%, $\approx$0 at 100%). Pretrain-seed variance on the fine-tuned endpoint is $\sigma \approx 10^{-3}$, an order of magnitude below the scratch-baseline seed variance ($\sigma \approx 0.017$), indicating the 35% gap is robust to pretrain-init randomness (the three pretraining runs are verifiably distinct initializations: epoch-0 train VRMSE $= 0.539, 0.545, 0.548$).

4.2 Wrong-domain pretraining does not help — and underperforms scratch

The most natural alternative explanation for the 35% advantage is that *any* pretraining regularizes the network — the source physics does not matter. To probe this, we pretrained an identical-architecture FNO3D on Polymathic `supernova_explosion_64` (non-MHD hydrodynamics with self-gravity and radiative cooling) and fine-tuned on the same M_A=0.7 1% target data (Figure 2). Wrong-domain pretraining *hurts* matched-architecture target performance by 14% relative to scratch (0.5285 ± 0.0177 vs 0.4654 ± 0.0175; 3 pretrain seeds). The $\approx$ 49% end-point gap between MHD- and

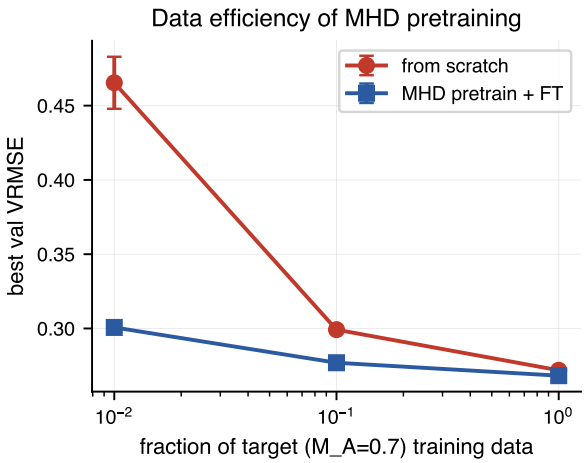

Figure 1: Data efficiency of MHD pretraining. Error bars are standard deviation across 3 seeds for {10%, 1%} configurations (single-seed at 100%). Baseline point at 1% reflects the best-tuned configuration from a coarse hyperparameter sweep; raw untuned baseline (not shown) is 0.557.

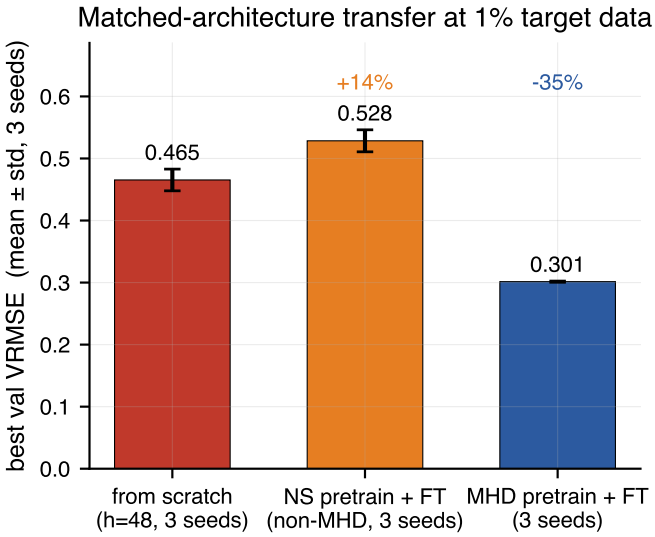

Figure 2: Matched-architecture transfer at 1% target data (`hidden=48` throughout); 3 pretrain seeds per condition. Non-MHD pretraining on supernova hydrodynamics underperforms the matched scratch baseline by 14%; MHD pretraining beats scratch by 35%. The $\approx 49\%$ gap between the two pretrain choices (relative to scratch) is consistent with the source corpus's domain content mattering, but the NS and MHD corpora also differ on non-physics axes (BC, dt, channel count, SPH origin); see Section 5.

NS-pretrain (relative to scratch) is inconsistent with the architecture-regularization-only hypothesis. Welch's t-test on the 3-seed samples (two-tailed, Welch approximation for degrees of freedom) gives $t = 13.2$ ($p \approx 0.006$) for scratch vs. MHD pretrain, $t = 3.58$ ($p \approx 0.023$) for scratch vs. NS pretrain, and $t = 18.1$ ($p \approx 0.003$) for NS vs. MHD pretrain; all three gaps are significant at the 5% level despite $n = 3$. We do not claim this cleanly attributes the gap to *physics content alone*: the two source corpora differ on multiple axes (governing equations and source terms, open vs. periodic boundaries, channel-set and normalization mismatch, SPH-to-grid substrate origin), and isolating physics-as-cause requires a matched-substrate control (same grid, channels, boundaries, cadence; only equations differ) that we have not run.

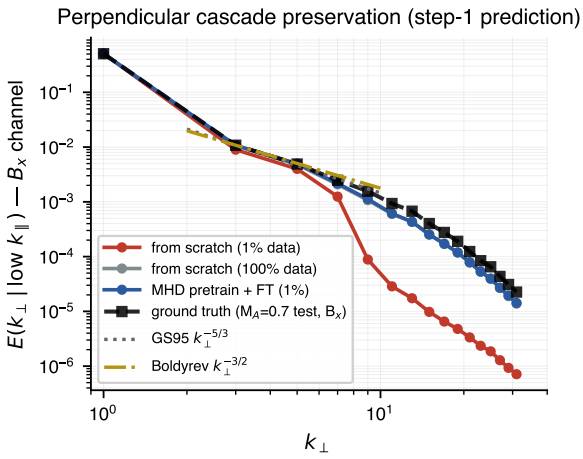

Figure 3: Perpendicular turbulent cascade of B_x at step-1 prediction, averaged over the lowest-$k_\parallel$ slab on 10 held-out M_A=0.7 test trajectories. "Ground truth" is the cascade computed directly from the true simulated B_x on the same trajectories — not a fit. Reference slopes for Goldreich-Sridhar $k_\perp^{-5/3}$ and Boldyrev $k_\perp^{-3/2}$ are anchored to ground truth at $k_\perp$=3 and shown over the inertial range $k_\perp \in [2, 10]$. From-scratch at 1% data fails to preserve high-$k_\perp$ structure; MHD pretrain + fine-tune matches ground truth across 1.5 decades of $k_\perp$ and closely tracks the 100%-data baseline.

4.3 What specifically transfers: spatial statistics, not temporal dynamics

Perpendicular cascade (transfers). Figure 3 plots $E(k_\perp|k_\parallel \to 0)$ — the perpendicular spectrum for the raw B_x field — for one-step predictions on held-out M_A=0.7 trajectories. At 1% target data, the from-scratch baseline loses high-$k_\perp$ power by nearly two orders of magnitude beyond $k_\perp \geq 8$. The MHD-pretrained fine-tuned model tracks ground truth across 1.5 decades of $k_\perp$. We include reference slopes for Goldreich-Sridhar Goldreich and Sridhar [1995] $k_\perp^{-5/3}$ and Boldyrev Boldyrev [2006] $k_\perp^{-3/2}$ as visual context; we do not subtract the mean guide field nor fit exponents with uncertainty bands, and so we make no claim of validating either scaling prediction from this evidence. Rigorous MHD-turbulence anisotropy analysis should be done in the local mean-field frame Cho and Vishniac [2000], Schekochihin [2022], which we defer to follow-up work.

Long-horizon rollout (fails, and differently). Figure 4 plots autoregressive rollout VRMSE versus step over 10 held-out M_A=0.7 test trajectories. The from-scratch 1%-data model drifts smoothly to ≈2.0 and stabilizes there — it has collapsed to a blurry, low-variance state that is stable because it barely evolves. The pretrained fine-tuned model is much better at short horizons (step 1 VRMSE 0.31 vs 0.56) but diverges past step 20, reaching ∼10.8 mean by step 50, with all 10 test trajectories inflating (per-trajectory range $[6.0, 15.6]$; divergence is not outlier-driven, see Fig. 7).

OOD control: zero-FT rollout on the target. As an additional control, we rolled the M_A=2.0-pretrained checkpoint directly on M_A=0.7 with no fine-tuning. The guide-field channel $\langle|B_x|\rangle$ collapses from 0.43 at step 1 to 0.04 at step 50, confirming that inability to preserve the imposed background field on OOD data is a property of any M_A=2.0-trained model, not a fine-tuning artifact. Step-50 VRMSE is 3.0 (vs 10.8 for ft_01), so fine-tuning does rewrite step-1 behavior — it pushes $\langle|B_x|\rangle$ to 3.68 at step 50 instead of collapsing it — but introduces a *new* late-stage magnetic-energy growth. Long-horizon failure is therefore not inherited from pretraining but emerges specifically during fine-tuning, consistent with autoregressive-stability theory McCabe et al. [2023a], Brandstetter et al. [2022]. Table 1 quantifies the guide-field trajectory across all four conditions.

Equipartition drift. Equipartition between magnetic and kinetic energy depends on the regime and the driving. As reference levels we use the empirically-measured $E_B/E_K \approx 2.1$ on our M_A=0.7 target test trajectories (mean over 10 trajectories × 51 rollout steps of the ground-truth ratio) and the theoretical $1/M_A^2 = 0.25$ for the M_A=2.0 source regime (M_A=2.0 test data was out of scope for this evaluation). Figure 5 tracks this ratio during autoregressive rollout. The from-scratch base-

Table 1: Mean $\langle B_x \rangle$ per field cell during autoregressive rollout; ground truth is ≈ 1.0 at all steps. Scratch mildly decays; the OOD zero-FT model collapses the guide field; the fine-tuned model over-inflates. The two pretraining-lineage models fail the guide-field preservation test in opposite directions.

Condition	step 1	step 5	step 10	step 25	step 50
Ground truth	1.001	1.001	1.001	1.001	1.001
Scratch (1%-data)	1.004	0.990	0.932	0.887	0.885
MHD pretrain + FT (1%)	0.999	0.992	1.006	1.312	3.680
MHD pretrain zero-FT (OOD on target)	0.428	0.104	0.089	0.069	0.044

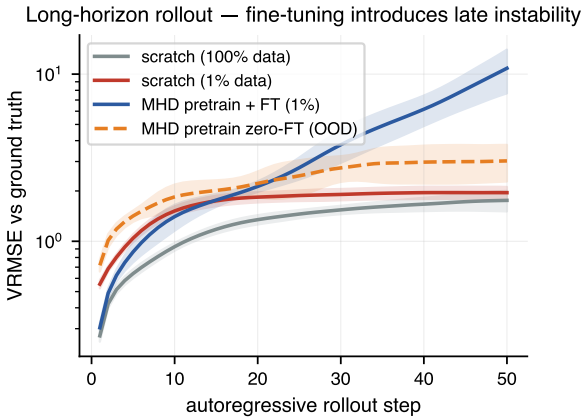

Figure 4: Autoregressive rollout VRMSE vs step. Scratch at 1% data (red) collapses to a low-variance blurry attractor at VRMSE $\approx$ 2. The pretrained fine-tuned model (blue) is much better at short horizons but diverges past step 20 to VRMSE $\approx$ 11. An OOD zero-FT control (pretrain checkpoint rolled directly on the target regime, no FT) reaches step-50 VRMSE $\approx$ 3 via guide-field collapse rather than inflation (see text), so the late-stage inflation is a fine-tuning artifact rather than inherited from pretraining.

line at 100% data tracks the truth curve near 2.0 modulo noise; the from-scratch 1%-data baseline undershoots slightly. The fine-tuned model drops rapidly from the target value, passes through the source-regime value, and overshoots — consistent with the fine-tuned model migrating toward a different attractor than either the source or target regime.

Conservation violation. Figure 6 shows that the pretrained fine-tuned model violates $\nabla \cdot B = 0$ approximately $4.3\times$ more than the from-scratch 1%-data baseline (floor-excess at step 50, annotated on the right panel), and that its magnetic energy drift accelerates past step 30 when the scratch baseline is stable. Ground-truth data itself has a non-zero numerical $\|\nabla \cdot B\|$ floor from the underlying finite-volume solver, so we report all monopole curves relative to that measured floor. FNO3D is not architecturally conservation-preserving Liu et al. [2024], Richter-Powell et al. [2022]; pretraining does not mitigate this weakness, and on these runs appears to worsen it. One plausible mechanism, consistent with the autoregressive-stability literature Lippe et al. [2023], Brandstetter et al. [2022], is that one-step-VRMSE-only fine-tuning allows off-solenoidal energy to accumulate without penalty and then amplify during the autoregressive rollout.

5 Discussion

Our results argue that, under matched architecture and tuning, **pretraining corpus composition has a real effect on downstream performance for this FNO3D/MHD setting, and that one-step accuracy does not imply long-horizon stability — in this setting, the one-step-best model is not the rollout-best**. Three actionable implications follow, phrased with the caveats above.

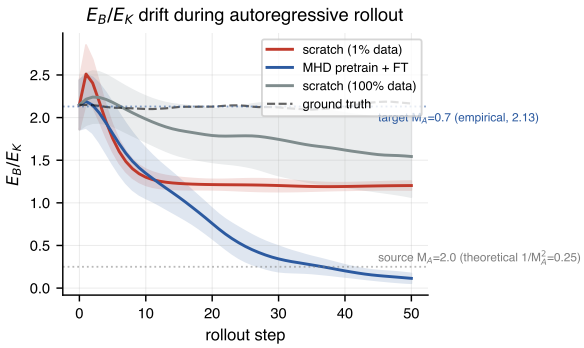

Figure 5: E_B/E_K during autoregressive rollout. Horizontal dashed lines mark the empirically-measured ratios on ground-truth trajectories for the M_A=0.7 target (≈2) and the M_A=2.0 source (≈0.25); we do not assume a universal $1/M_A^2$ scaling.

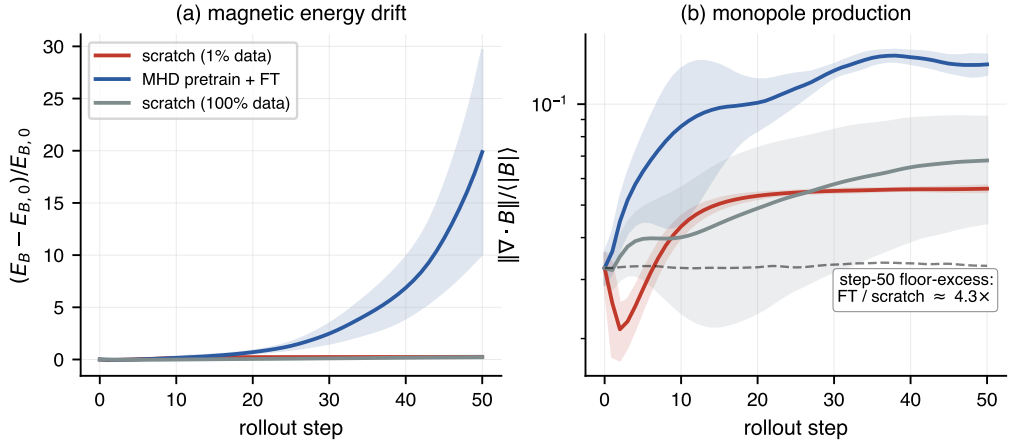

Figure 6: Conservation violations during rollout. (a) Relative drift in magnetic energy E_B. (b) Magnetic monopole production $\|\nabla \cdot B\|/\langle|B|\rangle$ (log scale). The pretrained model produces more monopoles and accelerating energy drift compared to the undertrained scratch baseline, despite being more accurate at short horizons.

Pretraining corpora matter, within our setting. The non-MHD control (Figure 2) shows that pretraining on a dataset far from the target domain can underperform the matched scratch baseline. This is consistent with *either* a physics-content effect *or* structural-substrate mismatch (BC, dt, channel count, solver family). We cannot separate these from a single control. The operational takeaway is conservative: pretraining benefits are not automatic with respect to corpus choice, so assume nothing and measure.

Long-horizon instability emerges during fine-tuning, not pretraining. The late-stage rollout failure (Figures 4 and 6) together with our OOD zero-FT control indicate that the fine-tuned model's step-50 divergence emerges during fine-tuning — pretraining alone does not produce it. We ran no architecture ablations, so we do not claim our evidence is architectural per se; but the autoregressive-stability literature McCabe et al. [2023a], Lippe et al. [2023], Brandstetter et al. [2022] already suggests that for emulators that need stable rollouts of tens of thousands of timesteps, explicit constraints (solenoidal projection for B, divergence-preserving update rules, conservation-aware losses, pushforward training) would likely be needed *in addition to* pretraining.

Equipartition drift as a rollout diagnostic. Figure 5 shows that the fine-tuned model's rollout drifts toward a different attractor than either the source or the target regime. For applications where regime-specific statistical accuracy matters (astrophysical surveys, plasma-astrophysics analyses), tracking an empirical E_B/E_K ratio during rollout is a cheap sanity check independent of VRMSE.

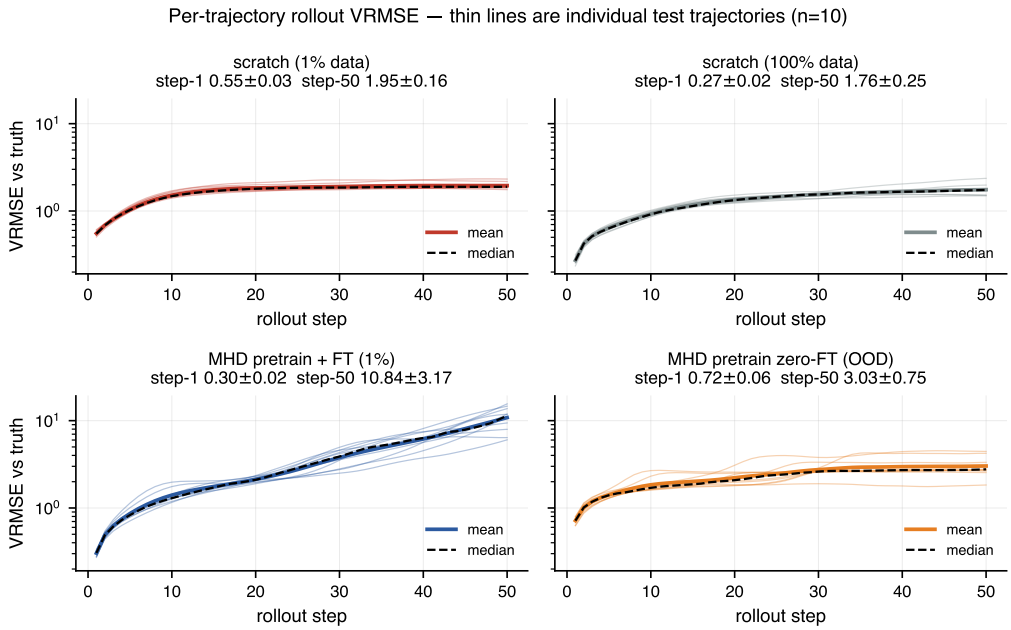

Figure 7: Per-trajectory rollout VRMSE for each of four conditions on 10 held-out M_A=0.7 test trajectories. Thin lines are individual trajectories; thick coloured lines are the mean, dashed black the median. All 10 ft_01 trajectories inflate to step-50 VRMSE $\geq$ 6, confirming that the mean divergence (10.8) is not outlier-driven. The OOD zero-FT panel (bottom right) shows a distinct failure mode — guide-field collapse producing a much lower step-50 VRMSE ($\approx$ 3) — not the late-stage inflation characteristic of fine-tuning.

6 Limitations

Our findings are conditional on a single architecture (FNO3D at 18.6 M parameters), a single dataset family (Polymathic *Well* MHD_64), and a single training recipe (one-step supervised adjacent-pair loss, full-network fine-tuning, no layer freezing or adapters). Headline numbers aggregate over three pretraining seeds per condition (MHD and NS), which captures pretrain-init randomness; they do not capture architectural or dataset-level randomness, so the quoted error bars are lower bounds on the true uncertainty. The non-MHD control is not a clean wrong-physics isolation: the NS and MHD corpora differ on governing equations, boundary conditions, channel set, and SPH-vs-grid origin, so the $\approx 49\%$ end-point gap is best read as *wrong-domain* rather than *wrong-physics-only*; a matched-substrate control (same grid, channels, boundaries, cadence) remains to be run. Other ablations we have not run and that would sharpen conclusions: (i) the Well's standard 4-timestep-history input instead of Markov adjacent pairs (which may itself affect rollout stability); (ii) layer-freezing / adapter / low-rank fine-tuning (our full-network FT is the most aggressive setting and not obviously the best choice for low-data transfer); (iii) pushforward Brandstetter et al. [2022], PDE-Refiner Lippe et al. [2023], or stability-regularized FNO McCabe et al. [2023a] during fine-tuning (the most likely candidates to close the late-stage rollout gap); (iv) fitted cascade exponents with uncertainty and a local-mean-field anisotropy analysis Cho and Vishniac [2000], Schekochihin [2022], which would turn Figure 3 from a qualitative into a quantitative transfer test. The target regime is a simulation proxy (sub-Alfvénic MHD with an imposed B_0), not a tokamak or astrophysical ground truth. We do not compare to Walrus McCabe et al. [2025] because (a) architecture mismatch (1.3 B transformer vs 18.6 M FNO3D), (b) Walrus was pretrained on the full MHD_64 train split, exposing it to orders of magnitude more target-regime data than our controlled condition, and (c) our experiment varies source-corpus composition as the controlled variable, which Walrus does not — a head-to-head would confound data volume with corpus composition. Cross-framework evaluation on data generated by different simulation codes (e.g., Athena++, Dedalus) is the natural follow-up setting for that comparison.

7 Conclusion

All results are conditional on a single architecture (FNO3D, 18.6 M params), a single dataset family (Polymathic *Well* MHD_64), and a single training recipe (one-step supervised, AdamW, cosine schedule). With that scope:

What the data supports.

- MHD pretraining at 1% target data reduces validation VRMSE by 35% versus the matched-architecture scratch baseline (0.3015 ± 0.0010 vs 0.4654 ± 0.0175; 3 pretrain seeds; Figs. 1, 2).
- A non-MHD (supernova hydrodynamics) pretrain *hurts* the matched scratch baseline by 14% (Fig. 2). Source-corpus composition matters within our setting, but the NS and MHD corpora differ on multiple non-physics axes (boundary conditions, channel set, SPH-to-grid substrate, normalization), so we cannot cleanly attribute the gap to physics content alone.
- Pretraining transfers spatial statistics but not long-horizon dynamics. The fine-tuned model matches the ground-truth perpendicular cascade across 1.5 decades of $k_\perp$ (Fig. 3) yet its step-50 rollout VRMSE reaches ≈ 10.8 versus ≈ 2.0 for scratch and ≈ 3.0 for the un-fine-tuned pretrain model (Figs. 4, 7).
- The late-stage instability is a fine-tuning artifact. The pretrained checkpoint rolled on the target with no fine-tuning fails *differently* — guide-field collapse ($\langle B_x \rangle : 1.0 \to 0.04$) — while fine-tuning rewrites step-1 behavior and introduces $\langle B_x \rangle$ over-inflation to 3.68 by step 50 (Table 1). Fine-tuning swaps one failure mode for another; it does not add instability to a stable baseline.

What our results do not establish. Our evidence is $n{=}1$ on architecture, dataset family, and training recipe; each of the following would require experiments we have not run:

- That fine-tuning is broken for plasma surrogates in general (different architecture or dataset could behave differently).
- That FNOs are the problem; we did not test known FNO stability fixes (pushforward training Brandstetter et al. [2022], PDE-Refiner Lippe et al. [2023], stability-regularized FNO McCabe et al. [2023a]).
- That transformer-based foundation models (e.g., Walrus McCabe et al. [2025]) would do better under the same fine-tuning protocol — we did not evaluate one.
- That the one-step-supervised loss specifically causes the late-stage pathology; this is consistent with the autoregressive-stability literature but not isolated in our setup.

Scope-appropriate takeaway. In this setting, one-step supervised fine-tuning of a vanilla FNO3D inherits spatial statistics from MHD pretraining but not long-horizon stability. Closing the stability gap plausibly requires either architectural constraints (e.g., divergence-free B parametrization Liu et al. [2024], Richter-Powell et al. [2022]) or training-recipe changes (pushforward, iterative refinement), which are the obvious next ablations before claiming what works for plasma foundation models.

References

S. Boldyrev. Spectrum of magnetohydrodynamic turbulence. *Physical Review Letters*, 96(11):115002, 2006.

J. Brandstetter, D. Worrall, and M. Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

J. Cho and E. T. Vishniac. The anisotropy of magnetohydrodynamic Alfvénic turbulence. *The Astrophysical Journal*, 539:273–282, 2000.

J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

P. Goldreich and S. Sridhar. Toward a theory of interstellar turbulence. II. Strong alfvénic turbulence. *The Astrophysical Journal*, 438:763–775, 1995.

V. Gopakumar, S. Pamela, L. Zanisi, Z. Li, A. Gray, D. Brennand, N. Bhatia, G. Stathopoulos, M. Kusner, M. P. Deisenroth, et al. Plasma surrogate modelling using Fourier neural operators. *Nuclear Fusion*, 2024. arXiv:2311.05967.

M. Herde, B. Raonic, T. Rohner, R. Kappeli, R. Molinaro, E. de Bezenac, and S. Mishra. Poseidon: Efficient foundation models for PDEs. *arXiv preprint arXiv:2405.19101*, 2024.

J. Kates-Harbeck, A. Svyatkovskiy, and W. Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568(7753):526–531, 2019.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

P. Lippe, B. Veeling, P. Perdikaris, R. E. Turner, and J. Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2308.05732.

N. Liu, Y. Fan, X. Zeng, M. Klöwer, Y. Yu, and Others. Harnessing the power of neural operators with automatically encoded conservation laws. In *International Conference on Machine Learning*, 2024.

M. McCabe, P. Harrington, S. Subramanian, and J. Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023a. arXiv:2306.10619.

M. McCabe, B. Regaldo-Saint Blancard, L. H. Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, S. Golkar, G. Krawezik, F. Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023b.

M. McCabe, P. Mukhopadhyay, T. Marwah, et al. Walrus: A cross-domain foundation model for continuum dynamics, 2025.

R. Ohana, M. McCabe, L. Meyer, R. Morel, F. Agocs, M. Beneitez, M. Berger, B. Burkhart, S. Dalziel, D. Fielding, et al. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 44989–45037, 2024.

J. Richter-Powell, Y. Lipman, and R. T. Q. Chen. Neural conservation laws: A divergence-free perspective. In *Advances in Neural Information Processing Systems*, 2022.

A. A. Schekochihin. MHD turbulence: a biased review. *Journal of Plasma Physics*, 88(5):155880501, 2022.

J. Seo et al. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 2024.

S. Subramanian, P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. W. Mahoney, and A. Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.00258.