Physics-constrained generative models of a parametric MHD dynamo

M2 internship report · race pilot, team claude · a comparative study of constraints imposed by design, loss and sampling

Table of contents

Abstract **5**

1 Introduction **7**
1.1 Context: the internship offer and the MilaDy project . . . 7
1.2 Question and approach . . . 7
1.3 What the offer asks and what this report covers . . . 9
1.4 Contributions . . . 10
1.5 Outline . . . 11

2 State of the art **12**
2.1 Diffusion and flow matching . . . 12
2.2 Generative models of flows and of PDE solutions . . . 12
2.3 Three places for a physical constraint . . . 12
2.4 Helical dynamos in periodic boxes . . . 13
2.5 Metrics . . . 13
2.6 What was not found . . . 13

3 Methods **14**
3.1 The physical system . . . 14
3.2 Stationarity identities . . . 14
3.3 Latent representation and generative model . . . 15
3.4 Constraint mechanisms . . . 16
3.5 The linear part of the velocity field . . . 19
3.6 Metrics . . . 19

4 Experiments **22**
4.1 Reference data . . . 22
4.2 Protocol . . . 24

5 Results **27**
5.1 The registered protocol: every first-generation model is far from the reference . . . 27
5.2 Diagnosis: the models do worse than a Gaussian field . . . 29
5.3 Giving the field its linear part . . . 30
5.4 Two Gaussian laws: what the covariance between modes holds . . . 31
5.5 Beyond a Gaussian law: three tests . . . 35
5.6 Where the third-order part comes from, and what the right covariance changes . . . 38
5.7 Loss and sampling mechanisms on a model close to the reference . . . 43
5.8 Topology: null points and the small scales . . . 45
5.9 Polarity . . . 45
5.10 Complements: longer trainings, replication, objectives . . . 46

6 Comparative study **48**

7 Toolbox **52**

8 Data and models **55**

9 Discussion **59**
9.1 What each place of constraint buys 59
9.2 Negative results and refuted claims 60
9.3 Limits 61
9.4 What the turbulent regime, the dynamo threshold and the data of the project would need 63

10 Conclusion **65**

11 References **67**

12 Reproducibility **70**
12.1 Commands 70
12.2 Result files and commits 70
12.3 The catalogue of the release 72
12.4 Figures of this report 72

A Further tests on the origin of the non-Gaussian part **73**

B Longer trainings and wider networks **79**

C Replication with other data **84**

D Objectives of the race **86**

E Register of the hypotheses **88**

F Compute **94**

G Workflow of the experiment and organisation of the agents **95**

H History of the corrections **97**

I What the incidents of the pilot teach **98**

J GPU jobs and costs **99**

Status of this document This report is the deliverable of team claude in the race pilot of this repository (`race/DELIVERABLES.md`). It is written in the form of the M2 internship report that the MilaDy offer asks for, by an AI agent working alone for one session, and it reports one experiment, E4, designed, run and reviewed during that session. Nothing in it has been read or validated by the supervisors of the internship, and no data of the MilaDy consortium was used.

Three statuses are kept apart throughout. What is **measured** comes from result files written by committed scripts of this repository (`experiments/results/`, summarised in `experiments/results/digest.md`); every table of numbers below is generated from those files by `experiments/e4_report_tables.py`. What is **published** is cited with a reference taken from the verified source files `docs/specs/sources/`. What is **conjectured** is said to be so. The system studied is small (a 16^3 grid, 665 Fourier modes, Reynolds number 10, a smooth flow, a magnetic Prandtl number of 0.6 to 1.7, one conditioning parameter): the conclusions are about the mechanisms, not about turbulent dynamos. What the internship offer asks and what this report covers of it is in Section 1.3; what a larger system would need is in Section 9.4.

The HTML version of this report and an interactive explorer of the results are published with the site of the repository, under `teams/claude/report/` and `teams/claude/explorer/`.

Abstract

The MilaDy internship offer asks for generative models (diffusion or flow matching) of three-dimensional magnetohydrodynamic fields conditioned on physical parameters, with physical constraints (energy balance, zero divergence) imposed through model design, training losses or sampling procedures, evaluated on energy spectra, helicity, divergence control and topology preservation, and delivered as a comparative study with code in the project toolbox. This report carries out that comparison on a system small enough to be simulated and learned many times: a helically forced dynamo in a periodic box (16^3 grid, 665 Fourier modes, $Re = 10$), conditioned on the magnetic diffusivity η alone (Rm from 6 to 17 and a magnetic Prandtl number from 0.6 to 1.7, in the saturated regime of a smooth flow), with training, validation and held-out values, five seeds per model, and 38 hypotheses (H4 to H41), each registered with a refutation criterion before its computation and decided after a review (the checks of two of them, H35 and H36, were amended after a first cycle that was invalid by their letter, and the amendments are reported).

All results are **measured** in this repository. *Constraints.* Zero divergence is exact by design (vector potentials) or by projection at sampling time (relative divergence of order 10^{-8}); a penalty in the loss does not enforce it. The energy balance is a property of the law, not of each sample: a stationarity loss lowers its residual at parameter values without data but degrades a model close to the reference; exponential tilting at sampling time improves such a model on most metrics; and a linear map applied to the samples, which gives them the mean and covariance fitted on the data, gives them the energy balance and the spectrum of the data. *Methods.* None of the divergence mechanisms decides the quality of the law. What does is an explicit mode-wise affine term in the velocity field of the flow-matching model, which brings it from more than a hundred noise floors of the reference law to nine; at equal cost, flow matching interpolates in η better than the one diffusion model tried, both being more than 150 floors from the reference. Two Gaussian laws without a network are closer still (3.5 and 2.5 floors): the comparison ranks constraints on learned models that a Gaussian baseline beats in energy distance. What the best learned model holds beyond a Gaussian law is part of the third-order magnetic production of the reference (0.6 and 0.4 of it at two of the three held-out values), which the same model does not produce from Gaussian training data. Recoloured by the linear map, it is at 2.1 floors, keeps that production and, restarted in the solver, passes half as far above the reference energy as Gaussian draws with the same moments. *Metrics.* Spectra and helicity follow the second moments. Topology is not preserved by any law sampled without the solver: the laws close to the reference have two to three times its magnetic null points. From states with the second moments of the data, four time units of the solver bring the count within 0.05 to 0.13 of that of the reference; the solver alone, from the initial condition of the data, reaches a null-count error of at most 0.2 with a magnetic energy within 5 % after 64 time units and not after 32.

What the offer asks and this report does not cover is listed in Section 1.3. One conditioning parameter out of the families the offer names, at one Reynolds number and one resolution. No statistics of polarity reversals: the models draw independent snapshots, which fix at most a mean rate of sign changes (not computed), and a descriptive diagnostic shows that the sign of the large-scale magnetic modes of this system is lost in 17 to 80 time units. No reduced-order model or data assimilation fed with the generated fields; the restart in the solver is the nearest test. And a network on the flattened vector of 5320 coefficients, which does not scale to the turbulent

regime, to the dynamo threshold or to the finite-element data of the project (Section 9.4). The code (sixteen tested registry entries: eleven constraint mechanisms, the unconstrained baseline, a diffusion model and three Gaussian laws; a null-point finder with the Kac-Rice factors of the count), the data set in the data format of the project, the model files (the same models retrained on CPU on the data of the main run for the release, the weights of the cited runs not having been kept; the released recoloured law is not the cited one and is farther from the reference than the released Gaussian laws) and a catalogue are published with the repository (Section 8).

1 Introduction

1.1 Context: the internship offer and the MilaDy project

The internship offer that this report answers is entitled *Physics-Constrained Generative AI Modelling of Parametric Dynamical Systems*. It belongs to the ANR project MilaDy, on machine-learning models of turbulent dynamos, and asks for generative models (diffusion or flow matching) of parameter-dependent three-dimensional fields in which physical constraints are imposed "through model design, training losses or sampling procedures". The models are to be evaluated with energy spectra, helicity, divergence control and the preservation of topology, and the expected outcome is a comparative study of the strategies together with code integrated into the project toolbox (`docs/specs/sources/supervisors.md`, §1.2).

A dynamo is a flow of an electrically conducting fluid that sustains its own magnetic field. The equations are those of magnetohydrodynamics (MHD): the Navier–Stokes equations with the Lorentz force, coupled to the induction equation. Two control parameters matter here, the kinetic Reynolds number Re and the magnetic Reynolds number Rm, which measures induction against magnetic diffusion. Simulating a dynamo at one parameter value is expensive, and the scientific questions (thresholds, saturation levels, statistics of the saturated state) are questions about how the statistically stationary state changes with the parameters. A conditional generative model addresses exactly that object: it learns the law of the snapshots of the stationary state as a function of the parameter and can then draw new snapshots at parameter values that were never simulated.

The difficulty, and the subject of the offer, is that a generated field must be physically admissible. The magnetic field must be divergence-free. The energy injected by the forcing must, on average, be dissipated. The symmetries of the equations must be those of the generated law. The topology of the magnetic field, for instance its null points, should be that of the reference. The published literature offers three places where such requirements can enter a generative model, which are the three places named by the offer: the design of the model, the training loss, and the sampling procedure.

1.2 Question and approach

The question of this report is narrow and practical: **for a parametric 3D MHD dynamo, which constraints should be imposed where, and what does each mechanism buy on the metrics the offer names?** The approach is a controlled comparison on a system small enough to be simulated, learned and evaluated many times: a helically forced dynamo in a periodic box, on a 16^3 grid, with the magnetic diffusivity η as the conditioning parameter (Figure 1.1). On this system the reference law can be sampled at will, every quantity can be computed to round-off, and each comparison can be repeated over five model seeds.

The work follows the protocol of the repository (`AGENTS.md`): each hypothesis is written in a register with a quantitative refutation criterion before the computation; each quoted comparison uses five seeds; a reviewing role, the `encadrant`, recomputes the result and gives the verdict;

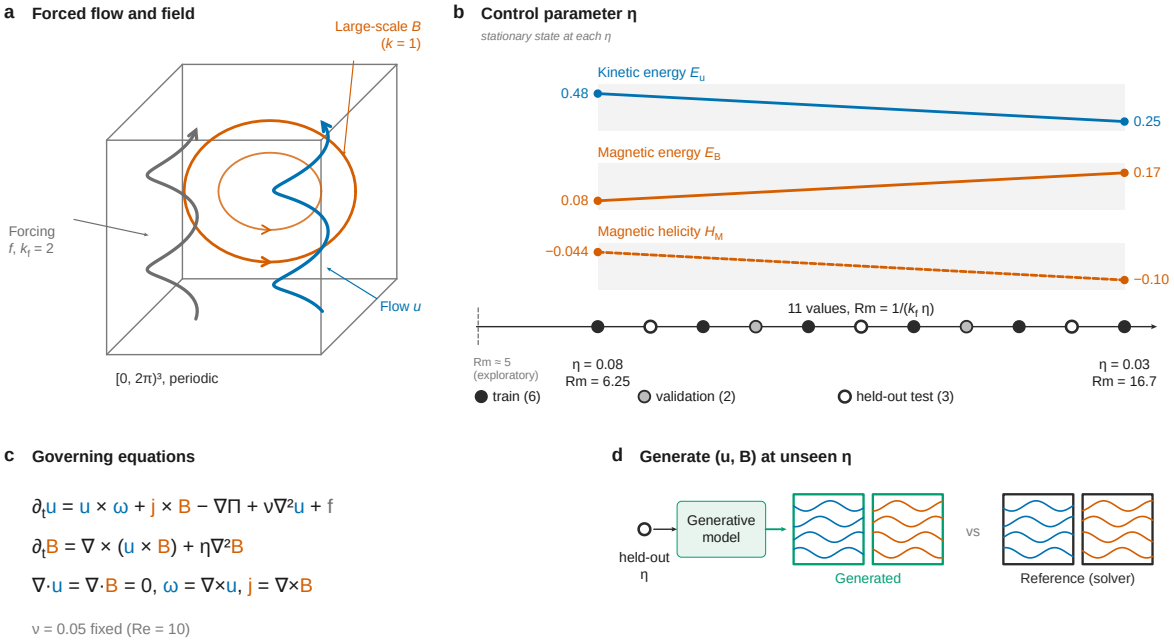

Figure 1.1: The parametric problem of experiment E4. (a) Periodic box with the helical ABC forcing at wavenumber $k_f = 2$, the flow it drives and the large-scale magnetic field at $k = 1$. (b) The eleven values of the magnetic diffusivity η, alternating between training, validation and held-out test values, with the trends of the kinetic energy, magnetic energy and magnetic helicity of the saturated state (end values measured, segments schematic). (c) The equations; the viscosity is fixed. (d) The task: draw snapshots (u, B) at a held-out η and compare their law with that of the solver. Schematic drawn with the `schematic-figure` skill; source in `figures/mhd-dynamo-problem/`.

negative results are kept. Forty-one hypotheses are in the register, thirty-eight of them in this experiment (H4 to H41), all decided; Section E lists them with their status and says where each group came from.

1.3 What the offer asks and what this report covers

Table 1.1 takes the requirements of the offer one by one (quotations from the offer, `docs/specs/sources/supervisors.md`, §1.2) and says what this report does with each. Three of them are not covered, four are covered only in part, and one, the preservation of topology, is measured without being achieved: the report is a comparative study of constraint mechanisms on a small system, and it is not a generator of turbulent dynamo fields.

Table 1.1: The requirements of the internship offer and what this report covers.

What the offer asks	Status	What this report does	Where
Generative models, "diffusion or flow matching"	Partly	Conditional flow matching throughout; one diffusion model, of the first generation only; two Gaussian baselines without a network; no learned model as trained is closer to the reference than a Gaussian law	Section 3.3, Table 6.1
Conditioning on "Reynolds number, magnetic Reynolds number, material properties"	Partly	One parameter, the magnetic diffusivity η ($Rm = 6.25$ to 16.7 at $Re = 10$), in the saturated regime, with validation and held-out values	Section 4.1, Section 9.3
Constraints "through model design, training losses, or sampling procedures"	Covered	Eleven constraint mechanisms (four by design, two by the loss, five at sampling time), with an unconstrained baseline, a diffusion model and three Gaussian laws, compared by the place where they act	Section 3.4, Comparative study
"Divergence-free conditions"	Covered	Exact by design or by projection (relative divergence of order 10^{-8}); not enforced by a penalty	Section 5.1
"Energy balance"	Partly	Constrained at the level of the law only, by a loss or by tilting; the best balance is that of second moments copied from the data at sampling time, not a trained constraint	Section 5.7, Section 5.6
Metrics: "energy spectra, helicity, divergence control"	Covered	Columns of the comparative table, with the production split, the restart in the solver and the energy distance to the reference law	Section 3.6, Table 6.1
Metric: "topology preservation"	Measured, not achieved	Number of magnetic null points: two to three times the reference for every law close to it; within the thresholds of the pilot (Table D.1) only after 4 time units of the solver	Section 5.8, Section 5.6

What the offer asks	Status	What this report does	Where
"3D fields that can serve as input for reduced-order models or data assimilation"	Not covered	No reduced-order model and no data assimilation was built. The nearest test is the restart of generated fields in the solver, which accepts none of them as stationary. The reference fields are published in the data format of the project, with the models and a script that draws generated fields into the same format	Section 5.5, Section 5.6, Section 8
Context: "abrupt reversals of the magnetic field polarity"	Not covered	The models draw independent snapshots of the stationary law. Measured: the symmetry of the generated law under $B \to -B$ and, descriptively, the loss of the sign of the large-scale modes along the reference trajectories	Section 5.9
Context: "large 3D datasets that are difficult to store, interpret, and reuse"	Not covered	A 16^3 grid and 2 GiB of data; what a larger system would need is listed, not done	Section 9.4
Deliverable: comparative study	Covered, with a restriction	One table of mechanisms and metrics, one table of hypotheses with their verdicts; twelve of its sixteen lines are models more than 100 floors from the reference	Comparative study
Deliverable: code "integrated into the MilaDy project toolbox"	Partly	The package `src/milady/` of this repository, with tests, the data set in its data format and a catalogue; the toolbox of the consortium was not available	Toolbox

1.4 Contributions

All of the following are measured in this repository, on the system described above.

1. **A data set and a metric suite for a parametric 3D dynamo** (Experiments): 11 values of η, 24 trajectories each, with a convergence check on finer grids, and metrics covering the four families of the offer (spectra, helicity, divergence, topology), the energy budgets, and two tests that go beyond statistics chosen in advance, a restart test in the solver and a polarity test.
2. **A comparison of constraint mechanisms by the place where they act** (Comparative study): by design (vector potentials, equivariance under $B \to -B$, an explicit affine term, an anchor on a Gaussian law), by the loss (divergence penalty, law-level stationarity identities) and by the sampling procedure (Leray projection, projected flow, exponential tilting, per-sample rescaling, recolouring of the samples onto given moments), against an unconstrained baseline, a diffusion model and a Gaussian baseline without a network.
3. **A diagnosis of why the first generators fail, a remedy, and its limit** (Results): networks narrower than the latent dimension do worse than the trivial Gaussian velocity

field; giving the velocity field its linear part explicitly reduces the distance to the reference law by a factor of about twenty; two Gaussian baselines without a network remain closer to the reference in energy distance, and the second, which keeps the covariance between Fourier modes, has the magnetic production that the first lacks; a split of the production and a Gaussian law fitted to the samples of the best learned model show what that model holds beyond its first two moments (part of a third-order production) and how little this changes its null count and its energy distance; surrogate training sets show that this part is not produced by the model from Gaussian data; and a linear recolouring of its samples at sampling time joins it to the second moments of the Gaussian law (Section 5.6).

4. **A stationarity identity without parameters**, $\mathbb{E}[\langle j \cdot B \rangle] = 0$, that follows from the magnetic-helicity budget and can be imposed at parameter values without data (Methods), and the measurement of what a law-level loss built on it and on the energy identity does and does not change.
5. **Code and data in the project toolbox** (Toolbox): sixteen tested registry entries (eleven constraint mechanisms, the unconstrained baseline, a diffusion model, three Gaussian laws), a null-point finder with the Kac-Rice factors of the count, the tooling to run and account for GPU jobs, and the data set in the data format of the project, the model files and a catalogue, published as a release (Section 8).

The report also states what was refuted, by the data or by the reviewing roles, including its own first interpretation of the results (Discussion).

1.5 Outline

The section State of the art places the work in the literature. The section Methods defines the system, the generative model, each constraint mechanism and each metric. The section Experiments gives the data and the protocol. The section Results reports the measurements, the section Comparative study gathers them in one table with the verdict of each hypothesis, the section Toolbox describes the code, the section Data and models the published files, the section Discussion discusses negative results and limits and lists what the turbulent regime, the dynamo threshold and the data of the project would need (Section 9.4), and the section Reproducibility lists the commands and commits. The appendices hold the detail of what the text summarises and what has no bearing on the requirements of the offer: the detail of the tests on the non-Gaussian part (Section A), longer trainings and wider networks (Section B), the replication on other data (Section C), the objectives of the race (Section D), the register of the hypotheses (Section E), the compute path (Section F), the workflow and the organisation of the agents (Section G), the history of the corrections (Section H), the incidents of the pilot (Section I) and the GPU jobs and costs (Section J).

2 State of the art

Everything in this section is **published** work, cited from the verified source files of the repository (`docs/specs/sources/literature.md`, `dynamo.md`, `data-metrics.md`); the full references are in the section References.

2.1 Diffusion and flow matching

Diffusion models learn to reverse a noising process, either as a discrete chain (Ho, Jain, and Abbeel 2020) or as a stochastic differential equation whose reverse drift involves the score of the noised law (Song et al. 2021). Flow matching (Lipman et al. 2023) and the closely related rectified flows (Liu, Gong, and Liu 2023) and stochastic interpolants (Albergo and Vanden-Eijnden 2023) learn instead the velocity field of an ordinary differential equation that transports a simple law to the data law; with a straight path between a noise sample and a data sample, the regression target is the difference of the two. The parameterisation of a diffusion model by a "velocity" $v = \alpha_t \varepsilon - \sigma_t x$ was introduced by Salimans and Ho (2022) (their Section 4) and is the one used for the diffusion baseline of this report. Both families extend to function spaces (Kerrigan, Migliorini, and Smyth 2024; Lim et al. 2025).

2.2 Generative models of flows and of PDE solutions

Generative models have been applied to turbulent flows to draw turbulent states directly, without rolling out a simulation (Lienen et al. 2024), to PDE solutions in function space with constraints built into the architecture (Wang et al. 2026), to reconstruction from sparse data (Shu, Li, and Barati Farimani 2023; Huang et al. 2024), and to latent representations of flows (Du et al. 2024). FunDiff (Wang et al. 2026) is the main reference of the offer: a generative model over function spaces, with a latent model trained by rectified flow and a decoder that can enforce constraints such as zero divergence through a stream function.

2.3 Three places for a physical constraint

By design. A field written as the curl of a potential is divergence-free whatever the potential; Richter-Powell, Lipman, and Chen (2022) build divergence-free networks on this principle, and FunDiff uses a stream function. Symmetries can also be built in: the flow of an equivariant vector field transports an invariant base law to an invariant law (Köhler, Klein, and Noé 2020).

By the training loss. Physics-informed diffusion models add the residual of the governing equation, evaluated on a denoised estimate, to the training loss (Bastek, Sun, and Kochmann 2025). The estimate is a conditional mean, so for a nonlinear residual the penalty does not vanish at the optimum (the Jensen gap discussed in that paper). Baldan et al. (2026) report a trade-off between the distributional accuracy and the physical accuracy of flow-matching models trained with a residual term.

By the sampling procedure. Constraints can be imposed during sampling by guidance with the gradient of a residual (Chung et al. 2023; Huang et al. 2024), by projection of each iterate onto the constraint set (Christopher, Baek, and Fioretto 2024), or by corrections interleaved with the integration of a flow-matching model, without retraining (Cheng et al. 2025; Utkarsh et al. 2025). For incompressible flows, the Leray projection has been used inside diffusion samplers (Genuist et al. 2026; Li et al. 2026).

2.4 Helical dynamos in periodic boxes

A flow forced with helicity at an intermediate scale in a periodic box generates a strong magnetic field at the scale of the box, and the time needed to build it grows with Rm (Brandenburg 2001); that paper attributes the underlying inverse cascade of magnetic helicity to Pouquet, Frisch, and Léorat (1976). The magnetic helicity $\langle A \cdot B \rangle$ is an invariant of ideal MHD (Woltjer 1958) and evolves, in a periodic domain, as $\mathrm{d}\langle A \cdot B \rangle / \mathrm{d}t = -2\eta \langle j \cdot B \rangle$ [the review of Brandenburg and Subramanian (2005), their Section 3.4, as recorded in the source file; in this work the identity is also checked numerically against the solver by a test]. The forcing used here is the ABC (Arnold–Beltrami–Childress) flow, whose kinematic dynamo properties have been studied in detail (Alexakis 2011; and Galloway and Frisch 1986, known here from its bibliographic record only). Dealiased pseudo-spectral codes are a standard tool for such periodic boxes (for instance Ponty et al. 2005).

2.5 Metrics

The energy distance of Székely and Rizzo (2013) is a distance between two laws computable from samples; it is zero only when the laws are equal. Magnetic null points, where $B = 0$, organise the topology of a magnetic field (Parnell et al. 1996); the trilinear method of Haynes and Parnell (2007) locates them on a grid.

2.6 What was not found

A literature search made for this repository on 6 November 2026 and extended on 12 November found no generative model trained on dynamo simulations, no use of the magnetic-helicity budget as a constraint of a generative model, no use of null-point counts as an evaluation metric of a generative model, and no restart test of generated MHD snapshots. These are negative statements: an absence in a search is not a proof of novelty, and the source files mark them as unconfirmed.

The repository's own theory brief (`docs/specs/theory-brief.md`) gives the propositions used below. Its Propositions 1 to 7 are elementary and probably known in other forms; in particular the use of weak stationarity identities as the loss of a sample generator is published (Cai et al. 2026), and the present work only claims their use for a field PDE, at the level of the law of a parametric conditional model.

3 Methods

3.1 The physical system

The fields are the velocity u and the magnetic field B (in Alfvén units) of an incompressible conducting fluid in the periodic box $[0, 2\pi)^3$:

$$\partial_t u = u \times \omega + j \times B - \nabla \Pi + \nu \nabla^2 u + f, \qquad \partial_t B = \nabla \times (u \times B) + \eta \nabla^2 B, \qquad \nabla \cdot u = \nabla \cdot B = 0, \tag{3.1}$$

with the vorticity $\omega = \nabla \times u$, the current density $j = \nabla \times B$ and the total pressure Π. The forcing is a constant ABC force at wavenumber $k_f = 2$,

$$f = \nu k_f^2 \, (\sin k_f z + \cos k_f y, \ \sin k_f x + \cos k_f z, \ \sin k_f y + \cos k_f x),$$

which is helical ($\nabla \times f = k_f f$) and whose amplitude makes the unit ABC flow the laminar solution without a magnetic field. The viscosity is fixed, $\nu = 0.05$, so that $Re = 1/(k_f \nu) = 10$. The control parameter is the magnetic diffusivity η, between 0.03 and 0.08, that is $Rm = 1/(k_f \eta)$ between 6.2 and 16.7.

The solver is the pseudo-spectral code of the toolbox (`milady.sim.mhd3d`) on a 16^3 grid with the 2/3 rule: the Fourier modes with $|m_i| \le 5$ in each direction are kept. Counting one mode for each pair $(m, -m)$ of a real field, that is $(11^3 - 1)/2 = 665$ independent complex modes.

3.2 Stationarity identities

A statistically stationary state has an invariant law μ_η. For any smooth observable φ of the state, the mean of its time derivative under μ_η is zero (Proposition 5 of the theory brief). Three quadratic observables give identities that involve only quantities computable on one snapshot (Figure 3.1):

$$\begin{aligned} \text{energy:} \quad & \mathbb{E}\big[\langle f \cdot u \rangle - \nu \langle |\omega|^2 \rangle - \eta \langle |j|^2 \rangle\big] = 0, \\ \text{magnetic helicity:} \quad & \mathbb{E}\big[\langle j \cdot B \rangle\big] = 0, \\ \text{cross helicity:} \quad & \mathbb{E}\big[\langle f \cdot B \rangle - (\nu + \eta) \langle \omega \cdot j \rangle\big] = 0, \end{aligned} \tag{3.2}$$

where $\langle \cdot \rangle$ is the spatial mean and $\mathbb{E}$ the mean over the law. The first is the balance between injection and dissipation. The second follows from the magnetic-helicity budget quoted above: in a periodic box the magnetic helicity changes only through $-2\eta \langle j \cdot B \rangle$, so the mean current helicity of a stationary state vanishes. It contains no parameter and no forcing term: it holds at every η, including values without data. The third is odd in B and is satisfied trivially by a law that is invariant under $B \to -B$.

A fourth identity is used as a metric and not imposed: the magnetic energy budget, $\mathbb{E}[P - \eta\langle|j|^2\rangle] = 0$ with the production $P = \langle j \cdot (u \times B)\rangle$, the balance between the work of the flow on the field (the dynamo term) and the Ohmic dissipation. P is cubic in the fields: it is a sum over triads of wavevectors $k+p+q=0$ of third-order moments $\mathbb{E}[\hat{u}(k)\hat{B}(p)\hat{B}(q)]$. For a Gaussian law with zero mean magnetic field these moments reduce exactly to $\mathbb{E}[\hat{u}(k)]\,\mathrm{Cov}(\hat{B}(p), \hat{B}(q))$: the mean production of a Gaussian law is the contraction of its mean flow with the covariance of the magnetic field between different Fourier modes. The constant forcing gives this system a strong mean flow at k_f and statistics that are not homogeneous, so that covariance exists. A law whose Fourier modes are independent has none of it and therefore no mean production, whatever its spectra. A Gaussian law that keeps the covariance between modes has the part of the production carried by second-order statistics; what it cannot have is the third-order cumulant of the true law. This elementary remark separates the two Gaussian baselines used below.

These are constraints on the law, not on each sample: the injection and the dissipation of one snapshot differ, and their difference is the instantaneous $\mathrm{d}E/\mathrm{d}t$. Forcing each sample to satisfy the balance sets the variance of $\mathrm{d}E/\mathrm{d}t$ to zero and cannot give the right law (Corollary 6 of the theory brief).

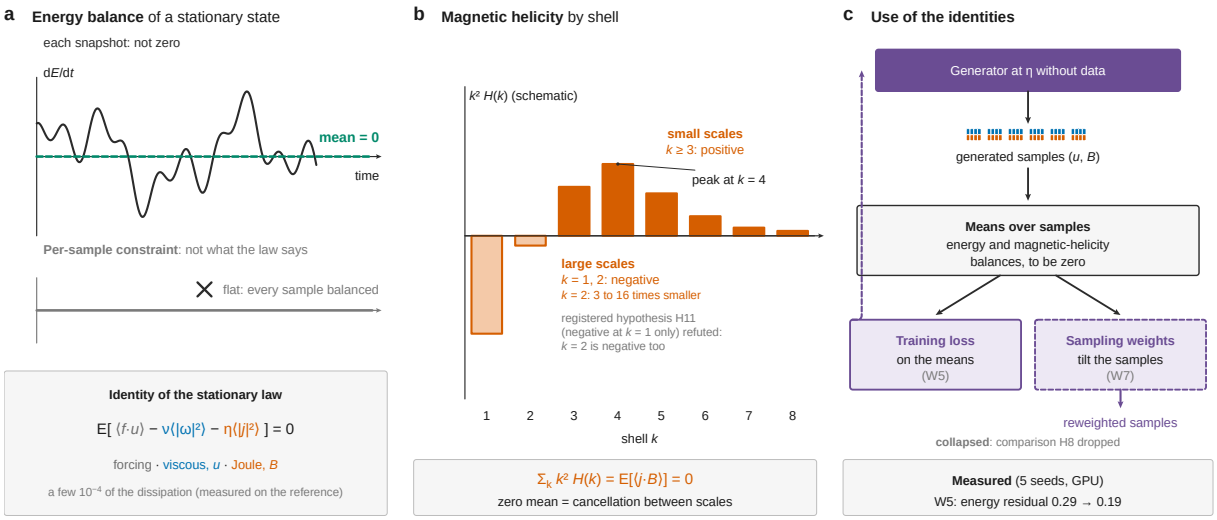

Figure 3.1: The three stationarity identities used in E4. Each is the mean, over the stationary law, of the time derivative of a quadratic invariant of ideal MHD; each term is computable on a single snapshot. The magnetic-helicity identity has no parameter in it. Schematic; source in `figures/stationarity-identities/`.

3.3 Latent representation and generative model

A snapshot (u, B) is transformed to Fourier space and restricted to the 665 modes of the solver. Two representations are used (Figure 3.2).

Free components. The three complex components of $\hat{u}$ and $\hat{B}$ at each mode: $665 \times 12 = 7980$ real numbers. Nothing forces $k \cdot \hat{u} = 0$ or $k \cdot \hat{B} = 0$.

Vector potentials. The potentials in Coulomb gauge, $\hat{A} = \mathrm{i}k \times \hat{B}/|k|^2$ and the same for u, written in the Craya–Herring basis (e_1, e_2) of the plane perpendicular to k: two complex numbers per mode and per field, $665 \times 8 = 5320$ real numbers. Decoding, $\hat{B} = \mathrm{i}k \times \hat{A}$, gives a divergence-free field whatever the latent (Proposition 1 of the theory brief). Because the modes are exactly

those of the solver, every quadratic quantity of Equation 3.2 computed from the latent is the one the solver would compute.

Both latents are normalised mode by mode. The generative model is conditional flow matching on the straight path (Lipman et al. 2023): with $x_0 \sim \mathcal{N}(0, I)$, a data latent x_1 at the condition c (the normalised $\log \eta$) and $x_t = (1-t)x_0 + tx_1$, a network $v_\theta(x, t, c)$ is trained by regression,

$$\mathcal{J}(\theta) = \mathbb{E}\left\|v_\theta(x_t, t, c) - (x_1 - x_0)\right\|^2, \tag{3.3}$$

and a sample is obtained by integrating $\mathrm{d}x/\mathrm{d}t = v_\theta(x, t, c)$ from $t = 0$ to 1 with Heun's scheme (50 steps). The network is a multilayer perceptron acting on the flattened latent: a linear map from the latent to a hidden space of width 512, four residual blocks in that space, and a linear map back to the latent, with Fourier embeddings of t and c.

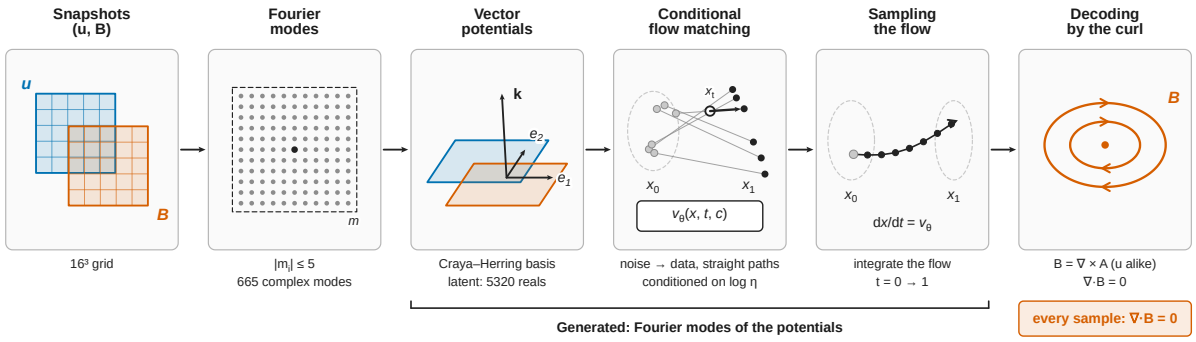

Figure 3.2: The generative pipeline: snapshots, Fourier modes of the solver, vector potentials in the Craya–Herring basis, conditional flow matching, sampling, decoding. Every decoded sample is divergence-free. Schematic; source in `figures/generative-pipeline/`.

3.4 Constraint mechanisms

The mechanisms are grouped by the place where they act (Figure 3.3); the names W0 to W11 and G0 are those of the result files.

Baseline (W0). Flow matching on the free components, no constraint.

Model design.

- *Vector potentials* (W4): flow matching on the potential latent.
- *Equivariance under* $B \to -B$ (W4s). Equation 3.1 is invariant under $B \to -B$, so the stationary law is too (when it is unique). With S the map that changes the sign of the magnetic channels, the field $v_{\mathrm{sym}}(x) = \frac{1}{2}\left[v(x) + Sv(Sx)\right]$ is equivariant and the generated law is exactly invariant (Proposition 4 of the theory brief; Köhler, Klein, and Noé (2020)).
- *Explicit affine term* (W9): $v(x, t, c) = G(t, c)\, x + b(t, c) + \mathrm{network}(x, t, c)$, with one 8×8 matrix G and one offset b per Fourier mode, both read linearly from an embedding of (t, c) and initialised at zero. The reason for this term is given in Section 3.5. W9 also uses the smooth embedding of the condition of W4c (frequencies up to half a cycle over the range of $\log \eta$), so it differs from W4s in two ways; the matched control, the same network with the smooth embedding and no affine term, is the variant of width 512 of the width scan.

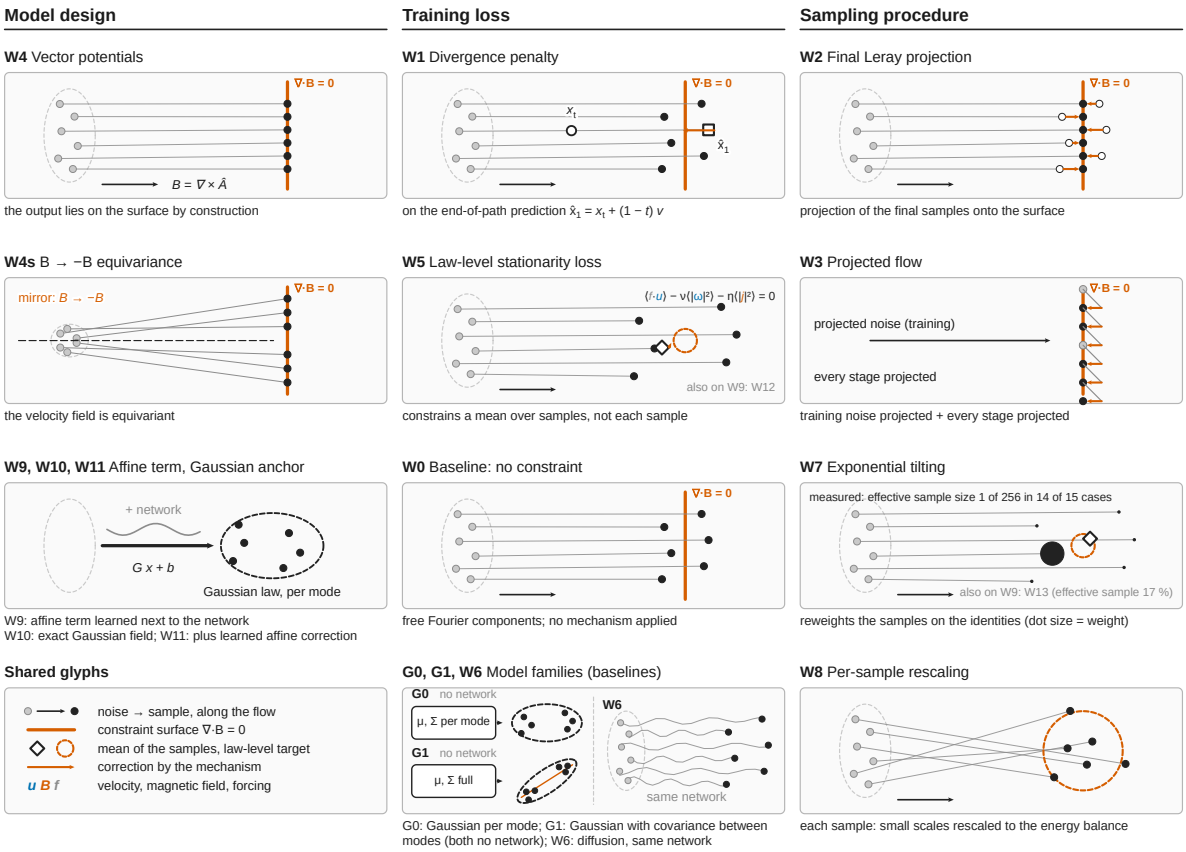

Figure 3.3: The constraint mechanisms compared in E4, by the place where they act: model design, training loss, sampling procedure. Dot positions are illustrative; no result is shown. Schematic; source in `figures/constraint-mechanisms/`.

- *Gaussian anchor* (W10) and *anchor with affine correction* (W11): $v = v_{\text{Gauss}}(x, t, c) +$ network, where v_{Gauss} is the exact flow-matching field of a mode-wise Gaussian law fitted to the data (Equation 3.4); W11 adds the learnable affine term of W9 to the correction.

Training loss.

- *Divergence penalty* (W1): on the free components, the loss Equation 3.3 plus 10 times the squared relative divergence of the end-of-path prediction $\hat{x}_1 = x_t + (1-t)v_\theta$. The constraint is linear, so the penalty vanishes at the optimal field and does not bias it (Proposition 3 of the theory brief).
- *Law-level stationarity loss* (W5): the equivariant potentials model is fine-tuned (400 iterations) with an additional term that draws, at four values of η taken at random in the whole range (with or without data), 16 samples through a differentiable 6-step integration, and penalises the squared normalised means of the energy and magnetic-helicity residuals of Equation 3.2. The square of a mean is estimated without bias by the product of the means of two half-batches; the naive square of one batch mean would add the variance of the residual and penalise the spread of the law. A control (W5_control) receives the same fine-tuning without the term.

Sampling procedure.

- *Final projection* (W2): Leray projection $\hat{f} - k\,(k \cdot \hat{f})/|k|^2$ of the samples of W0.
- *Projected flow* (W3): the noise is projected, the model is trained with projected noise, and every stage of Heun's scheme is projected, so that the whole path stays in the divergence-free subspace (Proposition 2 of the theory brief).
- *Exponential tilting* (W7): the samples of a trained model receive weights $w_i \propto \exp(\lambda \cdot g_i)$, where g_i are the per-sample residuals of the identities and λ is chosen so that the weighted means vanish; this is the law closest to the generated one, in relative entropy, that satisfies the identities (Proposition 7 of the theory brief; the existence and characterisation of this projection are due to Csiszár (1975)).
- *Per-sample rescaling* (W8): the modes above the forcing scale of each sample are multiplied by a factor that makes its own injection equal its own dissipation. This is the per-sample version of a law-level identity, included as the counter-example of Corollary 6.

Loss and sampling on a model close to the reference. The stationarity loss and the tilting were first applied to a model far from the reference (W4s). They were applied again to the model with the affine term: W12 is W9 fine-tuned with the stationarity loss (and a control with the same fine-tuning without the term), and W13 is the tilting of the control samples.

Model families.

- *Diffusion* (W6): the same network and latent as W4, trained by v-prediction on a variance-preserving path (Salimans and Ho 2022) and sampled with the ancestral sampler (100 steps, the same number of network evaluations as 50 Heun steps).
- *Mode-wise Gaussian law* (G0): no network. For each training η, one mean and one 8×8 covariance per Fourier mode are estimated from the data, made invariant under $B \to -B$, and interpolated linearly in $\log \eta$. This law reproduces the second-order statistics of each mode (hence the spectra, the helicities and the mean quadratic budgets) and nothing else; in particular it has no covariance between different modes.
- *Gaussian law with the full covariance* (G1): no network either. For each training η, the mean and the full sample covariance of the 5320-dimensional latent are kept (as a centred data matrix: the covariance has rank at most the number of snapshots), made invariant under $B \to -B$ by adding the image of every snapshot, and interpolated in $\log \eta$ as for G0: a sample at an intermediate η is a Gaussian draw with the convex combination of the

two neighbouring covariances. G1 differs from G0 only by the covariance between modes. It was added after the external review of a first version of this report, to separate what is Gaussian from what is not (Section 5.4). G1 is the point of comparison for any claim that a learned model holds non-Gaussian structure.

3.5 The linear part of the velocity field

Proposition (flow-matching field of a Gaussian law). Let $x_0 \sim \mathcal{N}(0, I)$ and $x_1 \sim \mathcal{N}(\mu, \Sigma)$ be independent and $x_t = (1-t)x_0 + tx_1$. The minimiser of Equation 3.3 is affine in x:

$$v_{\text{Gauss}}(x,t) = \mu + (t\,\Sigma - (1-t)\,I)(t^2\Sigma + (1-t)^2 I)^{-1}(x - t\mu). \tag{3.4}$$

Proof. The minimiser is the conditional mean $\mathbb{E}[x_1 - x_0 \mid x_t = x]$. The pair $(x_1 - x_0, x_t)$ is jointly Gaussian, with $\mathbb{E}[x_1 - x_0] = \mu$, $\mathbb{E}[x_t] = t\mu$, $\mathrm{Cov}(x_1 - x_0, x_t) = t\Sigma - (1-t)I$ and $\mathrm{Cov}(x_t) = t^2\Sigma + (1-t)^2 I$. The conditional mean of a Gaussian vector is the linear regression. □

This is elementary and certainly known. Two consequences matter here. First, for independent standard Gaussian coordinates ($\mu = 0$, $\Sigma = I$) the field is $v = (2t-1)x/(2t^2 - 2t + 1)$ and the minimum of the loss Equation 3.3 per coordinate, averaged over t, is $\int_0^1 \mathrm{d}t/(2t^2 - 2t + 1) = \pi/2 \approx 1.571$: a model whose loss on normalised data is above $\pi/2$ does worse than this trivial field. Second, the network of Section 3.3 maps the latent to a hidden space of width 512 and back, so its Jacobian with respect to x has rank at most 512, whereas the linear part of Equation 3.4 is a full-rank, time-dependent matrix in a latent of 5320 dimensions. That the plain network cannot represent this linear part is exact; that this is the cause of its failure is a **conjecture**, tested by a width scan in the section Results (hypothesis H21). The remedy that does not rely on the conjecture is to give the model the term explicitly (W9) or exactly (W10, W11); Figure 3.4 summarises the argument.

3.6 Metrics

All metrics compare an ensemble of generated snapshots at one η with the reference ensemble at the same η (Figure 3.5).

- **Energy-distance ratio.** Each snapshot is reduced to 21 physical features (logarithms of the shell spectra of u and B for $k = 1, \ldots, 8$, relative kinetic, magnetic and cross helicities, logarithms of the two energies), standardised with the reference. The energy distance (Székely and Rizzo 2013) between 512 generated and 512 reference snapshots is divided by a **noise floor**: the same distance between two reference subsamples of 512 snapshots taken from disjoint halves of the 24 trajectories (mean of 10 splits). The ratio is the number of floors separating the generated law from the reference. A perfect generator would not give 1 but less, because its samples are independent whereas the reference trajectories are correlated in time (an estimate from the two kinds of floors of Figure 4.5 gives about 0.6; it was not measured with a generator); the floor is a scale, not a bound, and it is itself uncertain: its standard deviation over the splits is 35 to 74 % of its mean.
- **Divergence.** $\|\nabla \cdot B\|/\|\nabla B\|$ in the spectral norm, largest value over the ensemble.
- **Energy balance.** $|\mathbb{E}[\langle f \cdot u\rangle - D_\nu - D_\eta]|/\mathbb{E}[D_\nu + D_\eta]$ with the two dissipations $D_\nu = \nu\langle|\omega|^2\rangle$ and $D_\eta = \eta\langle|j|^2\rangle$.

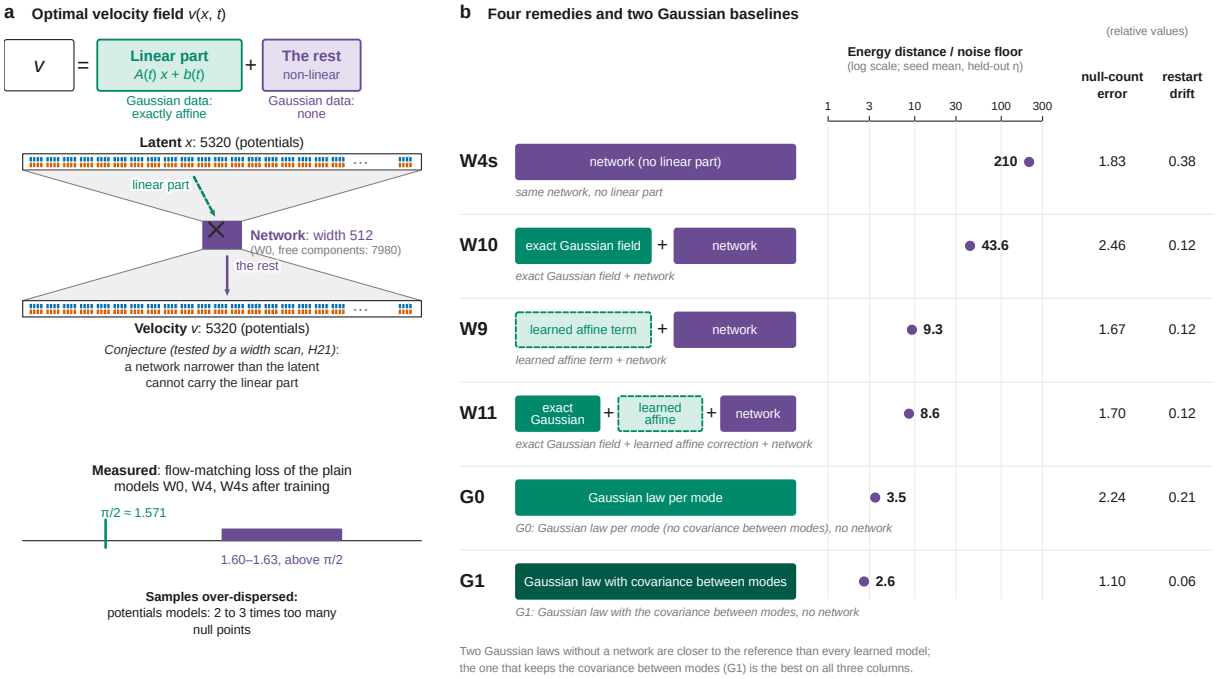

Figure 3.4: Why an explicit linear part. (a) The optimal velocity field of a Gaussian law is affine in x, with a full-rank matrix that depends on time; the plain networks end above the loss of the trivial Gaussian field; the statement on the width is a conjecture. (b) The remedies, the same network without them and the two Gaussian laws without a network: energy distance over the noise floor, null-count error and restart drift of the magnetic energy, seed means over five seeds on the held-out η (measured; values of Table 6.1 and, for G1, Table 5.1). Schematic; source in `figures/linear-part/`.

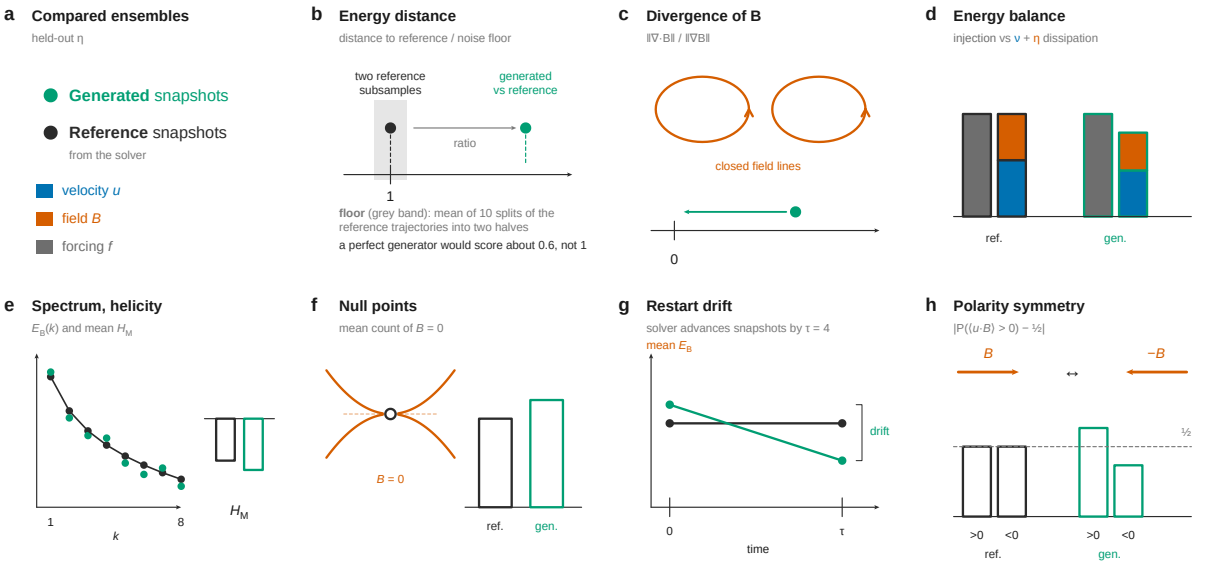

Figure 3.5: The metrics of E4: energy distance over its noise floor, divergence of B, energy balance, spectra and magnetic helicity, magnetic null points, restart drift, polarity. Schematic; source in `figures/metrics/`.

- **Magnetic energy balance.** $|\mathbb{E}[P - D_\eta]|/(|\mathbb{E}[P]| + \mathbb{E}[D_\eta])$ with the production $P = \langle j \cdot (u \times B) \rangle$: zero when the dynamo term balances the Ohmic dissipation, one when there is no production at all.
- **Spectrum and helicity.** Relative L^2 error of the mean shell spectrum $E_B(k)$; relative error of the mean magnetic helicity.
- **Null points.** Relative error of the mean number of magnetic null points per snapshot. The field is evaluated on a 48^3 grid, and the nulls of its trilinear interpolant are found by the method of Haynes and Parnell (2007): cells where the three components change sign, then Newton iterations from nine starting points. On a periodic box the numbers of nulls of the two signs of $\det \nabla B$ must be equal (Poincaré–Hopf), which is used as a test of the implementation.
- **Restart drift.** The generated snapshots are used as initial conditions of the solver and advanced over $\tau = 4$ time units; the metric is the relative change of the ensemble-mean magnetic energy (and, reported with it, of the kinetic energy). It is zero for samples of the true stationary law, and it is the only metric here that tests the generated law against the dynamics rather than against statistics chosen in advance. This is the metric of the objectives, written on the first day, and it has a defect that the external review found late: being relative to the ensemble's own initial energy, it charges a law that starts with a biased energy and relaxes towards the reference for its bias, and a law that starts at the right energy and leaves it for its excursion, without telling the two apart. The last runs therefore also record the course anchored on the reference, $a(\tau) = E_B(\tau)/E_{B,\mathrm{ref}} - 1$ at $\tau = 0.125, 0.25, \ldots, 16$, and its **overshoot** $\max_\tau |a(\tau)| - |a(0)|$: how far the ensemble moves away from the reference beyond its initial deviation (Section 5.5).
- **Polarity.** $|P(\langle u \cdot B \rangle > 0) - \frac{1}{2}|$, zero for a law invariant under $B \to -B$.

4 Experiments

4.1 Reference data

For each of the 11 values of η (logarithmically spaced from 0.08 to 0.03), 24 independent trajectories are started from a small random magnetic seed and a perturbed ABC flow, integrated over 200 time units to let the dynamo saturate, then sampled every 2 time units over 200 time units: 2400 snapshots of (u, B) per value (`experiments/e4_data.py`). Table 4.1 gives the statistics of the saturated state, all **measured**: the kinetic energy falls and the magnetic energy rises as η decreases; the magnetic helicity is negative and its magnitude grows; about 40 % of the magnetic energy sits in the largest scale $k = 1$; the number of magnetic null points per snapshot more than doubles across the range (Figure 4.1).

Table 4.1: Reference data of E4 (measured; `experiments/results/e4_data.json` and `e4_reference.json`). E_u, E_B: kinetic and magnetic energies; $H_M = \langle A \cdot B \rangle$; null points per snapshot (mean $\pm$ standard error over 400 snapshots); $k_\lambda = (\langle |j|^2 \rangle / \langle |B|^2 \rangle)^{1/2}$; energy residual: mean of injection minus dissipation, over the mean dissipation; floor: energy distance between two halves of the reference trajectories.

η	Rm	Split	E_u	E_B	H_M	$E_B(k{=}1)/E_B$	Null points	k_λ	Energy residual	Floor of the energy distance
0.0800	6.2	training	0.491	0.0783	-0.0433	0.436	21.1 $\pm$ 0.47	2.5	-3.0×10^{-5}	0.117
0.0725	6.9	held-out test	0.44	0.0967	-0.056	0.434	21.4 $\pm$ 0.53	2.53	4.4×10^{-6}	0.0703
0.0658	7.6	training	0.402	0.111	-0.0646	0.426	21.9 $\pm$ 0.49	2.57	-4.1×10^{-4}	0.0363
0.0596	8.4	validation	0.373	0.121	-0.0719	0.422	24.2 $\pm$ 0.56	2.61	1.3×10^{-4}	0.0425
0.0540	9.3	training	0.345	0.131	-0.0778	0.413	24 $\pm$ 0.62	2.65	2.2×10^{-4}	0.037
0.0490	10.2	held-out test	0.321	0.14	-0.0841	0.408	27.6 $\pm$ 0.65	2.69	-4.6×10^{-5}	0.0394

η	Rm	Split	E_u	E_B	H_M	$E_B(k{=}1)/E_B$	Null points	k_λ	Energy residual	Floor of the energy distance
0.0444	11.3	training	0.309	0.145	-0.0849	0.396	31.7 ± 0.76	2.74	-2.7×10^{-4}	0.0409
0.0403	12.4	validation	0.288	0.153	-0.0904	0.389	32.5 ± 0.77	2.79	-2.1×10^{-4}	0.0514
0.0365	13.7	training	0.276	0.16	-0.0947	0.384	38.1 ± 0.98	2.84	6.6×10^{-6}	0.0479
0.0331	15.1	held-out test	0.265	0.165	-0.0962	0.371	44.1 ± 1.1	2.9	1.7×10^{-4}	0.0733
0.0300	16.7	training	0.253	0.172	-0.0968	0.364	52 ± 1.2	2.95	2.3×10^{-4}	0.04

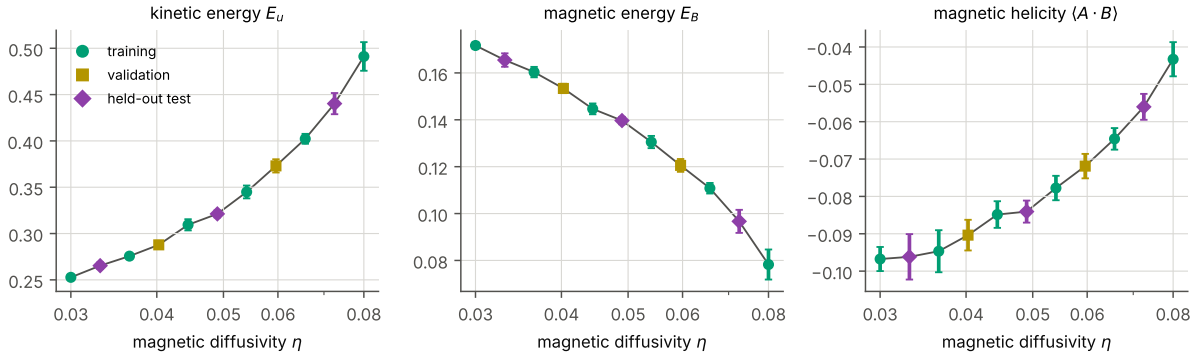

Figure 4.1: Reference data: kinetic energy, magnetic energy and magnetic helicity against η (mean over 24 trajectories; bars: ± 2 standard errors), with the training, validation and held-out values. Data plot from `experiments/results/e4_data.json` (`experiments/e4_plots.py`).

Two properties of the reference were checked before any model was trained.

The stationarity identities hold on the data. The energy residual of Table 4.1 is at most a few 10^{-4} of the dissipation at every η. The mean current helicity $\langle j \cdot B \rangle$ is zero within its standard error: its shell spectrum changes sign, negative at $k \leq 2$ and positive for $k \geq 3$ (Figure 4.2). A hypothesis registered beforehand (H11) predicted a simpler picture, a magnetic helicity spectrum negative at $k = 1$ only and positive for every $k \geq 2$; it was **refuted** by the data, the shell $k = 2$ being negative as well.

The data are converged for the purpose. The same statistics were recomputed at the three held-out η with half the time step and on a 24^3 grid, and at one η on a 32^3 grid (Figure 4.3): energies and helicity agree with those of the 16^3 data within one to two standard errors. The flow at

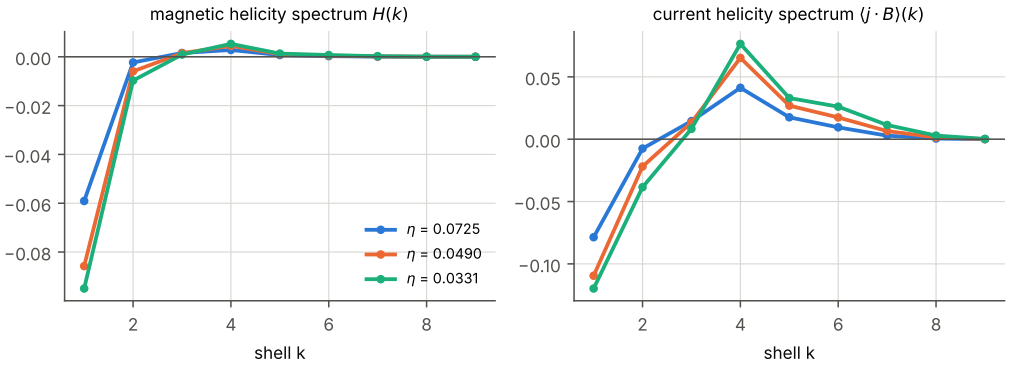

Figure 4.2: Reference data: shell spectra of the magnetic helicity and of the current helicity $\langle j \cdot B \rangle$ at the three held-out η. The mean of $\langle j \cdot B \rangle$ is zero: the negative shells $k \leq 2$ cancel the positive shells $k \geq 3$. Data plot from `experiments/results/e4_reference.json`.

$Re = 10$ is smooth; the small grid is a choice of scale for the comparison, not an under-resolved simulation.

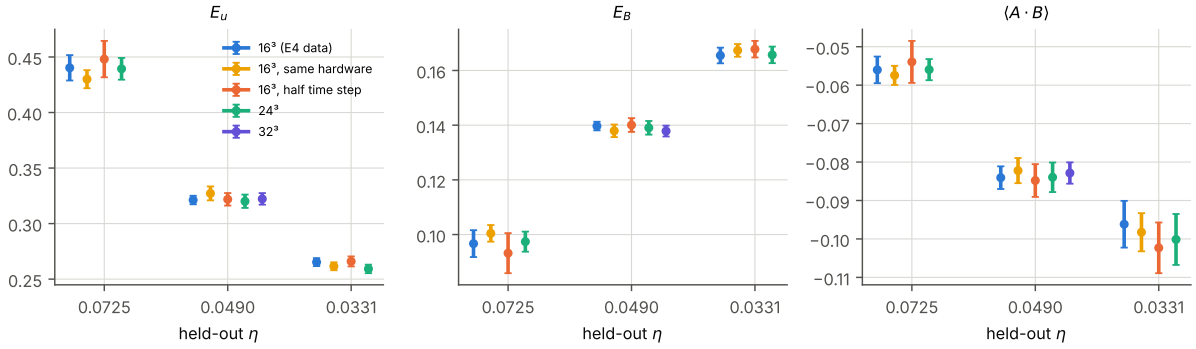

Figure 4.3: Convergence of the reference data: energies and magnetic helicity at the three held-out η for the 16^3 grid of E4, the same grid recomputed on the hardware of the check, half its time step, 24^3 and 32^3 (one η); bars: ± 2 standard errors over 24 trajectories. Data plot from `experiments/results/replications/e4_convergence/`.

4.2 Protocol

Split of the parameter values. The 11 values alternate (Figure 4.4): six training values (the two ends and every second value), two validation values and three held-out test values. The test values are used neither to train nor to choose a model. Every number quoted "on the held-out η" is the mean over the three test values.

Seeds. Each variant is trained with five seeds. A comparison between two variants is the paired difference per seed, quoted as mean $\pm$ standard error over the five seeds (rule 4 of `AGENTS.md`).

Noise floor. The floor of the energy distance (Section 3.6) is computed per η from disjoint halves of the trajectories and not from a random split of the snapshots: snapshots of one trajectory are correlated in time, and a random split gives a floor three to six times lower that no independent run of the solver would reach (Figure 4.5).

Training and sampling. 4000 iterations of Adam, batch 128, learning rate 5×10^{-4}, exponential moving average of the weights; 512 samples per validation and test value, 256 per training value.

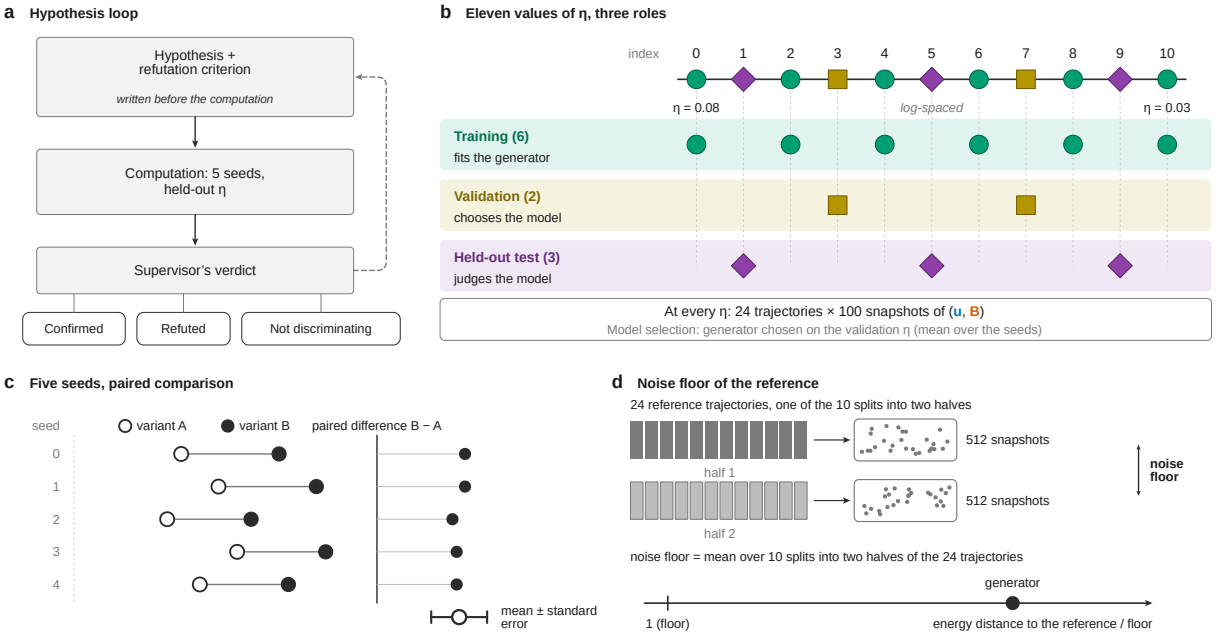

Figure 4.4: The protocol of E4: values of η for training, validation and held-out test; five model seeds; noise floor of the energy distance from two halves of the reference trajectories; hypotheses registered before the computation and reviewed afterwards. Schematic; source in `figures/protocol/`.

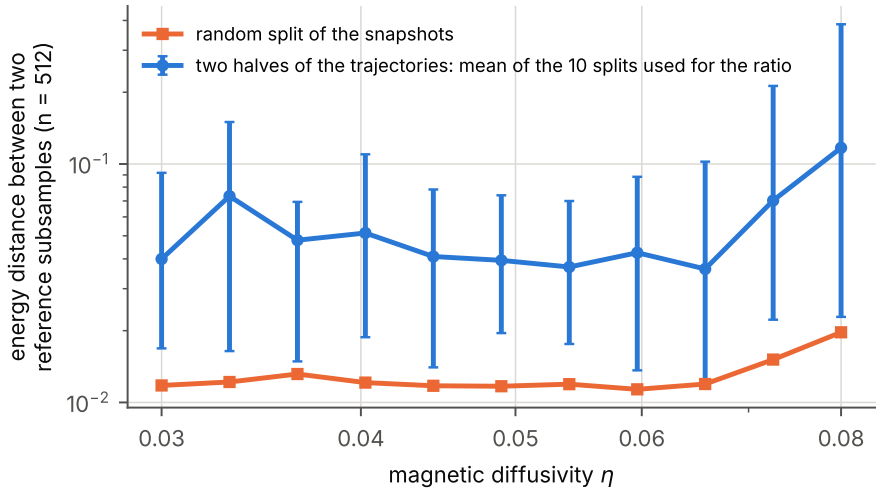

Figure 4.5: Noise floors of the energy distance on the reference data: between two subsamples from disjoint halves of the 24 trajectories (mean of the 10 splits that define the floor; bars: smallest and largest of 20 splits) and between two random subsamples. Data plot from `experiments/results/e4_floor.json`.

The "long" variants use 8 times the iterations. The null points are counted on 200 samples per value and the restart test uses 128 samples.

Register and review. Each hypothesis is written in `research/teams/claude/hypotheses.json` with its refutation criterion before the corresponding run, and the commit of the registration precedes the commit of the results. The `encadrant` role then recomputes the comparison from the result files and gives one verdict per criterion; its reports are in `research/teams/claude/reviews/`. The selection of the generator that is measured against the objectives of the race uses the validation values only.

Compute. The data, the references and the first seeds were computed on the four CPU cores of the session; the five-seed runs were made on rented GPUs through a job tool that clones the pushed commit, runs the command under a time limit and records its cost. The compute path, the workflow of the experiment and the organisation of the agents are in the appendices (Section F, Section G), the jobs and their costs in Section J.

5 Results

Everything in this section is **measured**, on five model seeds unless a table says otherwise; the numbers are in Table 6.1 of the section Comparative study and in the tables below.

The subsections follow the questions of the offer (Table 1.1). Which generative method and which constraint mechanism give a law close to the reference: Section 5.1 to Section 5.4, and Section 5.7 for the loss and sampling mechanisms on a model close to the reference. What a learned model holds that a Gaussian law does not, and how the second moments, and with them the energy balance and the spectrum, can be imposed at sampling time: Section 5.5 and Section 5.6. Topology: Section 5.8. Polarity: Section 5.9. Three complements are summarised in Section 5.10 and detailed in the appendices.

5.1 The registered protocol: every first-generation model is far from the reference

The first protocol, registered before any training, compared the baseline W0, the mechanisms W1 to W5, the diffusion model W6 and the sampling-time corrections W7 and W8, all with the same residual network of width 512 (Figure 5.1).

The outcome is negative for all of them: the generated laws are between about one hundred and five hundred floors away from the reference on the held-out values, and not much closer on the training values. Within this failure the mechanisms can still be told apart, and the registered hypotheses were decided as follows (verdicts of the `encadrant`, Table E.1).

- **Divergence.** The potentials (W4), the final projection (W2) and the projected flow (W3) bring the relative divergence of B to round-off, about 10^{-8}; the free model stays at order one and the penalty (W1) only halves it (H5 confirmed). The divergence of the free model does not change when the number of integration steps is multiplied by four, from 50 to 200 (H14 confirmed): it is an error of the learned field, not of the integrator, as Proposition 2 of the theory brief predicts for a field that lacks the contraction towards the constraint set.
- **Potentials against free components.** Generating potentials does not bring the law closer to the reference than generating free components: the paired difference has the wrong sign in all five seeds (H4 refuted). The projected flow is not better than a projection of the final samples (H6 refuted); the two are the best of this group.
- **The penalty costs accuracy.** The model trained with the divergence penalty is about twice as far from the reference as the baseline. The penalty is unbiased at the optimum, but the network is far from the optimum, and there the penalty competes with the regression.
- **Stationarity loss.** Fine-tuning with the law-level identities lowers the residual of the mean energy balance by about a third against the control that receives the same fine-tuning without the term (H7 confirmed), at parameter values where no data were used. This is the effect on the energy identity only. The residual of the magnetic-helicity identity, the second term of the loss, does not move (-0.004 ± 0.009), and the loss has side effects that the registered criterion did not look at: against the control, the residual of the magnetic

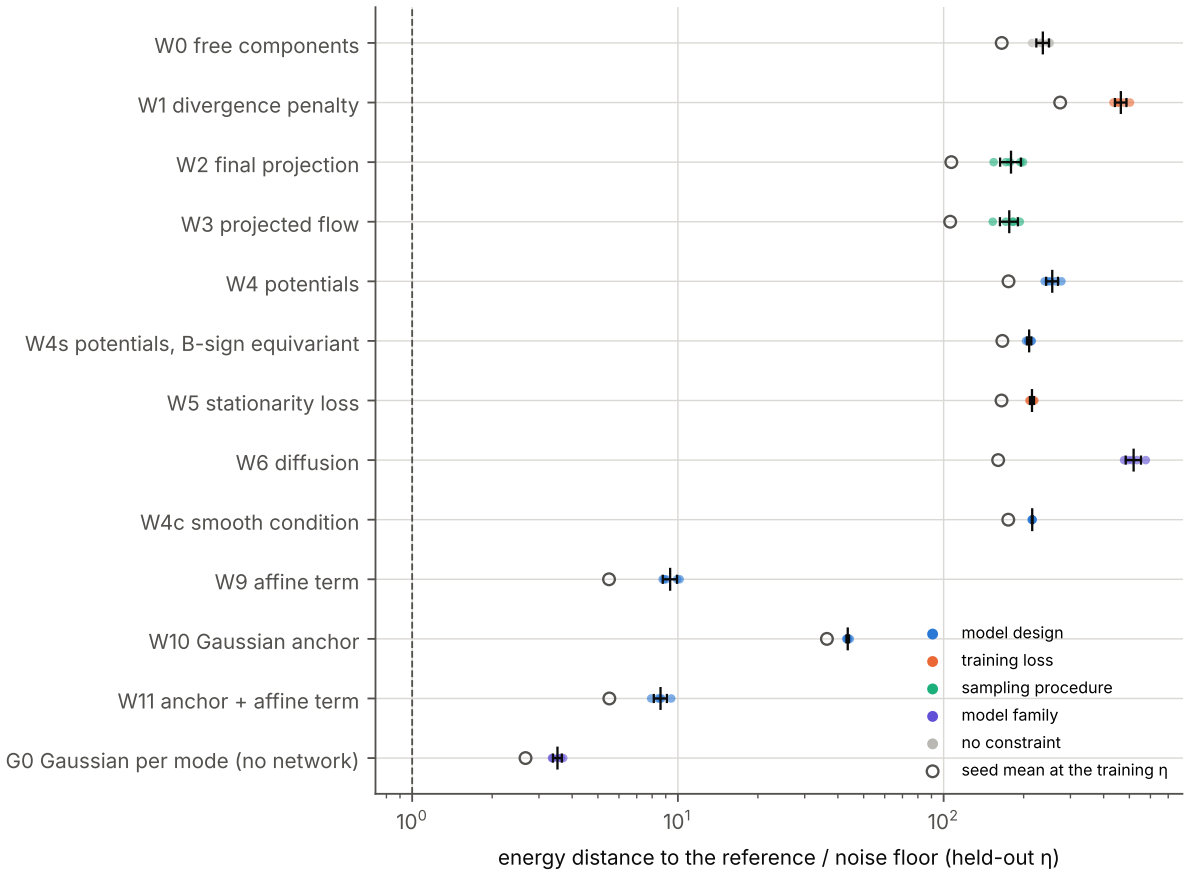

Figure 5.1: Energy distance to the reference divided by the noise floor on the held-out η, per variant: one dot per seed, bar: seed mean ± 2 standard errors (drawn with at least five seeds), open circle: seed mean at the training η; dashed line: the floor. Data plot from `experiments/results/e4_mhd3d_s<k>.json`.

energy balance doubles (0.25 to 0.52), the error on the null count rises by 0.56, the error on the magnetic spectrum falls by 0.56, and the distance to the reference law is unchanged. The comparison between imposing the identities in the loss and by tilting was dropped as not discriminating (H8): the tilted ensemble is one weighted sample, and the registered criterion could be met by its sampling noise.

- **Which identity constrains more.** The registered conjecture that the magnetic-helicity identity is violated more than the energy identity was refuted (H12): on the samples of the equivariant model the normalised residual of $\mathbb{E}[\langle j \cdot B \rangle] = 0$ is about half that of the energy balance.
- **Equivariance.** The polarity asymmetry of the non-equivariant model is already at the sampling noise of 512 samples; the equivariant model cannot lower it (H9 refuted on its first clause). The symmetry is exact by construction in W4s, but in this system it is not what the unconstrained model gets wrong.
- **Flow matching against diffusion.** At the same number of network evaluations, flow matching is about twice closer to the reference than the diffusion model on the held-out η (H10 confirmed as registered). On the training η the order is reversed: the diffusion model is the closer one (160 against 175 floors, in every seed). The advantage of flow matching is in the interpolation in η (Figure 5.2), not in the family at a fixed parameter value.
- **Per-sample rescaling.** W8 satisfies the energy balance of each sample exactly, and has the lowest distance of this group; its law is still about a hundred floors from the reference, and by construction the variance of $\mathrm{d}E/\mathrm{d}t$ of its samples is zero.

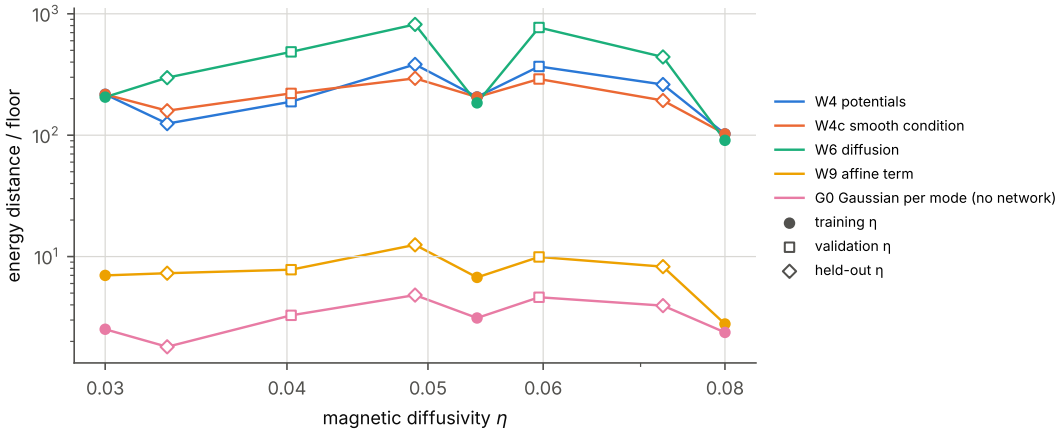

Figure 5.2: Energy-distance ratio against η, seed mean: at the training values (filled circles) and between them (open squares: validation, open diamonds: held-out test). Data plot from `experiments/results/e4_mhd3d_s<k>.json`.

5.2 Diagnosis: the models do worse than a Gaussian field

Three measurements point to the same cause. The flow-matching loss of the plain flow-matching models (W0, W4, W4s and the control of W5; the other variants have another loss or another noise and are not comparable), evaluated on training samples, ends between 1.60 and 1.63, above $\pi/2 \approx 1.571$, the loss of the exact field for independent standard Gaussian coordinates (Section 3.5): after training, these networks fit the normalised data worse than a field that ignores every correlation. The samples of the potentials models are over-dispersed: about two to three times too many magnetic null points (the penalty model W1 and the diffusion model W6 err the other way, with too few) and a fluctuation of the energy budget several times that of the reference. The losses quoted here come from the training logs of the first runs (seeds 0

to 2); the result files store the final loss only for the later variants. And the error is of the same order on the training values as on the held-out ones, so it is not an error of interpolation in η.

Two explanations were registered and tested separately.

The embedding of the condition (H15). The Fourier embedding of the condition used frequencies of up to 8 cycles over the range of $\log \eta$, above what six training values can resolve. A model with a smooth embedding (W4c, frequencies up to half a cycle) was compared with W4.

The linear part of the field (H16, H17). The Gaussian baseline G0 has no network at all; the affine term of W9 gives the network of W4s the linear part of the field explicitly.

The measurements (five seeds, Table 6.1; paired differences in the digest; verdicts of the `encadrant` in Table E.1) separate the two explanations.

- *The embedding matters for interpolation only.* With the smooth embedding the ratio on the held-out η falls from 256 to 215 (paired difference 40.9 ± 6.8, positive in the five seeds), and on the training η nothing changes (175 for both). The registered criterion asked for a factor of two: H15 is **refuted**, the embedding is a real but minor part of the error.
- *A Gaussian law without a network is fifty times closer.* G0 is at 3.5 floors on the held-out η, against 177 for the best trained model of the first protocol, the projected flow W3 (H16 **confirmed**).
- *The affine term is what was missing.* With the explicit mode-wise affine term (W9) the equivariant network goes from 210 floors (W4s, the comparison registered for H17, **confirmed**) to 9.3. W4s also differs from W9 by its embedding of the condition; against the matched control of the width scan (205 floors, Section B) the factor is 22, in every seed, and the flow-matching loss on training samples falls from 1.61 to 1.18, below $\pi/2$.

5.3 Giving the field its linear part

Four ways of giving the velocity field its linear part were compared (Table 6.1, Figure 5.1, Figure 5.3, Figure 5.4).

- **Sampling the mode-wise Gaussian law itself (G0)** is closer to the reference than any learned model in energy distance (3.5 floors), with a residual of the total energy balance of 0.0015 and spectra within a few per cent. It is also an illustration of what a distance built on second-order features of single modes does not see. Its magnetic energy balance is violated completely (residual 0.998): the law has the right Ohmic dissipation and no production, as the remark of Section 3.2 predicts for independent Fourier modes. It has more than three times the reference number of magnetic null points (relative error 2.24). And when the solver restarts from its snapshots the energies move 15 to 20 % away from their stationary values, against less than 1 % for reference snapshots.
- **Learning the affine term next to the network (W9)** gives 9.3 floors, and **anchoring the field on the exact mode-wise Gaussian field with a learned affine correction (W11)** gives 8.6; the two are indistinguishable within the seed noise. Both are about 2.5 times farther from the reference than G0 in energy distance, with a residual of the total energy balance of 0.05 to 0.06, a magnetic spectrum within 15 % and a magnetic helicity within 20 %. But they have what G0 lacks: the residual of the magnetic energy balance is 0.055 and 0.060 (the first-generation models are at 0.15 to 0.5), the error on the null count falls from 2.24 to 1.7, and the restart drift of the magnetic energy at $\tau = 4$ from 0.21 to 0.12 (42 % less), in every seed; what this single-time drift measures for each law is taken apart in Section 5.5.

- **Anchoring on the exact mode-wise Gaussian field with a plain network (W10)** is worse than the Gaussian law it starts from: 43.6 floors (H19 refuted). The registered explanation (H20) was that the network's own linear response spoils the second-order statistics that the anchor had right, and that a learnable affine correction would restore them. The correction does bring the model from 43.6 to 8.6 floors, but not below G0, and H20 was refuted on that clause; its two other clauses, on the restart drift and the null count, are met.

A first version of this report concluded here that the flow-matching models with an affine term had learned "the third-order correlations that make the field a dynamo". The external review refuted that reading, and the next section reports the measurement that replaces it.

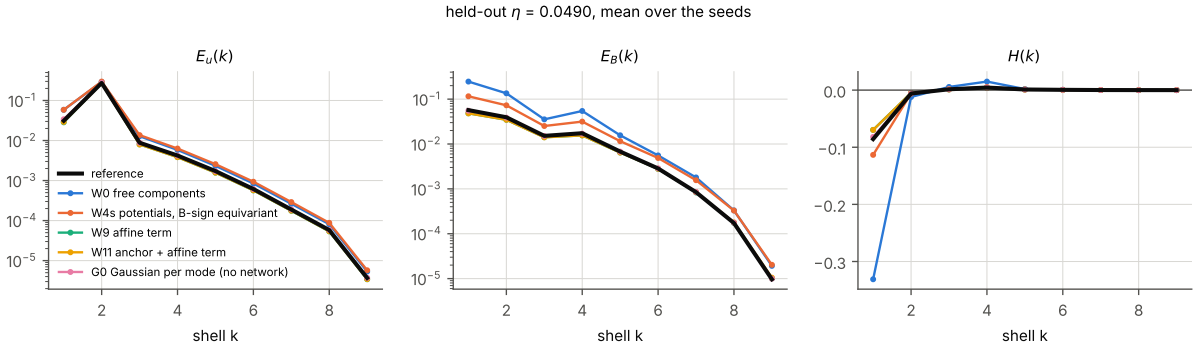

Figure 5.3: Shell spectra of the velocity and of the magnetic field and magnetic helicity spectrum at one held-out η: reference (black) and generated ensembles (mean over the seeds). Data plot from `experiments/results/e4_mhd3d_s<k>.json` and `e4_reference.json`.

5.4 Two Gaussian laws: what the covariance between modes holds

What do W9 and W11 have that G0 lacks? The remark of Section 3.2 gives a candidate that is not third-order: for a Gaussian law, the mean magnetic production is the product of the mean flow and the covariance of B between different modes, and G0 has no such covariance. The hypothesis registered after the review (H22) is that a Gaussian law that keeps this covariance, G1, fitted on the training η only, has a magnetic energy balance residual below 0.2 and beats W9 on the restart drift of the magnetic energy at $\tau = 4$ and on the null count. Table 5.1 gives the measurement, from a job in which G0, W9 and W11 were retrained next to G1 (they reproduce the main run).

H22 is **confirmed**, its three clauses in every seed (its statement, "and not from its Gaussianity", goes beyond its criterion: G1 keeps half of the excess of null points and, as Section 5.5 shows, overshoots the reference energy by 9 % at restart):

- the magnetic energy balance residual of G1 is 0.072, against 0.998 for G0;
- its error on the null count is 1.10, against 2.24 for G0 and 1.66 for W9;
- its restart drift of the magnetic energy at $\tau = 4$ is 0.064, against 0.21 for G0 and 0.12 for W9;
- and its energy-distance ratio on the held-out η is 2.55, against 3.5 for G0 and 9.3 for W9: G1 is the closest law of this report in that distance.

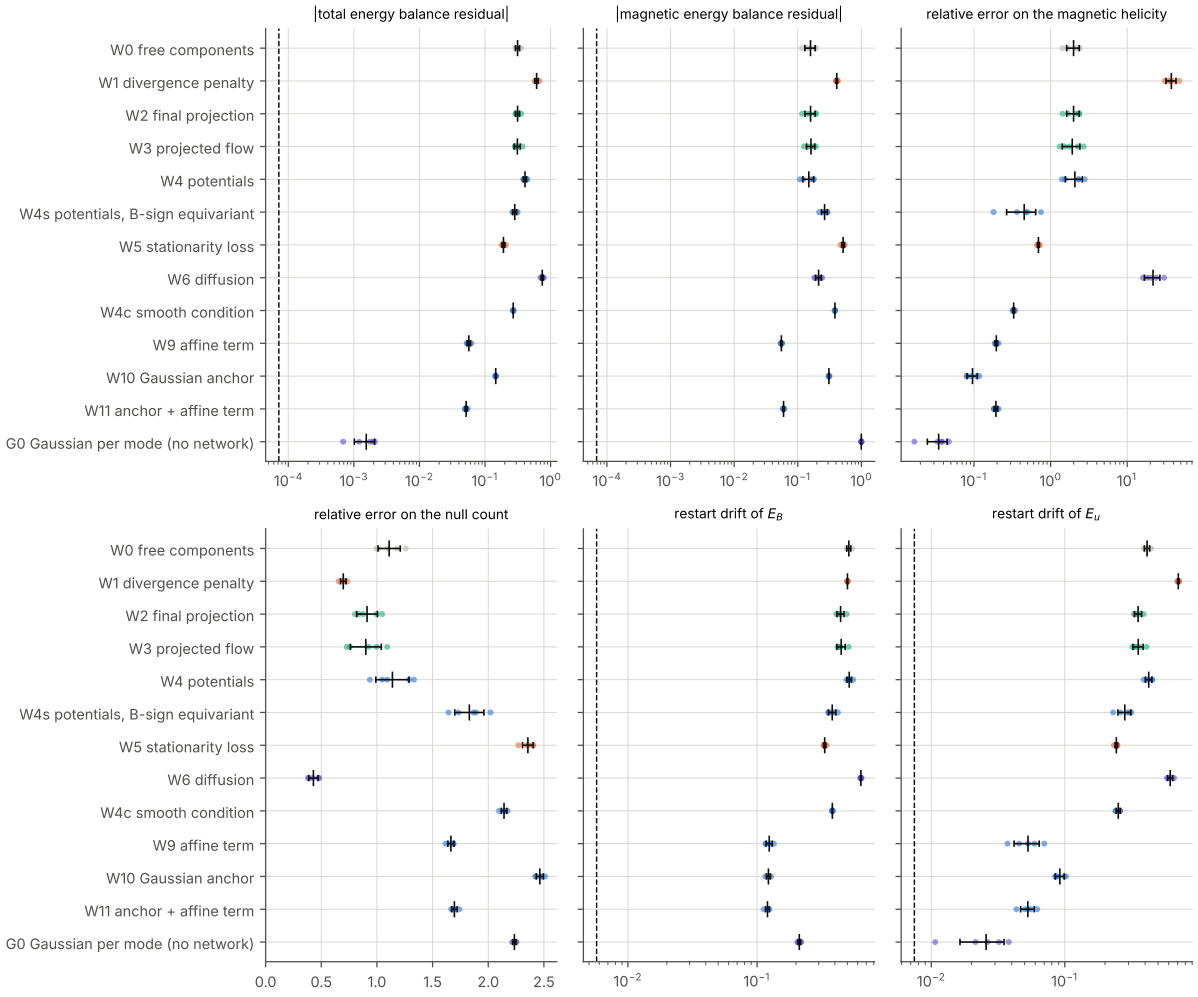

Figure 5.4: Physical metrics on the held-out η, one dot per seed and seed mean ± 2 standard errors: residuals of the mean total and magnetic energy balances, errors on the mean magnetic helicity and on the number of magnetic null points, drifts of the magnetic and kinetic energies when the solver restarts from the generated snapshots. Dashed lines: the same quantity on reference snapshots. Data plot from `experiments/results/e4_mhd3d_s<k>.json`.

G1 uses the training values only: its samples at a held-out η are Gaussian draws from an interpolated covariance, not copies of snapshots. Its five seeds are five draws from one fitted law; the comparison was therefore repeated on the second realisation of the reference data, that of Section C, with the same outcome on each point: 2.0 floors for G1 against 8.1 for W9 and 2.9 for G0, a magnetic balance residual of 0.072 (W9: 0.050), a null-count error of 1.00 against 1.56, a restart drift of the magnetic energy at $\tau = 4$ of 0.06 against 0.13 (`experiments/results/replications/e4_g1indep`).

What this establishes is limited and precise. The covariance between Fourier modes is enough to give a Gaussian law a magnetic production of the right order (0.93 to 1.28 of its Ohmic dissipation, where the mode-wise law has none) and half the spurious null points of the mode-wise law. The advantage of W9 and W11 over G0 on these metrics is therefore not evidence that they hold non-Gaussian structure, which is what the first version of this report claimed.

It does not establish that the learned models hold none, and G1 is not better than them everywhere:

- *Magnetic energy balance.* On the mean over the held-out η, W9 (0.055), the tilted ensemble W13 (0.038) and the widest plain network (0.037) are better than G1 (0.072), on both data realisations for W9; at their worst η, W9 and G1 are comparable. The production of G1 over its Ohmic dissipation is 0.93, 1.12 and 1.28 at the three held-out η, by decreasing η; that of W9 is 0.79, 0.93 and 1.02. A Gaussian law has no third-order cumulant, and at the smallest η G1 over-produces by 28 %. The lower residual of W9 is not an advantage, however: the split of the production in Section 5.5 shows that it results from two errors that cancel.
- *Restart test.* The drift at $\tau = 4$ ranks G1 (0.064) before W9 (0.124) for the magnetic energy and W9 (0.053) before G1 (0.095) for the kinetic energy. This ranking does not survive the anchoring on the reference energy: the drift of W9 is the recovery of a deficit of magnetic energy and that of G1 an excursion away from the right energy (Section 5.5, Figure 5.6). G0 is far behind on both readings.
- *Fluctuations of the energy budget.* The standard deviation of $\mathrm{d}E/\mathrm{d}t$ over that of the reference is 3.0 for G1, 1.6 for W9 and 1.4 for G0.
- *Null points.* Among the laws within 60 floors of the reference, G1 has the lowest error; several laws far from the reference have a lower one (the diffusion model, the rescaled ensemble, the widest plain network), for reasons that have nothing to do with a correct topology.

So: no learned model beats G1 in energy distance, and several are better than both Gaussian laws on the magnetic energy balance. Whether such advantages are non-Gaussian features that the models learned from the reference law could not be said from these runs; it is the question of the next section.

Table 5.1: Gaussian laws and mechanisms on the model with the affine term (measured; `experiments/results/replications/e4_mech`, five seeds, held-out η; the models G0, W9 and W11 are retrained in this job). G1: Gaussian law with the full covariance of the latent. W12: W9 fine-tuned with the stationarity loss, and its control. W13: exponential tilting of the second half of the control samples, to compare with the untilted half. Columns and notation as in Table 6.1.

Variant	Acts through	ED / floor, training η	ED / floor	div B	energy residual	magnetic balance	spectrum of B	magnetic helicity	null points	restart drift E_B	restart drift E_u	polarity	Seeds
G0 Gaussian per mode (no network)	family	2.67(8)	3.52(7)	1.5×10^{-8}	0.0015(3)	0.9981(9)	0.024(3)	0.035(5)	2.235(7)	0.212(3)	0.026(5)	0.018(3)	5
G1 Gaussian, full covariance (no network)	family	1.90(5)	2.55(4)	1.5×10^{-8}	0.0034(6)	0.0720(5)	0.0265(19)	0.037(4)	1.097(11)	0.0644(19)	0.095(3)	0.0148(19)	5
W9 mode-wise affine term	design	5.51(9)	9.3(3)	1.5×10^{-8}	0.0567(17)	0.0553(6)	0.145(3)	0.196(4)	1.665(14)	0.124(3)	0.053(6)	0.019(4)	5
W11 anchor + affine term	design	5.52(13)	8.6(2)	1.5×10^{-8}	0.0515(9)	0.0598(6)	0.143(4)	0.193(5)	1.696(12)	0.120(2)	0.053(3)	0.0169(10)	5
W12 control (same fine-tuning, no identity)	loss	5.46(5)	9.3(4)	1.5×10^{-8}	0.0567(18)	0.0573(8)	0.143(4)	0.195(6)	1.683(11)	0.123(4)	0.053(6)	0.019(2)	5
W12 affine term + stationarity loss	loss	9.5(5)	16.4(7)	1.5×10^{-8}	0.0292(11)	0.113(4)	0.218(7)	0.305(6)	1.981(20)	0.080(5)	0.139(6)	0.015(3)	5
W12 control, half ensemble	loss	4.59(11)	7.3(3)	1.5×10^{-8}	0.0564(19)	0.0564(5)	0.142(4)	0.192(7)	1.669(14)	–	–	0.030(5)	5
W13 affine term + exponential tilting	sampling	3.3(3)	4.6(3)	1.5×10^{-8}	0.011(2)	0.038(4)	0.081(8)	0.088(12)	1.25(2)	–	–	0.083(17)	5

5.5 Beyond a Gaussian law: three tests

The review of the third version of this report asked the question that the comparison with G1 left open, and showed that the restart drift read at one time could not answer it. Three hypotheses were registered (H25 to H27), their protocol attacked by the `jury-externe` role before the run, and two jobs of five seeds were run, one on each realisation of the reference data (`experiments/results/replications/e4_nongauss` and `e4_nongauss_indep`). G0, G1, W9 and W11 are retrained in these jobs and reproduce the earlier runs. Table 5.2 gives the measurements on the data of the main run; the second realisation gives the same verdicts.

The production split (H25). The mean magnetic production splits exactly into two parts,

$$\mathbb{E}\langle j \cdot (u \times B)\rangle \;=\; \mathbb{E}\langle j \cdot (\bar{u} \times B)\rangle \;+\; \mathbb{E}\langle j \cdot ((u-\bar{u}) \times B)\rangle, \qquad \bar{u} = \mathbb{E}[u].$$

The first is carried by the mean flow and the second-order statistics of B; the second is a third-order moment of the fluctuations and vanishes for any Gaussian law. In the reference (Figure 5.5), the third-order part is $+0.054 \pm 0.008$, -0.137 ± 0.004 and -0.270 ± 0.007 of the Ohmic dissipation at the three held-out η, by decreasing η (jackknife over the trajectories, the mean flow recomputed each time). Over the eleven values of η it decreases steadily from $+0.14$ at $\eta = 0.08$ to -0.31 at $\eta = 0.03$ and changes sign near $\eta = 0.066$. The reference law is measurably non-Gaussian in the very term that makes it a dynamo, and at the smallest held-out η the fluctuations remove about a fifth of what the mean flow produces. This is why G1, which has the mean-flow part right within 2 % and no third-order part (below 0.002 for the four Gaussian laws of the table), over-produces by 28 % there.

The hypothesis was that W9 holds none of it. It is **refuted**: the third-order part of W9 is -0.082 and -0.118 at the two lower η, with the sign of the reference in the five seeds, that is 0.60 and 0.44 of the reference value (0.63 and 0.41 on the second realisation); W11 gives the same figures. The Gaussian law fitted to the samples of W9 (G2, below) has none, as it must, which checks the estimate: this is structure of the generated law beyond its first two moments. It is partial: at the largest held-out η, where the third-order part of the reference is positive, that of W9 is small and of the opposite sign (-0.006), and the ratio to the reference is between 0.5 and 1.5 at one held-out value only. This is not an error of interpolation in η: at the three training values where the models are evaluated ($\eta = 0.08$, 0.054 and 0.03), the third-order part of W9 is $+0.008$, -0.069 and -0.121, that is 0.06, 0.73 and 0.39 of the reference. And it comes with a second-order error: the mean-flow part of W9 is 10 to 16 % below that of the reference at every held-out η. The magnetic energy balance of W9 (production over dissipation 0.79, 0.93 and 1.02) is therefore the sum of two errors: a mean-flow part too small by 0.15, 0.13 and 0.13 of the dissipation, and a third-order part off by -0.06, $+0.06$ and $+0.15$. At the two lower η they have opposite signs and cancel, in part at $\eta = 0.049$ (-0.07 is left) and almost exactly at $\eta = 0.033$ ($+0.02$); at the largest they add, and W9 under-produces by 21 %. Its lower residual on the mean over the held-out η (0.055 against 0.072 for G1) comes from this cancellation and is not an advantage.

The restart anchored on the reference (H26). With $a(\tau) = E_B(\tau)/E_{B,\mathrm{ref}} - 1$, reference snapshots stay within about 1 % of their stationary energy over 16 time units (Figure 5.6). W9 starts 12.3 % below it and comes back to it: -8.5 % at $\tau = 1$, -1.4 % at $\tau = 4$, -0.7 % at $\tau = 16$, with an overshoot of 0.0045 on the mean over the held-out η (0.014 at the largest). G1 starts at the right energy (-0.3 %), rises to $+10$ % at $\tau = 2$ and comes back to $+1.2$ % at $\tau = 16$ (overshoot 0.091). G0 starts at -2 %, falls to -18 % at $\tau = 0.5$, rises to $+19$ % at $\tau = 4$ and is still at $+3$ % at $\tau = 16$ (overshoot 0.173). H26 is **confirmed** on its registered criterion, each clause in every seed and on both realisations of the data.

The review of the `encadrant` role adds that the six laws of this job that start on the reference energy or below it all gain 10 to 21 % of it after the restart, and that the overshoot of the criterion gives each law an allowance equal to its initial deviation: the criterion does not show that W9 is closer to the stationary state than G1 (Section A).

The Gaussian law with the moments of W9 (H27). G2 is the Gaussian law with the mean and the full covariance of 8192 samples of W9 at each η; the same construction applied to samples of G1 (G3) is the null control. The hypothesis was that G2 would overshoot more than W9 by at least 0.02, that is, that the course of W9 at restart needs its non-Gaussian part. The paired difference is -0.0051 ± 0.0005 on the data of the main run and -0.0056 ± 0.0005 on the second realisation, negative in every seed: G2 starts from the same deficit and shows no overshoot either. The null control passes on the second realisation (-0.0004 ± 0.0036 against a threshold of 0.02) and misses its threshold by 0.004 on the first (-0.024 ± 0.013; the review traced it to the sampling noise of the initial energy of the 128 restart samples, for laws that start near zero). By the letter of the criterion the test is invalid on the first realisation, and H27 is **refuted** on the second.

What this refutation allows is narrow: the registered measure cannot see an excursion beyond the reference smaller than the initial deficit of W9 (Section A).

Comparing W9 with G2 isolates what its non-Gaussian part changes on the other metrics. It changes the third-order production (lost in G2), the fluctuation of the energy budget (the standard deviation of $\mathrm{d}E/\mathrm{d}t$ over that of the reference is 1.63 for W9 and 2.45 for G2) and the course of both energies at restart. It changes the null count little: the error of G2 is higher by 0.04 ± 0.03 (on the second realisation 0.04 ± 0.02), out of 1.7. And the energy distance barely sees it: samples of G2 and of W9 are 0.43 reference floors apart on the 21 features, about twice the value expected for two samples of one law at this size and below the resolution of the objective (one floor), and G2 is 0.6 to 0.9 floors farther from the reference than W9.

So the answer to the question of this section has two halves. The model with the affine term is not a Gaussian law: it holds a third-order production of the sign of the reference and of 0.6 and 0.4 times its size at two held-out values out of three. But the first protocol could hardly show it. On the null count, one of the two metrics in which the model had seemed to stand out, its non-Gaussian part changes little (0.04 out of 1.7). On the other, the restart, the overshoot registered for the test does not separate the model from its Gaussianisation; the drift at $\tau = 4$ does (0.124 against 0.170 for G2, in the five seeds on both realisations), as does the excursion above the reference computed after the fact (0.024 ± 0.004 higher for G2; a **conjecture** until registered, Section A). The converse holds for the reference. The Gaussian law with the mean and covariance of the reference itself (G1) has 2.1 times its null points on the mean over η, at the training η as at the held-out ones (2.5 times at the largest η, 1.6 at the smallest), and at the held-out η, the only ones where the restart was run and where G1 is interpolated, it leaves the stationary energy by 9 to 11 %. The low null count of the reference is a non-Gaussian property, and so is its stationarity under the solver as far as G1 at interpolated η shows; no model of this report has learned either.

Table 5.2: Production split and restart anchored on the reference energy (measured; `experiments/results/replications/e4_nongauss`, five seeds; notation as in Table 6.1). Mean-flow and third-order parts of the mean magnetic production over the mean Ohmic dissipation, at the three held-out η. $a_B(0)$: initial deviation of the magnetic energy from the stationary energy of the reference; overshoot: $\max_\tau |a_B(\tau)| - |a_B(0)|$ over $\tau = 0.125, \dots, 16$; last column: $|a_B(16)|$; means over the held-out η. G2 and G3: Gaussian laws with the mean and covariance of 8192 samples of W9 and of G1. Reference row: jackknife standard errors for the production, and for the restart the mean curve of 4 draws of 128 snapshots.

Law	mean-flow part, $\eta =$ 0.0725	mean-flow part, $\eta =$ 0.0490	mean-flow part, $\eta =$ 0.0331	third-order part, $\eta =$ 0.0725	third-order part, $\eta =$ 0.0490	third-order part, $\eta =$ 0.0331	$a_B(0)$	overshoot of E_B	$\lvert a_B \rvert$ at $\tau = 16$
reference (SE: jackknife over trajectories; 4 restart draws)	0.946(8)	1.137(4)	1.270(7)	0.054(8)	-0.137(4)	-0.270(7)	0.00388	0.00108	0.00375
G0 Gaussian per mode (no network)	-0.0010(18)	-0.0022(5)	-0.002(3)	2.9×10^{-4}	2.3×10^{-4}	-0.0014(6)	-0.020(4)	0.173(3)	0.028(3)
G1 Gaussian, full covariance (no network)	0.9281(14)	1.1174(17)	1.2797(15)	-4.0×10^{-4}	-7.5×10^{-4}	0.0018(8)	-0.003(3)	0.091(8)	0.014(4)
G3 Gaussian law fitted to G1 samples (null control)	0.9235(12)	1.1243(15)	1.270(3)	2.2×10^{-4}	0.0013(7)	0.0017(8)	-0.011(4)	0.068(8)	0.0175(18)
W9 mode-wise affine term	0.7946(12)	1.007(3)	1.1368(19)	-0.0055(7)	-0.0817(10)	-0.1180(13)	-0.123(6)	0.0045(3)	0.0102(9)
G2 Gaussian law fitted to W9 samples	0.791(5)	1.010(4)	1.134(4)	-7.1×10^{-4}	-4.4×10^{-4}	-4.7×10^{-5}	-0.129(9)	-6.0×10^{-4}	0.016(3)
W11 anchor + affine term	0.7769(9)	1.008(2)	1.141(6)	-0.0096(3)	-0.0813(13)	-0.122(2)	-0.1220(14)	0.0060(2)	0.0096(20)
W10 anchored on the Gaussian field	0.448(3)	0.593(2)	0.670(3)	-0.0120(12)	-0.0426(15)	-0.0637(7)	0.211(3)	-0.0449(7)	0.0142(9)

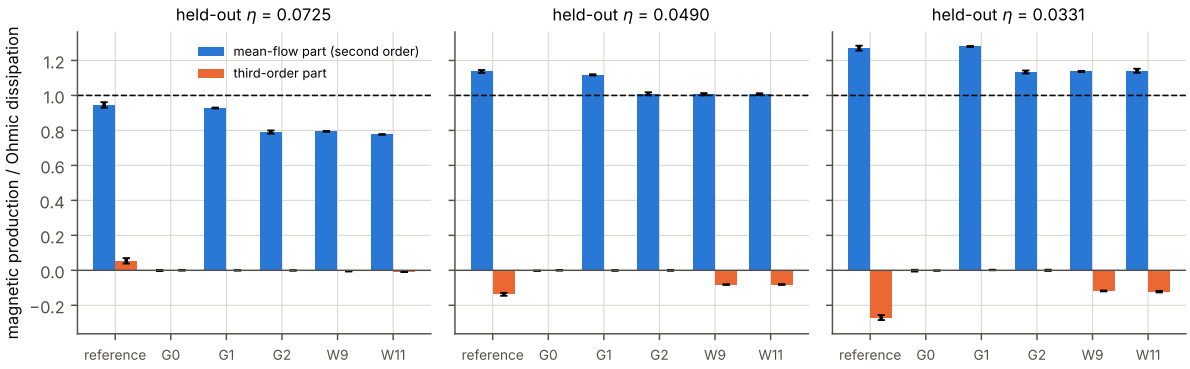

Figure 5.5: Mean magnetic production over the Ohmic dissipation, split into the part carried by the mean flow and the third-order part, at the three held-out η: reference (± 2 jackknife standard errors over the trajectories), the Gaussian laws G0, G1 and G2 and the learned models W9 and W11 (seed means, ± 2 standard errors over the seeds); the two parts of a stationary law sum to one (dashed line). Data plot from `experiments/results/replications/e4_nongauss`.

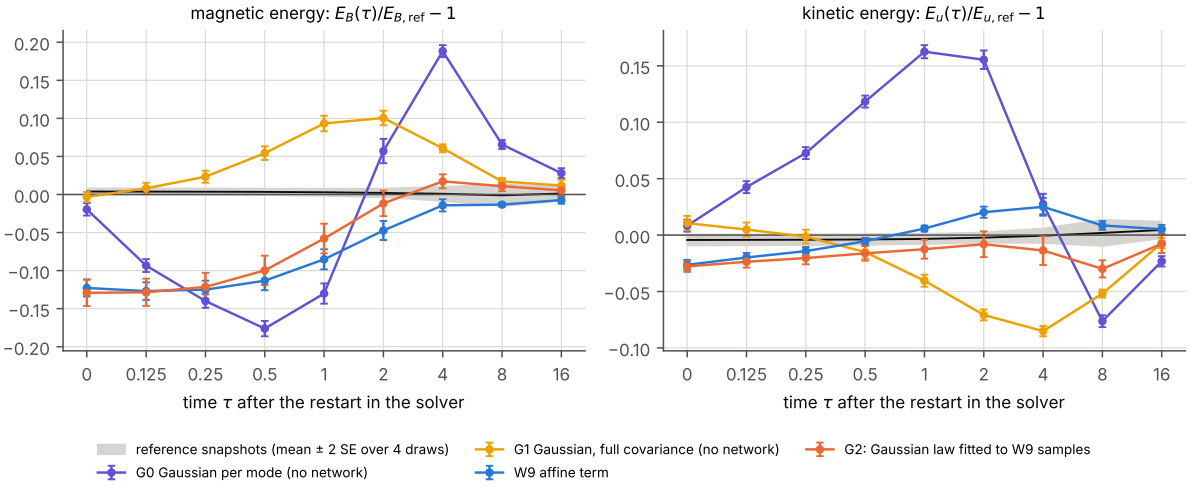

Figure 5.6: Restart test anchored on the reference: relative deviation of the ensemble-mean magnetic and kinetic energies from the stationary energies of the reference, when the solver restarts from generated snapshots, up to 16 time units; mean over the held-out η and the seeds (bars: ± 2 standard errors over the five seeds); grey band: reference snapshots, mean ± 2 standard errors over 4 draws of 128. Data plot from `experiments/results/replications/e4_nongauss`.

5.6 Where the third-order part comes from, and what the right covariance changes

The section above leaves a learned model (W9) that holds part of the non-Gaussian production with wrong second moments, and a Gaussian law (G1) with the right second moments and nothing else. Fourteen further hypotheses (H28 to H41) ask whether the two can be joined, where the third-order part and the second-order error of W9 come from, what the Gaussian law gets wrong, and what a short relaxation in the solver repairs. Each was registered before its run and read by the `encadrant` role; each job has five seeds, and most were run on two realisations of

the data (H38 and H41 on the data of the main run only; H39 on one set of ensembles compared with the two references; H36 on two new data sets). This section reports those that bear on the questions of the offer; the side tests (H34, H36, H39), the dropped hypothesis H32 and the repair of the measurement of H35 are summarised at the end and detailed in the appendix (Section A).

In this budget, a network anchored on a Gaussian law with inter-mode covariance acquired no third-order part (H28, H38, H40, H41). A flow-matching model whose velocity field is the exact field of the Gaussian law G1 plus a network that starts at zero stays at G1 over the whole training (Figure 5.7): at no checkpoint does it hold more than 5×10^{-3} of the third-order production of the reference on the seed means (0.023 in one seed) (H28, **refuted** on both realisations). The failure belongs to the anchored construction, or to its optimisation in this budget and dimension, and says nothing about the data: trained on samples of W9, whose third-order production is known, the same anchored model recovers none of it (H38, **refuted**), and with an anchor of full rank (one block per Fourier mode plus the 300 leading inter-mode directions) the network trains, the law moves from 2.5 to 16 floors from the reference and no third-order production appears (H40, **refuted**). The code shared by these models is cleared for a block-diagonal anchor (H41, not refuted); the anchors of H28, H38 and H40 are dense, and a fault specific to a dense covariance is not excluded. Among the anchored models, a third-order part is measured only when the anchor is the mode-wise Gaussian law; the cause is not established. The full account is in Section A.

A model trained on Gaussian data has no third-order part (H33). If a network left with only the non-Gaussian part learns nothing, the third-order production of W9 might not come from the data at all, but from the way a network narrower than the latent represents the Gaussian part. The test: train the same model, with the same seed and budget, on a surrogate training set that has the first two moments of the data and nothing else (independent draws of G1, as many as there are snapshots at each training η), and compare it with the model trained on the data in the same job. The hypothesis would be refuted if the surrogate-trained model held 0.3 or more of the third-order part of the other. It holds none (Figure 5.9): its third-order part is below 0.001 of the Ohmic dissipation in absolute value at the three held-out η ($+0.0001$ and $+0.0006$ at the two lower ones, against -0.082 and -0.118 for the model trained on the data; -0.0002 and -0.0007 against -0.086 and -0.113 on the second realisation), under 1 % of the other on the seed means and at most 4 % in any seed. The three checks pass (G1 and the surrogate sets have no third-order part; the model trained on the data reproduces the values of H25), one of them on three of the six training η only. H33 is **supported** in the narrow sense written in the register: this model does not produce a third-order production from independent Gaussian data with these two moments. An effect of the time correlation of the 24 trajectories of each η, which the independent draws do not have, is not excluded by this test; it is the object of H34 (Section A), with the same outcome. Trained on Gaussian draws, the same model has no such deficit of magnetic energy (at most 5.3 %) and still misses about half of the mean-flow production that the model trained on the data misses (a descriptive result, in Section A).

Recolouring the model at sampling time (H29). The two results above suggest a division of labour: take the second moments from the Gaussian law, which has them right, and the rest from the learned model. At each η the mean and the covariance of the model are estimated on 131 072 of its samples, and every sample is sent through the linear map that carries these moments onto those of G1 (the optimal-transport map between two Gaussian laws, $A = \Sigma_s^{-1/2}(\Sigma_s^{1/2}\Sigma_t\Sigma_s^{1/2})^{1/2}\Sigma_s^{-1/2}$): the result, W16, has the mean and covariance of G1 by construction and whatever the map leaves of the higher-order structure of W9. W15 does the same mode by mode, onto the moments of G0. A null control, G4, sends Gaussian draws with the moments of W9, taken from a second independent estimate, through the map of W16: it measures what the estimation error of the moments costs. Table 5.3 and Figure 5.8 give the results

on the data of the main run; H29 is **confirmed** on its four clauses on both realisations. W16 has the mean-flow part of the production within 0.017 of the reference and the magnetic energy within 1.3 % (2.6 % on the second realisation), where W9 was 15 % low on both; it holds 0.74 and 0.58 of the third-order production of the reference at the two lower held-out η (0.79 and 0.59), more than W9 before the map (0.60 and 0.43), with the sign of the reference in every seed; its residual of the magnetic balance is 0.034 against 0.072 for G1; and its energy distance is not above that of G1 (2.07 floors against 2.55; 1.61 against 1.98). The control passes: G4 is G1 within 0.03 floors (0.23 on the second realisation) and 1.1 % of magnetic energy, and four times the integration steps change nothing (0.07 floors).

Three restrictions come with this result. At the largest held-out η, where the third-order part of the reference is small and positive, W16 holds none of it and is farther from the reference than G1 by 0.20 floors (0.13): the lower energy distance is a mean over η, and it was not a registered claim (the clause only required W16 not to be farther than G1 by more than half a floor). Recolouring mode by mode gives the same energy distance (W15 against W16: 0.01 ± 0.02 floors), so the covariance between modes is not what the energy distance rewards; what it brings is the mean-flow part of the production (-0.12 of the dissipation for W15, -0.017 for W16), the null count of G1 (error 1.41 against 1.00) and a restart drift twice lower (0.040 against 0.021). And the control understates the estimation error for a source that is not Gaussian: the role found, on a synthetic case, a covariance error 1.7 times larger for a heavy-tailed source.

The restart at equal moments (H30, H31). H27 could not tell whether the non-Gaussian part of W9 matters at restart, because W9 and its Gaussianisation both start 12 % below the reference energy. W16 and G4 start on it, with the same first and second moments up to the estimation error, and differ only by what W9 holds beyond them. Restarted in the solver (512 states), the Gaussian control passes above the reference energy by 0.095 and the recoloured model by 0.046 (0.114 and 0.063 on the second realisation): the paired difference is 0.049 ± 0.005 (0.051 ± 0.003), positive in every seed, and the control behaves as G1 (difference of excursion 0.003). H30 is **confirmed**: at equal moments, the structure of W9 beyond its second moments halves the passage above the reference energy. It cannot be attributed to the third-order production: the difference is the same at the three held-out η (0.055, 0.049, 0.043) while the third-order part of W16 goes from zero to -0.16 of the dissipation. And W16 still leaves the reference energy by 4.6 %, five times more than restarted reference snapshots (at most 0.9 %). Gaussian states with the sample moments of the data, without interpolation in η, overshoot the reference energy too, by 0.08 to 0.13 at three training values (H31, **confirmed**; Section A).

Is that learned from the data? (H37.) H30 compares the recoloured model with Gaussian draws; it does not exclude that a flow-matching model of this family, sent through the map, passes less far above the reference whatever it was trained on. The control asked by the `jury-externe` role trains the same model, with the same seed and budget, on Gaussian surrogate data (the independent draws of H33) and recolours it the same way (W17). H37 is **confirmed** on both realisations: W17 passes above the reference energy by 0.101 (0.121), W16 by 0.047 (0.063), a paired difference of 0.054 ± 0.002 (0.059 ± 0.002), positive in every seed and at each held-out η; W17 is at the level of G1 (0.092; 0.113) or slightly above. The quantities reported with it tell the same story: the restart drift of W17 is 0.061 against 0.021 for W16 and 0.065 for G1, and its magnetic balance residual 0.071 against 0.034 and 0.072. What the recoloured model has that Gaussian draws have not at restart (half the excursion, a third of the drift, half the residual of the magnetic balance) therefore comes from what the data hold beyond the mean and covariance of G1, not from the model or the map. Of the 0.4 to 0.5 floor that W16 gains over G1 in energy distance, 0.3 is lost by the model trained on surrogates. The null count is nearly untouched: training on the data lowers its error by 0.04 to 0.05 out of about 1.0. Reference snapshots sent through the same kind of map keep about 0.8 of their third-order production at the two lower held-out η and get a null-count error of about 0.7, where they had

0.04: a linear map onto the moments of G1 is not neutral for real states at this sample size (the map is estimated on the snapshots it moves, fewer than dimensions; Section A).

Topology and the solver (H32, H35, H39). Three hypotheses on the number of null points are detailed in Section A; their results are these. At equal spectrum, and so at equal rms gradient, the Gaussian law G1 has 1.4 to 1.6 times the mean $|\det \nabla B|$ of the reference over all the cells (1.4 to 1.5 at the two lower held-out η) and about twice its null points (H32 was dropped: its measurement of the factors where the field is weak is invalid as registered at the largest held-out η). From states with the second moments of the data, 4 time units of the solver bring the null count within 0.05 to 0.13 of that of the reference, while the energy distance and the magnetic energy first degrade, with a maximum after one time unit (Figure 5.10). Followed by that relaxation, the recoloured model W16 is inside the five thresholds of the pilot that a relaxation could help or break, the null count included, and its residual relaxation time (the first duration at which the null-count error is at most 0.2 and the magnetic energy within 5 %) is 4 time units against 8 for G1 and for W9 (H35, **confirmed** on the second realisation of the data, the only one that counts as a test after a repair of the reference count). This is a statement about the solver acting on states with the right second moments more than about the learned model: G1 relaxed by 4 time units is inside the same thresholds on one of the two realisations. The solver alone, from the initial condition of the data, meets the two conditions at 64 time units and not at 32 (H39, **confirmed**): under the same conditions and on these grids of durations, the recoloured model followed by the solver qualifies at least 8 times sooner, the cost of the training data and of the training left out. A generator followed by a few time units of the solver is the only construction of this report that is within the thresholds on the null count; it is not shown that the relaxed ensemble samples the stationary law.

Two side tests on the second-order error of the learned model (H34, H36). Trained on Gaussian trajectories that have the time correlation of the data, the model has no third-order production either (at most 0.003 of the dissipation) and not the deficit of magnetic energy of the model trained on the data (+1, +8 and +1 % against the reference; +3, +10 and −5 % on the second realisation): the excess of 8 to 10 % at one held-out η refutes the second clause of H34 by its letter. The deficit of magnetic energy of the model trained on the data is not halved when the same number of snapshots comes from four times as many trajectories (0.16, 0.10 and 0.14 against 0.15, 0.12 and 0.11; H36, confirmed on a second cycle). Both are in Section A.

Table 5.3: The model with the affine term recoloured at sampling time onto the moments fitted on the data (measured; `experiments/results/replications/e4_recolour`, five seeds, 512 restarted states; notation as in Table 6.1). W9: the model as trained, with 50 and with 200 integration steps; W15: its samples recoloured mode by mode onto the moments of G0; W16: recoloured onto the mean and full covariance of G1; G4: Gaussian draws with the moments of W9, from an independent estimate, sent through the map of W16 (null control). Energy distance over the noise floor on the validation and the held-out η; third-order part of the magnetic production over that of the reference at each held-out η; then, on the held-out η: deviation of largest absolute value of the mean-flow part and of the magnetic energy from the reference, residuals of the magnetic and of the total energy balance, spectrum error, null-count error, and the restart anchored on the reference energy (initial deviation $a_B(0)$, excursion $\max(0, \max_\tau a_B)$, overshoot, drift at $\tau = 4$).

Law	ED / floor, validation η	ED / floor	third-order part / reference, $\eta = 0.0725$	third-order part / reference, $\eta = 0.0490$	third-order part / reference, $\eta = 0.0331$	mean-flow part − reference	$E_B/E_{B,\mathrm{ref}} - 1$	magnetic balance	energy residual	spectrum of B	null points	$a_B(0)$	excursion of E_B	overshoot of E_B	drift E_B, $\tau = 4$	Seeds
G1 Gaussian, full covariance (no network)	2.50(11)	2.55(4)	-0.007(14)	0.005(8)	-0.007(3)	-0.0194	-0.0249	0.0720(5)	0.0034(6)	0.0265(19)	1.033(11)	-0.012(2)	0.0920(17)	0.071(3)	0.065(3)	5
G4 Gaussian draws with the moments of W9, through the map of W16 (null control)	2.63(9)	2.57(10)	0.011(11)	-0.0113(16)	-0.007(4)	-0.0156	-0.0142	0.0704(15)	0.0050(6)	0.0169(17)	1.048(9)	-0.009(3)	0.095(4)	0.082(5)	0.0625(17)	5
W9 mode-wise affine term (base of the recolouring, same job)	9.1(5)	9.3(2)	-0.073(14)	0.600(7)	0.432(6)	-0.153	-0.155	0.0547(9)	0.0566(15)	0.146(3)	1.599(18)	-0.127(3)	8.7×10^{-5}	0.00510(16)	0.125(3)	5
W9, 200 Heun steps instead of 50 (same noise)	9.1(5)	9.3(2)	-0.073(14)	0.601(7)	0.432(6)	-0.153	-0.155	0.0547(9)	0.0565(16)	0.146(3)	1.599(18)	-0.127(3)	7.3×10^{-5}	0.00512(15)	0.125(4)	5
W15 W9 recoloured mode by mode onto the moments of the data	2.07(9)	2.08(6)	-0.108(14)	0.676(7)	0.499(6)	-0.119	-0.0212	0.0478(7)	0.0022(3)	0.0202(18)	1.411(17)	-0.0129(18)	0.040(3)	0.025(4)	0.0398(8)	5
W16 W9 recoloured onto the full covariance of the data	2.02(8)	2.07(5)	-0.021(15)	0.737(9)	0.582(6)	-0.0171	-0.013	0.0339(6)	0.0059(6)	0.016(2)	1.001(18)	-0.0035(19)	0.046(2)	0.035(3)	0.0213(8)	5

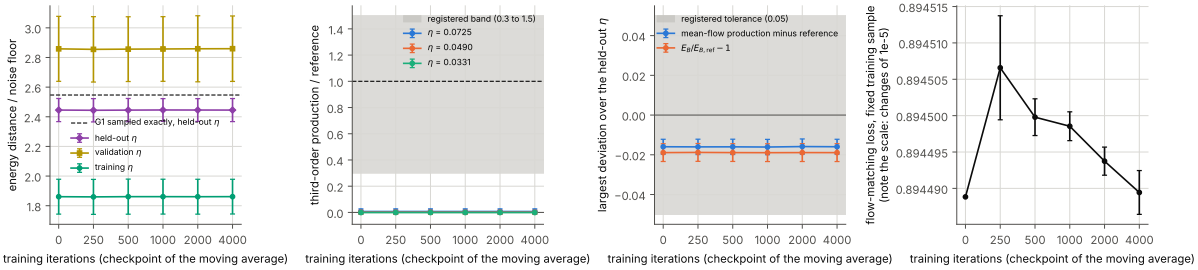

Figure 5.7: Flow matching anchored on the Gaussian law with the full covariance (G1), read along one training of 4000 iterations (H28). Iteration 0 is the untrained model, which is G1 integrated like any model. Data plot from `experiments/results/replications/e4_anchor`, five seeds.

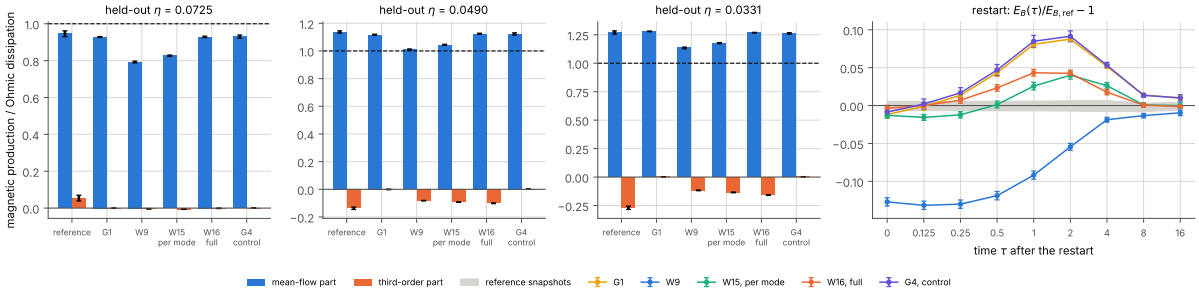

Figure 5.8: The model with the affine term recoloured at sampling time onto the moments fitted on the data (H29, H30): mode by mode (W15) or with the full covariance of G1 (W16); G4: Gaussian draws with the moments of W9 sent through the map of W16 (null control). Left: mean magnetic production over the Ohmic dissipation at the three held-out η, split into its mean-flow and third-order parts (reference: ± 2 jackknife standard errors). Right: magnetic energy after a restart in the solver, relative to the stationary energy of the reference, mean over the held-out η (512 states). Seed means; bars: ± 2 standard errors over the five seeds. Data plot from `experiments/results/replications/e4_recolour`.

5.7 Loss and sampling mechanisms on a model close to the reference

The loss and sampling mechanisms of the first protocol were tried on a model two hundred floors from the reference, where they could only trade one error for another. Applied to the model with the affine term (Table 5.1), they behave differently from each other.

The stationarity loss still buys its target and now costs the law (H23). Fine-tuning W9 with the loss halves the residual of the total energy balance against the control (0.029 against 0.057, in every seed), and lowers the restart drift of the magnetic energy at $\tau = 4$ (0.080 against 0.123), which is not an improvement: anchored on the reference energy, the fine-tuned model starts 17 % below it and is still 11 % below at $\tau = 4$, where the control goes from -12 % to -2 % (the smaller drift is a slower recovery from a larger deficit). The energy-distance ratio rises from 9.3 to 16.4, the residual of the magnetic energy balance doubles (0.113), the residual of the magnetic-helicity identity, the second target of the loss, gets worse (0.024 against 0.014), the error on the null count rises (1.98 against 1.68) and the drift of the kinetic energy nearly triples. The registered criterion allowed a degradation of 10 % of the distance: H23 is **refuted**

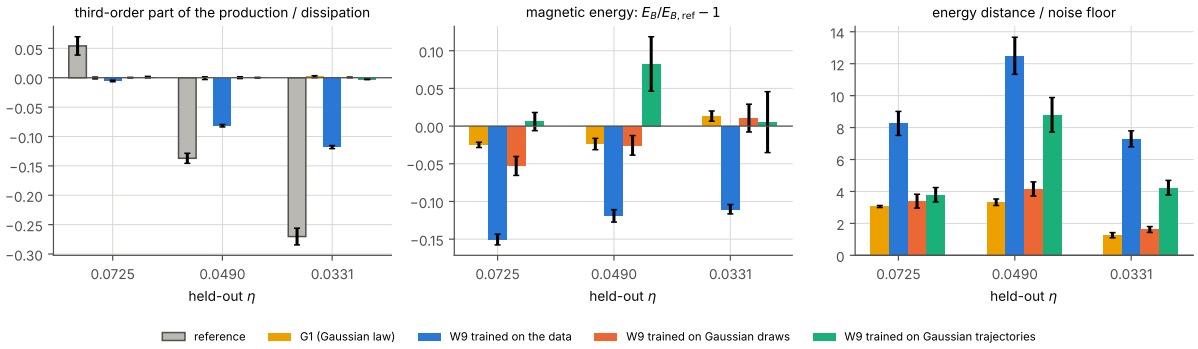

Figure 5.9: The model with the affine term (W9) trained on the data and on independent Gaussian draws with the mean and covariance of the data (H33), with the Gaussian law G1: third-order part of the magnetic production over the Ohmic dissipation (reference: ± 2 jackknife standard errors), magnetic energy against that of the reference, and energy distance to the reference over the noise floor, at the three held-out η; seed means, bars ± 2 standard errors over the five seeds. Data plot from `experiments/results/replications/e4_surrogate`.

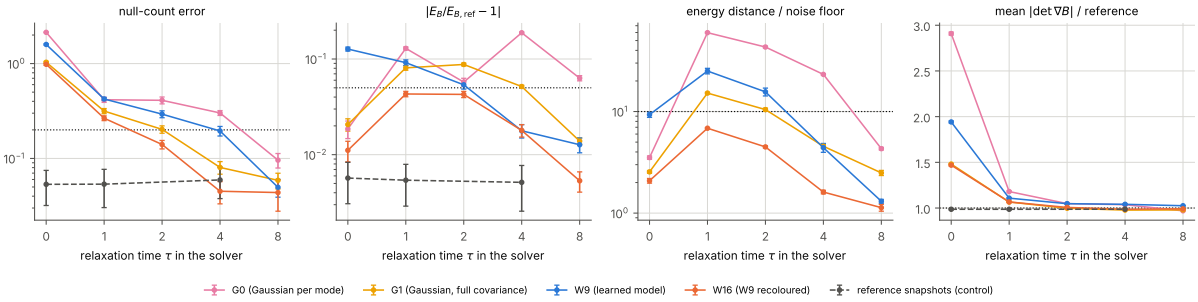

Figure 5.10: Laws followed by a relaxation of duration τ in the solver of the data (H35): the Gaussian laws G0 and G1, the flow-matching model W9, its samples recoloured onto the moments of G1 (W16), and reference snapshots through the same code (control). Null-count error, deviation of the magnetic energy from the reference, energy distance to the reference over the noise floor, and mean of $|\det \nabla B|$ over all the cells over that of the reference; means over the held-out η and the seeds, bars ± 2 standard errors over the five seeds; dotted lines: the thresholds of the objectives (0.2, 0.05, 10) and the reference value. The null-count error is against the count of the reference on 400 snapshots of this job, 7 % above the count on all the snapshots at the smallest η. Data plot from `experiments/results/replications/e4_relax2` (data of the main run, second set of jobs).

on that clause. A penalty on two moments of the generated law, at the weight chosen for the first protocol, moves the law away from the data.

The tilting works when the law is close (H24). On the first-generation model the weights collapsed onto one sample. On the samples of the model with the affine term the effective sample size is 17 % of the sample on average (it is below 10 % in 5 of the 15 cases, with a minimum of 5 %), and H24 is **confirmed** as registered: the residual of the energy balance of the tilted ensemble is 0.011 against 0.056 for the untilted half, in every seed. The tilted ensemble is also better on the energy-distance ratio (4.6 against 7.3, and 3.3 once the bias of a weighted sample is estimated and removed), on the magnetic spectrum (0.08 against 0.14), the helicity (0.09 against 0.19), the null count (1.25 against 1.67) and the magnetic balance (0.038 against 0.056). It is worse on five quantities: the residual of the magnetic-helicity identity, one of its two targets (0.027 against 0.016), the kinetic spectrum (0.04 against 0.02), the kinetic energy, the fluctuation of the energy budget, and the polarity asymmetry, which a weighted sample of some forty effective members cannot keep at the sampling floor.

So the place of a constraint and the quality of the model interact: the same identities that could not help a wrong model improve most statistics of a nearly right one when they are imposed at sampling time, by reweighting, and hurt it when they are imposed through the loss at this weight. In both places the energy identity is the one that responds; the magnetic-helicity identity does not.

5.8 Topology: null points and the small scales

The number of magnetic null points is the most demanding of the metrics: no generator has it right. The penalty model and the diffusion model have too few (0.3 and 0.6 times the reference count); all the others have too many. A conjecture of the `physicien` role (H13) was that the count scales as the cube of the magnetic Taylor wavenumber k_λ, as it would for a Gaussian random field, so that the error on the count would be three times the relative error on k_λ. The fit over all generators, held-out values and seeds gives a slope close to five and not three (Figure 5.11): H13 was **refuted** on the slope. The correlation is strong, but it does not explain the excess: the fitted line does not pass through the origin, and a generator with exactly the reference k_λ would still have about 2.8 times the reference number of null points. The excess of nulls is a property of the generated fields beyond their small-scale energy.

What the second moments leave out of the topology, and what a short relaxation in the solver repairs, is in Section 5.6: the laws with the covariance of the data still have twice the null points of the reference, and 4 time units of the solver bring the count within 0.05 to 0.13 of it.

5.9 Polarity

The offer places the project in the context of abrupt reversals of the polarity of the magnetic field, as observed in the von Kármán sodium experiment (Berhanu et al. 2007). Two things can be said of polarity here, and neither is a model of reversals.

The first is a symmetry. The equations are invariant under $B \to -B$, and so must be the generated law. The polarity test of the comparison measures the asymmetry of the sign of the cross helicity over the generated samples, $|P(\langle u \cdot B \rangle > 0) - \frac{1}{2}|$. It is at the sampling noise for every model, with or without the equivariance built into the velocity field (H9, refuted: the equivariance did not lower it), and 0.011 for the recoloured model, for a target of 0.03 (**measured**).

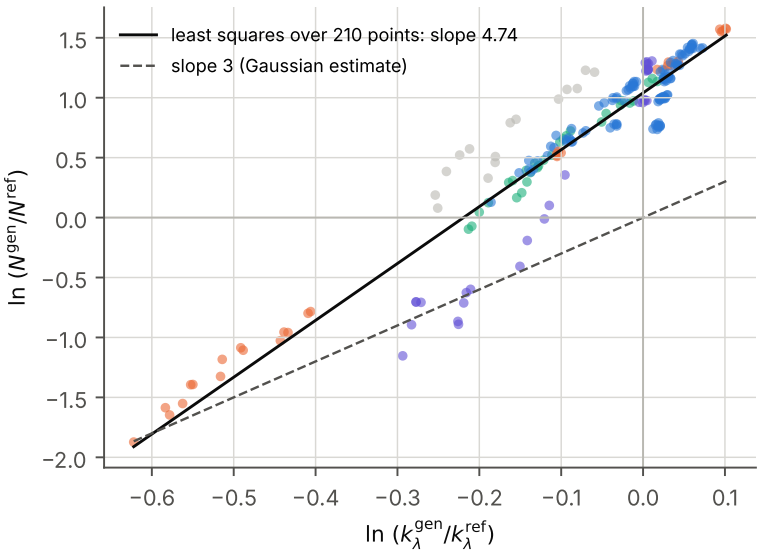

Figure 5.11: Number of magnetic null points against the magnetic Taylor wavenumber, generated over reference, one point per variant, held-out η and seed; the fit is the one of the digest (hypothesis H13). Data plot from `experiments/results/e4_mhd3d_s<k>.json`.

The second is a description of the reference data, with no hypothesis and no threshold (`experiments/e4_polarity_diag.py`, on the data of the session container; digest, section "diagnostic: polarity"). The magnetic field of largest scale lives on the three wavevectors of modulus 1, which hold 0.20 to 0.24 of the magnetic energy (the shell $k = 1$ of Table 4.1, which holds 0.37 to 0.44 of it, also counts the six wavevectors of modulus $\sqrt{2}$). The law of each of these modes is symmetric under a change of sign: by the invariance of the equations under $B \to -B$, and also because a shift of π along an axis leaves the force at $k_f = 2$ unchanged and changes the sign of the modes of wavenumber 1 along that axis. Along the 24 trajectories of each η, the autocorrelation of their complex amplitude falls under $1/e$ after 17 time units at the largest η and after 66 to 80 at the three smallest (not monotonic in η; no uncertainty was computed, and each trajectory holds about three correlation times there; the Ohmic diffusion time at that scale, $1/\eta$, goes from 12 to 33), and at the end of the window of 198 time units 0.40 to 0.60 of the (trajectory, mode) pairs have a projection of opposite sign to their initial one. The sign of the large-scale field of this system is thus lost over a few tens of time units. Whether this happens by abrupt reversals between long-lived states or by a slow drift of the phase was not examined, and each trajectory holds only 2.5 to 12 of these correlation times.

The dynamics of the polarity are a property of trajectories. A generator of independent snapshots of the stationary law, which is what every model of this report is, has the symmetry and, of the dynamics, only what the law and the equations fix together: the mean rate of sign changes of a large-scale amplitude (Rice formula, Proposition 8 of the theory brief), which was not computed here; it gives neither the waiting times nor the course of a reversal. What a study of reversals would need is in Section 9.4.

5.10 Complements: longer trainings, replication, objectives

Three sets of results have no bearing on the requirements of the offer beyond what is said here, and are in the appendices.

Longer trainings and wider networks (Section B). With eight times the iterations, the flow-matching loss of the models with the affine term goes down and the generated law moves away from the reference (18.8 and 19.5 floors on the held-out η against 9.3 and 8.6). A network without the affine term is closer to the reference when it is wider, up to a point, and the widest one, with the lowest loss, is farther again (H21, refuted).

Replication (Section C). The whole first protocol was run once more on reference data regenerated with other seeds of the initial conditions: every conclusion of this section is unchanged, and so is every verdict recomputed on it. The later hypotheses were run on both realisations unless said otherwise.

Objectives of the race (Section D). The pilot set nine quantified objectives on the first day, measured on one generator selected on the validation values among the learned models. Five are met (distance to the reference law, divergence, number of decided comparisons, toolbox, polarity symmetry); the energy balance misses its target narrowly; spectra and helicity, topology and the restart drift miss theirs by factors of two to eight. The recoloured model, built later, is not the selected generator.

6 Comparative study

Table 6.1 is the comparison the offer asks for: one line per mechanism or model family, one column per metric, on the three held-out values of η, seed mean and standard error. The status of each registered hypothesis after review is in Table E.1, in the appendix (Section E). Both are generated from the result files and the register; the sortable version of the first is in the explorer.

Table 6.1: Comparative table of E4 (measured; `experiments/results/e4_mhd3d_s<k>.json`). ED / floor: energy distance to the reference over the noise floor, at the training and at the held-out η; div B: largest relative divergence of B; energy residual: absolute residual of the mean energy balance over the mean dissipation; magnetic balance: absolute residual of the mean magnetic energy balance (0: production equals Ohmic dissipation; 1: no production); spectrum of B: relative L^2 error; magnetic helicity and null points: relative errors of the means; restart drift: relative drift of the magnetic and of the kinetic energy after $\tau = 4$; polarity: asymmetry of the sign of the cross helicity. Lower is better in every column except the two restart drifts, which do not rank the laws (Section 5.5). Notation: 217(14) is the seed mean 217 with a standard error of 14 on its last digits; values below 10^{-3} are given without error. The tilted ensemble W7 is a weighted sample whose effective size is about one sample: its line describes that sample. No restart test for W7 and for the rescaled ensemble W8.

Variant	Acts through	ED / floor, training η	ED / floor	div B	energy residual	magnetic balance	spectrum of B	magnetic helicity	null points	restart drift E_B	restart drift E_u	polarity	Seeds
W0 free components	none	165.4(5)	236(6)	0.454(2)	0.314(11)	0.159(15)	2.37(13)	1.99(19)	1.11(5)	0.513(9)	0.413(9)	0.015(3)	5
W1 divergence penalty	loss	275(4)	464(12)	0.261(6)	0.613(18)	0.410(5)	23.0(17)	38(3)	0.698(12)	0.5008(14)	0.709(4)	0.020(4)	5
W2 final projection	sampling	106.9(6)	179(8)	3.1×10^{-8}	0.314(11)	0.159(15)	2.04(13)	1.99(19)	0.91(5)	0.443(14)	0.354(11)	0.015(4)	5
W3 projected flow	sampling	105.8(6)	177(7)	2.6×10^{-8}	0.312(16)	0.161(12)	2.00(18)	1.9(2)	0.90(7)	0.448(17)	0.355(16)	0.015(2)	5
W4 vector potentials	design	175.4(5)	256(7)	1.4×10^{-8}	0.408(12)	0.149(14)	2.45(19)	2.1(3)	1.14(7)	0.517(10)	0.425(12)	0.013(3)	5
W4s potentials, B $\to$ -B equivariant	design	166.3(3)	209.8(19)	1.4×10^{-8}	0.284(10)	0.264(14)	0.97(8)	0.45(9)	1.83(7)	0.382(12)	0.281(16)	0.013(3)	5
W5 stationarity loss	loss	165.1(5)	215.0(19)	1.4×10^{-8}	0.191(5)	0.516(13)	0.473(15)	0.693(9)	2.36(2)	0.334(3)	0.243(3)	0.013(2)	5
W5 control (same fine-tuning, no identity)	loss	166.4(3)	210(2)	1.4×10^{-8}	0.290(10)	0.251(15)	1.03(9)	0.50(11)	1.79(7)	0.390(12)	0.289(16)	0.011(2)	5
W6 diffusion (VP, v-prediction)	family	160.4(6)	519(17)	1.3×10^{-8}	0.747(14)	0.213(11)	15.8(16)	22(2)	0.43(2)	0.637(2)	0.616(14)	0.022(6)	5
W7 exponential tilting	sampling	161(9)	195(11)	1.4×10^{-8}	0.204(8)	0.43(4)	0.51(6)	0.41(11)	2.24(19)	–	–	0.47(3)	5
W8 per-sample balance rescaling	sampling	92.9(3)	117(2)	1.4×10^{-8}	$< 10^{-12}$	0.195(19)	0.86(8)	0.52(11)	0.71(8)	–	–	0.0134(18)	5
W4c potentials, smooth condition	design	175.1(8)	215.4(3)	1.4×10^{-8}	0.2694(6)	0.3843(16)	0.637(6)	0.330(5)	2.142(12)	0.3826(11)	0.252(4)	0.0139(19)	5
G0 Gaussian per mode (no network)	family	2.67(8)	3.52(7)	1.5×10^{-8}	0.0015(3)	0.9981(9)	0.024(3)	0.035(5)	2.235(7)	0.212(3)	0.026(5)	0.018(3)	5
W9 mode-wise affine term	design	5.51(9)	9.3(3)	1.5×10^{-8}	0.0567(17)	0.0553(6)	0.145(3)	0.196(4)	1.665(14)	0.124(3)	0.053(6)	0.019(4)	5
W10 anchored on the Gaussian field	design	36.37(10)	43.6(2)	1.5×10^{-8}	0.1458(9)	0.3092(17)	0.2081(19)	0.096(7)	2.464(16)	0.122(2)	0.092(4)	0.0135(9)	5

W11 anchor + affine term	design	5.52(13)	8.6(2)	1.5×10^{-8}	0.0515(9)	0.0598(6)	0.143(4)	0.193(5)	1.696(12)	0.120(2)	0.053(3)	0.0169(10)	5

Reading the two tables (Table 6.1 and its complement Table 5.1) by the place of the constraint:

- **By design.** Exact, and free at sampling time: the divergence is at round-off for every model on potentials. The design choice that changes the law is not the potential representation (W4 against W0) or the equivariance (W4s against W4, a gain of 46 floors out of 256) but the explicit linear part (W9, W11): a factor of twenty against the matched network without it.
- **By the loss.** The divergence penalty does not enforce its constraint and costs accuracy. The law-level stationarity loss lowers the residual of the energy identity, at parameter values without data, on a wrong model (0.29 to 0.19) and on a nearly right one (0.057 to 0.029); it does not lower the residual of the magnetic-helicity identity imposed with it; on the wrong model it leaves the distance to the reference unchanged, and on the nearly right one it increases it by three quarters.
- **By the sampling procedure.** Projections give exact zero divergence; projecting every stage brings nothing over projecting the final sample. The tilting collapses on a wrong model and improves a nearly right one on most metrics. The per-sample rescaling satisfies the balance exactly, by removing the fluctuations of the energy budget that the true law has. Two sampling-time constructions tested after these tables act on the law rather than on each sample (Section 5.6): a linear recolouring of the samples onto the mean and covariance fitted on the data, which gives a learned model the second moments of the Gaussian law and keeps its third-order production; and a relaxation of a few time units in the solver, the only constraint of this study that brings the null count close to that of the reference, at the price of a transient in which the other metrics first degrade.
- **Model families.** At equal cost, flow matching interpolates in η better than the diffusion model, which is the better one at the training values. The two Gaussian laws are the baselines to beat. In energy distance they are not beaten by any model as trained: G1 is the closest law of these tables and G0 the second (the recoloured model of Section 5.6, which takes the moments of G1, is at 2.1 floors). Among the laws close to the reference, G1 also has the lowest null-count error, while the learned models with the affine term (W9, W11) have a lower residual of the magnetic energy balance on the mean over the held-out η, through two errors that cancel at two values and add at the third, and are, among the laws close to the reference, those with a third-order part of the magnetic production (the anchored model W10 has one of 0.3 and 0.2 times that of the reference at the two lower η; the other variants were not measured; Section 5.5). The restart drifts of these tables are relative to each law's own initial energy and do not rank the laws: anchored on the reference energy, the six laws tested that start on it or below it gain magnetic energy after the restart, the Gaussian laws fitted to the data overshooting the reference by 9 and 17 % and the learned models with the affine term coming back to it from a 12 % deficit.

7 Toolbox

The offer asks for "reproducible code integrated into the project toolbox". The toolbox of this repository is the Python package `src/milady/`; what this work added to it is drawn in the code-backed diagram code-architecture.html (Figure 7.1) and listed in Table 7.1.

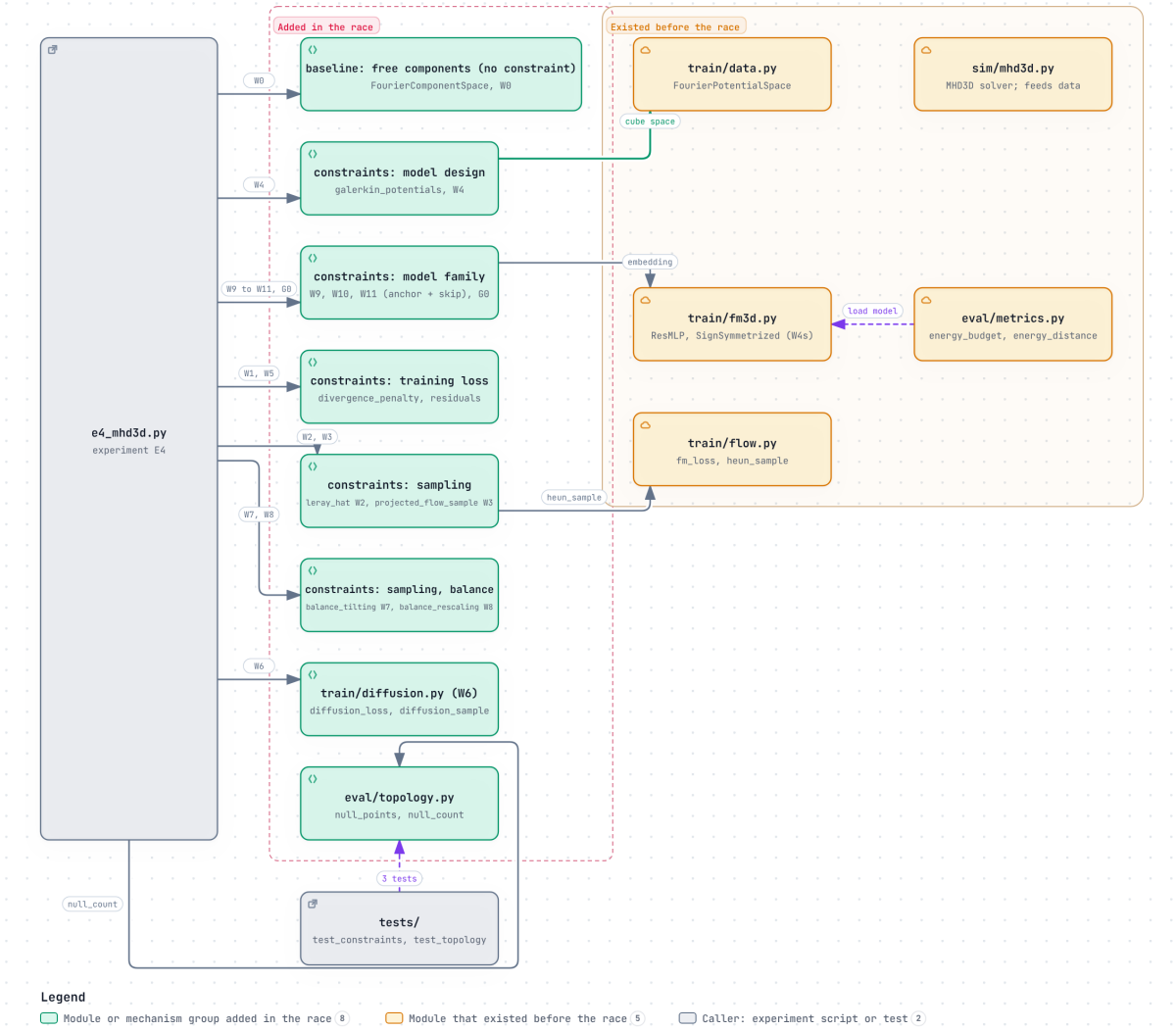

Figure 7.1: Code architecture of what E4 added to `src/milady/`: the constraint mechanisms by group, the diffusion module, the topology module, and the modules they build on. Static export of the Archify diagram `diagrams/code-architecture.html`.

Table 7.1: What entered the toolbox, with its tests.

Module	What it provides	Tests
`milady.train .constraints`	`REGISTRY` of 16 entries (eleven constraint mechanisms, the unconstrained baseline, a diffusion model, three Gaussian laws): `FourierComponentSpace` (free components), `leray_hat`, `projected_flow_sample`, `divergence_penalty`, `budget_terms_hat` and `stationarity_residuals` (identities), `half_batch_product` (unbiased squared mean), `balance_tilting`, `balance_rescaling`, `ModewiseAffineSkip`, `ModewiseGaussian`, `FullGaussian`, `LowRankModewiseGaussian`, `GaussianAnchoredField`, `FullGaussianAnchoredField`, `moment_recolouring`	`tests/ test_constraints.py`: at least one test per mechanism, and the save and reload of the Gaussian laws (19 tests)
`milady.train .data`	`FourierPotentialSpace` with `truncation="cube"`: potentials on exactly the modes of the solver	`test_galerkin_potentials` in the same file
`milady.train .diffusion`	`diffusion_loss` (v-prediction, variance-preserving path), `diffusion_sample` (ancestral or deterministic)	`test_diffusion` in the same file
`milady.train .fm3d`	`ResMLP` and `ConditionEmbedding` with a `cond_scale` argument (largest frequency of the embedding of the condition)	existing tests
`milady.eval .topology`	`null_points`, `null_count`: null points of a periodic 3D field by the trilinear method	`tests/ test_topology.py` (4 tests, with the Poincaré–Hopf balance of signs and the Kac-Rice factors)
`milady.data .convert`	`convert_e4`, `convert_e4_file`: the E4 snapshots to the data format of the project (MFF 1.0, readable as a data set of The Well (Ohana et al. 2024)); `e4_latents`: the potentials back from the stored fields	`test_e4_converter_synthetic` in `tests/ test_format.py`
`tools/ runpod_job.py`	`run`, `collect`, `exec`, `cleanup`, `spend`: GPU jobs with a time limit, a ledger and per-team pod names	`tests/ test_pilot_tools.py`

How to call it, for a latent space `space` and a trained field `model`:

```
from milady.train import constraints as C
from milady.train.data import FourierPotentialSpace
from milady.eval.topology import null_count
```

```
# latent space: 665 modes, 8 channels (vector potentials of u and B)
space = FourierPotentialSpace(n=16, kmax=5, truncation="cube")
fh = space.field_hat(space.denormalize(x))      # (batch, 2 fields, modes, 3)
# stationarity identities: energy, magnetic helicity, cross helicity
terms = C.budget_terms_hat(space, fh, nu=0.05, eta=0.05, forcing_hat=f_hat)
res = C.stationarity_residuals(terms, nu=0.05, eta=0.05)
# unbiased estimate of (E<j.B>)^2 over 4 conditions
loss = C.half_batch_product(*res["magnetic_helicity"], n_groups=4)
# Gaussian baselines and the explicit affine term of W9
gauss = C.ModewiseGaussian(sign=space.sign).fit(x_train, c_train)   # G0
full = C.FullGaussian(sign=space.sign).fit(x_train, c_train)        # G1
field = C.ModewiseAffineSkip(model, space.n_modes, space.n_channels, n_cond=1)
# null points of fields of shape (n, nx, ny, nz, 3)
counts = null_count(b_grid)
```

The tests check each mechanism against an independent computation: the identities against the time derivative of the solver, the Gaussian anchor against the closed form of Equation 3.4, the tilting against its moment conditions, the null finder against fields with a known number of nulls. `make test` runs them (the test suite, the linter and the check of the generated agent files pass on the branch).

The data, the cited laws and their catalogue are published as a release of the repository, described in Section 8.

The experiment itself is driven by scripts in `experiments/`: `e4_data.py` (data), `e4_mhd3d.py` (variants, by group), `e4_floor.py`, `e4_digest.py` and `digest.py` (digest), `race_metrics.py` (objectives), `e4_assemble_scan.py` (the width scan from its three jobs), `e4_plots.py` (the data plots of this report), `e4_report_tables.py` (its tables), `export_explorer.py` (the data of the explorer), `e4_polarity_diag.py` (the diagnostic of Section 5.9), and `e4_release_build.sh`, `e4_release_models.py`, `e4_release.py` and `e4_release_example.py` (the release of Section 8 and its reload example).

8 Data and models

The offer asks for fields that other parts of the project can reuse. The reference data of this study, the two Gaussian laws that its conclusions cite, a CPU retraining of the two learned ones and a catalogue are published as the release `e4-data-v1.0` of the repository. Table 8.1 lists the files, with a link to each; `SHA256SUMS` holds their checksums and `manifest.json` their sizes and identifiers.

Table 8.1: Files of the release `e4-data-v1.0` (data, models, catalogue, checksums).

File	Content	Format	Size
`e4_eta<η>.h5` for $\eta = 0.08000$, 0.07253, 0.06575, 0.05961, 0.05404, 0.04899, 0.04441, 0.04026, 0.03650, 0.03309, 0.03000	Reference data, one file per η (24 trajectories × 100 snapshots): fields u and B on the 16^3 grid, time series of the solver, provenance; splits train, val and test in the metadata	MFF 1.0: HDF5 with the layout of The Well	11 files, 2.08 GiB in all
`e4_sample_eta0.04899.h5`	Sample of the data: $\eta = 0.04899$, 2 trajectories × 25 snapshots	MFF 1.0	4.1 MiB
`e4_latent_space.pt`	Latent space of the models and its normalisation	PyTorch dictionary of tensors	0.1 MiB
`e4_gaussian_G0.pt`	Parameters of the Gaussian law G0 (mean and covariance per Fourier mode at each training η)	PyTorch dictionary of tensors	2.2 MiB
`e4_gaussian_G1.pt`	Parameters of the Gaussian law G1 (mean and factor of the full covariance at each training η)	PyTorch dictionary of tensors	584.6 MiB
`e4_model_W9_skip_sym_s<k>.pt` for k = 0, 1, 2, 3, 4	Weights of the flow-matching model W9, one file per seed	PyTorch dictionary of tensors	5 files, 339.2 MiB in all
`e4_W16_map_s<k>.pt` for k = 0, 1, 2, 3, 4	Linear map of W16 at the three held-out η, one file per seed of W9	PyTorch dictionary of tensors	5 files, 1.58 GiB in all
`catalog.sqlite`, `catalog.json`	Catalogue: data set `e4_mhd3d` 1.0 with its splits, 20 released models with file, hyperparameters, commit, data set and evaluations, and the cited runs as records without files	SQLite (`milady.catalog`), JSON export	–

File	Content	Format	Size
`e4_release_eval_s<k>.json` for k = 0, 1, 2, 3, 4, `e4_release_w16.json`	Evaluations of the released models as written by the pipeline: G0, G1 and W9 per seed, and W16	JSON	–
`SHA256SUMS`, `manifest.json`, `README.md`	Checksums, list of the files with sizes and identifiers, notes	text, JSON	–

What the data are. One file per value of η in the data format of the project (MFF 1.0, `docs/specs/data-format.md`: an HDF5 file with the layout of The Well (Ohana et al. 2024); a test reproduces the metadata reading of its loader, `tests/test_format.py`, and the loader itself was not run), with the fields u and B on the 16^3 grid, the time series of the solver (transient included), the physical parameters, the split of that η (training, validation, held-out) and the provenance. Each file passes the full validation of the format. The fields are band-limited, so the vector potentials on which the models are trained are recovered from them by one transform (`milady.data.convert.e4_latents`), to float32 rounding. The files were generated by the build of the release, from the committed script and seeds, on the CPU of a hosted runner, and they are the realisation of the data of the main run of this report: the mean magnetic energy of each of the 11 files equals, to the last digit, the value recorded when the data of the main run were generated on the CPU of a rented GPU machine (`experiments/results/e4_data.json`; the manifest of the release records the check). The solver is chaotic, and some processor types give other trajectories from the same seeds; the second realisation of the data, on which most of the later hypotheses were also run, uses other seeds and is not published. A Zarr copy, which the format allows, is not published; `milady.data.to_zarr` writes one from each file.

What the models are. The weights of the runs cited by this report were not kept: each rented GPU was released with its files at the end of its job, and the option that brings the weights back was added only for this release. What is published is therefore the same code, seeds, width and number of iterations, **retrained by the build of the release on the published data, which are those of the main run, on CPU instead of a GPU**, and evaluated by the same pipeline against the reference statistics of those data: the two Gaussian laws G0 and G1 as parameter files (they are closed-form fits and have no weights), the flow-matching model W9 as one weight file per seed, and the recoloured law W16 as the linear map that sends the samples of W9 of each seed onto the moments of G1, at the three held-out η. Table 8.2 puts their metrics next to those of the cited runs. These numbers are **measured** by the build on five seeds, on the same data with another arithmetic (CPU). The Gaussian laws are closed-form fits and reproduce the cited numbers that depend on the data and the draws alone (for G1, the null-count error and the restart drift differ: another draw of the reference count, and 128 restarted fields against 512); the flow-matching model is another training of the same model, and its numbers are close to the cited ones; the null-count errors differ more, as they do between two draws of the reference count (Section 9.3). The released W16 is not the cited law: its map is estimated on four times fewer samples of W9 (the sampler is slow on CPU) and from moments that are not symmetrised under the sign symmetry, as they are in the cited run; which of the two explains the gap was not measured (the release has no null control). It is at 2.9, 5.7 and 5.4 floors against 0.8, 2.2 and 3.3, farther from the reference than the released G1 and G0 at the three held-out η. The ranking of the conclusions, in which W16 is the closest law, is not reproduced by the released files. It was not restarted in the solver. These numbers were reviewed by the `encadrant` role on 14 November 2026: both columns were recomputed from the committed and released files; the cause of the gap of the released W16 was not tested. They are a reproduction of the cited laws, not a result of the study. The catalogue (`milady.catalog`) records each released model with

its file, its hyperparameters, the commit, the data set and its evaluations, and keeps the cited runs as entries without files, commit or data set, with the evaluations of the committed result files.

The models of the release were trained at commit `240b56cc1863` on the CPU of a GitHub Actions runner. The map of W16 is estimated there on 32768 samples of W9 per held-out η (131 072 in the report, on a GPU), and W16 is evaluated on 512 samples outside that estimate, without the restart in the solver.

Table 8.2: Released models against the cited runs: seed means at the three held-out η (0.0331, 0.0490, 0.0725).

Law	Metric	Runs cited by the report (GPU)	Released models (CPU build)
G0	energy distance over the noise floor	1.81, 4.82, 3.94	1.81, 4.82, 3.94
G0	relative error of the magnetic energy	0.00795, 0.0215, 0.0263	0.00795, 0.0215, 0.0263
G0	third-order part of the production over the dissipation	-0.00136, 0.000231, 0.000288	-0.00136, 0.000231, 0.000288
G0	residual of the energy balance	0.000241, 0.0013, -0.0021	0.000241, 0.0013, -0.0021
G0	null-count error	1.73, 2.43, 2.64	1.73, 2.43, 2.64
G0	restart drift of the magnetic energy	0.179, 0.2, 0.258	0.179, 0.2, 0.258
G1	energy distance over the noise floor	1.26, 3.32, 3.06	1.26, 3.32, 3.06
G1	relative error of the magnetic energy	0.0133, 0.0238, 0.0249	0.0133, 0.0238, 0.0249
G1	third-order part of the production over the dissipation	0.00184, -0.000746, -0.000396	0.00184, -0.000746, -0.000396
G1	residual of the energy balance	-0.00281, 0.000677, -0.00426	-0.00281, 0.000677, -0.00426
G1	null-count error	0.635, 1.03, 1.43	0.801, 1.12, 1.43
G1	restart drift of the magnetic energy	0.0329, 0.063, 0.0979	0.0375, 0.061, 0.0947
W9	energy distance over the noise floor	7.54, 12, 8.5	7.49, 12, 8.23
W9	relative error of the magnetic energy	0.113, 0.114, 0.155	0.113, 0.113, 0.152
W9	third-order part of the production over the dissipation	-0.117, -0.0822, -0.00394	-0.116, -0.0817, -0.00539
W9	residual of the energy balance	0.0564, 0.0518, 0.0618	0.0564, 0.0511, 0.0618
W9	null-count error	1.03, 1.58, 2.19	1.24, 1.69, 2.16
W9	restart drift of the magnetic energy	0.114, 0.112, 0.148	0.114, 0.103, 0.147
W16	energy distance over the noise floor	0.809, 2.15, 3.26	2.94, 5.66, 5.37
W16	relative error of the magnetic energy	0.0125, 0.00794, 0.013	0.0412, 0.0224, 0.0197
W16	third-order part of the production over the dissipation	-0.157, -0.101, -0.00114	-0.156, -0.0988, -0.00325
W16	residual of the energy balance	-0.00479, -0.00507, -0.0079	-0.0348, -0.0295, -0.0272
W16	null-count error	0.56, 1.02, 1.42	0.81, 1.17, 1.51
W16	restart drift of the magnetic energy	0.0134, 0.0164, 0.0341	not evaluated

Reload a model and draw samples. From a checkout of the repository with its pinned dependencies (`infra/env/requirements.txt`), the example script downloads what a law needs,

checks the checksums, draws fields at one η and can write them in the data format of the project:

```
python3 experiments/e4_release_example.py --dir e4_release --download \
    --law W16 --seed 0 --eta 0.04899 --n 256 --out samples.h5
```

The map of W16 is stored at the three held-out η only. At another η the script recomputes it from 8192 samples of W9 (five minutes on four cores, plus the file of G1), which gives a magnetic energy about 20 % too high: pass `--n-recolour 131072`, or use G1 or W9. Both blocks run from the root of the repository, with its pinned dependencies; they create the folder `e4_release/` and the file `samples.h5` there (git-ignored). The command above downloads 392 MiB and takes about half a minute; the Python lines below download 653 MiB (the file of G1) and take about a minute. The script compares the energies of the samples with those of the data only if the data file of that η is in the folder.

The same in a few lines of Python, for the flow-matching model and for a Gaussian law:

```
import sys; sys.path[:0] = ["src", "experiments"]
import e4_release_example as R
for name in (R.SPACE, R.GAUSS.format(law="G1"), R.WEIGHTS.format(seed=0)):
    R.fetch("e4_release", name)                # download, check the SHA-256
space = R.load_space("e4_release")             # latent space, normalisation
w9 = R.load_w9("e4_release", space, seed=0)    # flow-matching model
z = R.sample_w9(w9, space, eta=0.04899, n=64)  # latents (64, 665, 8)
fields = R.to_fields(space, z)                 # u, B: (64, 16, 16, 16, 3)
g1 = R.load_gaussian("e4_release", "G1")       # Gaussian law, full covariance
import torch
z1 = g1.sample(R.cn(0.04899), 64, generator=torch.Generator().manual_seed(0))
```

What the review of the release measured on such samples (64 per law at $\eta = 0.0490$; `research/teams/claude/reviews/2026-11-14-jury-externe-release.md`): the stated shapes, a divergence at float32 rounding, and a mean magnetic energy within the errors that the catalogue records. The samples match the data on the mean energy and not on its spread: the standard deviation of the magnetic energy per sample is 0.021 to 0.038 for the released laws against 0.014 for the data.

A data file is read with `milady.data.read_mff(path)` (fields, coordinates, parameters, diagnostics) and checked with `python -m milady.data.convert validate --full`; the training script is `experiments/e4_mhd3d.py`. The release is rebuilt from the code by the workflow `.github/workflows/e4-release.yml` (`experiments/e4_release_build.sh`), and `research/teams/claude/release/` holds a copy of its notes, manifest and checksums.

9 Discussion

9.1 What each place of constraint buys

The three places named by the offer do not play the same role, and the measurements suggest a division of labour.

Design is the place for what is exactly known. A linear constraint that holds for every sample (zero divergence), an exact symmetry of the law ($B \to -B$), and the second-order structure of a homogeneous field (the mode-wise affine part of the velocity field) can all be built into the model. They cost nothing at sampling time and they cannot be violated. The first two did not change the quality of the law in this experiment; the third decided it.

The loss is the place for what is known about the law and not about each sample. The stationarity identities are of this kind. Imposed at the level of the law, with an unbiased estimator, the energy identity moved the residual it targets, at parameter values without data; the magnetic-helicity identity did not, and the term changed statistics it was not aimed at. The identities are necessary conditions, not a substitute for a model that fits the data: a model a hundred floors from the reference stays there, and a loss on one moment of a wrong law can only trade one error for another.

The sampling procedure is the place for linear per-sample constraints, and a dangerous place for the others. Projection onto a linear constraint is harmless. Enforcing a law-level identity on each sample is a different operation: it produced the best-looking energy balance of the comparison and a law without the fluctuations of the real one.

Four results go beyond the mechanisms.

The first is methodological. Two baselines without a network, cheap to compute, are closer to the reference in energy distance than every learned model of this report as trained (the recoloured model, which is at 2.1 floors, takes its second moments from the second baseline). Without the first, the comparison of constraints would have ranked models that were all wrong for the same reason. Without the second, which exists only because an external reviewer asked for it, the report would have credited the learned models with non-Gaussian physics on the strength of metrics that a Gaussian law improves just as much. A comparative study of generative models of a physical system needs its Gaussian baselines, including the one with the full covariance, before any claim about what a network has learned.

The second concerns evaluation. The energy distance on second-order features of single modes ranks a law without magnetic production (G0) above models that have it, and barely separates the best learned model from the Gaussian law with its first two moments (0.4 floors). The restart drift at one time, relative to the ensemble's own energy, confounds the bias of a law with its dynamics, and ranked laws that the anchored curves show to be comparably far from the reference energy over the course (0.13 and 0.10 for W9 and G1; 0.11 and 0.12 on the second realisation). The quantities that did discriminate are the split of the magnetic production into a mean-flow part and a third-order part (the residual of the balance alone hid two errors of W9 that cancel), the null count, and the restart curve anchored on the reference. The objectives of

this pilot, written on the first day, rewarded the energy distance and a restart drift at one time; they would have been better with the production split and the anchored curve.

The third is about where the non-Gaussian part of the problem lies, and it is now partly **measured**. Three properties of the reference are out of reach of the Gaussian law with its own mean and covariance (G1): the third-order part of its magnetic production, a sink of more than a quarter of the Ohmic dissipation at the smallest η, which changes sign over the range; its number of null points, of which G1 has 2.1 times as many on the mean over η; and its stationarity under the solver, since the states of G1 at the held-out η, where it is interpolated, move 9 to 11 % away from the stationary magnetic energy. The best learned model holds about half of the first at two held-out values. On the second, its non-Gaussian part changes little: its null count differs from that of its own Gaussianisation by 0.04 out of 1.7. On the third, the overshoot registered for the restart does not separate the two, while the drift at $\tau = 4$ does (0.124 against 0.170). It is also behind G1 on the second-order part, by 10 to 16 % on the mean-flow production and 12 % on the magnetic energy. What a better generative model should learn is therefore, in this order, the covariance between modes to the accuracy of G1, and then the non-Gaussian part, for which the production split, the null count against the Gaussianised model and the anchored restart curve give three measurable targets.

The fourth follows from the tests of Section 5.6 and changes how the two parts can be obtained. The second moments do not have to be learned by the network: a linear map applied to the samples of the model gives them the covariance of G1 and keeps, even raises, the third-order production that the model holds (H29), and the two recoloured laws (mode by mode and with the full covariance) are the only laws sampled without the solver that are within the thresholds of the pilot (Table D.1) on the energy distance, the energy balance, the spectrum-plus-helicity error and the restart drift together. The first three are those of the Gaussian law, which meets them too: the energy balance and the spectrum are functions of the mean and covariance that the map copies from the data (5320 means and 1.4×10^7 covariances per training η, interpolated in between). What the learned model brings to the construction is the restart drift (0.021 against 0.065 for G1), the lower passage above the reference energy and the third-order production. This is a constraint at sampling time of a kind the first protocol did not have: linear, exact on the first two moments of the law, and cheap. The non-Gaussian part, on the other hand, was not obtained by giving the network the Gaussian part for free: anchored on the exact Gaussian field of the sample covariance, the network learned nothing in the same budget (H28), and nothing either from samples of the learned model itself (H38: the failure is that of this construction), while the model that learns both parts together holds half of the third-order production, and holds none when its training data are Gaussian (H33, H34). At equal moments, what W9 holds beyond its second moments halves the passage above the reference energy at restart (H30); this is not the third-order production, and it comes from the data: the same model trained on Gaussian surrogates and recoloured the same way passes as far above the reference as the Gaussian law (H37). What the construction does not touch is the topology: the recoloured model has the null count of the Gaussian law, twice that of the reference. At equal spectrum, and so at equal rms gradient, the Gaussian field has 1.4 to 1.6 times the mean $|\det \nabla B|$ of the reference, over all the cells and not only where the field is weak (H32): the gradient tensor of the reference is closer to singular than that of a Gaussian field with its covariance. This is a property beyond the second moments that none of the mechanisms compared here constrains.

9.2 Negative results and refuted claims

The register keeps what failed. Of the eleven hypotheses of the first protocol, written before any model was trained, six were refuted and one was dropped:

- *Potentials are not closer to the reference than free components* (H4 refuted). The constraint by design gives exact zero divergence at no cost, but the law is not better for it; the theory brief already says that with a perfect field the free model, the projections and the potentials give the same law, so any difference is an approximation effect, and here it has the opposite sign to the one expected.
- *The projected flow is not better than a final projection* (H6 refuted).
- *Equivariance does not lower the polarity asymmetry* (H9 refuted): the asymmetry was already at the sampling noise.
- *The helicity spectrum is not bipolar at* $k = 1$ *against* $k \geq 2$ (H11 refuted): the shell $k = 2$ is negative.
- *The magnetic-helicity identity is not the more violated one* (H12 refuted).
- *The null count does not scale as* k_λ^3 (H13 refuted on the slope).
- *Loss against tilting* (H8) was dropped: the criterion could be met by the sampling noise of one weighted ensemble.

The later hypotheses added five refutations. The embedding of the condition is not the cause of the failure of the first models (H15). Nor is the width of the network alone: a wider network is closer up to a point, and the widest one, with the lowest loss, is farther again (H21). Anchoring the velocity field on the mode-wise Gaussian field does not give a model at least as close as that law, with or without a learned affine correction (H19, H20). And the stationarity loss, applied to a model close to the reference, degrades the law beyond the margin that was registered (H23). Of the three hypotheses of the last tests, two were refuted, in opposite directions: the model with the affine term does hold a third-order part of the magnetic production (H25 predicted none), and, on the measure registered for the test (the largest deviation from the reference energy beyond the initial one), the Gaussian law with its first two moments does not move farther than the model does (H27 predicted that it would; refuted on one realisation of the data, the null control of the test failing on the other). That measure cannot see a passage above the reference smaller than the 12 % initial deficit; computed after the fact, the Gaussian law passes 2.4 % higher (2.6 % on the second realisation).

The tests of Section 5.6 added four refutations (H28, H38, H40 and, by its letter, the second clause of H34), one dropped hypothesis (H32) and two measurements that were invalid by the letter of their checks before being repeated (H35, H36); the account is in Section 5.6, Section A and Section H.

Two readings of the main result were refuted by the `jury-externe` role and withdrawn (non-Gaussian structure as the reason why the learned models beat the mode-wise Gaussian law; the ranking of the laws by the restart drift), two claims made during the work were corrected by the reviewing role, and one run labelled as a replication was not one: the account is in the appendix (Section H).

9.3 Limits

Every result of this report is obtained on one system: a 16^3 grid, $Re = 10$, one conditioning parameter, a periodic box. The limits below say what this restricts; Section 9.4 lists what the objects of the offer would need.

- **Scale.** The system has 665 Fourier modes and $Re = 10$. The flow is smooth, the statistics of the Fourier modes are close to Gaussian at second order, and that is why a mode-wise Gaussian law is such a strong baseline here. Nothing guarantees this at the Reynolds numbers of the MilaDy simulations, where the non-Gaussian part of the law carries the

physics (intermittency, coherent structures). The ranking of the mechanisms may change with scale; the method of comparison does not.

- **Architecture.** The network acts on the flattened vector of modes. This does not scale: at 64^3 the latent has more than 10^5 dimensions. The explicit affine term is mode-wise and scales linearly with the number of modes, but the nonlinear part needs an architecture with locality in Fourier or physical space (a convolutional or attention-based field, as in the references of the section State of the art).
- **Homogeneity.** The mode-wise form of the affine term and of the Gaussian baseline G0 treats different Fourier modes as uncorrelated, which is true of a statistically homogeneous field. It is not true here: the forcing has a fixed phase, the mean flow is strong, and the covariance between modes carries the mean magnetic production (Section 3.2). In a bounded domain (the von Kármán geometry of the project) the departure from homogeneity is larger still, and the natural basis is different.
- **Metrics.** The energy distance is computed on 21 physical features, all second-order quantities of one snapshot. A law can be close to the reference in this distance and wrong in its higher-order structure: the distance barely separates the best learned model from its Gaussianisation (0.43 floors, about twice the sampling value; the Gaussianised law is 0.6 to 0.9 floors farther from the reference; Section 5.5). The production split sees the difference clearly, and the restart drift at $\tau = 4$ by 0.045; the null count and the overshoot registered for the restart hardly do. The reference null count itself is measured on 400 snapshots drawn at random among 2400, and two draws on the same data differ by 4 and 10 % at the two lower held-out η (27.6 and 28.7 null points; 42.7 and 47.1): an uncertainty of every null-count error of this report that the standard errors over the seeds do not include. Counted on all the snapshots, the reference has 21.1, 27.1 and 44.0 null points at the held-out η, known to 3 to 5 % (standard error over the 24 trajectories); the null-count errors of the tables are against the draws of 400 of their own jobs.
- **One conditioning parameter.** Only η varies, over a range in which the dynamo is always saturated and the law changes smoothly. The dynamo threshold, where the law changes fastest (Proposition 9 of the theory brief), is outside the range. A diagnostic on the CPU of the session (six trajectories per η over 800 time units; not a result of the study) places it between $\eta = 0.09$ and 0.095: below, the magnetic energy varies by a factor of ten between trajectories; above, it decays, slowly and unevenly near the threshold. A stationary law is not defined there on the times of this protocol (200 time units of transient, 200 kept), and a study across the threshold needs runs of thousands of time units.
- **Statistics.** Five seeds per variant; one realisation of the reference data for the main run and one independent realisation for the replication (a third one, by accident, for the second job of H34). For the Gaussian laws and the null controls the five seeds are five draws, not five trainings. The null control of the recolouring measures the estimation error of the moments for a Gaussian source and understates it for a non-Gaussian one. The objective on the distance to the reference law is met inside the uncertainty of the noise floor.
- **Restart test.** The metric of the objectives measures the drift at one time, $\tau = 4$, relative to the ensemble's own initial energy; it is confounded with the bias of the law on the energy. The anchored curves to $\tau = 16$ exist for the Gaussian laws, the two models with the affine term and the anchored model W10 only; the anchored values quoted for two other variants come from older runs and an approximate initial energy, good to ± 0.01; for the other variants the tables give the single-time drift, which should not be read as a ranking. The rise of the magnetic energy at restart of the six laws tested that start on the reference or below it is measured and not explained; the overshoot registered for H26 and H27 gives a law an allowance equal to its initial deviation, and the null control of H27 failed on one of the two realisations of the data, which the review traced to the sampling noise of the initial energy on 128 restart samples.

- **Downstream use and reversals.** No reduced-order model and no data assimilation was fed with generated fields, and no statistics of polarity reversals were measured or modelled (Section 5.9). The restart in the solver is the only test of a generated field as an initial condition.

9.4 What the turbulent regime, the dynamo threshold and the data of the project would need

None of what follows was done. It is the list of what separates this study from the objects of the offer, with what is **measured** here to size each step; the statements about larger systems are **conjectured** unless a reference is given.

The turbulent regime. Measured here: at 16^3 a snapshot is 5320 real coefficients, the data of one η are 2400 snapshots (194 MiB in the format of the project), a training takes about one minute on a rented GPU, and the GPU cost of the whole study is given in Section J. Three components do not carry over. The network acts on the flattened vector of modes; with the same truncation the latent has 3×10^5 real coefficients at 64^3 and 2.5×10^6 at 128^3, and the nonlinear part needs an architecture with locality in Fourier or physical space. The second moments were taken from a full sample covariance, 1.4×10^7 numbers per training η here and 5×10^{10} at 64^3, out of reach of any number of snapshots; the recolouring then needs a structured covariance. One was built here (one block per Fourier mode plus the 300 leading inter-mode directions; its mean-flow production is 0.11 of the dissipation below the reference) and failed as the anchor of a network (H40), but it was not tested as the target of the recolouring. And the Gaussian baselines are strong here because the flow is smooth: at higher Reynolds numbers the non-Gaussian part of the law carries more of the physics, and the ranking of the mechanisms may change. What carries over is the method: the Gaussian baselines before any claim about a network, the split of the production, the restart curve anchored on the reference, the null count on all the snapshots, and the residual relaxation time in the solver. The first step is a 64^3 box at higher Reynolds number with the same forcing, for which the solver, the data format and the job tooling exist. Its cost can be sized from one measurement: the solver takes 4.9 ms per step at 64^3 on a rented GPU (float32, one trajectory; `research/teams/claude/journal.md`), so 24 trajectories of 400 time units at an assumed time step of 0.005 (scaling the step of the 16^3 runs with the grid gives 0.0125) are about 2.6 hours of GPU if run one after the other, less in batches, for each value of the parameter; the training cost at that size is not measured. Two criteria could be registered for that step (**proposals**, not registered): that a Gaussian law with a structured covariance stays within a few noise floors of the reference there, as G1 does here, or not; and that the recolouring onto that covariance still lowers the restart drift of a learned model below that of the Gaussian law. Published three-dimensional MHD turbulence data exist at 64^3 and 256^3 in The Well (Ohana et al. 2024) (72 GB and 4.6 TB; ideal MHD, compressible and isothermal, so without a magnetic diffusivity to condition on), whose format the tools of the toolbox read. Storage is the cost that changes first: 100 snapshots of one trajectory at 128^3 are about 5 GiB before compression.

The dynamo threshold. Known here only from a diagnostic on the CPU of the session, which is not a result of the study and has no committed result file (six trajectories per η over 800 time units; recorded in `research/teams/claude/journal.md` on 13 November): the magnetic energy had not decayed after 800 time units at $\eta = 0.09$, where it varies by a factor of ten between trajectories, and it decays at $\eta = 0.095$. Three things are needed. Runs of thousands of time units per trajectory near the threshold, before a stationary law exists to be compared with (200 time units are kept here). A model whose law can change fast with the parameter, down to the non-magnetic state: the Gaussian laws of this study interpolate a mean and a covariance

linearly between training values, and the conditioning of the network was tested only where the law changes smoothly (Proposition 9 of the theory brief gives a lower bound on the error of a model that depends smoothly on the parameter when the law jumps). And metrics that remain defined when the magnetic energy vanishes: every ratio to the Ohmic dissipation or to the reference magnetic energy used here diverges there.

Polarity reversals. A generator of independent snapshots cannot reproduce their course or their waiting times; their mean rate can be estimated from snapshots and the equations (Proposition 8 of the theory brief), on the reference data first, which was not done. The rest needs either a generative model of trajectories or of transitions (the law of the state after a time lag given the present state and the parameters), or the route that the offer names: generated fields as inputs of a reduced-order model that carries the dynamics. Measured here: the solver accepts none of the generated ensembles as stationary without a relaxation (the recoloured model passes 4.6 % above the reference magnetic energy, 6.3 % on the second realisation of the data), so a model fed with such fields starts with a transient of a few time units; and each reference trajectory holds 2.5 to 12 correlation times of the large-scale modes, too few for statistics of reversals. The quantities to measure are then those of the dynamics, as functions of the parameters: the waiting times between reversals and the law of the amplitude of the large-scale field. In the experiment (**published**), the reversal regime was observed with the two impellers driven at different frequencies, with mean waiting times of 120 and 50 s between reversals that last about 5 s, for an Ohmic time of about 0.4 s (Berhanu et al. 2007): hundreds of Ohmic times between reversals, where a trajectory of this study holds 6 to 16.

The simulations of the project. The nearest published study of the project simulates the reversals of the von Kármán sodium experiment at $Re = 1594$ and $Rm = 531$ and analyses 2000 snapshots (**published**: Bousquet et al. (2026); data not public): a kinetic Reynolds number 160 times that of this study in nominal value, with a number of snapshots smaller than that of one value of η here. The experiment has also been simulated with the code SFEMaNS in its bounded geometry, with soft-iron impellers of relative magnetic permeability of order 50; the magnetic Prandtl number of sodium is of order 10^{-5} (**published**: Nore et al. (2018)); the magnetic Prandtl number of this study is 0.6 to 1.7. Three things change. The representation: there is no Fourier basis in which zero divergence is diagonal; the constraint by design becomes a discrete curl of a vector potential, or a projection with the boundary and interface conditions, and the mode-wise affine term and Gaussian baseline, which rest on statistical homogeneity, need another basis. The data: the format of the project has an extension that stores Fourier coefficients in azimuth on a finite-element mesh of the meridian plane without interpolation, with tools tested on synthetic fields; the reader of the native files is not written, the data not being public (`docs/specs/data-format.md`, §9). And the parameters: the Reynolds number, the magnetic Reynolds number and the material properties vary together there; the conditioning on several parameters was not tested here.

10 Conclusion

This report compared, on a small parametric MHD dynamo (16^3 grid, $Re = 10$, the magnetic diffusivity as the only parameter), the three places where a physical constraint can enter a generative model. All statements below are **measured** on this system: five seeds for the models, a jackknife over the 24 trajectories for the reference, and held-out parameter values unless said otherwise.

1. Zero divergence is obtained exactly by design (potentials) or by projection at sampling time; a penalty in the loss does not enforce it and degrades the law.
2. None of these divergence mechanisms decides the quality of the generated law. What decides it is whether the velocity field of the flow-matching model can carry the linear, second-order part of the transport: an explicit mode-wise affine term brings the same network from 205 noise floors to about 9.
3. In energy distance, the learned models do not beat two Gaussian baselines without a network. A Gaussian law per Fourier mode is at 3.5 floors but has no magnetic production, because it has no covariance between modes. A Gaussian law with that covariance, fitted on the training values and interpolated, is the closest law of the first protocols (2.5 floors; the recoloured model of item 6, which takes its second moments, is at 2.1 on the mean over the held-out values), with a magnetic energy balance residual of 0.07 and half the null-count error of the first. The learned models with the affine term have a lower residual of the magnetic energy balance on the mean (0.055 against 0.072), through two errors that cancel at the two lower values and add at the largest (item 5).
4. A law-level stationarity loss can be imposed at parameter values without data. It lowers the residual of the energy identity, on a wrong model and on a nearly right one, and it is not free: on the nearly right model it moves the law away from the reference. Exponential tilting on the same identities, at sampling time, collapses on a wrong model and improves a nearly right one on most metrics. Imposing a balance on each sample satisfies it exactly and removes real fluctuations.
5. The reference law is non-Gaussian in its magnetic production: the third-order part goes from $+0.14$ to -0.31 of the Ohmic dissipation over the range of η. The model with the affine term holds 0.6 and 0.4 of it at two held-out values and a small part of the opposite sign at the third; the Gaussian law fitted to its samples holds none, and the energy distance separates the two by 0.4 floors only. The same model trained on Gaussian data with the first two moments of the data, independent or with the time correlation of the trajectories, holds none either; and a network anchored on the exact Gaussian field of the data learns none in the same budget, nor any from samples of the model itself, so that this last failure says nothing about the data. Among the anchored models, a third-order part is measured only when the anchor is the mode-wise Gaussian law; with 30 or 300 inter-mode directions in the anchor, or the full sample covariance, none is (seed means below 0.002 of the dissipation), and the cause is not established.
6. The second moments do not have to come from the network. A linear map applied to the samples of the model at sampling time gives them the mean and covariance fitted on the data and keeps 0.74 and 0.58 of the third-order production of the reference. The recoloured law has the energy balance and the spectrum of the Gaussian law, an energy distance of 2.1 floors, a restart drift of 0.021 and, at equal moments, half the passage of

Gaussian draws above the reference energy at restart; the same model trained on Gaussian data and recoloured the same way has the passage, the drift and the magnetic balance of the Gaussian law, so this comes from what the data hold beyond their second moments. It is within the thresholds that the pilot set (Table D.1) on the energy distance, the energy balance, the spectrum and the restart drift; the first three are those of the copied moments.

7. The model trained on the data misses 11 to 16 % of the magnetic energy and 0.13 to 0.15 of the dissipation in mean-flow production; the tests of where this error comes from are in Section A.
8. No law sampled without the solver reproduces the magnetic topology: Gaussian or learned, recoloured or not, the laws close to the reference have two to three times its null points, and the Gaussian law with the spectrum of the reference has 1.4 to 1.6 times its mean $|\det \nabla B|$. Nor does any produce states that the solver accepts as stationary: restarted, the Gaussian laws fitted to the data overshoot the reference magnetic energy by 9 and 17 % at the held-out values (the first by 8 to 13 % at the training values, H31), the learned model comes back to it from a 12 % deficit, and the recoloured model passes 5 % above it. From states with the second moments of the data, 4 time units of the solver bring the null count within 0.05 to 0.13 of that of the reference, while the energy distance and the magnetic energy first degrade (maximum after one time unit); followed by that relaxation, the recoloured model is within the thresholds on the null count too, with a residual relaxation time (the first duration at which the null-count error is at most 0.2 and the magnetic energy within 5 % of the reference) of 2 to 4 time units against 8 for the Gaussian law and 64 for the solver alone from the initial condition of the data.

What remains **conjectured**: that the excess of null points of the Gaussian and recoloured laws is a missing alignment of the small scales of the field by the flow, which the solver rebuilds in a few time units (the mechanism is not tested; only the relaxation is measured); that the degradation of the wide and long-trained networks is memorisation of the training snapshots; why the model trained on the data loses second moments that the same model keeps on Gaussian data; and that the ranking of the mechanisms carries over to turbulent regimes.

The next steps for the internship follow from these results. Keep the Gaussian law with the full covariance as the baseline, and report every model next to its Gaussianisation and to its recoloured version. Treat the second moments as a constraint at sampling time: at 64^3 a full sample covariance cannot be estimated, so the map needs a structured covariance (mode-wise, plus the low-rank part that carries the mean-flow production), which was built here and remains to be tested as the target of the map. Use an architecture with locality for the non-Gaussian part, since the flattened perceptron does not scale and a network anchored on the Gaussian law learned nothing here. Count the null points of the reference on all its snapshots, and put among the objectives the two parts of the production split, the largest deviation from the reference energy over the anchored restart curve, and the residual relaxation time in the solver, which measures in time units what a generator saves over a simulation started from scratch. Use the solver as the last stage of the sampler where the topology matters, and measure what a few time units cost and repair at higher resolution. Look for a training signal aimed at the determinant of the gradient of the field, where every law without the solver fails. Find out why the model trained on the data loses magnetic energy. Repeat the protocol on a 64^3 box at higher Reynolds numbers, where the law is less Gaussian; a range of η across the dynamo threshold needs runs of thousands of time units near the threshold, where the field is intermittent, before a stationary law can be compared there.

What the offer asks and this report leaves open is in Table 1.1 and Section 9.4: parameters other than η, the turbulent regime, polarity reversals, the use of generated fields by a reduced-order model or a data assimilation, and the simulations of the project.

11 References

Albergo, M. S., and E. Vanden-Eijnden. 2023. "Building Normalizing Flows with Stochastic Interpolants." In *ICLR 2023*.

Alexakis, A. 2011. "Searching for the Fastest Dynamo: Laminar ABC Flows." *Physical Review E* 84: 026321. https://doi.org/10.1103/PhysRevE.84.026321.

Baldan, G., Q. Liu, A. Guardone, and N. Thuerey. 2026. "Physics Vs Distributions: Pareto Optimal Flow Matching with Physics Constraints."

Bastek, J.-H., W. Sun, and D. M. Kochmann. 2025. "Physics-Informed Diffusion Models." In *ICLR 2025*.

Berhanu, Monchaux, Fauve, Mordant, Pétrélis, Chiffaudel, Daviaud, et al. 2007. "Magnetic Field Reversals in an Experimental Turbulent Dynamo." *EPL* 77: 59001. https://doi.org/10.1209/0295-5075/77/59001.

Bousquet, R., Y. Ponty, V. Botez, N. Plihon, and C. Nore. 2026. "Magnetic Field Reversals in Numerical Simulations of the von Kármán Sodium Experiment." *Physical Review Fluids* 11 (6): 063703. https://doi.org/10.1103/9tx2-1m13.

Brandenburg, A. 2001. "The Inverse Cascade and Nonlinear Alpha-Effect in Simulations of Isotropic Helical Hydromagnetic Turbulence." *Astrophysical Journal* 550: 824–40. https://doi.org/10.1086/319783.

Brandenburg, A., and K. Subramanian. 2005. "Astrophysical Magnetic Fields and Nonlinear Dynamo Theory." *Physics Reports* 417: 1–209. https://doi.org/10.1016/j.physrep.2005.06.005.

Cai, Z., Y. Cao, Y. Huang, and X. Zhou. 2026. "Weak Generative Sampler to Efficiently Sample Invariant Distribution of Stochastic Differential Equation." *SIAM Journal on Scientific Computing* 48 (4): C708–35. https://doi.org/10.1137/24M1665271.

Cheng, C., B. Han, D. C. Maddix, A. F. Ansari, A. Stuart, M. W. Mahoney, and Y. Wang. 2025. "Gradient-Free Generation for Hard-Constrained Systems." In *ICLR 2025*.

Christopher, J. K., S. Baek, and F. Fioretto. 2024. "Constrained Synthesis with Projected Diffusion Models." In *NeurIPS 2024*. https://doi.org/10.52202/079017-2834.

Chung, Kim, McCann, Klasky, and Ye. 2023. "Diffusion Posterior Sampling for General Noisy Inverse Problems." In *ICLR 2023*.

Csiszár, I. 1975. "I-Divergence Geometry of Probability Distributions and Minimization Problems." *Annals of Probability* 3 (1): 146–58. https://doi.org/10.1214/aop/1176996454.

Du, P., M. H. Parikh, X. Fan, X.-Y. Liu, and J.-X. Wang. 2024. "Conditional Neural Field Latent Diffusion Model for Generating Spatiotemporal Turbulence." *Nature Communications* 15: 10416. https://doi.org/10.1038/s41467-024-54712-1.

Galloway, D., and U. Frisch. 1986. "Dynamo Action in a Family of Flows with Chaotic Streamlines." *Geophysical and Astrophysical Fluid Dynamics* 36 (1): 53–83. https://doi.org/10.1080/03091928608208797.

Genuist, W., É. Savin, F. Gatti, and D. Clouteau. 2026. "Divergence-Free Diffusion Models for Incompressible Fluid Flows."

Haynes, A. L., and C. E. Parnell. 2007. "A Trilinear Method for Finding Null Points in a Three-Dimensional Vector Space." *Physics of Plasmas* 14 (8): 082107. https://doi.org/10.1063/1.2756751.

Ho, J., A. Jain, and P. Abbeel. 2020. "Denoising Diffusion Probabilistic Models." In *NeurIPS 2020*.

Huang, J., G. Yang, Z. Wang, and J. J. Park. 2024. "DiffusionPDE: Generative PDE-Solving Under Partial Observation." In *NeurIPS 2024*, 130291–323. https://doi.org/10.52202/079017-4140.

Kerrigan, Migliorini, and Smyth. 2024. "Functional Flow Matching." In *AISTATS 2024, PMLR 238*, 3934–42.

Köhler, J., L. Klein, and F. Noé. 2020. "Equivariant Flows: Exact Likelihood Generative Learning for Symmetric Densities." In *ICML 2020, PMLR 119*, 5361–70.

Li, Buzzicotti, Bonaccorso, and Biferale. 2026. "Physics-Constrained Diffusion Model for Synthesis of 3D Turbulent Data."

Lienen, Lüdke, Hansen-Palmus, and Günnemann. 2024. "From Zero to Turbulence: Generative Modeling for 3D Flow Simulation." In *ICLR 2024*.

Lim, J. H., N. B. Kovachki, R. Baptista, C. Beckham, K. Azizzadenesheli, J. Kossaifi, V. Voleti, et al. 2025. "Score-Based Diffusion Models in Function Space." *Journal of Machine Learning Research* 26 (158): 1–62.

Lipman, Chen, Ben-Hamu, Nickel, and Le. 2023. "Flow Matching for Generative Modeling." In *ICLR 2023*.

Liu, X., C. Gong, and Q. Liu. 2023. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow." In *ICLR 2023*.

Nore, Castanon Quiroz, Cappanera, and Guermond. 2018. "Numerical Simulation of the von Kármán Sodium Dynamo Experiment." *Journal of Fluid Mechanics* 854: 164–95. https://doi.org/10.1017/jfm.2018.582.

Ohana, R., M. McCabe, L. Meyer, R. Morel, et al. 2024. "The Well: A Large-Scale Collection of Diverse Physics Simulations for Machine Learning." In *NeurIPS, Track on Datasets and Benchmarks*.

Parnell, C. E., J. M. Smith, T. Neukirch, and E. R. Priest. 1996. "The Structure of Three-Dimensional Magnetic Neutral Points." *Physics of Plasmas* 3 (3): 759–70. https://doi.org/10.1063/1.871810.

Ponty, Y., P. D. Mininni, D. C. Montgomery, J.-F. Pinton, H. Politano, and A. Pouquet. 2005. "Numerical Study of Dynamo Action at Low Magnetic Prandtl Numbers." *Physical Review Letters* 94: 164502. https://doi.org/10.1103/PhysRevLett.94.164502.

Pouquet, A., U. Frisch, and J. Léorat. 1976. "Strong MHD Helical Turbulence and the Nonlinear Dynamo Effect." *Journal of Fluid Mechanics* 77 (2): 321–54. https://doi.org/10.1017/S0022112076002140.

Richter-Powell, Lipman, and Chen. 2022. "Neural Conservation Laws: A Divergence-Free Perspective." In *NeurIPS 2022*, 38075–88. https://doi.org/10.52202/068431-2759.

Salimans, T., and J. Ho. 2022. "Progressive Distillation for Fast Sampling of Diffusion Models." In *ICLR 2022*.

Shu, D., Z. Li, and A. Barati Farimani. 2023. "A Physics-Informed Diffusion Model for High-Fidelity Flow Field Reconstruction." *Journal of Computational Physics* 478: 111972. https://doi.org/10.1016/j.jcp.2023.111972.

Song, Sohl-Dickstein, Kingma, Kumar, Ermon, and Poole. 2021. "Score-Based Generative Modeling Through Stochastic Differential Equations." In *ICLR 2021*.

Székely, G. J., and M. L. Rizzo. 2013. "Energy Statistics: A Class of Statistics Based on Distances." *Journal of Statistical Planning and Inference* 143 (8): 1249–72. https://doi.org/10.1016/j.jspi.2013.03.018.

Utkarsh, Cai, Edelman, Gómez-Bombarelli, and Rackauckas. 2025. "Physics-Constrained Flow Matching: Sampling Generative Models with Hard Constraints." In *NeurIPS 2025*. https://doi.org/10.52202/085713-5354.

Wang, S., Z. Dou, S. Shan, T.-R. Liu, and L. Lu. 2026. "FunDiff: Diffusion Models over Function Spaces for Physics-Informed Generative Modeling." *Nature Communications* 17: 5749. https://doi.org/10.1038/s41467-026-72292-0.

Woltjer, L. 1958. "A Theorem on Force-Free Magnetic Fields." *Proceedings of the National Academy of Sciences* 44 (6): 489–91. https://doi.org/10.1073/pnas.44.6.489.

12 Reproducibility

12.1 Commands

Everything is in the branch `claude/modest-newton-zx8fqs` of the repository. From the root:

```
make test                                    # tests of src/milady and of the tools
cd experiments
python3 e4_data.py --quick                   # reference data (--jobs N: full data)
python3 e4_mhd3d.py --reference --quick      # reference statistics and floors
python3 e4_mhd3d.py --quick --seed 0 --group all   # groups A-E
python3 e4_mhd3d.py --quick --seed 0 --group D,G,M # G: Gaussian law G1; M: loss and
                                             # tilting on W9; L: long; S: width scan
python3 e4_floor.py                          # noise floors
python3 e4_assemble_scan.py                  # width scan from its three jobs
python3 digest.py > results/digest.md        # digest of E1-E4
python3 race_metrics.py --all                # objectives of the race
python3 e4_plots.py                          # data plots of this report
python3 e4_report_tables.py                  # tables of this report
python3 e4_export_explorer.py --team claude     # data of the explorer
python3 e4_polarity_diag.py                  # diagnostic of the polarity section
# the release (data in MFF, models, catalogue) is built by the workflow
# .github/workflows/e4-release.yml: bash e4_release_build.sh all <out> <url>
# on a GPU pod (5 seeds), from the root:
python3 tools/runpod_job.py run --hours 2 \
    --outputs experiments/results/replications/e4_final \
    -- bash experiments/e4_gpu_job.sh full "0 1 2 3 4" e4_final all
python3 tools/render.py teams/claude/report teams/claude/explorer
python3 tools/check_deliverables.py --team claude --built
```

Without `--quick`, the data take about 25 minutes on 11 CPU cores and one seed of the groups A to E about 35 minutes on a GPU. Quick outputs are not versioned and are never quoted.

12.2 Result files and commits

Table 12.1: Result sets of the hypotheses H4 to H27, with the commit of the code that produced them and the commit that added them; for H28 to H41 the result sets and their commits are in the fields `results` and `commit` of the register (`research/teams/claude/hypotheses.json`) and in the digest. The registrations precede the results of the hypotheses they concern.

Result set	Location	Code commit of the run	Commit of the results
Objectives of the team (day 1)	`race/teams/claude/objectives.toml`	—	`90e64af`
Register: H4–H14, H15, H16–H18, H19–H20, H21	`research/teams/claude/hypotheses.json`	—	`68e2fd4`, `9fd7272`, `fc04247`, `f3868f5`, `5027f7c`
Main run: groups A–E, 5 seeds (GPU job `e4_final`)	`experiments/results/e4_*.json`	`973fe3a`	`991d682`
First GPU run of groups A and B (reviewed for H4–H14)	`experiments/results/replications/e4_gpu`	`d0e4018`	`fb0786b`
Bitwise re-run of groups A and B	`experiments/results/replications/e4_main`	`62e31ca`	`8184bc9`
Convergence of the data	`experiments/results/replications/e4_convergence`	`d0e4018`	`56dbb91`
CPU seeds 0 and 1 (container)	`experiments/results/replications/e4_cpu`	—	`991d682`
Long trainings (H18): groups D, E, L	`experiments/results/replications/e4_final_long`	`973fe3a`	`bf07893`
Independent replication (other data seeds): groups A–E	`experiments/results/replications/e4_indep`	`d7aa2bd`	`d7aa2bd`
Register: H22, H23, H24	`research/teams/claude/hypotheses.json`	—	`a0cd918`
Gaussian law G1, loss and tilting on W9 (H22–H24): groups D, E, G, M	`experiments/results/replications/e4_gauss`	`6074cbd`	`cc1da7b`
G1 on the second data realisation: groups D, G	`experiments/results/replications/e4_gauss_indep`	`81a0e0c`	`133934a`
Width scan (H21), 5 seeds on three jobs: groups D, S	`experiments/results/replications/e4_scan2` (parts `e4_scan2a`, `b`, `c`)	`6074cbd`, `d7aa2bd`, `830a40a`	`81a0e0c`, `133934a`
First width scan, cut by its time limit	`experiments/results/replications/e4_final_scan`	`973fe3a`	`d7aa2bd`
Register: H25, H26, H27	`research/teams/claude/hypotheses.json`	—	`bad65ae`
Production split, anchored restart, Gaussianised W9 (H25–H27): groups D, E, G, N, on both data realisations	`experiments/results/replications/e4_nongauss`, `e4_nongauss_indep`	`2e87cbd`	`245c783`

The main run was launched with a file-naming bug (Section I): its result files were written under

another suffix, copied back by a loop started on the pod, and renamed; the `encadrant` checked the renamed files against the per-seed logs of the job (`research/teams/claude/gpu-logs/e4-final/`).

12.3 The catalogue of the release

The hyperparameters and the evaluations per seed are in the SQLite file only; the JSON export holds the seed means. The hyperparameters are the configuration of the run: its `cond_scale` of 8 is that of the first-generation models, the condition embedding of W9 having the scale 0.5 set in the code.

12.4 Figures of this report

The eight schematics were drawn with the `schematic-figure` skill (sources in `site/teams/claude/figures/<n` rendered to SVG and PDF), the three diagrams with Archify (`site/teams/claude/diagrams/`), and the data plots by `experiments/e4_plots.py` from the result files named in each caption (`site/teams/claude/plots/plots.json` lists them).

A Further tests on the origin of the non-Gaussian part

This appendix holds the full account of what Section 5.5 and Section 5.6 summarise: the remarks of the reviews on the restart tests (H26, H27), the chain of the anchored models (H28, H38, H40, H41), the map applied to reference snapshots, a descriptive result of H33, the restart of the Gaussian law at the training values (H31), and five further hypotheses: two side tests on what in the data produces the second-order error of the learned model (H34, H36), the dropped hypothesis on the Kac-Rice factors of the null count (H32), and the relaxation in the solver with its baseline (H35, H39), including the repair of its measurement. All of it is **measured**, with five seeds.

The restart anchored on the reference: what the criterion does not show (H26). The review of this result by the `encadrant` role adds what the criterion does not show. The overshoot gives each law an allowance equal to its own initial deviation, and the six laws of this job that start on the reference energy (within 0.04) or below it (G0, G1, G2, G3, W9, W11) all end up gaining magnetic energy, by 0.10 to 0.21 of the reference energy: the rise $\max_\tau a(\tau) - a(0)$ is 0.12 for W9, 0.11 for G1, 0.16 for G2 and 0.21 for G0, which first falls by 16 % (W10, which starts 21 % above the reference, relaxes downwards). W9 starts 12 % low, so its rise lands on the reference; G1 starts on the reference, and a rise of 0.11 takes it away. That the two rises have one cause is not established: W9 may simply relax to the stationary energy, as W10 does from above. The older runs stop at $\tau = 4$, where the six laws above have gained 0.06 to 0.21: there, with an approximate initial energy, the model fine-tuned with the stationarity loss and the network of width 2048 have gained 0.07 and are still rising (0.11 and 0.03 below the reference), which does not contradict the range; the network of width 5632, which starts 34 % low, has gained 0.29, above it. The largest distance to the reference energy over the whole course is comparable for the two (0.13 for W9, at the start; 0.10 for G1, at $\tau = 2$; 0.11 and 0.12 on the second realisation). The result therefore does not show that the states of W9 are closer to the attractor than those of G1. It shows that the restart drift of the objectives, relative to each ensemble's own initial energy, cannot be read as a ranking: its value for W9 (0.124 at $\tau = 4$) is almost entirely the recovery of the initial deficit, and its value for G1 (0.064) is an excursion on its way back. Neither law is stationary under the solver. The excursion of G1 is present at each held-out η, including the largest, where its production is below its dissipation: its instantaneous magnetic balance does not explain it, and no explanation was tested.

The Gaussian law with the moments of W9: what the refutation allows (H27). What this refutation allows is narrow. The registered measure does not separate W9 from the Gaussian law with its first two moments, because it cannot see an excursion beyond the reference smaller than the initial deficit; it does not show that the two behave alike. Computed after the fact, and therefore a **conjecture** until it is registered and tested, the excursion of G2 above the reference energy is larger than that of W9 by 0.024 ± 0.004 (0.026 ± 0.003 on the second realisation, in the five seeds of each): G2 rises faster and passes 2.4 % higher (2.6 % on the second realisation). A missing third-order sink of production would give this at the two lower η (production over dissipation 1.01 and 1.13 for G2, 0.93 and 1.02 for W9). The differences per η do not support that reading: the difference decreases as the missing third-order part grows (0.039, 0.021 and

0.011 for production gaps of 0.00, 0.08 and 0.12) and is not significant at the smallest η (1.5 standard errors, 3 of 5 seeds).

A network anchored on G1 learns nothing more (H28). The velocity field of a flow-matching model is written as the exact field of the Gaussian law G1 at the condition η, plus a network of the size of W9 whose last layer starts at zero: at iteration 0 the model is G1, and the network has only the non-Gaussian part to learn. The model is read along one training, at 250, 500, 1000, 2000 and 4000 iterations (Figure 5.7). The hypothesis was that some checkpoint, chosen on a validation η, would hold between 0.3 and 1.5 of the third-order part of the reference while staying as close to it as G1. It is **refuted** on both realisations by its first clause: at no checkpoint does the model hold more than 5×10^{-3} of the third-order part of the reference (seed means -3×10^{-4} on the data of the main run, $+4.5 \times 10^{-3}$ on the second realisation). The check of the measurement passes (iteration 0 reproduces G1 sampled exactly). The generated law stays that of G1, within 0.2 % on the energy distance, over the 4000 iterations, and the flow-matching loss on a fixed training sample does not move (0.89449 and 0.89157 on the two realisations, changes of at most 3×10^{-5}). The review checked with its own replica that the network did train, and a unit test shows that the same anchored model learns a skewed law in low dimension. What is measured is narrow: in this budget and with this optimiser, a network given the exact Gaussian field learns nothing more from 14 400 snapshots in dimension 5320. It is not shown that the non-Gaussian part cannot be learned this way. A positive control at full size, asked by the `jury-externe` role (H38), shows where the failure lies. The training set is replaced by 14 400 samples of W9 (2400 per training η), whose third-order production is known, and the anchor by the Gaussian law fitted to these samples. The hypothesis that the anchored model would then recover at least 0.3 of the third-order part of W9 is **refuted**: it recovers 0.006 and -0.002 of it at the two lower held-out η (between -0.04 and $+0.02$ in every seed), from iteration 0 to iteration 4000, and its loss on a fixed training sample stays at 1.0370; W9 retrained from scratch on the same samples recovers 0.32 and 0.44 of it. The negative result of H28 therefore belongs to the anchored construction, or to its optimisation in this budget and dimension, and says nothing about the data. Its cause is not identified. The review checked on a small case that the anchored network receives a gradient and that the evaluated copy is the trained one. The anchor is the Gaussian law of a covariance estimated with fewer samples than dimensions at each training η, and its loss on the training set is below that of a model that generalises (1.04 against 1.22 for the retrained W9 on the same sample), which suggested that it leaves the network nothing to learn; a small case of the review with many more samples than dimensions did not learn either. The conjecture was tested at full size with an anchor of full rank (H40): a Gaussian law whose covariance keeps the 300 leading eigen-directions of the sample covariance at each training η and one block per Fourier mode for the rest (the block of every mode is exact, the inter-mode covariance is that of the leading directions; with 300 of them the mean-flow production of the law is 0.11 of the dissipation below the reference). The three clauses of H40 are **refuted** on both realisations. With this anchor the loss on a fixed training sample falls (1.208 to 1.151) and the generated law changes, which never happened with the sample covariance; but it moves away from its anchor (from 2.5 to 16.4 floors, the magnetic energy 11 to 15 % too high, the energy residual from 0.005 to 0.07 and more) and no third-order production appears: 0.001 of that of the reference for the model trained on the data, 0.002 and 0.004 of that of W9 in the positive control. The rank deficiency of the sample covariance is therefore not a sufficient explanation, and giving a network without the affine term the Gaussian part of the law by design did not work at rank 300 in this budget. The four constructions without a third-order production share one piece of code, which had never produced one. Its control (H41) passes: with an anchor of rank 0, which is the mode-wise Gaussian law, that code reproduces the model anchored on the mode-wise law through the earlier code (W10) within 1.3×10^{-4} of the dissipation on the third-order part (limit 0.02; five seeds, data of the main run). The control clears the shared code for an anchor whose covariance is block-diagonal; an anchor with dense

eigenvectors is not exercised in training by it, and the floor of the small eigenvalues, used only by the anchors of H28 and H38, is not exercised at all, so a fault specific to those cases is not excluded. With one more job at rank 30, the measurements read as follows, without a causal claim. A third-order part (0.31 and 0.24 of that of the reference at the two lower held-out η; at the largest one W10 has the opposite sign to the reference) is measured only with the mode-wise anchor; with 30 or 300 inter-mode directions in the anchor, or the full sample covariance, none is (seed means below 0.002 of the dissipation, at most 0.005 in any seed). The measurement at rank 30 is one realisation of the data at one intermediate rank: the ranks between 1 and 29 were not run, and no threshold in the rank is claimed. Neither anchored model is a good law: the models of rank 0 and of rank 30 end 44 floors from the reference, against 2 to 4 for their anchors, with a magnetic energy 10 to 28 % too high, and the total production of W10 is 0.43 to 0.60 of the dissipation against 1 for the reference.

The map applied to reference snapshots. The same jobs give the measurement that the jury called the ceiling of the construction: 512 reference snapshots per η, sent through the same kind of map from the moments of the 2400 snapshots of that η onto the moments of G1. They keep 0.81 and 0.76 (0.86 and 0.78) of their third-order production at the two lower held-out η, and they get a null-count error of 0.73 (0.68) where the same snapshots without the map have 0.04, an excursion at restart of 0.027 (0.050) and a restart drift of 0.023 (0.020). This is not a ceiling, and not an explanation: the map is applied to the snapshots from which its source moments were estimated, with fewer snapshots than dimensions, and the review showed on a toy case that such a map moves the states by itself; W16 is below this "ceiling" on the drift at two values of η and 0.26 above it on the null count. What it does show is that a linear map onto the moments of G1 is not neutral for real states: at this sample size, the second moments fitted on the training values and interpolated are not close enough to those of the reference at a held-out η for their imposition to leave its topology alone.

The model trained on Gaussian draws: a descriptive result (H33). The same job gives a second, descriptive, result, with no registered threshold. Trained on Gaussian draws, the model lands near the law of its training set: its energy distance to the reference is 1.6 to 4.2 floors against 1.3 to 3.3 for G1 (higher than G1 by 0.2 to 0.8 floors, in 27 of 30 seed pairs), and its magnetic energy is within 5.3 % of that of the reference. Trained on the data, with the same seed, it is at 4.3 to 12.5 floors with a deficit of magnetic energy of 11 to 15 %. On independent Gaussian data of this dimension and covariance, the same architecture and budget thus show no such deficit of magnetic energy (at most 5.3 %), but they still miss 0.06 to 0.10 of the dissipation in the mean-flow production (G1 is within 0.02), about half of what the model trained on the data misses (0.13 to 0.15): part of the second-order error of W9 is there without any non-Gaussian data. What in the data produces the rest (their non-Gaussian part, the time correlation of their trajectories, or something else) is not established by this comparison.

The rise of the Gaussian law at restart without interpolation (H31). H31 closes a question left by H26: G1 is interpolated in η at the held-out values, so its rise at restart could have been an effect of the interpolated covariance. Restarted at three training η, where it has the sample mean and covariance of the data and no interpolation, G1 overshoots the reference energy by 0.13, 0.08 and 0.09 (0.12, 0.10, 0.09), in every seed above the 0.05 that the hypothesis allowed. H31 is **confirmed**: the interpolation is not needed to produce the rise; Gaussian states with the second moments of the data gain magnetic energy when the solver takes them over. (The reference itself was not restarted at the training η; at the held-out ones it stays within 0.9 %.)

Gaussian trajectories with the time correlation of the data (H34). The surrogate of H34 keeps the structure of the training set: at each training η, 24 Gaussian trajectories of 100 snapshots with the mean and covariance of the data and, along each eigen-direction of that

covariance, the lag-one autocorrelation measured on the data (0.75 to 0.79 on the mean weighted by the variances, reproduced within 0.002). The hypothesis had two clauses. The first holds on both realisations: trained on these trajectories, the model has a third-order part of at most 0.003 of the dissipation, under 3 % of that of the model trained on the data at the two lower held-out η (at most 4 % in any seed). With H33, this model does not produce a third-order production from Gaussian data, independent or with the per-direction time correlation of the data (correlations in time between directions are not reproduced by the surrogate). The second clause said that the deficit of magnetic energy of W9 would not appear either, and was written on absolute values: the deviation of the surrogate-trained model had to stay below half that of the model trained on the data. It fails by its letter at one held-out η on both realisations, and H34 is **refuted** on that clause; but what is measured there is an excess, not a deficit. Trained on Gaussian trajectories the model has $+1 \pm 1$, $+8 \pm 2$ and $+1 \pm 2$ % of magnetic energy against the reference ($+3 \pm 1$, $+10 \pm 2$ and -5 ± 2 % on the other realisation), where the model trained on the data has -16, -11 and -11 % (-12, -11, -15 %), and its seed scatter is up to five times that of the model trained on independent draws. On the data of the main run, the energy distance to the reference is 3.8, 8.8 and 4.2 floors for the model trained on correlated Gaussian trajectories, 3.4, 4.2 and 1.6 on independent Gaussian draws, and 8.7, 12.0 and 7.3 on the data (G1: 3.1, 3.3, 1.3). The deficit of magnetic energy of W9 is therefore not reproduced by Gaussian data with the time correlation of the trajectories: with few independent states the second moments of the model become less accurate, and its magnetic energy is then too high more often than too low (+8 to +15 % at five of the eight η where it is evaluated on each realisation, validation and training values included; -5 % at one). Its mean-flow production is still 0.02 to 0.16 of the dissipation below the reference at the held-out η. The second job of this test ran on a pod whose processor gave other trajectories from the same seeds: it is another realisation of the data than that of the replication, which the digest now checks for every run.

The same number of snapshots from four times as many trajectories (H36). The snapshots of one trajectory are correlated (a diagnostic of the `physicien` role on the CPU, not a result of the study, gives decorrelation times of 17 to 75 time units for the large-scale field), so the 2400 snapshots of each η hold far fewer independent states. If the deficit of W9 were an effect of that, taking the same number of snapshots from more trajectories should reduce it. A data set of 96 trajectories of 25 snapshots per η was generated (same solver, transient and sampling interval), and G0, W9 and G1 trained on it with the seeds and budget of the main run. A first data set gave a measurement invalid by the letter: one check, written without the sampling error of the reference, required the reference magnetic energy to be within 5 % of that of the main run at every η, and it is 5.1 % away at one training value. The check was amended before a second cycle on new data (it also passes there as first written), and H36 is **confirmed** on that cycle: the deficit of magnetic energy of W9 is 0.159 ± 0.006, 0.097 ± 0.004 and 0.141 ± 0.005 at the held-out η (threshold 0.09; 0.167, 0.125 and 0.139 on the first data set; 0.151, 0.119 and 0.110 with 24 trajectories). The margin is thin at the middle value, where one seed is below the threshold; what is solid is that the deficit is not halved. The third-order part, the mean-flow part, the null count and the restart drift of W9 do not change either. The number of independent states was not measured: with these decorrelation times, four times the trajectories over a quarter of the duration multiply it by one to three, not four. The deficit of second moments of the model trained on the data is thus not reduced by spreading the same number of snapshots over four times as many trajectories; it is not shown to be independent of the amount of data.

What the Gaussian law gets wrong in the null count (H32). The mean number of null points of a field is, by the Kac-Rice formula, the number of cells times the probability density of B at zero times the mean of $|\det \nabla B|$ where B vanishes. The hypothesis was that G1 has about twice the null points of the reference because its field varies faster where it is

weak (Jacobian factor at least 1.5 times that of the reference) and not because it is weak more often (density at most 1.25 times), with a clause that the excess be specific to the weak places (mean of $|\det \nabla B|$ over all the cells below 1.25 times that of the reference). The two factors were estimated with a threshold of 0.1 of the rms of $|B|$, and their product is within 4 % of the counted number for the reference and for G1. The measurement is nevertheless **invalid as registered** on both realisations: at the largest held-out η the factors of the reference change with the threshold by more than the 20 % allowed (the Jacobian factor by up to 35 %, 53 % on the second realisation; the density at zero by 24 % and 34 %). The clause without a threshold goes against the statement everywhere: over all the cells, the mean of $|\det \nabla B|$ of G1 is 1.61 ± 0.02, 1.49 ± 0.01 and 1.39 ± 0.02 times that of the reference (1.55, 1.46, 1.36 on the second realisation), so the excess is not specific to the weak places. A second measurement could only refute the hypothesis, which is dropped and reported as a negative result. What the data support is descriptive and holds at the two lower held-out η, where every factor is stable within 13 % across the thresholds: G1 has the density of the reference at zero within 4 to 13 % and 2.2 and 1.6 times its Jacobian where B is weak, for 2.1 and 1.8 times its null points; and the mean of $|\det \nabla B|$ over all the cells is 1.4 to 1.5 times that of the reference, not only where the field is weak. G1 has the spectrum of the reference, and so its rms gradient, by construction (the ratio of the Taylor wavenumbers is within 0.7 % of one): its field does not vary faster. The determinant of its gradient is larger: the gradient tensor of the reference is closer to singular than that of a Gaussian field with the same covariance, a property of its higher-order statistics that the second moments do not fix, and a target that a training signal could be given.

A short relaxation in the solver (H35). The `physicien` role proposed a reading of this excess: a Gaussian law draws the velocity and the magnetic field independently (their covariance vanishes by the symmetry $B \to -B$), whereas the flow aligns the small scales of the field within a fraction of a time unit; the excess of null points would then be a fast transient of the dynamics. The test: advance 512 states of each law in the solver of the data, read them at $\tau = 1$, 2, 4 and 8, and evaluate each relaxed ensemble like any generated ensemble, with reference snapshots sent through the same code as a control (Figure 5.10). The hypothesis had three clauses: the null-count error of G1 and of the recoloured model W16 falls below 0.6 of its initial value within one time unit; W16 relaxed by 4 time units is inside the five thresholds of the pilot that a relaxation could help or break (null count, energy residual, spectrum and helicity, energy distance, restart drift); and the residual relaxation time τ^* of W16 (the first τ with a null-count error of at most 0.2 and a magnetic energy within 5 %) is shorter than that of G1.

The measurement had to be repaired on the way, and the repair is part of the result. On the data of the main run, a first set of jobs failed a check for a reason found in the reference: the null count of the reference was taken on 400 snapshots drawn at random among 2400, and two draws differed by 4 and 10 % at the two lower held-out η. A second set of jobs used the draw of the earlier job it was compared with; the jury then pointed out that this draw was the higher one and that the outcome was computable in advance, so the first realisation does not count as a test. The reference was finally counted on all its snapshots, on the pods (21.1, 27.1 and 44.0 null points at the held-out η of the main run, with standard errors over the trajectories of 3 to 5 %; 22.0, 28.1 and 44.1 on the second realisation): the draw first used for the comparison was 7 % high at the smallest η. The second realisation, run once with the protocol fixed, is the test, and H35 is **confirmed** on it: the null-count error at $\tau = 1$ is at most 0.58 of its initial value (0.46 for W16); W16 relaxed by 4 time units has a null-count error of 0.070, an energy residual of 0.003, a spectrum-plus-helicity error of 0.117, an energy distance of 1.8 floors and a restart drift of 0.017; and τ^* is 4 time units for W16 against 8 for G1 in every seed and for W9 on the seed means (4 in one seed) (none on the grid for G0). The first realisation agrees (errors of 0.05 to 0.10, 0.002, 0.05, 1.6 floors and 0.02; τ^* of 2 or 4 against 8, depending on the reference count). Against the count on all the snapshots the three clauses read the same on every run.

What this establishes is narrower than the clauses suggest, and the two reviews set the limits. The second clause is a statement about the solver acting on states with the right second moments, not about the learned model: G1 relaxed by 4 time units is inside the five thresholds too on the first realisation, and misses only the spectrum on the second (0.17). What the recoloured model adds at $\tau = 4$ is quantitative: 1.6 to 1.8 floors against 4.6 to 5.1, an error of magnetic energy of 2 to 4 % against 5 to 7 %, and a relaxation time half as long or less. The first clause is marginal for G1 at the smallest η (0.57 to 0.59, two seeds of five above 0.6), inside the uncertainty of the reference count. And the relaxation is not monotone: the energy distance, the magnetic energy and the energy residual first degrade, with a maximum at $\tau = 1$ (W16: 2.1, 6.8, 4.5, 1.6 and 1.1 floors at $\tau = 0$, 1, 2, 4, 8; G1: 2.6, 15, 10, 4.6, 2.5; G0: 3.5, 60, 43, 23, 4.3), so a relaxation shorter than a few time units makes a law worse on those metrics while it repairs its topology. Within these limits: from states with the second moments of the data, the solver brings the null count to within 0.05 to 0.13 of that of the reference in 4 time units; the mean of $|\det \nabla B|$ is back to that of the reference by $\tau = 2$. The baseline for these times was measured afterwards (H39, **confirmed**): five ensembles of 512 trajectories started from the initial condition of the data (a laminar flow with a perturbation and a seed field of amplitude 0.1) first meet the same two conditions at 64 time units, in every ensemble, and not at 32 (null-count error 0.19, magnetic energy 6 to 7 % off; energy distance 23 floors at 16 time units, 3.6 at 32, 0.7 at 64). Under the same conditions, thresholds and number of states, the recoloured model followed by the solver qualifies at 2 to 4 time units: at least 8 times shorter on these grids. The transient of a simulation depends on its initial condition, and the comparison leaves out the cost of the training data and of the training. A generator followed by a few time units of the solver is the first construction of this report that is within the thresholds on the null count. It is not shown that the relaxed ensemble samples the stationary law (the large-scale field comes from the generator and does not decorrelate in that time), and the mechanism proposed by the role is not tested.

B Longer trainings and wider networks

Would the models with an affine term overtake the Gaussian law with more training? The two best models were retrained with eight times the iterations (W9L, W11L; Table B.1, Figure B.1), in a separate job whose base models reproduce those of the main run bit for bit.

The answer is no, and the result is a negative one. The flow-matching loss on training samples goes down (from 1.18 to 1.08), but the generated law moves away from the reference: 18.8 and 19.5 floors on the held-out η against 9.3 and 8.6, and 12 against 5.5 on the training η themselves. The effect is present in every seed on the mean over the held-out η; it is largest at the smallest η and not separable at the largest one for W11L. The longer training improves the topology (error on the null count 1.4 against 1.7) and keeps the magnetic energy balance (residual 0.05), while the residual of the total energy balance nearly doubles and the restart drift of the kinetic energy rises; the restart drift of the magnetic energy is not separable from that of the base models. Why a lower training loss gives a worse law was not established; a decrease of the loss that sharpens the velocity field beyond what 50 integration steps resolve is one **conjecture**, memorisation of the 14 400 training snapshots another.

The hypothesis registered for this run (H18) was narrower: that with the affine term and a long training, flow matching holds what a Gaussian law cannot, a lower restart drift of the magnetic energy and a lower error on the null count than G0. Both paired differences are positive in the five seeds (0.070 ± 0.011 and 0.86 ± 0.03): H18 is **confirmed** as registered. The credit does not go to the long training, though: the models at the base budget already meet both clauses, and the long training does not improve the drift. Nor does the first clause mean what the hypothesis says: the drift relative to the initial energy is lower for the learned models than for G0 partly because of their second moments (the Gaussian law fitted to the samples of W9 is at 0.170 against 0.212 for G0) and partly because of their non-Gaussian part (0.124 for W9); and a Gaussian law, G1, is lower still (0.064), so the clause does not measure what a Gaussian law cannot hold (Section 5.5).

Table B.1: Long trainings (measured; `experiments/results/replications/e4_final_long`, five seeds, held-out η unless said otherwise): Gaussian baseline, models with the affine term or the anchor at the base budget and with 8 times the iterations. Columns and notation as in Table 6.1; FM loss: flow-matching loss on a training sample ($\pi/2 = 1.571$ for the exact field of independent standard Gaussians).

Variant	Acts through	ED / floor, training η	ED / floor	div B	energy residual	magnetic balance	spectrum of B	magnetic helicity	null points	restart drift E_B	restart drift E_u	polarity	FM loss	Seeds
G0 Gaussian per mode (no network)	family	2.67(8)	3.52(7)	1.5×10^{-8}	0.0015(3)	0.9981(9)	0.024(3)	0.035(5)	2.235(7)	0.212(3)	0.026(5)	0.018(3)	–	5
W9 mode-wise affine term	design	5.51(9)	9.3(3)	1.5×10^{-8}	0.0567(17)	0.0553(6)	0.145(3)	0.196(4)	1.665(14)	0.124(3)	0.053(6)	0.019(4)	1.1760(15)	5
W9L affine term, 8× iterations	design	11.91(15)	18.8(14)	1.5×10^{-8}	0.096(5)	0.049(4)	0.112(15)	0.12(2)	1.38(3)	0.142(10)	0.077(11)	0.0147(14)	1.075(2)	5
W10 anchored on the Gaussian field	design	36.37(10)	43.6(2)	1.5×10^{-8}	0.1458(9)	0.3092(17)	0.2081(19)	0.096(7)	2.464(16)	0.122(2)	0.092(4)	0.0135(9)	1.2727(18)	5
W11 anchor + affine term	design	5.52(13)	8.6(2)	1.5×10^{-8}	0.0515(9)	0.0598(6)	0.143(4)	0.193(5)	1.696(12)	0.120(2)	0.053(3)	0.0169(10)	1.1742(14)	5
W11L anchor + affine, 8× iterations	design	12.0(5)	19.5(15)	1.5×10^{-8}	0.098(8)	0.043(3)	0.128(14)	0.14(2)	1.41(3)	0.154(17)	0.081(13)	0.018(5)	1.075(2)	5

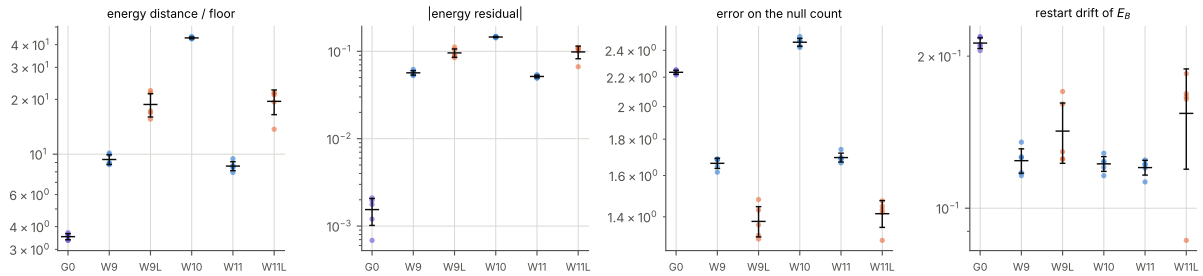

Figure B.1: Gaussian baseline (violet), flow matching with the explicit affine term or the Gaussian anchor at the base budget (blue) and with 8 times the iterations (orange), on the held-out η: one dot per seed, seed mean ± 2 standard errors. Data plot from `experiments/results/replications/e4_final_long/`.

Is the failure of the plain networks a matter of width, as the rank argument of Section 3.5 suggests (H21)? The equivariant network with the smooth embedding and no affine term was trained at three widths, 512, 2048 and 5632 (the last one above the latent dimension of 5320), and at 8 times the iterations for the first two (Table B.2, Figure B.2). The five seeds come from a second scan, split over three GPU jobs whose reference data are identical; a first scan on a single GPU was cut by its time limit and agrees with it on the seeds it has.

- **Width 512** is the matched control of W9, the same network without the affine term: 205 floors, with a final loss of 1.61, above $\pi/2$. The affine term is worth a factor of 22.
- **Width 2048** is at 25.6 floors, with a loss of 1.22: eight times closer. Capacity matters. This network is farther from the reference than W9 in energy distance but better on the total energy residual (0.012 against 0.057); its restart drift of the magnetic energy at $\tau = 4$ is lower (0.072 against 0.124), which is not an improvement: anchored on the reference energy, it starts 10 % below and is 3 % below at $\tau = 4$, a slower recovery than W9 from a smaller deficit.
- **Width 5632**, above the latent dimension, reaches a much lower loss (0.78) and a worse law: 104 floors. The order registered for H21 (the wider the closer, and the widest at less than a fifth of the narrowest) does not hold, in any seed: H21 is **refuted**. This network is not uniformly worse, though: its null-count error (0.96) is lower than those of W9 (1.66) and of width 2048 (2.01), its magnetic balance residual (0.037) is the lowest of the networks, and its restart drift of the magnetic energy (0.45) the largest.
- **More training does not help.** At width 512 the loss falls to 1.52, now below $\pi/2$, and the law is slightly worse (215 floors, $+10 \pm 1.6$); at width 2048 the loss falls to 0.95 and the distance more than doubles (59 floors against 25.6), the damage being concentrated at the smallest η.

So the width is part of the story and not the whole of it. For these plain networks the training loss and the distance of the generated law to the reference part ways: beyond width 2048 at the base budget, a lower loss on the training samples comes with a worse law, at the training η as well as between them. With 14 400 training snapshots in 5320 dimensions, fitting the training set more closely than the affine model does (its loss is 1.18) is not learning the law; that this is memorisation of the snapshots is a **conjecture**, not a measurement. In energy distance, the explicit affine term at width 512 does better than any width tried; on other metrics the ranking of the networks is not the same.

Table B.2: Width scan (measured; `experiments/results/replications/e4_scan2`, five seeds on three GPU jobs with the same data, held-out η): the equivariant potentials model with the smooth embedding and no affine term, for three widths and two training budgets, against the model with the affine term and the mode-wise Gaussian law. Columns and notation as in Table 6.1; FM loss as in Table B.1. In these result files the field `cond_scale` of the configuration is the default of the run (8); the builders of these variants and of W9 fix the embedding of the condition to 0.5.

Variant	Acts through	ED / floor, training η	ED / floor	div B	energy residual	magnetic balance	spectrum of B	magnetic helicity	null points	restart drift E_B	restart drift E_u	polarity	FM loss	Seeds
S width 512	family	165.8(4)	205.3(5)	1.4×10^{-8}	0.2737(9)	0.3284(19)	0.663(5)	0.184(6)	2.116(11)	0.3725(16)	0.268(3)	0.013(2)	1.613(3)	5
S width 2048	family	25.8(2)	25.6(4)	1.5×10^{-8}	0.0117(8)	0.0951(3)	0.151(2)	0.286(4)	2.007(19)	0.0719(14)	0.075(4)	0.011(2)	1.222(2)	5
S width 5632	family	70(5)	104(8)	1.5×10^{-8}	0.274(16)	0.037(5)	0.34(3)	0.37(3)	0.96(5)	0.45(4)	0.27(3)	0.022(4)	0.777(3)	5
S width 512, 8× iterations	family	173.4(11)	215.5(19)	1.4×10^{-8}	0.260(2)	0.400(5)	0.563(10)	0.349(14)	2.202(17)	0.366(5)	0.223(5)	0.010(2)	1.5247(14)	5
S width 2048, 8× iterations	family	110(11)	59(3)	1.5×10^{-8}	0.150(7)	0.138(6)	0.291(11)	0.441(12)	2.14(3)	0.264(10)	0.30(2)	0.026(3)	0.949(3)	5
W9 mode-wise affine term	design	5.47(12)	9.3(3)	1.5×10^{-8}	0.0566(17)	0.0551(4)	0.144(3)	0.195(4)	1.669(16)	0.124(3)	0.052(6)	0.016(3)	1.1773(10)	5
G0 Gaussian per mode (no network)	family	2.67(8)	3.52(7)	1.5×10^{-8}	0.0015(3)	0.9981(9)	0.024(3)	0.035(5)	2.235(7)	0.212(3)	0.026(5)	0.018(3)	–	5

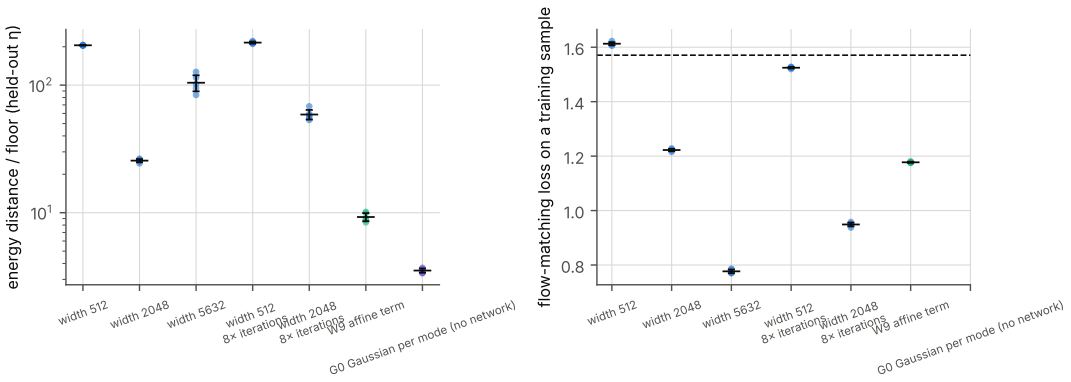

Figure B.2: Width scan (hypothesis H21): energy-distance ratio and flow-matching loss for three widths and two budgets, against the model with the explicit affine term and the Gaussian baseline. Dashed line: $\pi/2$. Data plot from `experiments/results/replications/e4_scan2/`.

C Replication with other data

The five-seed runs made on pods of the same type reproduce one another to the last digits, because the data and model seeds are the same: they check the reproducibility of the pipeline, not the robustness of the conclusions. For that, the whole protocol (groups A to E, five seeds) was run once more on reference data regenerated with other seeds of the initial conditions (Table C.1). The statistics of the two reference sets agree within their standard errors. The ratios of the replication differ from those of the main run by −18 % to +12 %; part of this comes from the noise floors (they differ by a factor of 1.6 at one held-out value) and part from the distances themselves, which are 8 to 13 % lower for the three best laws.

Every conclusion of this section is unchanged in the replication, and so is every verdict recomputed on it by the `encadrant`: the first-generation models are between 110 and 490 floors from the reference, in the same order (the only inversions are among the half-sample and tilted ensembles); the affine term brings the equivariant model from 191 to 8.2 floors (a factor of 23); the mode-wise Gaussian law is closer (2.9) and has no magnetic production; the anchored model without correction stays at 43 floors; the null-count errors and restart drifts keep their order.

Table C.1: Main run ("main") against a replication in which the reference data were regenerated with other seeds of the initial conditions ("other": `E4_DATA_SEED=100`; `experiments/results/replications/e4_indep`); five seeds each, held-out η, notation of Table 6.1.

Variant	ED / floor, main	ED / floor, other	magnetic balance, main	magnetic balance, other	null points, main	null points, other	restart drift E_B, main	restart drift E_B, other
W0 free components	236(6)	221(8)	0.159(15)	0.162(12)	1.11(5)	0.98(7)	0.513(9)	0.520(10)
W1 divergence penalty	464(12)	420(7)	0.410(5)	0.395(6)	0.698(12)	0.686(7)	0.5008(14)	0.4995(20)
W2 final projection	179(8)	169(9)	0.159(15)	0.162(12)	0.91(5)	0.79(8)	0.443(14)	0.453(16)
W3 projected flow	177(7)	164(8)	0.161(12)	0.165(13)	0.90(7)	0.82(8)	0.448(17)	0.452(19)
W4 vector potentials	256(7)	237(6)	0.149(14)	0.153(18)	1.14(7)	1.05(8)	0.517(10)	0.517(10)
W4s potentials, B → -B equivariant	209.8(19)	191(3)	0.264(14)	0.256(12)	1.83(7)	1.68(6)	0.382(12)	0.385(12)
W5 stationarity loss	215.0(19)	196.7(8)	0.516(13)	0.516(8)	2.36(2)	2.209(17)	0.334(3)	0.334(2)
W5 control (same fine-tuning, no identity)	210(2)	192(3)	0.251(15)	0.244(12)	1.79(7)	1.64(6)	0.390(12)	0.392(11)
W6 diffusion (VP, v-prediction)	519(17)	491(10)	0.213(11)	0.212(12)	0.43(2)	0.456(19)	0.637(2)	0.632(2)
W7 exponential tilting	195(11)	214(6)	0.43(4)	0.41(4)	2.24(19)	2.05(6)	–	–
W8 per-sample balance rescaling	117(2)	110(3)	0.195(19)	0.191(17)	0.71(8)	0.63(7)	–	–
W4c potentials, smooth condition	215.4(3)	205.8(7)	0.3843(16)	0.387(3)	2.142(12)	2.018(12)	0.3826(11)	0.3807(16)
G0 Gaussian per mode (no network)	3.52(7)	2.90(5)	0.9981(9)	0.9984(8)	2.235(7)	2.102(7)	0.212(3)	0.210(3)
W9 mode-wise affine term	9.3(3)	8.20(9)	0.0553(6)	0.0506(6)	1.665(14)	1.551(8)	0.124(3)	0.122(4)

Variant	ED / floor, main	ED / floor, other	magnetic balance, main	magnetic balance, other	null points, main	null points, other	restart drift E_B, main	restart drift E_B, other
W10 anchored on the Gaussian field	43.6(2)	43.3(2)	0.3092(17)	0.3122(19)	2.464(16)	2.323(15)	0.122(2)	0.125(4)
W11 anchor + affine term	8.6(2)	7.28(13)	0.0598(6)	0.0547(7)	1.696(12)	1.565(8)	0.120(2)	0.121(3)

D Objectives of the race

The objectives of the team were quantified on the first day, before any model was trained (`race/teams/claude/objectives.toml`), and are measured by `experiments/race_metrics.py` on one generator, chosen on the validation values (Table D.1).

Generator measured by C1–C5, T1 and T2: `W9_skip_sym`, the learned variant of the main run with the lowest energy-distance ratio on the validation η (file `experiments/results/e4_selection.json`).

Table D.1: Objectives of the race against the measured values (generated by running the command of each objective).

Objective	Target	Measured	Seeds	Met
C1-3d-generator	≤ 10	9.35	5	yes
C2-divergence	$\leq 1.0 \times 10^{-6}$	1.6×10^{-8}	5	yes
C3-energy-balance	≤ 0.05	0.0567	5	no
C4-spectra-helicity	≤ 0.15	0.341	5	no
C5-topology	≤ 0.2	1.66	5	no
C6-comparative-study	≥ 6	32	5	yes
C7-toolbox	≥ 6	16	0	yes
T1-restart-drift	≤ 0.05	0.124	5	no
T2-polarity-symmetry	≤ 0.03	0.0186	5	yes

Five of the nine objectives are met: the distance to the reference law (C1, below ten floors), the divergence (C2), the number of decided comparisons (C6), the toolbox (C7) and the polarity symmetry (T2) reach their targets; the energy balance (C3) misses its target narrowly; the spectra and helicity (C4), the topology (C5) and the restart drift (T1) miss theirs by factors of two to eight. The targets were set before any model existed and were not changed. Three caveats on this table. The selection between W9 and W11 on the validation values is inside the seed noise (paired difference -0.29 ± 0.60): the measured generator could as well have been W11, with the same outcome for every objective. The objective C1 is met inside the uncertainty of the floor: the ratio is 9.35 with the floor of the protocol and 9.96 with the floor averaged over 20 splits, and one seed of the five is above 10 with the first floor, two with the second. And the selected generator is the best *learned* one: the Gaussian baselines, which have lower ratios on the validation values (2.5 for G1 and 4.0 for G0, against 8.9 for W9), are not candidates, because the objective written on the first day names the model family (conditional flow matching on Fourier modes); the line of code that excludes them was written after the results of G0 were known. G0 would meet seven of the nine objectives, missing the topology and the restart drift. A fourth caveat concerns T1 itself: the drift relative to the ensemble's own initial energy, read at one time, is a poor measure of what it was written for. Its value for W9 (0.124) is almost entirely

the recovery of a 12 % deficit of magnetic energy, and a law with the right energy and the same rise would have the same drift, away from the reference instead of towards it: the number does not say which (Section 5.5). Two constructions made after the selection are reported beside this table and are not the selected generator (Section 5.6). The samples of W9 recoloured onto the moments fitted on the data (W16) have 2.07 floors (C1), an energy residual of 0.006 (C3), a spectrum-plus-helicity error of 0.040 (C4), a restart drift of 0.021 (T1) and a polarity asymmetry of 0.011 (T2), and miss the topology (C5: 1.00); C3 and C4 are properties of the copied moments, which the Gaussian law G1 has too. Followed by 4 time units of the solver, the same samples have a null-count error of 0.05 to 0.10 (0.07 on the second data realisation), inside the target of C5, with the other four still inside theirs; the Gaussian law relaxed the same way is inside the five on one data realisation and misses the spectrum on the other.

The recoloured model against the objectives. Measured with the definitions of the objectives of the pilot, W16 has an energy distance of 2.07 floors (C1, target 10), an energy residual of 0.006 (C3, target 0.05), a spectrum-plus-helicity error of 0.040 (C4, target 0.15), a restart drift of 0.021 (T1, target 0.05; on 512 states) and a polarity asymmetry of 0.011 (T2, target 0.03), and misses the null count (C5: 1.00 for a target of 0.2); the second realisation gives 1.61, 0.005, 0.058, 0.021, 0.017 and 0.95. It is not the selected generator of the pilot, which was chosen on the validation η among the learned models of the first protocol before this construction existed; W15 meets the same five thresholds and G1 misses the restart drift (0.065).

E Register of the hypotheses

Forty-one hypotheses are in the register, thirty-eight of them in this experiment (H4 to H41), all decided: six were registered after external reviews of the first and third versions of this report, seven (H28 to H34) after the reviews of the fourth, two (H35 and H36) from a consultation of the `physicien` role on the results of those seven, two (H37 and H38) are controls asked by the `jury-externe` role, H39 is the baseline that H35 lacked, H40 follows from H38, and H41 is a control of the code shared by the models of H28, H38 and H40, asked by the `encadrant` role.

Table E.1: Hypotheses of E4, registered with a refutation criterion before the computation (`research/teams/claude/hypotheses.json`), with their status after the review of the `encadrant` role.

Id	Hypothesis (shortened)	Status	Verdict of the `encadrant` role (beginning)
H4	In E4 (3D), generating vector potentials (W4_potentials, constraint by model design) gives a lower energy-distance ratio on the held-out η than generating free components (W0_free).	refuted	REFUTED (encadrant, 2026-11-12): W0_free - W4_potentials = -20.05 ± 3.21 (-6.2 SE), W4 below W0 in 0 of 5 seeds, negative at each held-out η; the two CPU seeds agree (-26.9).
H5	In E4, the divergence penalty on the end-of-path prediction (W1_penalty, weight 10; constraint by a training loss) does not enforce the constraint: […]	supported	CONFIRMED (encadrant, 2026-11-12): relative divergence of B 0.206 against 0.380 (ratio 0.544, not below one tenth); W0_free - W1_penalty = -228.2 ± 12.7, 0 of 5: the penalty does not enforce the constraint and degrades the law.
H6	In E4, the projected flow (W3_projected_flow: projected noise and projection of every integration stage) gives a lower energy-distance ratio on the held-out η than the Leray projection of the final samples of the free model (W2_final_projection).	refuted	REFUTED (encadrant, 2026-11-12): W2_final_projection - W3_projected_flow = 2.69 ± 3.72 (0.72 SE), 4 of 5 seeds: no difference at 2 SE; the CPU seeds show the same absence of effect.
H7	In E4, fine-tuning the equivariant potentials model with the law-level stationarity loss (W5_stationarity: energy identity and magnetic-helicity identity $E[\langle j \cdot B \rangle] = 0$, imposed at every η of the range) lowers the absolute residual of the mean energy balance on […]	supported	CONFIRMED (encadrant, 2026-11-12): |energy residual| W5_control - W5_stationarity = 0.0988 ± 0.0108 (9.1 SE), 5 of 5 seeds (0.290 → 0.191); ED ratio W5/control = 1.023.
H8	In E4, imposing the stationarity identities at training time (W5_stationarity) gives a lower energy-distance ratio on the held-out η than imposing the same identities at sampling time by exponential tilting of the samples of the control model (W7_tilting: […]	dropped	NOT DISCRIMINATING (encadrant, 2026-11-12): the registered criterion is formally met (W7_tilting - W5_stationarity_half = 30.0 ± 10.1, 5 of 5), but ESS/N is 1/256 in 14 of 15 cells: the tilted law is one sample and its energy distance is not a measurement of a law; with an estimate of the extra bias of a weighted sample the sign flips (-13.9 ± 9.5). […]

Id	Hypothesis (shortened)	Status	Verdict of the `encadrant` role (beginning)
H9	In E4, the B → -B equivariant velocity field (W4s_potentials_sym, constraint by model design) lowers the polarity asymmetry |P(⟨u·B⟩ > 0) - 1/2| of the generated law on the held-out η compared with the same model without symmetrization (W4_potentials), without […]	refuted	REFUTED (encadrant, 2026-11-12) on its first clause: asymmetry W4 - W4s = 0.0004 ± 0.0058, 3 of 5 seeds, both at or below the binomial floor 0.0176: the test could not discriminate at n = 512. The second clause holds: ED ratio W4s/W4 = 0.819 (-46.5 ± 4.8, 0 of 5).
H10	In E4, conditional flow matching (W4_potentials: linear path, Heun, 100 network evaluations) gives a lower energy-distance ratio on the held-out η than a variance-preserving diffusion with ancestral sampling (W6_diffusion: […]	supported	CONFIRMED as registered (encadrant, 2026-11-12): W6_diffusion - W4_potentials = 262.4 ± 12.1, 5 of 5, positive at each held-out η. Caveat of the encadrant: at the training η the diffusion model is better than flow matching (-15.06 ± 0.90, 0 of 5): the result is about interpolation in η with the aliased condition embedding, not about the model family at fixed η.
H11	In the E4 reference data, the magnetic helicity spectrum is bipolar at every η: H(k = 1) < 0 and Σ_{k ≥ 2} H(k) > 0, with a large-scale energy fraction E_B(k = 1)/E_B between 0.2 and 0.8, so that the stationarity identity E[⟨j·B⟩] = Σ k² H(k) = 0 is a […]	refuted	REFUTED (encadrant, 2026-11-12): recomputed from the raw snapshots at the 11 η, independently of the experiment code, every value equal to e4_reference.json. H(1) < 0 at 11 of 11 η (z from -19 to -58) and E_B(1)/E_B in [0.370, 0.438] hold; Σ_{k ≥ 2} H(k) > 0 by 2 SE fails at η = 0.0365 and 0.03309 (z = -1.4 and -1.2, sign undetermined) and at η = 0.03 (z = -5.5, negative). […]
H12	In E4, the magnetic-helicity identity E[⟨j·B⟩] = 0 constrains the generated magnetic law more than the energy identity: on the samples of W4s_potentials_sym at the held-out η its normalized residual is at least twice that of the energy identity, and tilting […]	refuted	REFUTED (encadrant, 2026-11-12): seed mean of r_b/r_a = 0.504, r_b > r_a in 0 of 5 seeds. The tilting clause is met (|err H_M| tilt a - tilt b = 0.238 ± 0.095, 5 of 5) against a degenerate comparator (tilt on the energy identity: ESS/N 0.0096); tilt b against no tilt gains 0.110 ± 0.049; the two CPU seeds do not agree in sign.
H13	In E4, the error of a generator on the number of magnetic null points follows its error on the magnetic Taylor wavenumber k_λ = (⟨|j|²⟩/⟨|B|²⟩)^{}1/2 as N ∝ k_λ³: […]	refuted	REFUTED on the slope (encadrant, 2026-11-12): slope 4.875, R² 0.902 over 135 points (seed-bootstrap interval [4.83, 4.92]); CPU 4.96. Protocol check passed (reference nulls change by 0.6 to 3.6 % between 48^3 and 64^3).
H14	In E4, the relative divergence of the free model W0_free is a representation error of the network and not an integration error (Proposition 2, consequence (c)): […]	supported	CONFIRMED (encadrant, 2026-11-12): relative divergence of B of W0_free 0.3792 at 50 Heun steps and 0.3788 at 200, ratio 0.998 to 1.000 per seed, above 0.1: integration error is excluded; 'representation error' is the remaining reading, not a separate measurement.
H15	In E4, the error of the generators at the held-out η comes from the embedding of the condition: the Fourier embedding of the pre-race code has frequencies up to 8 cycles over the normalized parameter range, above the Nyquist frequency of the 6 training values […]	refuted	REFUTED (encadrant, 2026-11-13): ratio of the seed-mean ED ratios W4c/W4 = 215.4/256.3 = 0.84 on the held-out η, above the 0.5 of the criterion; the paired difference ED(W4) - ED(W4c) = 40.9 ± 6.8 (6.0 SE, 5 of 5 seeds) is real, and absent at the training η (0.34 ± 0.32): the embedding of the condition explains part of the interpolation error, not the failure.

Id	Hypothesis (shortened)	Status	Verdict of the `encadrant` role (beginning)
H16	In E4, a conditional Gaussian model with one mean and one covariance per Fourier mode (G0_gaussian_modes: no network; moments estimated at the training η and interpolated linearly in log η, symmetrized under B → -B) has a lower energy-distance ratio on the […]	supported	CONFIRMED (encadrant, 2026-11-13): G0 3.52 against 177 to 519 for W0–W6 on the held-out η (W3 the lowest, 176.6); best of them per seed 153–194 against G0 3.36–3.70, 5 of 5 seeds.
H17	In E4, the over-dispersion of the flow-matching generators comes from the bottleneck of the network (width 512 for 5320 latent dimensions), which cannot represent the linear part of the velocity field: […]	supported	CONFIRMED (encadrant, 2026-11-13): ratio W9/W4s = 9.35/209.8 = 0.045, below 0.2; ED(W4s) - ED(W9) = 200.4 ± 2.2 (91 SE), 5 of 5 seeds; flow-matching loss of W9 on a training sample 1.18, below $\pi/2$. The factor is 22.4 at the held-out η and 30.2 at the training η.
H18	In E4, flow matching adds to the second-order baseline what a Gaussian law cannot hold: with the mode-wise affine term and a longer training (W9L_skip_long, 8 times the iterations of W9), the drift of the magnetic energy when the solver is restarted from the […]	supported	CONFIRMED as registered (encadrant, 2026-11-13): G0 - W9L on the held-out η, restart drift of E_B 0.070 ± 0.011 (6.4 SE, 5 of 5 seeds), null-count error 0.856 ± 0.035 (25 SE, 5 of 5). Caveats: W9 at the base budget already meets both clauses; […]
H19	In E4, a flow-matching field anchored on the exact field of the mode-wise Gaussian law, v = v_Gauss(x, t, η) + network (W10_anchored: […]	refuted	REFUTED (encadrant, 2026-11-13): ED(W10) - ED(G0) = +40.0 ± 0.26 on the held-out η (first clause not met); restart drift of E_B: G0 - W10 = 0.090 ± 0.003, 5 of 5 (met); null-count error: G0 - W10 = -0.229 ± 0.014, 0 of 5 (not met).
H20	In E4, the anchored model fails (W10_anchored on CPU seeds 0 and 1: 45 times the floor against 3.3 for G0) because the network's own linear component perturbs the second-order statistics that the fixed Gaussian field had right. […]	refuted	REFUTED (encadrant, 2026-11-13) on its first clause: ED(W11) - ED(G0) = +5.08 ± 0.22 (24 SE), 0 of 5 seeds; the two other clauses are met: restart drift of E_B G0 - W11 = 0.092 ± 0.003 (5 of 5), null-count error G0 - W11 = 0.539 ± 0.014 (5 of 5). Reported with it: […]
H21	In E4, the over-dispersion of the flow-matching generators without an explicit linear term comes from the width of the network and not from the training budget: […]	refuted	REFUTED as registered (encadrant, 2026-11-13) on its first two clauses, 5 seeds, held-out η: S_w512 205.3 ± 0.5, S_w2048 25.6 ± 0.4, S_w5632 104.4 ± 7.5 floors: the order is not monotonic (2048 - 5632 = -78.8 ± 7.6, 0 of 5 seeds) and S_w5632 is at 0.51 of S_w512, above one fifth. Third clause not triggered: S_w512_long 215.5 ± 1.9 (1.05 of S_w512). […]
H22	In E4, the failures of the mode-wise Gaussian law G0 on the magnetic energy balance, on the restart drift of the magnetic energy and on the number of null points come from the independence of its Fourier modes and not from its Gaussianity: […]	supported	CONFIRMED as registered (encadrant, 2026-11-13): |magnetic balance| of G1_gaussian_full 0.072 (per seed 0.071–0.074), below 0.2; restart drift of E_B at $\tau = 4$, W9 - G1 = 0.0598 ± 0.0034 (17.6 SE, 5 of 5 seeds); null-count error, W9 - G1 = 0.568 ± 0.020 (29 SE, 5 of 5). Reported with it: ED ratio 2.55 against 9.35 (W9) and 3.52 (G0); […]
H23	In E4, in the regime where the generator is close to the reference law (the model with the affine term, W9_skip_sym), fine-tuning with the law-level stationarity loss on the energy and magnetic-helicity identities (W12_skip_stationarity: […]	refuted	REFUTED on its second clause (encadrant, 2026-11-13): the first clause is met, |energy residual| control - stationarity = 0.0275 ± 0.0012 (22 SE, 5 of 5 seeds); the energy-distance ratio of W12_skip_stationarity is 16.37 against 9.31 for the control, a factor 1.76 where the criterion allowed 1.10. Reported with it: […]

Id	Hypothesis (shortened)	Status	Verdict of the `encadrant` role (beginning)
H24	In E4, exponential tilting on the energy and magnetic-helicity identities does not collapse when the generated law is close to the reference: […]	supported	CONFIRMED as registered (encadrant, 2026-11-13): ESS/N of W13_skip_tilting 0.172 on the held-out η (below 0.1 in 5 of the 15 cells, minimum 0.049); |energy residual| untilted half - tilted = 0.0449 ± 0.0030 (14.8 SE, 5 of 5 seeds). Reported with it: the tilted ensemble is better on the ED ratio (4.58 against 7.26; […]
H25	In E4, the flow-matching model with the affine term (W9_skip_sym) does not hold the third-order part of the mean magnetic production: […]	refuted	REFUTED (encadrant, 2026-11-13), on two realisations of the data: the third-order part of the production of W9 over D is -0.0055 ± 0.0007, -0.0817 ± 0.0010 and -0.1180 ± 0.0013 at η = 0.0725, 0.0490 and 0.0331 (same sign in 5 of 5 seeds; replication -0.0116, -0.0848, -0.1126), against +0.0542 ± 0.0078, -0.1371 ± 0.0043 and -0.2703 ± 0.0071 for the reference (jackknife): […]
H26	In E4, anchored on the stationary energy of the reference, the restart of the model with the affine term is a monotone relaxation and that of the Gaussian laws is not: […]	supported	CONFIRMED on the registered criterion (encadrant, 2026-11-13), each clause in 5 of 5 seeds on two realisations of the data: overshoot of W9 0.0045 (replication 0.0048), |a(16)| of W9 0.0102 (0.0133), overshoot of G1 0.0914 ± 0.0078 (0.1029 ± 0.0026), of G0 0.1732 (0.1986). Caveats of the review: the overshoot gives a law an allowance equal to its own |a(0)|; […]
H27	In E4, the monotone relaxation of the model with the affine term is not a property of its first two moments: the Gaussian law that has the mean and the full covariance of the samples of W9 (G2_gaussianised_W9, fitted at each η on 8192 fresh samples of the […]	refuted	REFUTED on the second realisation of the data; TEST INVALID as written on the first (encadrant, 2026-11-13). Paired difference of the overshoot G2 - W9N: -0.0051 ± 0.0005 (main data) and -0.0056 ± 0.0005 (second realisation), 0 of 5 seeds positive: the three refutation clauses are met on both. Null control |overshoot(G3) - overshoot(G1)| < 0.02: […]
H28	In E4, a flow-matching model anchored on the Gaussian law with the full covariance (W14_full_anchor: v = v_G1(x, t, η) + network, the network being zero at initialisation so that the untrained model is G1) acquires a third-order part of the magnetic […]	refuted	REFUTED by clause (a) on both realisations of the data (encadrant, 2026-11-13). At no checkpoint (250 to 4000 iterations) does the model anchored on G1 hold 0.3 of the third-order part of the reference at the validation η = 0.0403: seed means -3.2e-4 to -2.6e-4 of the reference on the data of the main run, +4.5e-3 on the second realisation; largest single seed 0.023. […]
H29	In E4, recolouring the samples of W9 at sampling time onto the first and second moments fitted on the data (those of G1: W16_recoloured_full) gives a law with the second-order accuracy of G1 that keeps the third-order magnetic production of W9, and whose […]	supported	CONFIRMED on the registered criterion: not refuted on both realisations of the data, null control and numerical-protocol check passed (encadrant, 2026-11-13). W16 (samples of W9 sent through the optimal-transport linear map from the moments of the model onto those of G1), main run; second realisation. Clause (a): […]

Id	Hypothesis (shortened)	Status	Verdict of the `encadrant` role (beginning)
H30	In E4, at equal first and second moments, the non-Gaussian part of W9 lowers the passage of the magnetic energy above the reference at restart: […]	supported	CONFIRMED on the registered criterion: not refuted on both realisations of the data, null control passed (encadrant, 2026-11-13). Excursion of E_B above the reference at restart (512 states), Gaussian draws with the moments of W9 sent through the same map (G4) minus W16, paired over the seeds, mean over the held-out η: […]
H31	In E4, the excursion of the magnetic energy when the solver restarts from draws of G1 is not an effect of the interpolation of the covariance in η: […]	supported	CONFIRMED on the registered criterion: not refuted on both realisations of the data (encadrant, 2026-11-13). G1 restarted at the three training η evaluated (η = 0.08, 0.054, 0.03), where it has the sample mean and covariance of the data and no interpolation (weights 0 or 1, checked by the review), overshoots the reference energy by 0.132, 0.084, 0.088 on the data of the main […]
H32	In E4, the Gaussian law with the covariance of the data (G1) has about twice the magnetic null points of the reference because its field varies faster where it is weak, not because it is weak more often: […]	dropped	MEASUREMENT INVALID as registered on both realisations of the data (encadrant, 2026-11-13): one cycle of rule 5; DROPPED and reported as a negative result, because a second cycle could only refute the statement. […]
H33	In E4, the third-order part of the magnetic production held by the flow-matching model with the affine term (W9: 0.60 and 0.44 of that of the reference at the two lower held-out η, H25) is learned from the data and not produced by the model on its own: […]	supported	CONFIRMED on the registered criterion, in its narrow sense: not refuted on both realisations of the data, the three checks of the measurement passed: SUPPORTED in the registered narrow sense (encadrant, 2026-11-13). Third-order part of the magnetic production over the Ohmic dissipation, seed means at the held-out η = 0.0725, 0.0490, 0.0331: […]
H34	In E4, the two properties that separate the model with the affine term trained on the data from the same model trained on independent Gaussian data (H33) are not effects of the time correlation of the 24 trajectories of the data: […]	refuted	REFUTED on clause (b) by the letter of the criterion on both realisations of the data, at η = 0.0490 only; clause (a) NOT REFUTED on both; checks of the measurement passed (encadrant, 2026-11-13). Clause (a): […]
H35	In E4, the excess of magnetic null points of states drawn from laws with the right second moments (the Gaussian law G1, and the flow-matching model with the affine term recoloured onto the moments of G1, W16) is a fast transient of the dynamics, and a short […]	supported	CONFIRMED on the registered criterion on the one realisation that counts as a test (encadrant, 2026-11-13; the jury-externe role asked that only the second realisation count). Second realisation (e4_relax_indep, 5 seeds, checks passed): (a) null-count error at τ = 1 over that at τ = 0 at most 0.58 (G1 at η = 0.0331; W16 at most 0.46); (b) W16 relaxed by 4 time units: […]
H36	In E4, the deficit of magnetic energy of the flow-matching model with the affine term trained on the data (W9: 11 to 16 % below the reference at the held-out η) is not an effect of the small number of independent states in the training set: […]	supported	CONFIRMED on the registered criterion on the second cycle alone (encadrant, 2026-11-13): with 2400 snapshots per η taken from 96 trajectories instead of 24 (e4_moretraj2, seed offset 300, 5 seeds, checks passed, also under the original check), the deficit of magnetic energy of W9 is 0.159 ± 0.006, 0.097 ± 0.004 and 0.141 ± 0.005 at the held-out η (threshold 0.09 at each; […]

Id	Hypothesis (shortened)	Status	Verdict of the `encadrant` role (beginning)
H37	In E4, what the recoloured model (W16) has that Gaussian draws with the same moments have not, a passage above the reference magnetic energy at restart about half as large (H30), comes from the training data and not from the model or the map: […]	supported	CONFIRMED on the registered criterion on both realisations of the data, the three checks passed (encadrant, 2026-11-13). Excursion of E_B above the reference at restart (512 states), the model trained on Gaussian surrogates and recoloured (W17) minus the model trained on the data and recoloured (W16), paired over the five seeds: […]
H38	In E4, the flow-matching model anchored on a full-covariance Gaussian law (the construction of H28) can learn a non-Gaussian part at this size when its training set holds one that a network of the same family represents: […]	refuted	REFUTED on the registered criterion (encadrant, 2026-11-13; data of the main run, 5 seeds, checks passed): the third-order part of the anchored model trained on 14 400 samples of W9, over that of W9, is 0.006 at $\eta = 0.0331$ and -0.002 at $\eta = 0.0490$ on the seed means (per seed -0.04 to +0.02), unchanged from iteration 0 to iteration 4000; […]
H39	In E4, the recoloured model followed by a short relaxation saves most of the transient of a simulation: started from the initial condition of the data (laminar flow with a perturbation and a seed magnetic field), an ensemble of trajectories of the solver […]	supported	CONFIRMED on the registered criterion against both references (encadrant, 2026-11-13; the same five ensembles of 512 trajectories per η evaluated against the reference of the main run and that of the second realisation: one realisation of the ensembles). […]
H40	In E4, the flow-matching model anchored on a Gaussian law learns a non-Gaussian part when the covariance of the anchor is structured (the 300 leading eigen-directions of the sample covariance at each training η plus one block per Fourier mode: […]	refuted	REFUTED on (a), (b) and (c) (encadrant, 2026-11-13; 5 seeds, checks passed on both realisations). (a) control, third-order part of the model anchored on the structured law of the samples of W9, over that of W9: 0.0015 and 0.0041 at $\eta = 0.0490$ and 0.0331 (per seed within $±$0.032; second realisation -0.003 and 0.009, supplementary). (b) trained on the data, over the reference: […]
H41	In E4, the code of the anchored models that hold no third-order production (FullGaussianAnchoredField: H28, H38, H40) is not at fault: […]	supported	CONFIRMED on the registered criterion: NOT REFUTED (encadrant, 2026-11-13; data of the main run, 5 seeds, check of the measurement passed: W10 of this job is within 5.1e-4 of its values in the run e4_nongauss). […]

F Compute

The session runs in a cloud container with four CPU cores and no GPU. The data, the references and the first two seeds were computed there; the five-seed runs were made on rented GPU pods through `tools/runpod_job.py`, which clones the pushed commit on the pod, runs the command under a time limit, brings the outputs back, deletes the pod and writes the cost in `research/teams/claude/compute-ledger.jsonl` (Figure F.1). A training of 4000 iterations takes 30 s to 2 minutes on the GPU, depending on the variant and on how many seeds share the card, against about 15 minutes on one CPU core; the evaluation (null points, restart test) dominates the run time. The costs are in Section J.

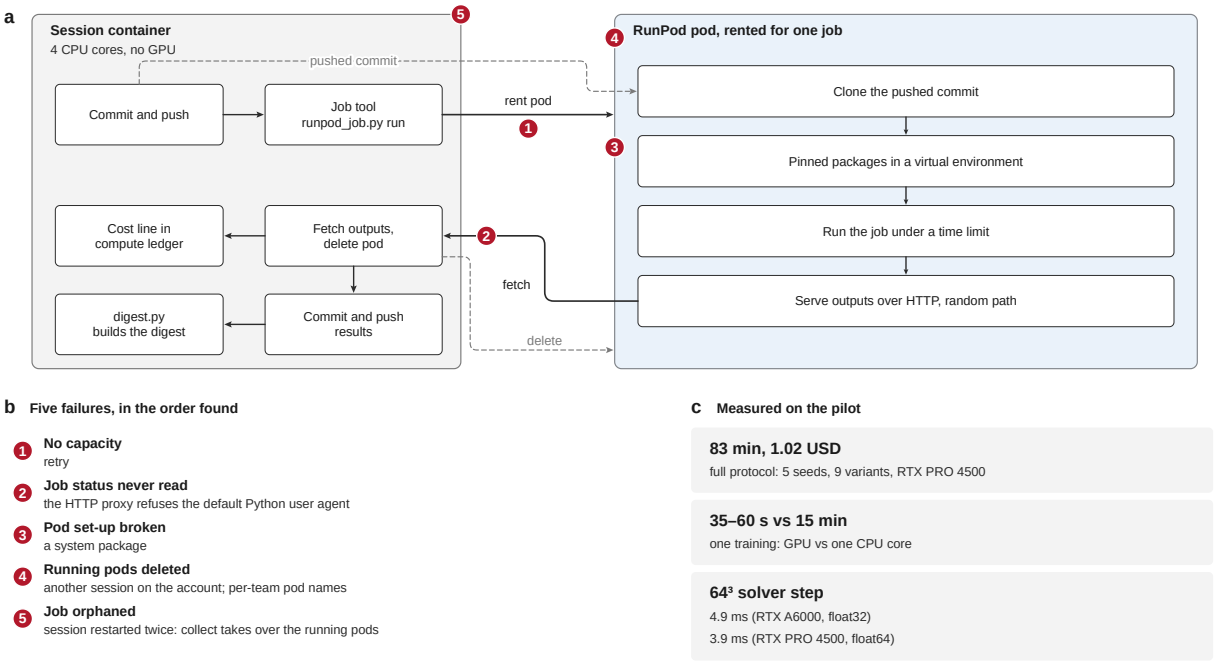

Figure F.1: The compute path of the pilot: session, GPU pod, results, ledger, and the failures met and repaired on the way. Schematic; source in `figures/compute-path/`.

G Workflow of the experiment and organisation of the agents

The workflow of the experiment and the organisation of the agents are drawn as two explorable diagrams, experiment-workflow.html and agent-organisation.html; their static exports are Figure G.1 and Figure G.2.

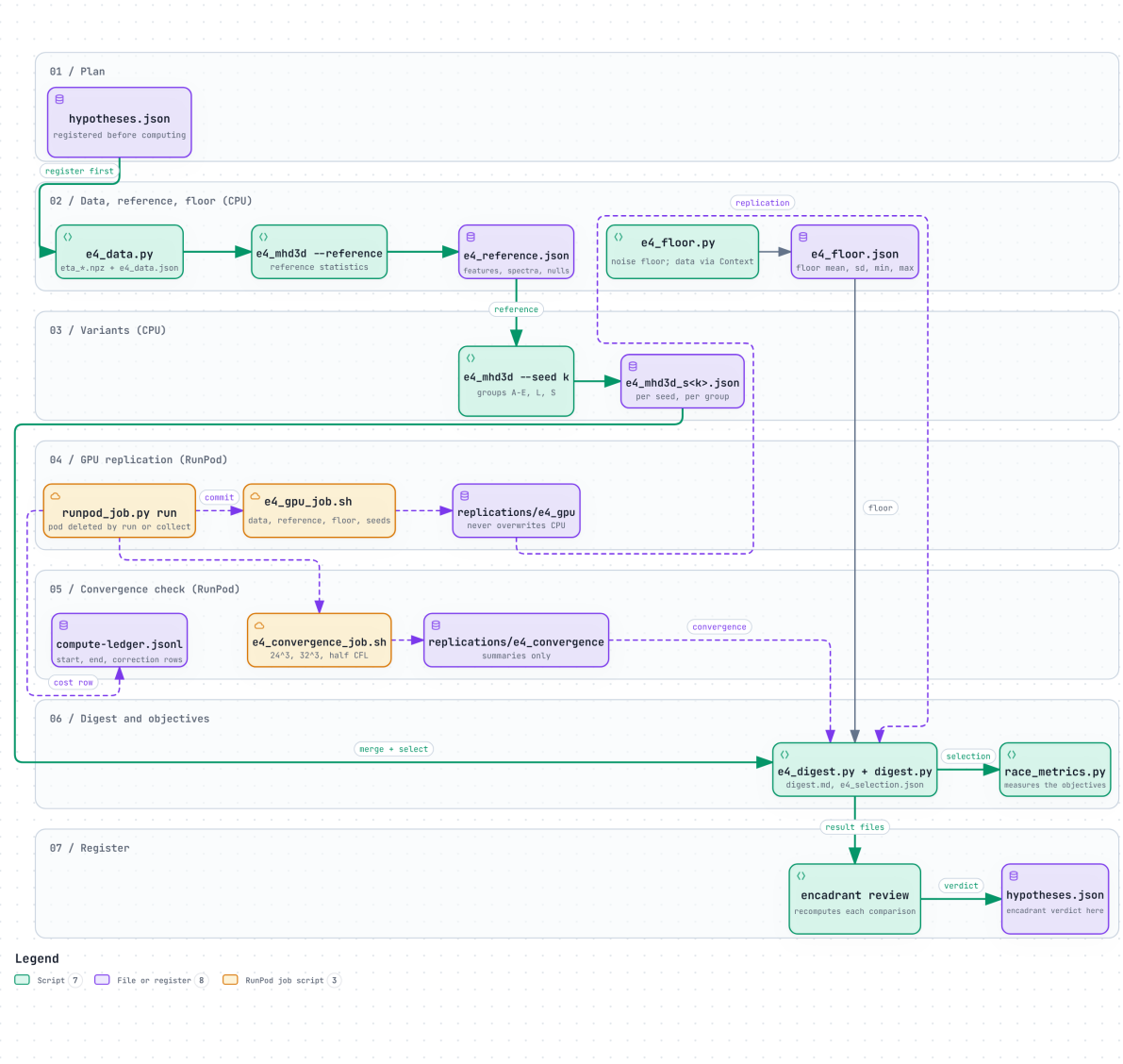

Figure G.1: Workflow of experiment E4: plan, data, reference and floors, variants, GPU runs, digest, objectives and register. Static export of the Archify diagram `diagrams/experiment-workflow.html`.

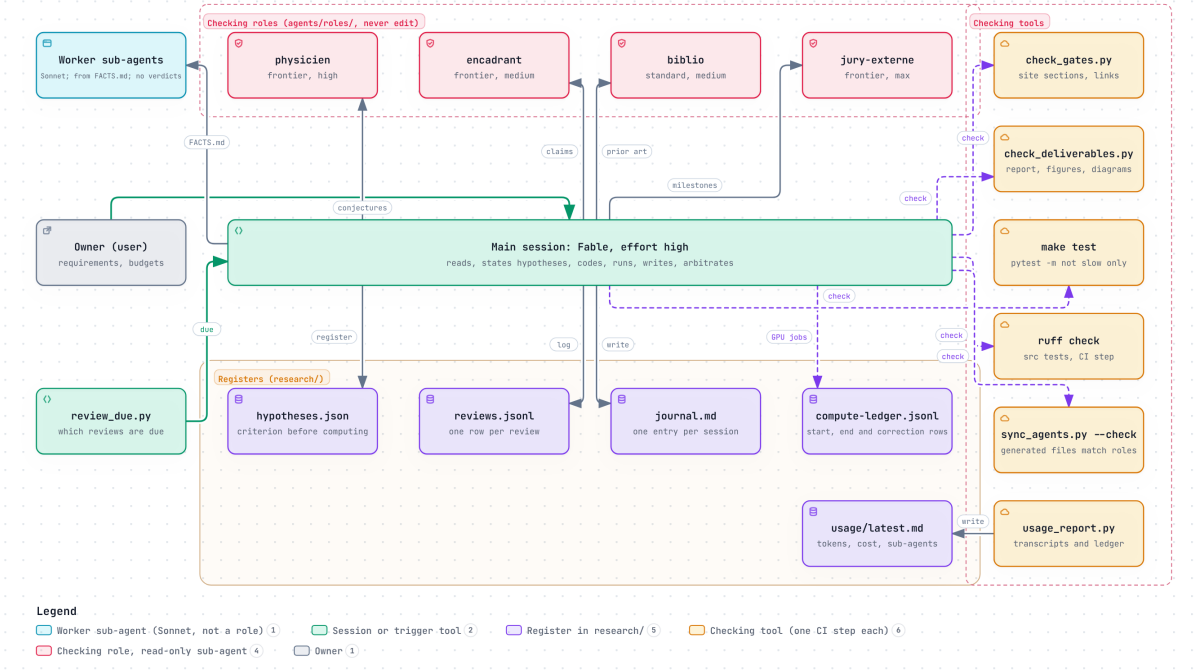

Figure G.2: Organisation of the team: the main session, the four roles with their model tier and reasoning level, the reviews named by `tools/review_due.py`, the registers and the checking tools. Static export of the Archify diagram `diagrams/agent-organisation.html`.

H History of the corrections

What the later tests refuted, and what the reviews refuted or corrected in the successive versions of this report.

The tests of Section 5.6 added four refutations and one dropped hypothesis (H38 is the positive control that the anchored model of H28 fails: trained on samples of the learned model, it learns nothing either; H40 gave the anchor a full-rank structured covariance: the network then trains, the law moves away from its anchor and no third-order production appears), and two measurements that were invalid by the letter of their checks before being repeated (H35, H36). A network anchored on the full-covariance Gaussian law does not learn the third-order production that the hypothesis expected of it (H28 refuted: it stays at the Gaussian law over the whole training). The hypothesis on the Kac-Rice factors (H32) was dropped: its measurement is invalid as registered at the largest held-out η, and its clause without a threshold goes against it everywhere (the determinant of the gradient of the Gaussian field is larger over all the cells, not specifically where the field is weak). And the second clause of H34 was refuted by its letter: the model trained on time-correlated Gaussian trajectories has no deficit of magnetic energy, as predicted, but an excess of 8 to 10 % at one held-out η, which a criterion written on absolute values counts against the hypothesis; the criterion was not amended after the run.

The interpretation of the main result was itself refuted. The first version of this report read the advantage of the flow-matching models over the mode-wise Gaussian law, on the magnetic energy balance, the null points and the restart test, as evidence that they had learned non-Gaussian, third-order structure. The `jury-externe` role showed with an oracle that a Gaussian law with the full covariance had the same advantage; the hypothesis was then registered, the baseline G1 built on the training values only and run, and the reading was replaced by the one of Section 5.4. The review that caught it is in `research/teams/claude/reviews/2026-11-13-jury-externe-report-v1.md`. The same role then showed that the restart drift of the objectives is confounded with the bias of each law on the magnetic energy (`research/teams/claude/reviews/2026-11-13-jury-externe-report-v3-and-plan.md`), and the statements of the second and third versions that ranked the laws by that drift were withdrawn (Section 5.5).

Two claims made during the work were corrected by the reviewing role and are reported as corrected. The first version of the objective on the energy distance said that a perfect generator would give a ratio of one; with a floor computed from trajectory halves it would give about 0.6, and the note was corrected before any model result existed. The statement "the cause of the failure is the width of the network" was refused as a conclusion from the first runs and registered as a hypothesis with a width scan (H21, Section B).

A run labelled as a replication was not one. The job `e4_main` reproduced the first GPU run bit for bit, because the data and model seeds were the same and the hardware identical; it is a reproducibility check. The replication with other data seeds is the one of Section C. The converse happened once: a job meant to run on the data of the replication regenerated other trajectories from the same seeds, on a pod with another processor (the solver is chaotic); the review of H34 found it, and the digest now states for every additional run whether its reference data are those of the realisation its seeds name (they are for every other GPU job).

I What the incidents of the pilot teach

Four incidents cost time and are part of the result, because the race is also a test of the method. Rented GPUs were refused for lack of capacity for about an hour. The first pod ran for 24 minutes without its status being read, because a proxy refused the default user agent of the HTTP client. Another session sharing the cloud account deleted two running pods with its clean-up command. The container of the session restarted twice and orphaned running jobs. Each was repaired in the shared tool with a test (user agent, per-team pod names, the `collect` command), and announced in the journal. One bug was in the experiment script itself: a loop variable overwrote the suffix of the result files after a certain variant, and one run brought back incomplete files; it was found by reading what a job returned instead of trusting its exit code, fixed, and covered by a static test.

J GPU jobs and costs

Every GPU job of the pilot is in `research/teams/claude/compute-ledger.jsonl` (Table J.1, Figure J.1); the pods were rented on demand and deleted at the end of each job. Jobs without an exit code or with a non-zero one are the capacity failures, the interrupted jobs, the scan cut by its time limit and the incidents described in Section I; the two negative entries are corrections made after another session deleted running pods, so that the ledger matches what was billed; one job (`e4_g1indep`) was launched twice by mistake.

Total: 82 ledger entries, 66.88 GPU hours, 59.79 USD.

Table J.1: GPU jobs of the pilot (generated from `research/teams/claude/compute-ledger.jsonl`).

End (UTC)	GPU	USD/h	Hours	USD	Exit	Commit	Command (shortened)
11-12 17:48	RTX PRO 4500 Blackwell	0.72	0.404	0.29	–	`36908cc`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 17:58	RTX PRO 4500 Blackwell	0.72	0.026	0.02	–	`40f6188`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 17:59	RTX PRO 4500 Blackwell	0.72	0.026	0.02	–	`40f6188`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 18:18	RTX PRO 4500 Blackwell	0.72	0.104	0.07	–	`d4f2135`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 18:25	RTX PRO 4500 Blackwell	0.72	0.026	0.02	0	`f099fa8`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 18:27	RTX PRO 4500 Blackwell	0.72	0.026	0.02	0	`f099fa8`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 18:40	RTX A6000	0.69	0.026	0.02	0	`47d057d`	python -m milady.sim.mhd3d bench –n 64 128 –device cuda – …
11-12 18:44	RTX PRO 4500 Blackwell	0.72	0.146	0.11	0	`8fd79a0`	e4_gpu_job.sh quick 0
11-12 19:03	RTX PRO 4500 Blackwell	0.72	0.172	0.12	0	`d0e018f`	e4_convergence_job.sh
11-12 20:16	RTX PRO 4500 Blackwell	0.72	1.415	1.02	0	`d0e018f`	e4_gpu_job.sh full “0 1 2 3 4”
11-12 22:21	RTX PRO 4500 Blackwell	0.72	0.590	0.42	–	`1c21afc`	e4_gpu_job.sh full “0 1 2 3 4” e4_gpu_de D,E
11-12 22:23	RTX PRO 4500 Blackwell	0.72	1.820	1.31	–	`a20bcb1`	e4_gpu_job.sh full “0 1 2 3 4” e4_gpu_all
11-12 22:24	RTX PRO 4500 Blackwell	0.72	0.031	0.02	–	`837c23a`	e4_gpu_job.sh full “0 1 2 3 4” e4_main all
11-12 22:24	RTX PRO 4500 Blackwell	0.72	-0.590	-0.42	–	`1c21afc`	CORRECTION of the entry of pod hd6qx71mm6jekm: the pod was d …
11-12 22:24	RTX PRO 4500 Blackwell	0.72	-0.628	-0.45	–	`a20bcb1`	CORRECTION of the entry of pod 00t6s3l481flvm: the pod was d …
11-12 23:24	A100-SXM4-80GB	2.19	0.126	0.28	–	`5027f7c`	e4_gpu_job.sh full “0 1 2 3 4” e4_scan S
11-12 23:24	RTX PRO 4500 Blackwell	0.72	0.126	0.09	–	`5027f7c`	e4_gpu_job.sh full “0 1 2 3 4” e4_long L
11-12 23:27	RTX PRO 4500 Blackwell	0.72	0.983	0.71	0	`62e31c9`	collected after a restart of the session: milady-claude-job- …

End (UTC)	GPU	USD/h	Hours	USD	Exit	Commit	Command (shortened)
11-13 00:17	RTX PRO 4500 Blackwell	0.72	0.879	0.63	0	`973fe3a`	collected after a restart of the session: milady-claude-job- …
11-13 00:55	RTX PRO 4500 Blackwell	0.72	1.508	1.09	0	`973fe3a`	collected after a restart of the session: milady-claude-job- …
11-13 01:25	GeForce RTX 4090	0.89	1.356	1.21	0	`dfc02a2`	env E4_DATA_SEED=100 e4_gpu_job.sh full “0 1 2 3 4” e4_indep …
11-13 01:27	A100-SXM4-80GB	2.19	2.034	4.46	124	`973fe3a`	collected after a restart of the session: milady-claude-job- …
11-13 02:09	RTX PRO 4500 Blackwell	0.72	0.762	0.55	0	`607ccbd`	e4_gpu_job.sh full “0 1 2 3 4” e4_mech D,E,G,M
11-13 03:01	RTX PRO 4500 Blackwell	0.72	1.636	1.18	0	`607ccbd`	e4_gpu_job.sh full “0 1” e4_scan2a D,S
11-13 03:16	RTX PRO 4500 Blackwell	0.72	1.472	1.06	0	`d7aa2bd`	e4_gpu_job.sh full “2 3” e4_scan2b D,S
11-13 03:35	RTX PRO 4500 Blackwell	0.72	0.505	0.36	0	`81a0e0c`	env E4_DATA_SEED=100 e4_gpu_job.sh full “0 1 2 3 4” e4_g1ind …
11-13 03:38	RTX PRO 4500 Blackwell	0.72	0.548	0.39	0	`81a0e0c`	env E4_DATA_SEED=100 e4_gpu_job.sh full “0 1 2 3 4” e4_g1ind …
11-13 03:38	L40S	1.09	1.096	1.19	0	`830a40a`	collected after a restart of the session: milady-claude-job- …
11-13 06:19	RTX PRO 4500 Blackwell	0.72	0.882	0.64	0	`2e87cbc`	e4_gpu_job.sh full “0 1 2 3 4” e4_nongauss D,E,G,N
11-13 06:21	RTX PRO 4500 Blackwell	0.72	0.918	0.66	0	`2e87cbc`	env E4_DATA_SEED=100 e4_gpu_job.sh full “0 1 2 3 4” e4_nonga …
11-13 08:18	RTX PRO 4500 Blackwell	0.72	0.019	0.01	–	`d4fc624`	env E4_N_RESTART=512 e4_gpu_job.sh full 0 1 2 e4_recolour_a …
11-13 08:18	RTX PRO 4500 Blackwell	0.72	0.025	0.02	–	`d4fc624`	env E4_DATA_SEED=100 e4_gpu_job.sh full 0 1 2 3 4 e4_anchor_ …
11-13 08:18	RTX PRO 4500 Blackwell	0.72	0.014	0.01	–	`d4fc624`	env E4_N_RESTART=512 e4_gpu_job.sh full 3 4 e4_recolour_b G, …
11-13 08:18	RTX PRO 4500 Blackwell	0.72	0.030	0.02	–	`d4fc624`	e4_gpu_job.sh full 0 1 2 3 4 e4_anchor G,F
11-13 09:04	RTX PRO 4500 Blackwell	0.72	0.729	0.53	0	`aeac08c`	env E4_DATA_SEED=100 e4_gpu_job.sh full ‘0 1 2 3 4’ e4_ancho …
11-13 09:05	RTX PRO 4500 Blackwell	0.72	0.743	0.54	0	`aeac08c`	e4_gpu_job.sh full ‘0 1 2 3 4’ e4_anchor G,F
11-13 10:00	RTX PRO 4500 Blackwell	0.72	0.914	0.66	–	`abac131`	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full ’3 …
11-13 10:00	RTX PRO 4500 Blackwell	0.72	0.921	0.66	–	`abac131`	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full ’0 …
11-13 10:07	RTX PRO 4500 Blackwell	0.72	1.776	1.28	124	`aeac08c`	env E4_N_RESTART=512 e4_gpu_job.sh full ‘0 1 2’ e4_recolour_ …
11-13 10:07	RTX PRO 4500 Blackwell	0.72	1.773	1.28	124	`aeac08c`	env E4_N_RESTART=512 e4_gpu_job.sh full ‘3 4’ e4_recolour_b …
11-13 10:46	GeForce RTX 4090	0.89	0.744	0.66	–	`e6df2bd`	collected after a restart of the session: milady-claude-job- …

End (UTC)	GPU	USD/h	Hours	USD	Exit	Commit	Command (shortened)
11-13 11:17	RTX PRO 4500 Blackwell	0.72	1.271	0.92	0	e6df2bd	collected after a restart of the session: milady-claude-job- …
11-13 11:52	GeForce RTX 5090	1.39	1.729	2.40	0	0722e02	collected after a restart of the session: milady-claude-job- …
11-13 12:03	RTX PRO 4500 Blackwell	0.72	1.924	1.39	124	0722e02	collected after a restart of the session: milady-claude-job- …
11-13 12:07	GeForce RTX 4090	0.89	1.344	1.20	0	e96b420	env E4_DATA_SEED=100 e4_gpu_job.sh full '0 1 2 3 4' e4_surro …
11-13 13:03	GeForce RTX 5090	1.39	1.754	2.44	0	df7109e	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full '0 …
11-13 13:13	RTX PRO 4500 Blackwell	0.72	1.153	0.83	0	c156486	env E4_N_RESTART=512 e4_gpu_job.sh full 0 e4_recolour_e0 R
11-13 13:30	GeForce RTX 4090	0.89	1.378	1.23	0	59532d5	env E4_N_RESTART=512 e4_gpu_job.sh full 1 e4_recolour_e1 R
11-13 13:31	RTX A6000	0.69	0.011	0.01	–	d5fa655	e4_gpu_job.sh full '0 1 2 3 4' e4_trajsur G,Y
11-13 13:45	GeForce RTX 4090	0.89	1.877	1.67	0	6e72225	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full '3 …
11-13 14:11	GeForce RTX 4090	0.89	0.669	0.60	0	d5fa655	e4_gpu_job.sh full '0 1 2 3 4' e4_trajsur G,Y
11-13 14:13	RTX PRO 4500 Blackwell	0.72	1.005	0.72	0	37a7af4	env E4_N_RESTART=512 e4_gpu_job.sh full 2 e4_recolour_e2 R
11-13 14:17	GeForce RTX 4090	0.89	1.241	1.10	0	092af2e	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full 2 e …
11-13 14:35	RTX A6000	0.69	0.821	0.57	0	1e00f23	env E4_DATA_SEED=100 e4_gpu_job.sh full '0 1 2 3 4' e4_trajs …
11-13 14:37	GeForce RTX 4090	0.89	0.319	0.28	–	d698812	e4_gpu_job.sh full 4 e4_relax_c X
11-13 15:09	L40S	1.09	0.919	1.00	0	97e2a25	e4_gpu_job.sh full '2 3' e4_relax_b X
11-13 15:10	RTX PRO 4500 Blackwell	0.72	0.940	0.68	0	97e2a25	e4_gpu_job.sh full '0 1' e4_relax_a X
11-13 15:11	RTX PRO 4500 Blackwell	0.72	0.022	0.02	–	461e998	env E4_DATA_SEED=100 e4_gpu_job.sh full '0 1 2' e4_relax_ind …
11-13 15:30	RTX PRO 4500 Blackwell	0.72	0.704	0.51	0	c09fedf	e4_gpu_job.sh full 4 e4_relax_c X
11-13 15:48	RTX PRO 4500 Blackwell	0.72	1.002	0.72	0	c09fedf	env E4_DATA_SEED=200 E4_N_TRAJ=96 E4_T_KEEP=50 e4_gpu_job.sh …
11-13 16:14	GeForce RTX 5090	1.39	1.043	1.45	0	2d050f2	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full '3 …
11-13 16:39	GeForce RTX 4090	0.89	1.461	1.30	0	2d050f2	env E4_DATA_SEED=100 E4_N_RESTART=512 e4_gpu_job.sh full '0 …
11-13 16:42	GeForce RTX 4090	0.89	0.907	0.81	0	481991f	env E4_N_RESTART=512 e4_gpu_job.sh full '3 4' e4_relax2_b X

End (UTC)	GPU	USD/h	Hours	USD	Exit	Commit	Command (shortened)
11-13 17:12	RTX PRO 4500 Blackwell	0.72	1.710	1.23	0	`ceeca07`	env E4_N_RESTART=512 e4_gpu_job.sh full '0 1 2' e4_relax2_a …
11-13 17:13	GeForce RTX 4090	0.89	0.982	0.87	0	`62d14b9`	env E4_DATA_SEED=300 E4_N_TRAJ=96 E4_T_KEEP=50 e4_gpu_job.sh …
11-13 17:31	A40	0.59	0.803	0.47	–	`681e5db`	env E4_N_RESTART=512 e4_gpu_job.sh full '2 3' e4_recsur_b V
11-13 17:57	GeForce RTX 4090	0.89	1.256	1.12	0	`4b42308`	env E4_N_RESTART=512 E4_REFNULLS=1 e4_gpu_job.sh full '0 1' …
11-13 17:58	RTX PRO 4500 Blackwell	0.72	0.008	0.01	–	`f34791a`	env E4_DATA_SEED=100 E4_N_RESTART=512 E4_SAME_DATA=1 e4_gpu_ …
11-13 17:58	GeForce RTX 4090	0.89	0.751	0.67	0	`3d63120`	env E4_N_RESTART=512 e4_gpu_job.sh full 4 e4_recsur_c V
11-13 18:01	GeForce RTX 5090	1.39	0.787	1.09	0	`1b459bc`	env E4_DATA_SEED=100 E4_N_RESTART=512 E4_REFNULLS=1 e4_gpu_j …
11-13 18:36	GeForce RTX 5090	0.69	0.569	0.39	0	`5833366`	env E4_DATA_SEED=100 E4_N_RESTART=512 E4_SAME_DATA=1 e4_gpu_ …
11-13 18:36	RTX PRO 4500 Blackwell	0.72	0.572	0.41	0	`5833366`	env E4_N_RESTART=512 E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 …
11-13 18:38	RTX PRO 4500 Blackwell	0.72	1.102	0.79	0	`39d4092`	env E4_N_RESTART=512 E4_SAME_DATA=1 e4_gpu_job.sh full '2 3' …
11-13 18:54	GeForce RTX 4090	0.89	0.863	0.77	0	`5833366`	env E4_DATA_SEED=100 E4_N_RESTART=512 E4_SAME_DATA=1 e4_gpu_ …
11-13 19:11	RTX PRO 4500 Blackwell	0.72	0.567	0.41	0	`d1d178d`	env E4_DATA_SEED=100 E4_N_RESTART=512 E4_SAME_DATA=1 e4_gpu_ …
11-13 20:06	GeForce RTX 5090	1.39	0.779	1.08	0	`6991c0d`	env E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 2 3 4' e4_anchorp …
11-13 20:46	RTX PRO 4500 Blackwell	0.72	1.287	0.93	0	`66af9e1`	env E4_DATA_SEED=100 E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 …
11-13 20:47	RTX PRO 4500 Blackwell	0.72	1.313	0.95	0	`66af9e1`	env E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 2 3 4' e4_cold K
11-13 21:52	RTX PRO 4500 Blackwell	0.72	0.953	0.69	0	`334e942`	env E4_DATA_SEED=100 E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 …
11-13 22:05	RTX PRO 4500 Blackwell	0.72	1.167	0.84	0	`334e942`	env E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 2 3 4' e4_struct …
11-13 23:15	GeForce RTX 5090	1.39	0.755	1.05	0	`107f40f`	env E4_STRUCT_RANK=30 E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 …
11-13 23:33	GeForce RTX 5090	1.39	1.046	1.45	0	`107f40f`	env E4_STRUCT_RANK=0 E4_SAME_DATA=1 e4_gpu_job.sh full '0 1 …

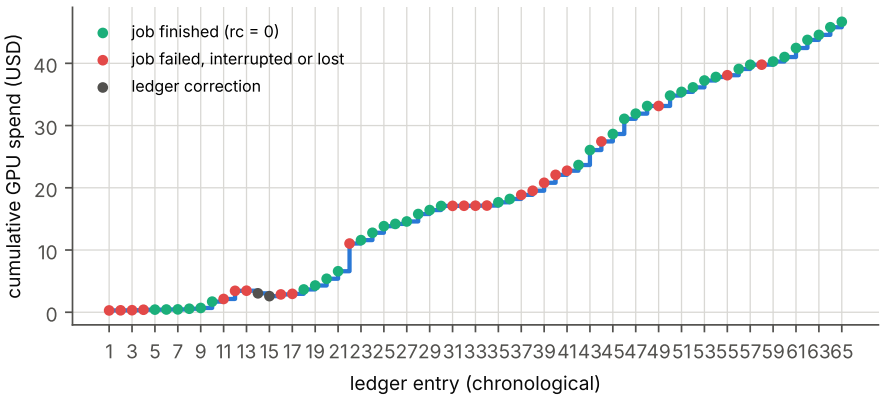

Figure J.1: Cumulative GPU spend of the pilot, one point per entry of the compute ledger. Data plot from `research/teams/claude/compute-ledger.jsonl`.

The cost of the session itself (tokens, sub-agent calls, CPU time) is reported by `tools/usage_report.py` in `research/teams/claude/usage/latest.md`.