A SHORTCUT TO STATISTICALLY STEADY-STATE TURBULENCE WITH FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many nonlinear physical systems exhibit an initial transient phase in which perturbations grow before nonlinear interactions lead to a statistically steady state. While this saturated regime is of primary interest, direct numerical simulations must resolve the full transient dynamics to reach it, incurring significant computational cost. In Computational Fluid Dynamics, reduced-order approaches such as Large Eddy Simulation mitigate computational cost by modeling small-scale dynamics, enabling tractable approximations of turbulent flows. In contrast, for systems such as gyrokinetics, comparably effective closures for the full dynamics are not generally available, and high-fidelity simulations remain necessary. Existing surrogate modeling approaches for these systems are autoregressive, hence they suffer from accumulating error. We instead propose to bypass explicit time evolution by directly modeling the distribution of saturated states under an ergodicity assumption, stating that ensemble averages over samples are equivalent to time averages of a single long simulation. We introduce GyroFlow, a latent generative model that directly estimates steady-state statistics of gyrokinetic turbulence in 5D phase space, without resolving the transient phase. At the same time, GyroFlow generates saturated snapshots from noise, conditioned on dimensionless operating parameters and matches the strongest autoregressive and tabular baselines, without requiring a saturated solver state, while outperforming other generative approaches. GyroFlow attains a $> 10000\times$ speedup compared to nonlinear gyrokinetics simulators and $> 1000\times$ speedup compared to autoregressive baselines. To evaluate generation quality we propose FGyD, a distributional metric computed in the latent space of a pretrained gyrokinetic model, and show that it correlates with downstream flux accuracy and the spectral divergence of warm-started solver runs. Finally, GyroFlow can be used to warm-start the numerical code used to produce the data, which shortens its heat-flux transient by about $3\times$.

1 INTRODUCTION

Numerical simulations of turbulent multiscale systems are a cornerstone of modern science and engineering, from aerospace and automotive design (Slotnick et al., 2014) to interstellar medium modeling (Federrath & Klessen, 2012) and energy systems, where they predict the turbulent transport that determines confinement in nuclear fusion power plant concepts such as STEP (Kennedy et al., 2023) and underpin the design of combustion engines and gas turbines (Pitsch, 2006). After an initial transient, these systems relax to a statistically steady regime from which quantities of interest for design or scientific analysis, such as heat and momentum transport in plasmas, are typically obtained via time averaging. The initial transient phase can be long and costly to compute, yet it is strictly necessary to simulate to reach the statistically steady state regime and therefore the quantities of interest.

In Computational Fluid Dynamics, substantial speedups can be obtained via Large Eddy Simulation, where dissipation at small scales is modeled instead of resolved explicitly, with many analytical closures available (Pope, 2000). However, such reductions collapse when no tractable closure is available or the observables of interest are themselves statistics of multiscale fluctuations. Gyrokinetics is a prominent example of this problem class, and one of the most challenging simulation problems in engineering. It models the turbulent transport that governs confinement in magnetic fusion plasmas, and therefore the viability of fusion as an energy source. Kinetic effects in the plasma

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

core and the low collisionality between particles require a 5D phase-space description, rather than a fluid approximation. The gyrokinetic equation (see Appendix A) evolves the distribution function of particle positions and velocities. An approach similar to LES in gyrokinetics is notoriously difficult, because of the bi-directional energy cascade where smaller scales drive an inverse transfer that generates large-scale structures such as zonal flows, which in turn suppress the small scales. The community-standard quasilinear models (Bourdelle et al., 2015; Staebler et al., 2007) are the only model order reduction currently available. Their formulation requires a fully converged linear simulation (i.e., the same linear transient that plagues nonlinear simulations) on top of which a saturation rule (a simplified analytical approximation of the nonlinear physics) is applied, but they are brittle in strongly turbulent and power plant-relevant regimes (Dimits et al., 2000; Bourdelle et al., 2008; Kiefer et al., 2021). Instead, reliable predictions for fusion power plants would require obtaining many long-running time-dependent nonlinear simulations, including a costly transient phase during which unstable modes grow before nonlinear couplings saturate the turbulence, resulting in millions of core hours spent in the transient phase during a simulation campaign.

Machine learning based surrogate models of gyrokinetics offer a fruitful alternative. Existing approaches sit on two opposite sides of a trade-off. Autoregressive neural surrogates of nonlinear gyrokinetics (Paischer et al., 2025) retain the 5D dynamics but accumulate rollout error as they proceed and they require traversing the transient before they can be deployed. Classical surrogates (Hornsby et al., 2024; van de Plassche et al., 2020) bypass the linear transient entirely, but discard the full nonlinear physics contained in the 5D distribution function.

In this work we attempt a different approach that brings together the best of both worlds: a fast, 5D-capable surrogate model that learns the nonlinear dynamics without traversing the transient and without simplifying analytic assumptions. We leverage the fact that gyrokinetic turbulence can be treated as approximately ergodic, as time-averaged observables converge to their ensemble averages over the underlying steady-state distribution after a sufficiently long time. These statistics are primarily determined by the dimensionless operating parameters. Therefore, we propose GyroFlow, a latent flow matching (Lipman et al., 2023; Dao et al., 2023) model trained to sample turbulent 5D snapshots conditioned on operating parameters. This enables estimating ensemble averages of an observable A (see Appendix A.4) in a single batched forward pass, without autoregressive rollouts or additional simplifying assumptions:

$$\underbrace{\mathbb{E}_{x\sim P}[A(x)]}_{\text{fast batched ensemble average: GyroFlow (ours)}} = \underbrace{\lim_{T\to\infty}\frac{1}{T}\int_0^T A(x(t;x_0))\,dt}_{\text{slow time averaging: gyrokinetic simulations / GyroSwin}}, \tag{1}$$

where the equivalence stems from the ergodicity property, P denotes the invariant probability measure of the statistically stationary turbulent dynamics at fixed operating parameters, and x denotes the turbulent 5D distribution function.

To evaluate generation of 5D plasma data we propose FGyD, a Fréchet distance in the latent space of a pretrained gyrokinetic surrogate (Paischer et al., 2025), whose features encode mode structure, amplitude, and zonal-flow content and verify that it correlates with downstream flux accuracy and the spectral divergence of warm-started solver runs.

Our contributions are summarized as follows.

- ❶ We introduce GyroFlow, a generative model of saturated 5D gyrokinetic turbulence which bypasses the costly transient, conditioned on operating parameters, with a $> 10000\times$ speedup compared to nonlinear gyrokinetics and a $> 1000\times$ speedup compared to autoregressive baselines.
- ❷ We introduce FGyD as a distributional metric defined in the latent space of GyroSwin (Paischer et al., 2025), and show that it correlates with downstream performance.
- ❸ We demonstrate that generated fields can be effectively used to warm-start numerical solvers, shortening their heat-flux transient by about $3\times$.

2 The (Un)avoidable Cost of Gyrokinetics

Gyrokinetic simulation codes, such as GKW (Peeters et al., 2009), evolve a 5D distribution function $f(k_x, k_y, s, v_\parallel, \mu;\, t)$ of charged particles in a magnetised plasma. Turbulence in a plasma is driven

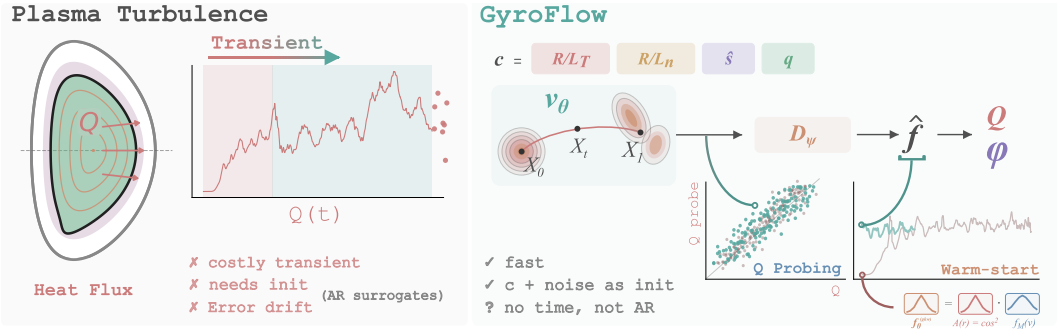

Figure 1: **Left:** In a Tokamak the heat flux $Q(t)$ is the radial transport of thermal energy across the nested magnetic flux surfaces, driven by turbulence. The transient of a gyrokinetic simulation traverses an initial ramp-up phase before converging to a statistically steady state. **Right:** GyroFlow assumes ergodicity and uses flow matching to sample directly from the saturated distribution, instead of traversing the transient. Downstream quantities such as average heat flux can be obtained either by latent probing or reconstruction of the 5D phase space.

by microinstabilities that draw free energy from the radial gradients of temperature and density, giving rise to, e.g., ion-temperature-gradient (Coppi et al., 1967, ITG) modes. These instabilities are responsible for the transport of heat and particles, which is the central quantity governing the performance of magnetically confined fusion devices.

A gyrokinetic simulation proceeds through two distinct phases. In the linear phase, perturbations to the distribution function grow exponentially as the microinstabilities amplify. Once the fluctuation amplitudes become sufficiently large, nonlinear mode coupling sets in and energy is redistributed across scales (Itoh et al., 2006). The system transitions into a saturated regime characterised by a dynamic equilibrium between instability drive and zonal flow regulation.

Simulations with identical parameters but different initial conditions converge to the same steady-state statistics (Dif-Pradalier et al. (2008); Villard et al. (2010), see also Appendix G). Thus, gyrokinetic turbulence is approximately ergodic, that is time averages are equivalent to ensemble averages (see Equation (1)). However, the linear buildup phase is so long and costly that the ensemble average is impractical, and therefore time averages along a single simulation are taken.

Moreover, the cost incurred by traversing the transient of a single gyrokinetic simulation is only the tip of the iceberg. Integrated modeling of plasma in a tokamak (Romanelli et al., 2014) further compounds this cost, as it requires hundreds or even thousands gyrokinetics evaluations per single design point (Citrin et al., 2023), rendering optimisation tasks unfeasible.

Based on the observation that the saturated distribution depends only on the operating parameters and is accessible through independent samples under the ergodicity assumption, we aim to replace temporal simulations with recent generative modeling approaches. That is, a model that maps operating parameters directly to realisations of the saturated state, bypassing both the ramp-up and the need for long time-averaging trajectories.

3 Related Works

Surrogates for gyrokinetics. There have been multiple attempts to develop surrogates for core turbulent transport as modelled by gyrokinetics. Quasilinear reduced-order models such as QuaLiKiz (Bourdelle et al., 2015; Citrin et al., 2017) and TGLF (Staebler et al., 2007; Staebler & Kinsey, 2010) evolve the linearized gyrokinetic equation, and approximate the nonlinear physics by means of analytical saturation rules fitted to nonlinear simulations. They are generally less accurate near stability boundaries and in strongly driven regimes (Dimits et al., 2000; Bourdelle et al., 2008). Classical surrogates, namely Gaussian process regression (Hornsby et al., 2024) and multilayer perceptrons (van de Plassche et al., 2020; Citrin et al., 2023; Zanisi et al., 2024), map operating parameters directly to scalar fluxes. They are efficient, but inherit the same limitations as reduced-order models.

Recent work targets full gyrokinetic surrogates. Narita et al. (2022) apply CNNs to 2D wavenumber slices for flux and time-to-saturation prediction, Honda et al. (2023) extend this with multimodal inputs, and Wan et al. (2025) study cross-fidelity transfer in a reduced 1D space. GyroSwin (Paischer et al., 2025) is the first surrogate operating on the full 5D phase space with orders-of-magnitude per-step speedup and state of the art accuracy. It still depends on an initially evolved saturated state from a numerical code.

Generative models for scientific simulation. Diffusion and flow-matching models (background in Section E) are commonly applied to scientific simulation.

Several works use diffusion to stabilise or correct autoregressive rollouts of spatiotemporal dynamics (Lippe et al., 2023; Kohl et al., 2026; Rühling Cachay et al., 2023) and to quantify uncertainty in under-resolved flow surrogates (Liu & Thuerey, 2024). Diffusion has also been used as the generative core of probabilistic ensembles for medium-range weather forecasting (Price et al., 2025) and as a posterior sampler for score-based data assimilation (Rozet & Louppe, 2023). Most similar to our work, Lino et al. (2025); Lienen et al. (2024); Gao et al. (2024) use diffusion models to learn the stationary distribution of developed fluid flows from transient Large Eddy or unsteady RANS simulations. The main difference is that our work does not make any assumption on ground-truth obtained by reduced-order models and extends to a domain where no closures are available, hence all turbulent scales are resolved. Our contribution is to extend this idea to the full 5D gyrokinetic phase space, while introducing a distributional metric tailored to gyrokinetics and an interface for warm-starting numerical solvers.

4 METHOD

GyroFlow is a two-stage latent generative model. A Swin5D autoencoder first compresses 5D snapshots of the saturated distribution function into a low-dimensional latent space. A Diffusion Transformer (DiT) (Peebles & Xie, 2023) is then trained with rectified flow matching to transport a Gaussian prior to the encoded data manifold, conditioned on the operating parameters. The two stages are trained independently, and the autoencoder weights are frozen during DiT training. Finally, to evaluate generative quality while respecting the distributional nature of turbulence, we propose FGyD (Section 4.5), a Fréchet distance in the latent space of a pretrained gyrokinetic surrogate.

4.1 PROBLEM SETUP

A gyrokinetic simulation is parameterised by dimensionless numbers entering Equation (6) through the equilibrium Maxwellian F_M (Equation (8)) and the magnetic geometry. GyroFlow conditions on four operating parameters that strongly affect ITG turbulence. R/L_T (normalised ion-temperature gradient) and R/L_n (normalised density gradient) set the free-energy source, with larger R/L_T pushing the simulation into strongly driven regimes and R/L_n modulating the contribution of density-gradient-driven modes. q (safety factor) and $\hat{s}$ (magnetic shear) enter the magnetic geometry through the parallel dynamics and the radial wavenumber along the field line. Jointly, $(R/L_T,\ R/L_n,\ q,\ \hat{s})$ specify the stationary distribution, independently of the initial condition, and are stacked into the conditioning vector

$$\boldsymbol{c} = (\ R/L_T\ ,\ R/L_n\ ,\ \hat{s}\ ,\ q\ ). \tag{2}$$

Let $f \in \mathbb{R}^{2\times v_\parallel \times \mu \times s \times x \times y}$ denote a saturated snapshot, with the two channels carrying the real and imaginary parts of the complex distribution function. A velocity network $v_\theta(z_t, t, \boldsymbol{c})$ is trained to transport $z_0 \sim \mathcal{N}(0, I)$ to $z_1 = E_\psi(f)$ along an interpolation parameter $t \in [0, 1]$, with no information about the physical time of Equation (6). At inference, $\hat{f} = D_\psi(z_1)$ recovers the full 5D state from the generated latent.

4.2 SWIN5D AUTOENCODER

E_ψ and D_ψ share hierarchical Swin5D layers, with n-dimensional Swin blocks functionally equivalent to GyroSwin (Paischer et al., 2025). Attention is applied to non-overlapping 5D windows of

size $M = M_{v_\parallel} \times M_\mu \times M_s \times M_x \times M_y$, alternating with shifted partitions (half-window cyclic shift) to recover approximate global attention at linear cost. The bottleneck is implemented via two Transformer layers (Vaswani et al., 2017) at the coarsest resolution after a linear projection to channel width d_z. The operating parameters $\boldsymbol{c}$ are not injected into either the encoder or the decoder to decouple compression from conditioning.Training uses a complex MSE reconstruction loss on f.

4.3 Latent diffusion model

We adhere to the stochastic-interpolant framework of Albergo et al. (2025) as a rectified flow (Liu et al., 2023). The prior is Gaussian, the data density is the latent representation of the saturated gyrokinetic distribution obtained via the pre-trained autoencoder, and the interpolant is the linear path of Equation (3). A full derivation is given in Section E in the appendix.

Flow-matching formulation Let $z_1 = E_\psi(f)$ denote a data latent and $z_0 \sim \mathcal{N}(0, I)$ a prior sample. For $t \in [0, 1]$, we define the linear probability path

$$z_t \;=\; t\, z_1 + (1-t)\, z_0, \qquad u^\star(z_t \mid z_0, z_1) \;=\; z_1 - z_0, \tag{3}$$

so that the target velocity is independent of t. We depart from the standard recipe in two respects to reduce path variance and linearize inference trajectories. First, instead of independent pairings, in-batch pairs are drawn from an optimal-transport coupling $\pi(z_0, z_1)$ that minimizes the expected path length between the noise and data distributions (Tong et al., 2024). Second, the integration time t is sampled from a logit-normal distribution $p_{\mathrm{LN}}(t)$ (Esser et al., 2024). The network $v_\theta(z_t, t, \boldsymbol{c})$ is trained to regress $u^\star$ by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta) \;=\; \mathbb{E}_{z_0, z_1 \sim \pi,\, t \sim p_{\mathrm{LN}},\, \boldsymbol{c}} \,\| \, v_\theta(z_t, t, \boldsymbol{c}) - u^* \, \|^2 . \tag{4}$$

At inference, we integrate $\dot{z} = v_\theta(z, t, \boldsymbol{c})$ from $z \sim \mathcal{N}(0, I)$ with an explicit Euler scheme on a uniform grid of N steps in $[0, 1]$. We use $N = 15$ throughout, chosen from the accuracy-runtime sweep in Section E.5.

Diffusion Transformer. The velocity network v_θ is a DiT (Peebles & Xie, 2023) that operates directly on the autoencoder bottleneck tokens, without further patching. The flow integration time t and the four operating parameters in $\boldsymbol{c}$ (Equation (2)) are each embedded via a sinusoidal expansion and mapped through a shared MLP to produce a joint conditioning vector. This vector modulates each transformer block via adaptive layer normalization (adaLN) applied prior to the self-attention and feed-forward layers. The residual gate is zero-initialised so DiT begins training as identity.

4.4 Implementation details

Prior work on compression (Galletti et al., 2026a) observes that training on 5D data presents instabilities, linked to the large dynamic range across velocity shells and the high dimensionality of the fields. We introduce three interventions to stabilize model training.

These are a per-μ-shell normalization of f, RMS normalization of queries and keys (Henry et al., 2020; Chameleon Team, 2025), and head-wise gated attention (Qiu et al., 2025); details are given in Section D.2.

4.5 Fréchet GyroSwin Distance (FGyD)

In vision it is common to evaluate generative models at a distributional level using Fréchet distance based on the latent space of pre-trained classifiers (Heusel et al., 2017; Szegedy et al., 2017). In our case, there is no pre-trained classifier for gyrokinetics on which we can base our evaluation. Furthermore, the latent space should ideally be sensitive to downstream quantities. To this end, we extract features $h = g_\xi(f, \boldsymbol{c})$ from a frozen and pre-trained encoder of GyroSwin (Paischer et al., 2025), fit Gaussians to generated and reference activations, and report

$$\mathrm{FGyD} \;=\; \|\mu_r - \mu_g\|_2^2 \;+\; \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2\,(\Sigma_r\,\Sigma_g)^{1/2}\Big). \tag{5}$$

We use the publicly available checkpoint from the huggingface hub (Wolf et al., 2020).GyroSwin is a UNet trained in a multitask manner to predict the time evolution of the 5D phase space, the

3D electrostatic potential fields, and the scalar heat flux. Therefore, it provides several choices for latents that could be used for computing the FGyD metric. Also, it allows us to evaluate on latents that encode different derived physical quantities. To provide a comprehensive evaluation, we select three different latents (bottleneck skip L=2, ϕ decoder, Q head) and report FGyD on each of them. For further details see Section I.

5 EXPERIMENTS

We evaluate GyroFlow model along three complementary axes: **(i)** downstream, time-averaged physical observables and parameter scans (Table 1 and Figure 2); **(ii)** distributional comparisons in a pretrained latent space (FGyD, Table 2); **(iii)** the convergence of the gyaradax solver (Galletti et al., 2026b), a GPU accelerated version of GKW (Peeters et al., 2009), warm-started from generated candidates (Table 2). GyroFlow is on par with GPR and autoregressive GyroSwin on the heat flux and outperforms other generative baselines.

5.1 SETUP

We evaluate GyroFlow on the dataset of Paischer et al. (2025), $\sim$250 saturated GKW trajectories in the 4D parameter space (Section B in appendix).Three trajectories form the validation split, 31 the in-distribution (ID) test split, and 28 the out-of-distribution (OOD) split, obtained by taking points outside the training convex hull.

Three baseline families are considered. Reduced-order models map linear growth rates to nonlinear saturated fluxes via saturation rules (Bourdelle et al., 2015, QuaLiKiz). Tabular regressors (Hornsby et al., 2024, GPR) map $\boldsymbol{c}$ to saturated transport scalars. Both neglect the nonlinear physics. Autoregressive surrogates such as *GyroSwin* (Paischer et al., 2025) evolve the 5D phase space forward in time. GyroSwin is trained such that the autoregressive rollout begins from the start of the nonlinear phase. Finally, we include three generative baselines following the same turbulence sampling perspective as GyroFlow. A VAE (Kingma & Welling, 2014) that decodes $z \sim \mathcal{N}(0, I)$, and two variants of a VQ-VAE (van den Oord et al., 2017) that differ in how code indices are drawn at inference. *VQ-VAE (rand)* samples them i.i.d. from the empirical codebook distribution, while *VQ-VAE (AR)* draws them from a Transformer prior trained autoregressively over the code sequences (Yan et al., 2021). See Section D in appendix for details.

Latent probing. A linear probe π_η is a linear regression model fit on GyroFlow latents. It maps directly to $\bar{Q}$, $W(k_y)$ and $Q(k_y)$, bypassing both D_ψ and the integrals from 5D to integrated quantities (Equation (13)). Concretely, we first apply PCA then perform Ridge regression on a training-set of (latents, $\bar{Q}$/$W(k_y)$/$Q(k_y)$) pairs. We call this method GyroFlow (probe), which trades the 5D output for lower cost compared to GyroFlow (decode).

5.2 DOWNSTREAM PHYSICS ACCURACY

We evaluate GyroFlow on three time-averaged transport quantities: $\bar{Q}$, the binormal potential spectrum $W(k_y)$, and the flux spectrum $Q(k_y)$. The latent nature of our method admits two routes, either decoding $\hat{f} = D_\psi(z_1)$ or probing the latent directly. Table 1 includes both, but unless reported differently we decode to the 5D space and perform integration to obtain the quantities of interest (see Section A.4). The same approach is used for all other generative baselines. The table is grouped in three parts, with reduced-order models at the top, autoregressive surrogates in the middle, and generative approaches at the bottom.

GyroFlow (probe) achieves the lowest $\bar{Q}_{\text{RMSE}}$ on the ID split, and GyroFlow (decode) has comparable RMSE to GPR and GyroSwin on both splits, while both requiring less than 100ms for a full evaluation, a fraction of the inference cost of GyroSwin which is dominated by the need to traverse the transient. On the spectra, the probed latent reaches the highest $W(k_y)$ Pearson correlation on the ID split, tied with VQ-VAE (AR). Conversely, GyroFlow (decode) is on-par with other methods on $W(k_y)$, while performing the best on the in-distribution $Q(k_y)$ correlation. Since GPR directly maps to the nonlinear heat fluxes, it cannot be evaluated on the spectra. We also report Pointwise RMSE and Wasserstein-distance in Table 10 in Section I.2 in the appendix. The community-standard

Table 1: Downstream accuracy on 31 in-distribution (**ID**) and 28 out-of-distribution (**OOD**) operating conditions. **5D**: reports 5D nonlinear phase space outputs. We report the RMSE of the time-averaged heat flux ($\bar{Q}$), Pearson correlation on $W(k_y)$, Pearson correlation on $Q(k_y)$, and wall-clock per saturated-phase sample on a single NVIDIA H100; $\pm$ denotes the standard deviation across operating conditions. GyroFlow (probe) uses latent-space probing without 5D decoding; GyroFlow (decode) uses the full decode path. GyroSwin predicts the ion heat flux, the electrostatic potentials and the flux spectrum. Best **bold**, second underlined.

Method	**5D**	$\bar{Q}_{\text{RMSE}}\downarrow$		$W(k_y)_{\text{PC}}\uparrow$		$Q(k_y)_{\text{PC}}\uparrow$		**Runtime**$\downarrow$ [in ms]
		ID	**OOD**	**ID**	**OOD**	**ID**	**OOD**	
Quasilinear	✗	$31.0_{\pm 19.2}$	$50.1_{\pm 24.7}$	$0.022_{\pm 0.066}$	$0.097_{\pm 0.171}$	$0.752_{\pm 0.365}$	$0.537_{\pm 0.342}$	3.54×10^5
GPR	✗	$\underline{16.9}_{\pm 12.3}$	$21.7_{\pm 16.4}$	—	—	—	—	<0.1
GyroSwin	✓	$22.0_{\pm 14.4}$	$\mathbf{13.7}_{\pm 10.6}$	$\underline{0.991}_{\pm 0.017}$	$\mathbf{0.888}_{\pm 0.179}$	$0.851_{\pm 0.187}$	$\mathbf{0.932}_{\pm 0.104}$	3.36×10^5
VAE	✓	$111.5_{\pm 24.8}$	$78.4_{\pm 44.7}$	$0.988_{\pm 0.025}$	$0.846_{\pm 0.253}$	$0.533_{\pm 0.148}$	$0.664_{\pm 0.209}$	**38.8**
VQ-VAE (rand)	✓	$71.6_{\pm 20.6}$	$54.1_{\pm 33.9}$	$0.989_{\pm 0.024}$	$0.864_{\pm 0.221}$	$0.621_{\pm 0.136}$	$0.703_{\pm 0.220}$	**38.8**
VQ-VAE (AR)	✓	$39.4_{\pm 16.9}$	$41.3_{\pm 26.7}$	$\mathbf{0.992}_{\pm 0.019}$	$\underline{0.883}_{\pm 0.192}$	$\underline{0.891}_{\pm 0.157}$	$0.835_{\pm 0.169}$	3.25^5
GyroFlow (probe)	✗	$\mathbf{14.1}_{\pm 9.2}$	$21.8_{\pm 16.2}$	$\mathbf{0.992}_{\pm 0.011}$	$0.876_{\pm 0.223}$	$0.680_{\pm 0.184}$	$0.691_{\pm 0.207}$	$\underline{41.6}$
GyroFlow (decode)	✓	$19.0_{\pm 11.2}$	$\underline{20.4}_{\pm 16.2}$	$0.989_{\pm 0.024}$	$0.859_{\pm 0.231}$	$\mathbf{0.938}_{\pm 0.105}$	$\underline{0.874}_{\pm 0.130}$	79.6

quasilinear baseline and the latent generative baselines (VAE, VQ-VAE) consistently perform worse on the time-averaged heat flux, with only the VQ-VAE (AR) being competitive on the spectra. The quasilinear model also incurs in a very long runtime dictated by the need to produce a fully developed linear simulation on top of which to apply the analytic saturation rule.

A nonlinear gyrokinetic run obtained with the gyaradax code (Galletti et al., 2026b) takes between 10 and 30 minutes on a single H100 card, depending on the operating parameters. GyroFlow provides a factor of $> 10000\times$ speedup compared to the ground truth, and $> 1000\times$ speedup compared to GyroSwin. It is the best compromise available between speed and accuracy on downstream quantities of interest.

Single-parameter sensitivity scans. We investigate whether the model has learned a smooth and physically consistent attractor manifold. To that end, we sweep each component of $\boldsymbol{c}$ independently across a wide ID and OOD support, with the other three fixed. See Section C in appendix for details. This generalization is non-trivial as the model is being pushed far outside of its training domain.

Figure 2 shows that GyroFlow reproduces the expected ITG thresholding behavior even in regimes absent or underrepresented in the dataset. This dataset is heavily skewed toward high R/L_T ranges, and thus highly turbulent regions. Figure 2 includes the training distribution across parameters and flux in gray, highlighting the gaps in training data and the robustness of our method.

GyroFlow captures near-zero fluxes at low R/L_T which are entirely unseen during training, whereas GPR returns unphysical negative Q. Additionally, GPR keeps climbing with q after the reference saturates. At negative $\hat{s}$, far from the training range, the reference inverts trend and decays. GyroFlow follows the turnover, while the GPR extrapolates monotonically upward. Sampling the distribution function and integrating the flux from it inherits constraints that a direct regression on the conditioning parameters does not. In other words, this shows that our method learns to approximate the manifold and not simply regress the flux.

5.3 Distributional quality and warm-start

Since draws from GyroFlow are i.i.d. samples of the saturated distribution rather than predictions of a specific reference, evaluating generative quality requires a distributional metric. We compare two complementary views, (i) warm-start dynamics in physical space (the solver is initialized from $\hat{f}$ and we measure how close the resulting stationary marginal is to the reference) and (ii) FGyD, introduced in Section 4.5. We find that both are cross-correlated as well as correlated to the downstream metrics.

Warm-starts and similarity to attractor. A generated $\hat{f}$ can replace the random initial condition of numerical solvers. If $\hat{f}$ already lies close to the attractor $\mathbb{P}_{\boldsymbol{c}}$, the linear-to-nonlinear transient is

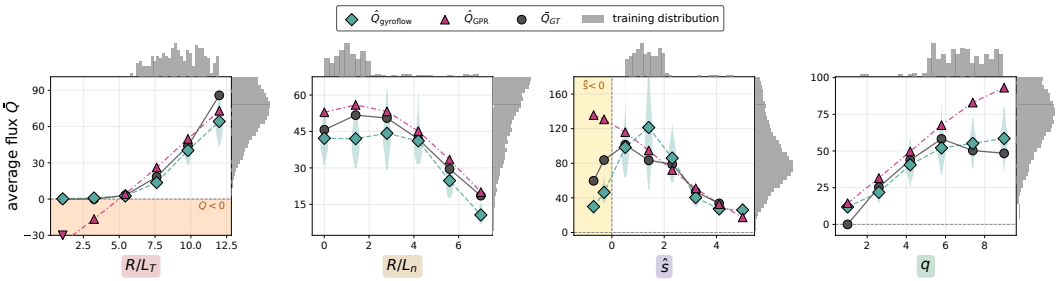

Figure 2: Single-param $\bar{Q}$ scans along each of the four conditioning axes. Green $\diamond$: mean of GyroFlow samples, violins represent distribution of K=64 draws. Rose $\triangle$: GPR. Dark $\circ$: reference. The region where $\hat{s} < 0$ is completely OOD.

Table 2: Distributional generation quality from both **(a)** warm-start Kolmogorov-Smirnov $\overline{D}_S$ and **(b)** FGyD at three GyroSwin U-Net depths. Lower is better, best in **bold** and second underlined.

Method	$\overline{D}_{W(k_y)}\downarrow$	$\overline{D}_{Q(k_y)}\downarrow$
Nearest $\boldsymbol{c}$	<u>0.363</u>	<u>0.267</u>
VAE	0.659	0.471
VQ-VAE (rand)	0.642	0.513
VQ-VAE (AR)	0.688	0.497
GyroFlow	**0.349**	**0.248**

(a) Warm-start KS statistic $\overline{D}_S$.

Method	skip L=2	flux head	ϕ decoder
VAE	3.6e+04	416.7	8.9e+05
VQ-VAE (rand)	1.9e+04	109.1	<u>1.2e+05</u>
VQ-VAE (AR)	<u>1.7e+04</u>	<u>55.9</u>	**8.0e+04**
GyroFlow	**1.1e+04**	**38.2**	2.6e+05

(b) FGyD per method at the three depths.

shortened or completely removed. To quantify this, for each held-out $\boldsymbol{c}$ we draw R=5 independent generations from GyroFlow, integrate each with gyaradax for $T \ll T_{\text{GT}}$ steps and pair all R warm-started rollouts ($\mathcal{S}^w_{\boldsymbol{c}}$) with the corresponding tail of a fully saturated reference run ($\mathcal{S}^r_{\boldsymbol{c}}$). At every binormal mode index $\ell \in \{1, \ldots, n_{k_y}\}$, the spectrum component $S_\ell(t)$ is a scalar time series (over the rolled, warm started integration). We compare the empirical distribution functions (CDFs) $\hat{F}^w_\ell, \hat{F}^r_\ell$ with the non-parametric two-sample Kolmogorov-Smirnov (KS) statistic (Massey, 1951) and mean across modes (Equations (20) and (21)).

The per-mode evaluation mitigates the bias introduced by the very high magnitude low-k_y range ($\sim 2-3$ orders of magnitude larger than the high-k_y). Details are provided in Section F in the appendix. We include the nearest neighbor in the training set according to operating parameters in the comparison to the warm-restart. The results are shown in Table 2a. GyroFlow attains the lowest $\overline{D}_S$ on both spectra across generative methods.

Time to saturation. To measure the saved solver time directly, we warm-start gyaradax from three GyroFlow samples at each of 33 in-distribution operating points (31 held-out and 2 validation points) and compare against a cold start. The reference is the last 80 heat-flux measurements of the GKW simulation at the same operating point. We call the heat flux saturated when its mean over the last 24 time units stays inside the reference 95% range for six consecutive checks (Section H). The heat flux in warm-started runs saturates after a median of 24 time units, the earliest time our criterion can report, against 90 time units for cold starts, a paired median speed-up of 2.95$\times$ (95% interval $[2.30, 3.75]$). Warm starts reach this criterion before the paired cold run in 92 of 99 cases. Over the first 96 time units, their mean relative flux error is 18.2%, compared with 47.9% for cold starts (Table 8).

Generation quality. For each held-out $\boldsymbol{c}$ we draw $K = 64$ samples $\{\hat{f}^{(k)}\}$ and extract features $h^{(k)} = g_\xi(\hat{f}^{(k)}, \boldsymbol{c})$ from the frozen GyroSwin encoder, with the reference set produced by passing ground-truth samples at the same $\boldsymbol{c}$ through g_ξ. Table 2b reports FGyD at three U-Net depths, namely

the deepest skip $L{=}2$ (coarsest 5D resolution level of GyroSwin), the flux head activation (before the final scalar flux projection), and the ϕ decoder (electrostatic potential reconstruction branch).

GyroFlow attains the best FGyD at skip $L{=}2$ and the flux head, in line with its performance on $\bar{Q}$ and $Q(k_y)$ in Table 1, indicating high quality generation of the density function. It is worse at the ϕ decoder, matching the weaker $W(k_y)$ spectra reconstruction (Table 1) which is itself a ϕ-derived quantity. The quantitative alignment between FGyD and the corresponding physical observable supports its use as a quantitative diagnostic.

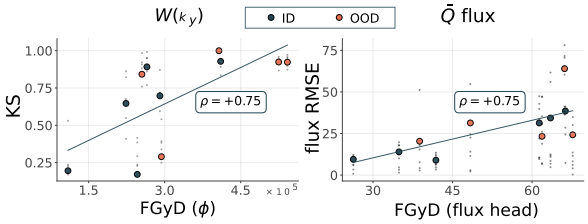

Figure 3: Correlation between FGyD and warm-start KS (left) and flux RMSE (right); $n{=}11$ held-out points.

To further drive this point and as a verification of both distributional metrics, we cross-validate them by pairing $\overline{D}_S$ and FGyD per held-out $\boldsymbol{c}$. Figure 3 (left) reports a positive correlation, suggesting that latent-space distance in the GyroSwin features is linked to attractor proximity of the generated warm start candidates. A similar pattern is observed when pairing FGyD on the flux head to the time-averaged flux RMSE (Figure 3, right). The feature extractor is the frozen public GyroSwin checkpoint from (Paischer et al., 2025).

6 Discussion and Conclusion

We introduced GyroFlow, a conditional latent flow-matching model that bypasses the transient of gyrokinetic simulations by leveraging the ergodicity assumption. Therefore it can directly sample from the statistical 5D steady state the simulation naturally converges to, and the related quantities of interest obtained from its integrals. On the cyclone-base case ITG benchmark, GyroFlow matches the strongest baselines for heat flux prediction without requiring a saturated solver state, and it outperforms community-standard methods. GyroFlow is $> 10000\times$ faster than the ground truth simulation and $> 1000\times$ faster than the strongest autoregressive baseline.

Furthermore, we demonstrate that GyroFlow can extrapolate along parameter regions it has not been trained on, indicating that it has learned to successfully traverse the latent attractor manifold. The single-axis OOD scans of Figure 2 show that the model behaves well in regions of parameter space where training data is sparse or absent: it exhibits textbook ITG threshold along R/L_T, has the expected monotone behavior in q and follows the trend inversion at $\hat{s} < 0$.

Finally, we propose FGyD, a distributional metric computed in the latent space of a pretrained gyrokinetic surrogate. By warm-starting the numerical solver with generated snapshots, we show that FGyD correlates with the spectral divergence of the resulting rollouts, validating it as a solver-free proxy for generation quality. Warm-starting shortens the heat-flux transient of the solver by about $3\times$, and initial-condition ensembles support the ergodicity assumption within the sampling noise of the reference simulations.

In summary, GyroFlow is the best compromise available between speed, generative performance and accuracy on downstream quantities and tasks of interest.

Limitations and future directions. The current dataset is electrostatic, single-species, and local flux-tube. GyroFlow can be easily generalised to the power plant-relevant high-β regimes where the contributions of multiple species, electromagnetic fluctuations and modes driven by the electron temperature gradient (ETG) can be important. This requires conditioning on additional dimensionless parameters and generating a larger training dataset, but minimal changes to GyroFlow. As these simulations are $10-100\times$ slower than those used here, while GyroFlow's inference time is expected to remain comparable to that reported in Table 1, we anticipate proportionally greater speedups.

A second axis is integration with transport modeling frameworks. Pairing GyroFlow with JINTRAC (Romanelli et al., 2014), TORAX (Citrin et al., 2024) or PORTALS (Rodriguez-Fernandez et al., 2024) would replace tens of thousands of numerical solver calls while retaining the nonlinear physics neglected by the currently used reduced-order alternatives (Hornsby et al., 2024; Zanisi et al., 2025; Citrin et al., 2024), taking a step toward real-time transport prediction in reactor design loops.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. Journal of Machine Learning Research, 26(209):1–80, 2025.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In International Conference on Learning Representations, 2018.

C. Bourdelle, A. Casati, X. Garbet, F. Imbeaux, J. Candy, F. Clairet, G. Dif-Pradalier, G. Falchetto, T. Gerbaud, V. Grandgirard, P. Hennequin, R. Sabot, Y. Sarazin, L. Vermare, and R. E. Waltz. Validity of quasi-linear transport model. In Proceedings of the 22nd IAEA Fusion Energy Conference, Vienna, Austria, 2008. International Atomic Energy Agency. Paper TH/P8-7.

C Bourdelle, J Citrin, B Baiocchi, A Casati, P Cottier, X Garbet, F Imbeaux, and JET Contributors. Core turbulent transport in tokamak plasmas: bridging theory and experiment with qualikiz. Plasma Physics and Controlled Fusion, 58(1):014036, November 2015. doi: 10.1088/0741-3335/58/1/014036.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In IEEE Conference on Computer Vision and Pattern Recognition, 2020.

J Citrin, C Bourdelle, F J Casson, C Angioni, N Bonanomi, Y Camenen, X Garbet, L Garzotti, T Görler, O Gürcan, F Koechl, F Imbeaux, O Linder, K van de Plassche, P Strand, G Szepesi, and JET Contributors. Tractable flux-driven temperature, density, and rotation profile evolution with the quasilinear gyrokinetic transport model QuaLiKiz. Plasma Physics and Controlled Fusion, 59 (12):124005, November 2017. doi: 10.1088/1361-6587/aa8aeb.

J. Citrin, P. Trochim, T. Goerler, D. Pfau, K. L. van de Plassche, and F. Jenko. Fast transport simulations with higher-fidelity surrogate models for iter. Physics of Plasmas, 30(6):062501, 06 2023. ISSN 1070-664X. doi: 10.1063/5.0136752.

Jonathan Citrin, Ian Goodfellow, Akhil Raju, Jeremy Chen, Jonas Degrave, Craig Donner, Federico Felici, Philippe Hamel, Andrea Huber, Dmitry Nikulin, David Pfau, Brendan Tracey, Martin Riedmiller, and Pushmeet Kohli. Torax: A fast and differentiable tokamak transport simulator in jax, 2024.

B. Coppi, M. N. Rosenbluth, and R. Z. Sagdeev. Instabilities due to temperature gradients in complex magnetic field configurations. The Physics of Fluids, 10(3):582–587, 03 1967. ISSN 0031-9171. doi: 10.1063/1.1762151.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), Advances in Neural Information Processing Systems, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

G. Dif-Pradalier, V. Grandgirard, Y. Sarazin, X. Garbet, Ph. Ghendrih, and P. Angelino. On the influence of initial state on gyrokinetic simulations. Physics of Plasmas, 15(4):042315, 04 2008. ISSN 1070-664X. doi: 10.1063/1.2904901.

A. M. Dimits, G. Bateman, M. A. Beer, B. I. Cohen, W. Dorland, G. W. Hammett, C. Kim, J. E. Kinsey, M. Kotschenreuther, A. H. Kritz, L. L. Lao, J. Mandrekas, W. M. Nevins, S. E. Parker, A. J. Redd, D. E. Shumaker, R. Sydora, and J. Weiland. Comparisons and physics basis of tokamak transport models and turbulence simulations. Physics of Plasmas, 7(3):969–983, 03 2000. ISSN 1070-664X. doi: 10.1063/1.873896.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the 41st International Conference on Machine Learning, ICML'24, 2024.

Christoph Federrath and Ralf S. Klessen. The star formation rate of turbulent magnetized clouds: Comparing theory, simulations, and observations. The Astrophysical Journal, 761(2):156, 2012. doi: 10.1088/0004-637X/761/2/156.

E. A. Frieman and Liu Chen. Nonlinear gyrokinetic equations for low-frequency electromagnetic waves in general plasma equilibria. The Physics of Fluids, 25(3):502–508, 1982. doi: 10.1063/1.863762.

Gianluca Galletti, Gerald Gutenbrunner, Sandeep S. Cranganore, William Hornsby, Lorenzo Zanisi, Naomi Carey, Stanislas Pamela, Johannes Brandstetter, and Fabian Paischer. Physics-informed neural compression of high-dimensional plasma data, 2026a.

Gianluca Galletti, Eric Volkmann, and Johannes Brandstetter. gyaradax: Local gyrokinetics jax code, 2026b.

Han Gao, Xu Han, Xiantao Fan, Luning Sun, Li-Ping Liu, Lian Duan, and Jian-Xun Wang. Bayesian conditional diffusion models for versatile spatiotemporal turbulence generation. Computer Methods in Applied Mechanics and Engineering, 427:117023, 2024. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2024.117023.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4246–4253. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

Mitsuru Honda, Emi Narita, Shinya Maeyama, and Tomo-Hiko Watanabe. Multimodal convolutional neural networks for predicting evolution of gyrokinetic simulations. Contributions to Plasma Physics, 63(5-6):e202200137, 2023. doi: https://doi.org/10.1002/ctpp.202200137.

W. A Hornsby, A. Gray, J. Buchanan, B. S. Patel, D. Kennedy, F. J. Casson, C. M. Roach, M. B. Lykkegaard, H. Nguyen, N. Papadimas, B. Fourcin, and J. Hart. Gaussian process regression models for the properties of micro-tearing modes in spherical tokamaks. Physics of Plasmas, 31 (1):012303, 01 2024. ISSN 1070-664X. doi: 10.1063/5.0174478.

K. Itoh, S.-I. Itoh, P. H. Diamond, T. S. Hahm, A. Fujisawa, G. R. Tynan, M. Yagi, and Y. Nagashima. Physics of zonal flows. Physics of Plasmas, 13(5):055502, 05 2006. ISSN 1070-664X. doi: 10.1063/1.2178779.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Advances in Neural Information Processing Systems, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. doi: 10.52202/068431-1926.

D. Kennedy, M. Giacomin, F.J. Casson, D. Dickinson, W.A. Hornsby, B.S. Patel, and C.M. Roach. Electromagnetic gyrokinetic instabilities in step. Nuclear Fusion, 63(12):126061, nov 2023. ISSN 1741-4326. doi: 10.1088/1741-4326/ad08e7.

C.K. Kiefer, C. Angioni, G. Tardini, N. Bonanomi, B. Geiger, P. Mantica, T. Puetterich, E. Fable, P.A. Schneider, ASDEX Upgrade Team , EUROfusion MST1 Team , and JET Contributors . Validation of quasi-linear turbulent transport models against plasmas with dominant electron heating for the prediction of iter pfpo-1 plasmas. Nuclear Fusion, 61(6):066035, may 2021. doi: 10.1088/1741-4326/abfc9c.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), 2nd International Conference on Learning Representations, ICLR, 2014.

Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. Neural Networks, 199:108641, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2026.108641.

John A. Krommes. The gyrokinetic description of microturbulence in magnetized plasmas. Annual Review of Fluid Mechanics, 44(1):175–201, 2012. ISSN 1545-4479. doi: 10.1146/annurev-fluid-120710-101223.

Marten Lienen, David Lüdke, Jan Hansen-Palmus, and Stephan Günnemann. From zero to turbulence: Generative modeling for 3d flow simulation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), International Conference on Learning Representations, volume 2024, pp. 5203–5220, 2024.

Mario Lino, Tobias Pfaff, and Nils Thuerey. Learning distributions of complex fluid simulations with diffusion graph networks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), International Conference on Learning Representations, volume 2025, pp. 77089–77119, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In Proceedings of the 11th International Conference on Learning Representations, 2023.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems, volume 36, pp. 67398–67433. Curran Associates, Inc., 2023. doi: 10.52202/075280-2946.

Qiang Liu and Nils Thuerey. Uncertainty-aware surrogate models for airfoil flow simulations with denoising diffusion probabilistic models. AIAA Journal, 62(8):2912–2933, 2024. doi: 10.2514/1.J063440.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In The Eleventh International Conference on Learning Representations, 2023.

Frank J. Massey. The Kolmogorov–Smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253):68–78, 1951. doi: 10.1080/01621459.1951.10500769.

E. Narita, M. Honda, S. Maeyama, and T.-H. Watanabe. Toward efficient runs of nonlinear gyrokinetic simulations assisted by a convolutional neural network model recognizing wavenumber-space images. Nuclear Fusion, 62(8):086037, jun 2022. doi: 10.1088/1741-4326/ac70e8.

Fabian Paischer, Gianluca Galletti, William Hornsby, Paul Setinek, Lorenzo Zanisi, Naomi Carey, Stanislas Pamela, and Johannes Brandstetter. Gyroswin: 5d surrogates for gyrokinetic plasma turbulence simulations. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4172–4182, 2023.

A.G. Peeters, Y. Camenen, F.J. Casson, W.A. Hornsby, A.P. Snodin, D. Strintzi, and G. Szepesi. The nonlinear gyro-kinetic flux tube code gkw. Computer Physics Communications, 180(12):2650–2672, 2009. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2009.07.001.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Heinz Pitsch. Large-eddy simulation of turbulent combustion. Annual Review of Fluid Mechanics, 38(1):453–482, 2006. doi: 10.1146/annurev.fluid.38.050304.092133.

Stephen B Pope. Turbulent flows. Cambridge University Press, 2000.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. Nature, 637(8044): 84–90, 2025. doi: 10.1038/s41586-024-08252-9. Preprint: GenCast: Diffusion-based ensemble forecasting for medium-range weather.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In Advances in Neural Information Processing Systems (NeurIPS 2025), volume 38, 2025. doi: 10.52202/085713-3345.

P. Rodriguez-Fernandez, N.T. Howard, A. Saltzman, S. Kantamneni, J. Candy, C. Holland, M. Balandat, S. Ament, and A.E. White. Enhancing predictive capabilities in fusion burning plasmas through surrogate-based optimization in core transport solvers. Nuclear Fusion, 64(7):076034, June 2024. doi: 10.1088/1741-4326/ad4b3d.

M. Romanelli, G. Corrigan, V. Parail, S. Wiesen, R. Ambrosino, P. Da Silva Aresta Belo, L. Garzotti, D. Harting, F. Köchl, T. Koskela, L. Lauro-Taroni, C. Marchetto, M. Mattei, E. Militello-Asp, M. F. F. Nave, S. Pamela, A. Salmi, P. Strand, G. Szepesi, and EFDA-JET Contributors. JINTRAC: A system of codes for integrated simulation of tokamak scenarios. Plasma and Fusion Research, 9:3403023, 2014. doi: 10.1585/pfr.9.3403023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

François Rozet and Gilles Louppe. Score-based data assimilation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems, volume 36, pp. 40521–40541. Curran Associates, Inc., 2023. doi: 10.52202/075280-1763.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems, volume 36, pp. 45259–45287. Curran Associates, Inc., 2023. doi: 10.52202/075280-1961.

Jeffrey Slotnick, Abdollah Khodadoust, Juan Alonso, David Darmofal, William Gropp, Elizabeth Lurie, and Dimitri Mavriplis. CFD Vision 2030 Study: A Path to Revolutionary Computational Aerosciences. Technical Report NASA/CR-2014-218178, NASA Langley Research Center, 2014.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In Proceedings of the 9th International Conference on Learning Representations, 2021.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. Advances in neural information processing systems, 33:12438–12448, 2020.

G. M. Staebler and J. E. Kinsey. Electron collisions in the trapped gyro-landau fluid transport model. Physics of Plasmas, 17(12):122309, 12 2010. ISSN 1070-664X. doi: 10.1063/1.3505308.

G. M. Staebler, J. E. Kinsey, and R. E. Waltz. A theory-based transport model with comprehensive physics. Physics of Plasmas, 14(5):055909, 05 2007. ISSN 1070-664X. doi: 10.1063/1.2436852.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017. doi: 10.1609/aaai.v31i1.11231.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. Transactions on Machine Learning Research, 2024. ISSN 2835-8856.

K. L. van de Plassche, J. Citrin, C. Bourdelle, Y. Camenen, F. J. Casson, V. I. Dagnelie, F. Felici, A. Ho, S. Van Mulders, and JET Contributors. Fast modeling of turbulent transport in fusion plasmas using neural networks. Physics of Plasmas, 27(2):022310, 02 2020. ISSN 1070-664X. doi: 10.1063/1.5134126.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 6000–6010. Curran Associates Inc., 2017. ISBN 9781510860964.

L. Villard, A. Bottino, S. Brunner, A. Casati, J. Chowdhury, T. Dannert, R. Ganesh, X. Garbet, T. Görler, V. Grandgirard, R. Hatzky, Y. Idomura, F. Jenko, S. Jolliet, S. Khosh Aghdam, X. Lapillonne, G. Latu, B. F. McMillan, F. Merz, Y. Sarazin, T. M. Tran, and T. Vernay. Gyrokinetic simulations of turbulent transport: size scaling and chaotic behaviour. Plasma Physics and Controlled Fusion, 52(12):124038, 2010. doi: 10.1088/0741-3335/52/12/124038.

Chenguang Wan, Youngwoo Cho, Zhisong Qu, Yann Camenen, Robin Varennes, Kyungtak Lim, Kunpeng Li, Jiangang Li, Yanlong Li, and Xavier Garbet. A high-fidelity surrogate model for the ion temperature gradient (itg) instability using a small expensive simulation dataset. Nuclear Fusion, 65(5):054001, apr 2025. doi: 10.1088/1741-4326/adc7c9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38–45, 2020.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. ACM Comput. Surv., 56(4), November 2023. doi: 10.1145/3626235.

L. Zanisi, A. Ho, J. Barr, T. Madula, J. Citrin, S. Pamela, J. Buchanan, F. J. Casson, V. Gopakumar, and JET Contributors. Efficient training sets for surrogate models of tokamak turbulence with active deep ensembles. Nuclear Fusion, 64(3):036022, February 2024. ISSN 1741-4326. doi: 10.1088/1741-4326/ad240d.

Lorenzo Zanisi, Aaro Järvinen, Adam Kit, Amanda Bruncrona, Anushan Fernando, Bhavin Patel, Catherine Siddle, Colin Roach, Daniel Jordan, Francis Casson, Frida Eriksson, Garud Snoep, Harry Dudding, James Buchanan, Jonathan Citrin, Luigi Quarantiello, Orso Meneghini, P. Hamel, Salvatore Correnti, Stanislas Pamela, T. Norman, Theodore Brown, Tom Neiser, and Vincenzo Lomonaco. Data-efficient digital twinning strategies and surrogate models of quasilinear turbulence in JET and STEP. In Proceedings of the 30th IAEA Fusion Energy Conference (FEC 2025), Chengdu, China, October 2025. International Atomic Energy Agency (IAEA). Paper ID: TH-C, Contribution 36059.

AI USE DISCLOSURE

We used coding agents and general AI assistants through the project. The main applications were preparing the experiment scripts (for instance, gyaradax solver handling and result collection for Figure 2 or Table 8 was done through agents). Paper writing was also in some parts aided by the same agents, with code and outcomes in context. In our case AI was not used in finding the problem, method design or ideation, but only in execution.

A GYROKINETICS

This appendix provides a term-by-term decomposition of the gyrokinetic Vlasov equation introduced in Equation (6). The presentation follows the standard δf formulation (Frieman & Chen, 1982; Krommes, 2012) in the local flux-tube limit, and the term numbering of Peeters et al. (2009), then the quantities of interest are presented.

Each species satisfies the gyrokinetic equation:

$$\underbrace{\frac{\partial f}{\partial t} + (v_\parallel \boldsymbol{b} + \boldsymbol{v}_D)\cdot\nabla f - \frac{\mu B}{m}\frac{\boldsymbol{B}\cdot\nabla B}{B^2}\frac{\partial f}{\partial v_\parallel}}_{\text{Linear}} + \underbrace{\boldsymbol{v}_E\cdot\nabla f}_{\text{Nonlinear}} \;=\; S, \tag{6}$$

where $\boldsymbol{b} = \boldsymbol{B}/B$ is the unit vector along the equilibrium magnetic field $\boldsymbol{B}$, $\boldsymbol{v}_D$ the magnetic drift, $\boldsymbol{v}_E$ the $\boldsymbol{E}\times\boldsymbol{B}$ drift, and S collects collisional, source terms and numerical dissipation. See Appendix A for a detailed description of the gyrokinetic equation. The quantities of interest that can be extracted from gyrokinetics for downstream use are the electrostatic potentials, $\phi(t)$, and heat flux, $Q(t)$, both defined as integrals of f (see Appendix A), as well as their spectra $W(k_y)$ and $Q(k_y)$ respectively.

A.1 THE δf DECOMPOSITION

The distribution function is split into a time-independent equilibrium Maxwellian and a fluctuating perturbation,

$$f \;=\; F_M + \delta f(v_\parallel, \mu, s, k_x, k_y;\, t). \tag{7}$$

The equilibrium Maxwellian encodes the background density n and temperature T through

$$F_M \;=\; \frac{n}{(2\pi T/m)^{3/2}}\,\exp\!\left(-\frac{m\,v^2}{2\,T}\right), \tag{8}$$

where m is the species mass. The normalised radial gradients of F_M yield the temperature and density gradient parameters R/L_T and R/L_n that enter the equilibrium drive and condition GyroFlow (Equation (11)).

Substituting Equation (7) into the full gyrokinetic Vlasov equation and retaining terms to first order in $\delta f/F_M$ yields a closed evolution equation for δf. The individual terms of the resulting right-hand side are detailed below.

A.2 RIGHT-HAND SIDE DECOMPOSITION

Substituting Equation (7) into Equation (6) and separating linear from nonlinear contributions, the right-hand side groups into four categories: kinetic dynamics, energy drives, nonlinear advection, and numerical dissipation. The $\boldsymbol{v}_E\cdot\nabla f$ term of Equation (6) splits into the equilibrium drive (V), which is linear in ϕ, and the nonlinear advection (III). In the local flux-tube representation with spectral perpendicular coordinates (k_x, k_y), the equation reads

$$\frac{\partial \delta f}{\partial t} \;=\; \underbrace{-v_\parallel\,\nabla_\parallel \delta f \;-\; i(\boldsymbol{k}_\perp\cdot\boldsymbol{v}_D)\,\delta f \;+\; \mu\,\nabla_\parallel B\,\frac{\partial \delta f}{\partial v_\parallel}}_{\text{Kinetic dynamics (I, II, IV)}} \;\underbrace{-\boldsymbol{v}_E\cdot\nabla F_M \;-\; \frac{Ze\,F_M}{T}\left(v_\parallel\,\nabla_\parallel \bar{\phi} + i(\boldsymbol{k}_\perp\cdot\boldsymbol{v}_D)\,\bar{\phi}\right)}_{\text{Energy drives (V, VII, VIII)}}$$

$$\underbrace{-\,\boldsymbol{v}_E\cdot\nabla_\perp \delta f}_{\text{Nonlinear advection (III)}} \;\underbrace{-\,\mathcal{D}(\delta f)}_{\text{Dissipation}} \tag{9}$$

where $\bar{\phi}$ denotes the gyro-averaged electrostatic potential, $\boldsymbol{v}_D$ the magnetic drift velocity, $\boldsymbol{v}_E$ the $\boldsymbol{E}\times\boldsymbol{B}$ drift velocity, Ze the species charge, and $\mathcal{D}$ a numerical dissipation operator. Each category is described below.

Kinetic dynamics (I, II, IV). These terms describe the collisionless motion of guiding centres in the equilibrium magnetic geometry. Parallel advection (I) represents streaming along the magnetic field. The magnetic drift (II) arises from curvature and ∇B drifts perpendicular to the field. The mirror term (IV) accounts for the parallel force exerted by gradients in the magnetic field magnitude, which leads to particle trapping in regions of low B. More explicitly, in s–α geometry,

$$\nabla_\parallel \;=\; \frac{1}{q\,R}\,\partial_s, \qquad k_\perp^2 \;\propto\; 1+\big(\,\hat{s}\;s-\alpha\sin s\big)^2. \tag{10}$$

so q rescales parallel dynamics and $\hat{s}$ modulates the radial wavenumber along the field line.

Energy drives (V, VII, VIII). The equilibrium drive (V) couples the $\boldsymbol{E}\times\boldsymbol{B}$ drift to the radial gradient of the Maxwellian, providing the free-energy source for micro-instabilities. It is through this term that the normalised gradients R/L_T and R/L_n enter the equation:

$$\boldsymbol{v}_E\cdot\nabla F_M \;=\; \boldsymbol{v}_E\cdot\hat{\mathbf{r}}\,F_M\left(\frac{R}{L_n}\;+\;\left(\frac{v^2}{v_{\mathrm{th}}^2}-\frac{3}{2}\right)\frac{R}{L_T}\right), \tag{11}$$

where v_{th} is the thermal velocity. The field drives (VII, VIII) represent the linear response of the background distribution to the fluctuating electrostatic potential, coupling parallel streaming and magnetic drift to $\bar{\phi}$.

Nonlinear advection (III). The $\boldsymbol{E}\times\boldsymbol{B}$ advection of δf by the fluctuating electric field is responsible for mode coupling, energy redistribution across spatial scales, and the development of saturated turbulence. It is the computationally dominant term and the one that distinguishes nonlinear from linear gyrokinetics.

Dissipation. A numerical dissipation operator $\mathcal{D}$ is added for stability, acting along the parallel and velocity coordinates and as spectral hyper-viscosity in the perpendicular plane. Its coefficients are chosen to damp grid-scale fluctuations without affecting the resolved physical scales.

Remark. Term VI in the convention of Peeters et al. (2009) is the neoclassical drive $-\boldsymbol{v}_D\cdot\nabla F_M$, which couples the magnetic drift to equilibrium gradients and is relevant only for rotating or neoclassical plasmas. It is omitted in this work.

A.3 Quasineutrality and the field solver

The electrostatic potential ϕ is determined self-consistently from δf through the gyrokinetic quasineutrality condition. In Fourier space, for each perpendicular wavevector $\boldsymbol{k}_\perp=(k_x,k_y)$ and parallel position s,

$$\sum_a \frac{Z_a^2\,n_a}{T_a}\,(\Gamma_0^a-1)\,\hat{\phi}(\boldsymbol{k}_\perp,s) \;=\; \sum_a Z_a\int J_0^a\;\delta f_a\;B\;\mathrm{d}v_\parallel\;\mathrm{d}\mu, \tag{12}$$

where the sum runs over species a with charge number Z_a, density n_a, and temperature T_a. The operator $J_0^a=J_0(k_\perp\rho_a)$ is the zeroth-order Bessel function that performs the gyro-average, corresponding to the operator $\mathbf{J_0}$ appearing in Equation (13). The quantity $\Gamma_0^a=I_0(b_a)\,e^{-b_a}$ with $b_a=k_\perp^2\rho_a^2$ is the velocity-space-integrated gyro-average, where I_0 is the modified Bessel function of the first kind and ρ_a is the species gyroradius.

Equation (12) is algebraic in $\hat{\phi}$ and is solved at each evaluation of the right-hand side. In the adiabatic electron approximation used for the dataset in this work, only the ion species is evolved kinetically and the electron response is replaced by a Boltzmann relation, $\delta n_e/n_e=e\,\phi/T_e$. This simplifies the left-hand side of Equation (12) and removes the need to resolve the fast electron time scales, while retaining the essential ion-temperature-gradient-driven turbulence.

A.4 INTEGRALS AND DIAGNOSTICS

The quantities of interest are the electrostatic potential $\phi(x, s, y)$ and the scalar heat flux $Q \in \mathbb{R}$. Following Peeters et al. (2009), both are velocity-space integrals of f,

$$\boldsymbol{\phi} \;=\; \mathbf{A} \int \mathbf{J_0}\, \boldsymbol{f}\; \mathrm{d}v_{\parallel}\mathrm{d}\mu, \qquad Q \;=\; \int \mathbf{C} \int v^2\, \boldsymbol{\phi}\, \boldsymbol{f}\; \mathrm{d}v_{\parallel}\mathrm{d}\mu\; \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}s, \tag{13}$$

where $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{x \times s \times y}$ collect geometric and operating coefficients, v^2 is the pointwise kinetic energy, and $\mathbf{J_0}$ is the zeroth-order Bessel envelope. Turbulence is diagnosed through binormal-direction wavespace spectra

$$W(k_y) \;=\; \sum_{s,\, k_x} \left|\hat{\boldsymbol{\phi}}(k_x, s, k_y)\right|^2, \qquad Q(k_y) \;=\; \sum_{v_{\parallel}, \mu,\, s,\, k_x} \boldsymbol{Q}(v_{\parallel}, \mu, s, k_x, k_y), \tag{14}$$

with $\hat{\boldsymbol{\phi}}$ the Fourier-space potential and $\boldsymbol{Q}$ the heat-flux field prior to the outer integral of Equation (13). In the saturated regime the amplitude of radial transport is regulated by zonal flows at $k_y = 0$ (Itoh et al., 2006). The quantities of practical interest are time-averaged: $\bar{Q}$, $Q(k_y)$, and $W(k_y)$.

B DATA GENERATION

We reuse the dataset introduced in Paischer et al. (2025). We summarise it here. All simulations are cyclone-base-case ion-temperature-gradient runs (Dimits et al., 2000) generated with GKW (Peeters et al., 2009), parameterised by four dimensionless numbers from Section 4.1, where the main ion is deuterium and electrons respond adiabatically (i.e. they are not evolved explicitly). Data collection is performed in two passes. The first pass samples $R/L_T \in [3, 12]$, $R/L_n \in [1, 7]$, $q \in [1, 9]$, $\hat{s} \in [0.5, 5]$, together with an initial-condition noise amplitude in $[10^{-5}, 10^{-3}]$ and initial shape in $\{\sin, \cos, \text{random}\}$. Out of 100 operating points, about half are linearly stable and do not develop turbulence. Following the sparsity of turbulent points in Figure 4, the second pass narrows the ranges to $R/L_T \in [6, 12]$, $R/L_n \in [0, 2]$, $q \in [5, 9]$, $\hat{s} \in [0.5, 2]$, reducing the stabilising factors; all 200 runs of the second pass develop turbulence. After filtering, the dataset consists of roughly 250 saturated nonlinear trajectories. We hold out three for validation and six as an in-distribution test split, and retain 241 trajectories for training. Linear counterparts of each configuration are generated in parallel for the reduced-order baselines.

C SINGLE-PARAMETER SENSITIVITY SCANS

This appendix specifies the protocol for the $\bar{Q}$ scans of Figure 2.

Operating points. Starting from a fixed strongly turbulent baseline (R/L_T=10.17, R/L_n=2.61, $\hat{s}$=3.08, q=4.57), we sweep one of $\{R/L_T,\ R/L_n,\ \hat{s},\ q\}$ at a time on a K=6-point linspace covering a wide range of that parameter (even outside the training regime), holding the other three fixed. The $\hat{s}$ axis carries two further points at negative shear (-0.30, -0.70), below the training range $\hat{s} \in [0.51, 5.00]$. The 26 scan points are disjoint from the training and test trajectories of Section B, so each scan is a single-axis OOD probe. Table 3 lists every operating point, the swept value, and the GT saturated heat flux.

Reference solver. GT $\bar{Q}$ at each scan point is obtained from gyaradax (Galletti et al., 2026b) (a GKW port) in the adiabatic-electron limit. Each run integrates for 318 time units from a $\cos^2$-shaped IC. We report the saturated mean as the average flux over the last 96 time units. Three interior scan points hit a numerical NaN under the exact baseline parameters (sensitive boundary-of-aliasing configurations); we re-ran those with $\Delta \sim 1-5\%$ jitter on the swept value (kept inside the linspace cell) and report the first stable run. One scan point ($\hat{s}$=5.00) failed to converge under any jitter inside the cell and is reported as NaN. The two negative-shear points follow the identical protocol and required no jitter.

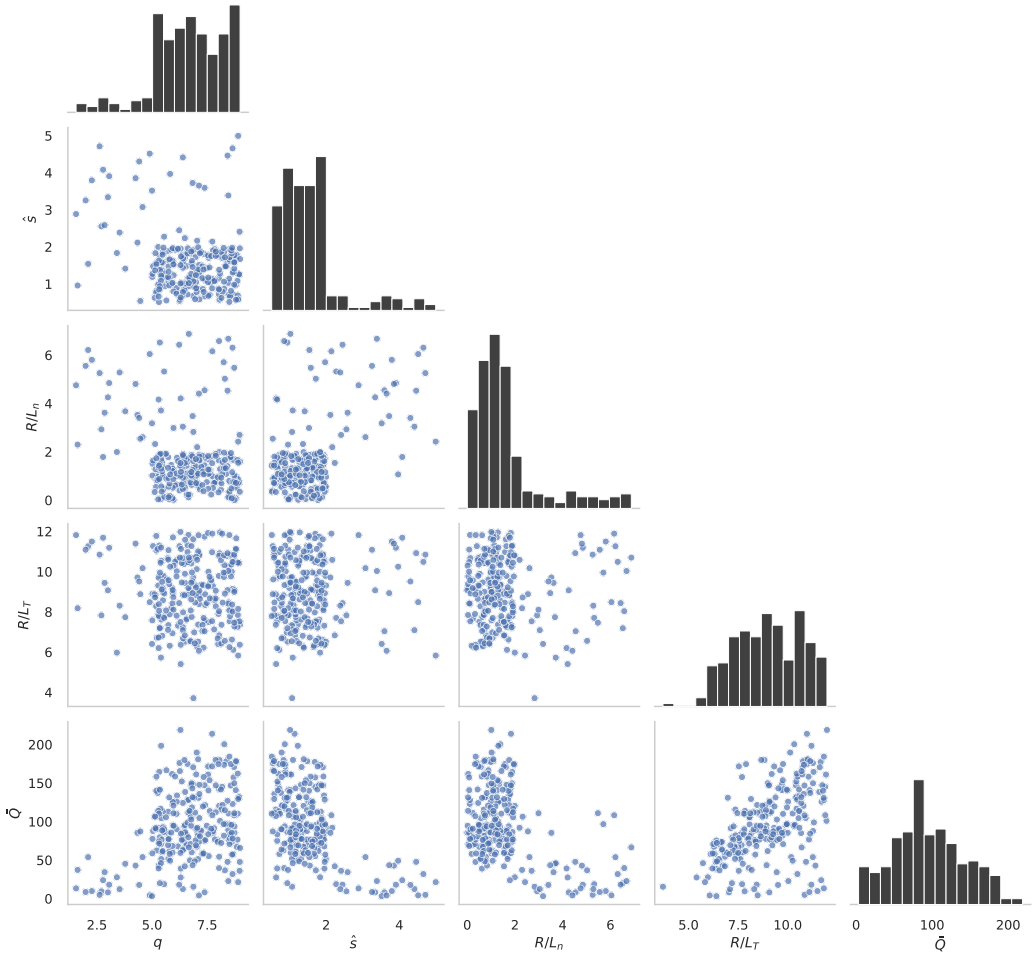

Figure 4: Distribution of the four operating parameters $\hat{s}$, q, R/L_n, R/L_T and the resulting mean heat flux $\bar{Q}$ across the dataset, after the two-pass sweep. Reproduced from Paischer et al. (2025).

Plot. Figure 2 reports the mean of the 64 model samples (green $\diamond$ markers) and the per-sample distribution as a faint cloud at the same x. GT is the saturated mean from gyaradax (dark $\circ$ markers). Rose $\triangle$ markers are a Gaussian-process surrogate regressed on $(R/L_T, R/L_n, q, \hat{s})$ from the same trajectories. With a gray shadow we report the histogram of training samples on flux and parameters. No value is clipped: the R/L_T axis is floored at -30 and the $\hat{s}$ axis capped at 180 for legibility, and a GPR point below the floor is drawn on the axis edge as $\triangledown$. Shading marks unphysical negative flux and, on the $\hat{s}$ axis, the negative-shear region.

D Training and architecture details

This appendix collects the architectural and optimisation hyperparameters of GyroFlow and the three generative baselines introduced in Section 5. Unless stated otherwise, all models share the same Swin5D backbone of Section 4 and differ only in their bottleneck and the way the operating parameters $\boldsymbol{c}$ enter. In Table 4 the reconstruction RMSE of the autoencoders is listed, as well as the count of the parameters.

Table 3: All 26 scan points and their gyaradax saturated heat flux $\bar{Q}_{\mathrm{GT}}$. Each block sweeps one operating parameter across the empirical min/max of the training corpus while the other three are held at the baseline; the $\hat{s}$ block is extended below the training range and so carries two more points than the others.

point	R/L_T		R/L_n		$\hat{s}$		q	
	swept value	$\bar{Q}_{\mathrm{GT}}$	swept value	$\bar{Q}_{\mathrm{GT}}$	swept value	$\bar{Q}_{\mathrm{GT}}$	swept value	$\bar{Q}_{\mathrm{GT}}$
0	1.060	0.00	0.004	45.69	−0.700	59.73	1.000	0.00
1	3.244	0.00	1.401	51.71	−0.300	83.72	2.598	25.35
2	5.428	3.12	2.798	50.52	0.510	101.33	4.196	43.75
3	7.612	18.38	4.196	41.88	1.408	83.35	5.794	58.23
4	9.796	43.98	5.593	29.55	2.306	78.97	7.392	50.33
5	11.980	85.81	6.990	18.56	3.204	47.05	8.990	48.41
6					4.102	33.57		
7					5.000	NaN		

D.1 SWIN5D AUTOENCODER

The encoder E_ψ and decoder D_ψ use a hierarchical Swin5D backbone of depth 8 with 16 attention heads per layer. Attention is computed locally over non-overlapping windows of size $M_{v_\parallel} \times M_s \times M_x \times M_y = 4 \times 4 \times 9 \times 4$, alternating standard and shifted-window partitions across consecutive blocks. The μ axis is folded into the channel dimension after a learned 1D positional embedding rather than attended over directly, which keeps attention 4-dimensional while still resolving the variable-amplitude μ shells. The bottleneck consists of two Transformer layers with four attention heads operating at the coarsest resolution after a linear down-projection to channel width $d_z = 256$. Both E_ψ and D_ψ are unconditional. Training uses an MSE reconstruction loss on the 2-channel real-space representation of f described in Section 4.

D.2 STABILISATION OF 5D TRAINING

Magnetic moment normalization. The magnetic moment coordinate μ indexes energy shells whose amplitudes span several orders of magnitude. A global z-score collapses small-μ contributions, whose amplitudes are orders of magnitude smaller than dominant shells. Therefore, instead of normalizing across channels as in Paischer et al. (2025) we normalize over μ via $\bar{f}_\mu, \sigma_{f_\mu}$ over all other axes. The resulting input to the model is then computed as $\tilde{f}(\cdot, \mu, \cdots) = (f - \bar{f}_\mu)/(\sigma_{f_\mu} + \varepsilon)$.

Query-Key normalization. Attention logits on 5D grids are prone to scale drift as the model size grows. Inspired by the language modelling community (Henry et al., 2020; Chameleon Team, 2025), RMS normalization is applied independently to the query and key tensors along the head dimension before the dot product, so that logit magnitudes are controlled independently of the feature norm.

Gated attention. Also following findings in Qiu et al. (2025), we apply a head-wise sigmoid gate to the attention output prior to the output projection. Let $o_h \in \mathbb{R}^{d_h}$ be the raw attention output of head h and q_h its query. We replace o_h with $\tilde{o}_h = \sigma\big(W_g \,\mathrm{ReLU}(q_h)\big) \odot o_h$, where $W_g \in \mathbb{R}^{d_h \times d_h}$ is a per-head learnable projection and $\sigma(\cdot)$ the sigmoid. The gate modulates each head contribution conditionally on its query. Qiu et al. (2025) report that this stabilizes attention logits and yields per-head sparsity.

D.3 LATENT GENERATIVE MODELS

All four generative models reuse the autoencoder of Section D.1; they differ in their bottleneck and in how $\boldsymbol{c}$ is injected.

VAE. A reparameterised Gaussian bottleneck with channel width $d_z = 256$. The KL weight is annealed cyclically from 0 to 0.2 over four cycles to mitigate posterior collapse, and the posterior log-variance is clamped to $[-20, 20]$ to avoid degenerate encodings. The encoder is unconditional;

Table 4: Reconstruction RMSE in physical (denormalised) units, evaluated for each autoencoder backbone underlying the latent generative models. Parameter counts in millions (M). ID and OOD entries denote the mean ± standard deviation across trajectories of the corresponding test split.

Method	Params (M)	$Recon_{\mathrm{RMSE}} \downarrow$		$\bar{Q}_{\mathrm{RMSE}} \downarrow$	
		ID	OOD	ID	OOD
AE	217.6	$0.753_{\pm 0.175}$	$0.786_{\pm 0.257}$	$4.76_{\pm 2.15}$	$7.12_{\pm 7.35}$
VAE	217.9	$0.839_{\pm 0.171}$	$0.869_{\pm 0.249}$	$3.82_{\pm 0.654}$	$6.92_{\pm 6.39}$
VQ-VAE	225.8	$0.888_{\pm 0.187}$	$0.924_{\pm 0.273}$	$21.6_{\pm 9.7}$	$23.5_{\pm 13.4}$

the decoder receives $\boldsymbol{c}$ through per-block adaptive layer normalization (adaLN) (Peebles & Xie, 2023). At inference, $z \sim \mathcal{N}(0, I)$ is decoded with the test-time $\boldsymbol{c}$.

VQ-VAE. A vector-quantised bottleneck with an EMA-updated codebook of 8192 entries, embedding dimension 256, and commitment loss weight 0.25. The conditioning structure mirrors the VAE (unconditional encoder, adaLN-conditioned decoder). At inference, code indices are drawn independently from the empirical codebook distribution estimated on the training set, and decoded jointly with $\boldsymbol{c}$. For the vector quantization we were using the implementation from `vector-quantize-pytorch`[1].

VQ-VAE + Transformer. The same VQ-VAE bottleneck, paired with a 12-layer, 16-head GPT-style causal Transformer with $d_{\text{model}} = 1024$ trained on the code sequences. The cross-entropy objective uses label smoothing 0.1, and $\boldsymbol{c}$ is injected per layer via FiLM (Perez et al., 2018). Sampling is autoregressive with KV-caching, and the resulting index sequence is decoded by the frozen VQ-VAE decoder under the same $\boldsymbol{c}$.

GyroFlow. A plain autoencoder (no KL, no quantisation) with an unconditional encoder and decoder, and a DiT that operates directly on the bottleneck tokens with no additional patching. The flow time t and the four entries of $\boldsymbol{c}$ are each lifted through a sinusoidal embedding and a shared MLP to a joint conditioning vector that modulates each DiT block via adaLN. Residual gates are zero-initialised so the DiT begins training as the identity. Architectural hyperparameters of the DiT are: depth 12, 16 attention heads, $d_{\text{model}} = 1024$, MLP ratio 2.0, token input/output dimension $d_z = 256$ (matching the AE bottleneck), and drop-path 0.1. The flow time t and the four entries of $\boldsymbol{c}$ are each lifted to 128 dimensions through a sinusoidal embedding ($\omega_k = 10^{-4k/K}$) followed by a SiLU MLP, concatenated, and broadcast to every block as the adaLN modulation signal.

D.4 Optimisation

The DiT and the autoregressive prior are trained with AdamW; the Swin5D autoencoders (AE, VAE, VQ-VAE) use Adam (PyTorch `weight_decay` added to the gradient, not decoupled). All runs use $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 10^{-6}, gradient clipping at norm 1.0, and bf16 mixed precision. Learning rates and schedules differ across components and reflect the regimes in which each was found stable (Table 5). Table 5 also lists batch sizes and total epochs.

Table 5: Optimisation hyperparameters for the autoencoder, the autoregressive prior over VQ-VAE codes, and the DiT used in GyroFlow. All runs use AdamW, weight decay 10^{-6}, gradient clipping at 1.0, and bf16 mixed precision.

Component	Batch size	Epochs	Peak lr	Schedule
Swin5D autoencoder (all variants)	512	400	$3{\cdot}10^{-4}$	cosine
VQ-VAE Transformer prior	128	1000	$1{\cdot}10^{-3}$	OneCycle
GyroFlow DiT	128	1000	$5{\cdot}10^{-4}$	cosine

For all autoencoders we observe early-training instabilities driven by large μ-shell amplitude imbalance; the per-shell normalization of Section 4 eliminates them. Without it, training of the larger

[1] https://github.com/lucidrains/vector-quantize-pytorch

Swin5D variants diverges within the first epoch. QK normalization and gated attention contribute smaller but consistent improvements at scale, in line with the LLM literature (Henry et al., 2020; Qiu et al., 2025).

D.5 Flow-matching specifics

The DiT in GyroFlow is trained with the rectified flow-matching objective of Equation (4). As detailed in Section E, we draw the integration time from a logit-normal distribution ($\tau \sim \mathcal{N}(0,1)$, $t = \sigma(\tau)$) and pair noise with data within each minibatch via the optimal-transport assignment of Equation (16), solved with the Hungarian algorithm at $\mathcal{O}(B^3)$ cost. Latents are rescaled by their average per-element standard deviation σ_z, estimated once over the training set, before flow matching. At inference we use $N = 15$ explicit-Euler steps on a uniform grid in $[0,1]$.

D.6 Compute

All models are trained on NVIDIA GH200 GPUs, using bf16 mixed precision throughout. Table 6 summarises the per-component compute budget. Inference cost per sample is reported in Table 1.

Table 6: Compute used for each training run, on NVIDIA GH200 120GB (bf16 mixed precision throughout). Wall-clock is end-to-end. The DiT and AR runs are single-GPU; the autoencoders use multi-node data parallelism (DeepSpeed ZeRO-2 for the AE, PyTorch DDP for the VAE/VQ-VAE).

Run	GPUs	Parallelism	Wall-clock	GPU-hours	Epochs
GyroFlow AE	32	DeepSpeed ZeRO-2	19 h	$\sim$600	400
VAE	64	DDP	11 h	$\sim$695	400
VQ-VAE	64	DDP	9 h	$\sim$580	400
VQ-VAE Transformer prior (AR)	1	–	19 h	$\sim$19	1000
GyroFlow DiT	1	–	24 h	$\sim$24	1000

E Generative model details

This appendix expands the flow-matching construction of Section 4 and gives the pragmatic variance-reduction choices that make training on 5D plasma latents stable.

Denoising diffusion models (Ho et al., 2020; Song & Ermon, 2020; Dhariwal & Nichol, 2021) sample from a target distribution by inverting a forward noising process and dominate content generation across images, video, and audio (Yang et al., 2023). Deterministic samplers (Song et al., 2021; Karras et al., 2022) and continuous-time ODE reformulations (flow matching (Lipman et al., 2023), stochastic interpolants (Albergo et al., 2025)) have cut integration steps and improved quality. The framework of Albergo et al. (2025) unifies flow and diffusion by bridging two densities in finite time through a time-indexed interpolant, recovering rectified flow (Liu et al., 2023) as a special case. Applying the generative process in the latent space of a pretrained autoencoder (Rombach et al., 2022; Dao et al., 2023) scales these models to high-resolution signals.

E.1 Stochastic interpolants and rectified flow

The stochastic-interpolant framework of Albergo et al. (2025) bridges two arbitrary densities ρ_0 and ρ_1 on $\mathbb{R}^d$ exactly in finite time through a time-indexed interpolant

$$x_t \;=\; \alpha(t)\, x_0 \;+\; \beta(t)\, x_1 \;+\; \gamma(t)\, z, \qquad t \in [0,1], \tag{15}$$

with $x_0 \sim \rho_0$, $x_1 \sim \rho_1$, and $z \sim \mathcal{N}(0, I)$ an independent latent, subject to $\alpha(0) = \beta(1) = 1$ and $\alpha(1) = \beta(0) = \gamma(0) = \gamma(1) = 0$. The time-dependent density of x_t satisfies a first-order transport equation together with a family of forward and backward Fokker–Planck equations with tunable diffusion, so the same marginal law can be realised either as an ODE or as an SDE of matching diffusion coefficient. The drift coefficients entering these equations are the unique minimisers of simple quadratic objectives estimable from samples of ρ_0 and ρ_1.

We instantiate the spatially linear one-sided interpolant of Albergo et al. (2025, §4.4), which takes $\rho_0 = \mathcal{N}(0, I)$, $\alpha(t) = 1 - t$, $\beta(t) = t$, $\gamma \equiv 0$, and selects the ODE branch with zero diffusion. Equation (15) then collapses to the straight-line path of Equation (3), and the target drift reduces to the rectified-flow velocity $u^\star(x_t \mid x_0, x_1) = x_1 - x_0$ of Liu et al. (2023). In the flow-matching formulation of Lipman et al. (2023), this is the probability path induced by a Gaussian prior and an independent-pair coupling; we extend it below with an OT coupling that tightens this pairing.

E.2 Logit-normal time schedule

Uniform sampling of $t \sim \mathcal{U}(0, 1)$ oversamples the endpoints, where the target velocity is already close to $x_1 - x_0$ (near t=1) or close to the prior mean (near t=0). The intermediate regime, where the network has to disambiguate mode structure from noise, is in contrast undersampled. Following Esser et al. (2024), we draw $\tau \sim \mathcal{N}(0, 1)$ and set $t = \sigma(\tau) = (1 + \mathrm{e}^{-\tau})^{-1}$, which concentrates training density in $t \approx 0.5$ and leaves the endpoints with lighter sampling. In practice we observe a clear reduction in training-loss variance and faster convergence at the same compute.

E.3 Optimal-transport minibatch coupling

The independent pairing $(z_0^{(i)}, z_1^{(i)})$ sampled within a minibatch induces paths that frequently cross and produce high-variance velocity targets. Following Tong et al. (2024), we permute prior samples against data latents by solving the assignment problem

$$\pi^\star \;=\; \arg\min_{\pi \in \mathfrak{S}_B} \sum_{i=1}^{B} \big\| \, z_0^{(\pi(i))} - z_1^{(i)} \, \big\|_2, \tag{16}$$

where $\mathfrak{S}_B$ is the symmetric group on B elements, and use the matched pairs $(z_0^{(\pi^\star(i))}, z_1^{(i)})$ in place of the independent ones. In practice we flatten each latent to a vector, build the Euclidean cost matrix, and solve Equation (16) with the Hungarian algorithm at $\mathcal{O}(B^3)$ cost, which is negligible compared to a forward pass of v_θ at small batch sizes. This pairing approximates the optimal-transport coupling between the empirical prior and empirical data measures, straightens the resulting paths, and shortens inference-time trajectories at fixed N.

E.4 Inference

At inference time we draw $z \sim \mathcal{N}(0, I)$ and integrate $\dot{z} = v_\theta(z, t, \boldsymbol{c})$ with an explicit Euler scheme,

$$z_{k+1} = z_k + \Delta t \, v_\theta(z_k, t_k, \boldsymbol{c}), \qquad t_k = k/N, \;\; \Delta t = 1/N, \tag{17}$$

on a uniform grid of N steps in $[0, 1]$. The combination of rectified-flow straight paths, logit-normal time sampling, and OT-coupled training is known to shorten inference trajectories at fixed sample quality (Liu et al., 2023; Esser et al., 2024; Tong et al., 2024).

E.5 Choice of integration steps

Figure 5 reports $\bar{Q}_{\mathrm{RMSE}}$ and per-sample wall-clock time as a function of the number of explicit-Euler steps N used to integrate $\dot{z} = v_\theta(z, t, \boldsymbol{c})$ at inference. Accuracy improves rapidly up to $N \approx 7$, reaches its minimum around N=15, and degrades slightly with larger N as the spread across samples grows, while runtime increases roughly linearly. We adopt $N = 15$ throughout Section 5: it sits at the elbow of the accuracy curve and keeps the cost of generation low.

F Warm starts and the Kolmogorov-Smirnov statistic

This appendix specifies the construction of the warm-start divergence reported in Table 2a and used as the y-axis of Figure 3.

F.1 Setup

Fix a held-out operating point $\boldsymbol{c}$ and let $S(t) \in \mathbb{R}^{n_{k_y}}$ denote one of the binormal spectra of Section A.4, either $W(k_y, t)$ or $Q(k_y, t)$. Saturated turbulence is statistically stationary, so along a

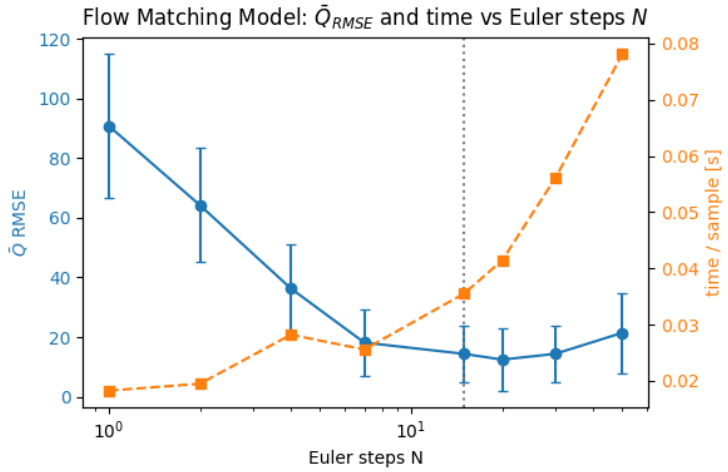

Figure 5: $\bar{Q}_{\mathrm{RMSE}}$ (blue, left axis) and per-sample wall-clock time (orange, right axis) versus the number of explicit-Euler integration steps N. Error bars denote standard deviation over the validation split. The dotted line marks our operating point at $N = 15$.

sufficiently long trajectory the empirical distribution of $\{S(t)\}$ converges to a $\boldsymbol{c}$-dependent stationary law $\mathbb{P}_{\boldsymbol{c}}$ on $\mathbb{R}^{n_{k_y}}$. We compare two empirical estimates of $\mathbb{P}_{\boldsymbol{c}}$:

- the warm-side sample $\mathcal{S}^w_{\boldsymbol{c}}$ obtained by initializing gyaradax from generated fields $\hat{f}^{(k)} \sim p_\theta(\cdot \mid \boldsymbol{c})$ drawn from GyroFlow and integrating for T steps;
- the reference sample $\mathcal{S}^r_{\boldsymbol{c}}$ obtained from the post-saturation tail of a long trajectory at the same $\boldsymbol{c}$.

Under the null hypothesis that $\hat{f}^{(k)}$ already lies on the attractor of $\mathbb{P}_{\boldsymbol{c}}$, the two samples are i.i.d. from the same law and any statistical test should fail to reject equality.

F.2 Two-sample Kolmogorov-Smirnov statistic

We score the agreement between $\mathcal{S}^w_{\boldsymbol{c}}$ and $\mathcal{S}^r_{\boldsymbol{c}}$ with the two-sample Kolmogorov-Smirnov (KS) statistic (Massey, 1951), the classical non-parametric measure of distributional discrepancy. Given samples $\mathcal{S}^w = \{x_i\}_{i=1}^n$ and $\mathcal{S}^r = \{y_j\}_{j=1}^m$ with empirical CDFs $F^w_n(x) = \frac{1}{n}\sum_i \mathbf{1}_{x_i \le x}$ and $F^r_m(x) = \frac{1}{m}\sum_j \mathbf{1}_{y_j \le x}$, the statistic is the supremum of their pointwise gap,

$$D_{n,m}(\mathcal{S}^w, \mathcal{S}^r) \;=\; \sup_{x \in \mathbb{R}} \left| F^w_n(x) \;-\; F^r_m(x) \right| \;\in\; [0, 1]. \tag{18}$$

$D_{n,m} = 0$ corresponds to identical empirical CDFs and $D_{n,m} = 1$ to disjoint supports. Under the null $\mathbb{P} = \mathbb{Q}$, the rescaled statistic $\sqrt{nm/(n+m)}\, D_{n,m}$ converges weakly to the Kolmogorov distribution and admits a distribution-free p-value; we report D rather than the p-value because p collapses below numerical resolution at our pooled sample sizes ($\sim 10^4$). D requires no kernel or bandwidth choice, is monotone in distributional discrepancy, and is computed in $\mathcal{O}((n+m)\log(n+m))$ via sorted-merge of the two samples. We use `scipy.stats.ks_2samp` for the per-mode evaluation.

F.3 Restart pooling

A single warm trajectory yields a strongly auto-correlated time-series whose empirical mode-marginal sits at one realisation of the attractor and may differ from $\mathbb{P}_{\boldsymbol{c}}$ by an amount comparable to the inter-trajectory variation we wish to measure. To reduce this within-condition variance we draw $R = 5$ independent generations $\{\hat{f}^{(r)}\}_{r=1}^R$ from GyroFlow at the same $\boldsymbol{c}$, integrate each for T steps

with gyaradax, and pool the post-saturation halves along the time axis,

$$\mathcal{S}_{\boldsymbol{c}}^{w} \;=\; \bigcup_{r=1}^{R} \left\{\, S^{(r)}(t) \,:\, t \in [\,T/2,\, T\,] \,\right\}, \qquad \left|\mathcal{S}_{\boldsymbol{c}}^{w}\right| \;=\; RT/2. \tag{19}$$

Pooling along time is justified by the stationarity of $\mathbb{P}_{\boldsymbol{c}}$. We additionally compute D once per restart against the same reference, and report the cross-restart standard deviation as a within-condition uncertainty for the scatter of Figure 3.

F.4 Per-mode aggregation

For vector spectra we treat $S(t) \in \mathbb{R}^{n_{k_y}}$ component-wise. At binormal mode index $\ell \in \{1, \ldots, n_{k_y}\}$ the scalar warm-side and reference samples are $\{S_\ell^w(t)\}$ and $\{S_\ell^r(t)\}$ with empirical CDFs $\hat{F}_\ell^w$, $\hat{F}_\ell^r$, and the per-mode KS statistic is

$$D_\ell \;=\; \sup_{x \in \mathbb{R}} \left|\hat{F}_\ell^w(x) - \hat{F}_\ell^r(x)\right| \in [0, 1], \tag{20}$$

the form of Equation (18) applied to a single mode. The headline reported in Table 2a is the arithmetic mean over modes,

$$\overline{D}_S(\boldsymbol{c}) \;=\; \frac{1}{n_{k_y}} \sum_{\ell=1}^{n_{k_y}} D_\ell. \tag{21}$$

Per-mode aggregation makes the statistic insensitive to the absolute amplitude scale across k_y, which would otherwise be dominated by the high-energy low-k_y end. Each D_ℓ is computed via `scipy.stats.ks_2samp`.

G Sensitivity to the initial condition

We evolve 20 operating points, spanning $R/L_T \in [3.0, 11.8]$, $R/L_n \in [0.1, 6.4]$, $\hat{s} \in [0.6, 3.1]$ and $q \in [1.1, 9.0]$, with gyaradax from three initial conditions: a $\cos^2$ profile at amplitude 10^{-4}, Gaussian noise at $3{\cdot}10^{-5}$ and a sine profile at 10^{-3}. Each run covers 40,080 steps (400.8 time units) and statistics are computed over the final 96 time units. One operating point is linearly stable and is excluded. Table 7 lists the remaining 19 points.

The cross-seed standard deviation of the mean heat flux is $10.6 \pm 6.2\%$. For comparison, the standard error of a single 96-unit average, measured from consecutive windows of the same run, is 13%, since the flux decorrelates over $\approx$28 time units. Differences between initial conditions do not persist into the next averaging window ($r = -0.17$). Fitting the spread at different window lengths gives an upper bound of 4% on a persistent difference between initial conditions. The time-averaged k_y spectra of the three runs differ by $3.2 \pm 1.2\%$.

An autocorrelation-based estimate on the reference series of Table 1 (223 time units; 77 for the resumed OOD points) gives a standard error of 9–10% for the ID and 6.5% for the OOD references. The OOD series are short, so the latter estimate is likely a lower bound. The experiment tests the sensitivity to the initial perturbation; it does not rule out bifurcations between distinct saturated states.

H Heat-flux saturation of warm-started runs

For each of 33 operating points (31 ID test and two validation points), we run gyaradax from three GyroFlow samples and from a cold start with a $\cos^2$ perturbation of amplitude 10^{-4}. Each run lasts 400.8 time units. We save the heat flux every 1.2 time units. As the reference, we use the last 80 saved heat-flux measurements of the GKW simulation at the same operating point, covering 96 time units. All times below are in R/v_{th}.

Heat-flux criterion. At each saved time t_j, we average the last 20 heat-flux measurements and check whether this mean lies inside the reference 95% band:

$$\widetilde{Q}_j = \frac{1}{20} \sum_{r=0}^{19} Q(t_{j-r}), \qquad q_{0.025}^{\mathrm{ref}} \le \widetilde{Q}_j \le q_{0.975}^{\mathrm{ref}}. \tag{22}$$

Table 7: Sensitivity of the saturated statistics to the initial condition: 19 turbulent operating points (dataset sweep ids of Section B), each evolved with gyaradax for 400 time units from three seeds ($\cos^2$ at 10^{-4}, Gaussian noise at $3{\cdot}10^{-5}$, sine at 10^{-3}). $\bar{Q}$ over the final 96-unit window per seed; cross-seed standard deviation of the window mean in % of the mean; standard error of one 96-unit window mean of a single run, measured from consecutive windows, in %; maximum relative log-L_2 difference of the seeds' time-averaged k_y spectra from their ensemble mean, in %. A 20th point (id 35) is linearly stable and produces no flux.

id	$\bar{Q}$, $\cos^2$ seed	$\bar{Q}$, noise seed	$\bar{Q}$, sine seed	cross-seed sd [%]	window SE [%]	k_y spectra [%]
84	16.4	16.2	18.2	6.3	8.4	4.1
7	16.7	21.5	22.0	14.5	11.8	6.0
51	28.1	28.7	29.7	2.8	5.3	3.0
13	48.2	48.4	50.1	2.1	2.7	1.9
107	56.2	54.6	51.8	4.0	17.4	2.2
200	77.3	66.9	81.7	10.1	7.1	3.0
120	76.4	71.4	79.4	5.3	8.6	5.6
221	75.2	98.0	82.7	13.6	11.1	3.1
222	106.3	80.8	85.8	14.8	11.9	3.2
207	93.6	95.6	100.9	3.9	19.2	2.5
171	85.1	112.9	97.5	14.1	24.9	1.7
154	117.7	93.8	122.4	13.8	23.2	3.1
198	114.2	134.5	103.8	13.3	18.8	2.3
129	175.4	111.6	115.9	26.5	23.3	4.3
205	123.2	134.2	158.8	13.2	9.5	2.9
229	144.2	142.9	148.9	2.2	7.3	2.4
218	170.4	130.0	138.8	14.5	12.9	4.1
143	168.2	147.6	128.9	13.2	6.3	3.2
295	134.8	172.8	142.1	13.5	8.5	2.4
mean $\pm$ sd over operating points				10.6 ± 6.2	12.5	3.2 ± 1.2

Here q_p^{ref} is the p-quantile of the reference flux, so the band contains its middle 95% of values. We require six consecutive checks to pass and report the first of those times. The first 20-measurement window ends at $t = 24$ and can be confirmed at $t = 30$. Thus 24 is the earliest time we can report, even if the flux agrees with the reference from the start.

Flux error. We compare the flux over 96-unit windows of 80 measurements, matching the reference sample count, using all runs. For each run, the absolute relative error of its mean flux is

$$\mathrm{Rel}_Q = 100\,\frac{|\langle Q\rangle_{\mathrm{window}} - \bar{Q}_{\mathrm{ref}}|}{|\bar{Q}_{\mathrm{ref}}|}. \tag{23}$$

We also compute KS_Q, the KS distance between the window's flux values and the reference values (Equation (18)). We average both quantities over the three warm runs at each operating point, then over the 33 points.

Results. All runs meet the flux criterion. Warm starts meet it before their paired cold run in 92 of 99 cases; 74 meet it at the first window, compared with none of the cold runs (Table 8(a)). The median cold-to-warm time ratio is 2.95 (95% interval $[2.30, 3.75]$), a 66.1% reduction in time ($[56.5, 73.3]$%). We compute the ratio for each pair before taking the median, so it differs from $90/24$.

Over the first 96 time units, warm starts reduce the mean relative flux error by 62.1% (95% interval $[53.2, 69.8]$%) and the flux KS distance by 44.0% ($[36.3, 50.7]$%). All intervals use 10,000 paired bootstrap resamples of operating points. Each point's three warm runs are kept together, and the reference data are held fixed.

Table 9 shows how the errors change over the run. In the final 96 time units, the mean relative flux error is 17.0% for cold starts and 22.8% for warm starts. The improvement is therefore in the initial transient.

Table 8: Warm and cold gyaradax runs at 33 operating points, with one cold and three warm runs per point. The flux criterion requires the mean of the last 20 measurements (24 time units) to stay inside the reference 95% band for six consecutive checks. We report the first of these times: the minimum is 24, confirmed at 30. Rel_Q is the mean absolute relative flux error and KS_Q the flux KS distance. Times are in R/v_{th}; lower errors and distances are better.

	Cold start	GyroFlow warm start
(a) Heat flux during the transient		
Flux criterion met at the first window ($t = 24$)	$0/33$	$74/99$
Median time to meet the flux criterion	90	24
Flux criterion first met by $t = 96$	$24/33$	$94/99$
Mean flux error Rel_Q, first 96 units	47.9%	18.2%
Flux KS distance KS_Q, first 96 units	0.735	0.412
(b) Late-time heat flux		
Mean flux error Rel_Q, final 96 units	17.0%	22.8%

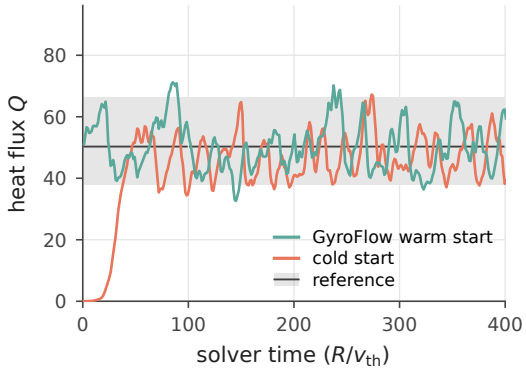

Figure 6: Heat flux of gyaradax runs at the baseline operating point of the parameter scans (Section C), started from a GyroFlow sample (warm start) and from a small initial perturbation (cold start). The reference is the last 80 saved heat-flux values (96 time units) of the GKW simulation at the same operating point. The grey band spans their 2.5th to 97.5th percentiles, and the horizontal line is their mean. The warm start begins near the saturated flux, while the cold start first passes through the linear growth phase. With the flux criterion of Table 8, the warm start is saturated after 24 time units (the earliest reportable time) and the cold start after 52; over all 33 operating points the medians are 24 and 90.

I Distributional metrics: Wasserstein and FGyD

This appendix lists the four GyroSwin U-Net depths we tracked during development, three of which are reported in the main paper, and addresses the statistical reliability of FGyD at the sample counts we use.

I.1 Wasserstein distance

We use the Wasserstein distance to compare the time-averaged spectra $W(k_y)$ across methods in Table 10. It measures the minimum cost of transporting probability mass between two distributions, with cost proportional to the distance the mass is moved, and stays well-defined when the two supports do not overlap, which is useful when generated and reference spectra peak at different k_y. We normalise each spectrum so that its total sum is one, ensuring it represents a probability distribution over k_y before computing the distance. Formally,

$$W_p(P,Q) \;=\; \left(\inf_{\gamma \in \Gamma(P,Q)} \int \|x-y\|^p \, d\gamma(x,y) \right)^{1/p},$$

Table 9: Mean absolute relative flux error Rel_Q and flux KS distance KS_Q in 96-unit windows, over all 33 cold and 99 warm runs. Four consecutive windows fit into the 400.8-unit run. The final window covers the last 96 units (304.8–400.8), overlapping the fourth. Lower is better.

Solver-time window	Rel_Q [%]		KS_Q	
	cold	warm	cold	warm
$(0, 96]$	47.9	18.2	0.735	0.412
$(96, 192]$	18.7	16.5	0.465	0.404
$(192, 288]$	22.5	18.9	0.457	0.454
$(288, 384]$	15.3	23.3	0.392	0.494
$(304.8, 400.8]$ (final window)	17.0	22.8	0.418	0.484

where $\Gamma(P, Q)$ is the set of couplings of P and Q. We use $p = 1$ on the 1D index k_y. FGyD (Section 4.5, Equation (5)) is itself a Wasserstein-2 distance between Gaussian fits of the GyroSwin activations.

I.2 Spectral RMSE and Wasserstein distance

Table 10 complements the Pearson-correlation columns of Table 1 with two additional spectral fidelity metrics: pointwise RMSE on the unnormalised amplitudes of $\langle W(k_y)\rangle$ and $\langle Q(k_y)\rangle$, and Wasserstein distance on each (computed after normalising every spectrum to a probability distribution over k_y, see Section I.1). RMSE penalises mismatched spectral magnitudes, while the Wasserstein distance is invariant to overall scale and instead penalises misplaced peaks.

Table 10: Spectral fidelity on the original six in-distribution (**ID**) and five out-of-distribution (**OOD**) operating conditions. GyroSwin (warm)* is the public checkpoint of Paischer et al. (2025) and GyroSwin (cold) an earlier retraining with a modified architecture; neither is the matched pair of Table 1. We report the Pearson Correlation and Wasserstein distance on the time-averaged binormal potential spectrum $W(k_y)$ and flux spectrum $Q(k_y)$. Quasilinear and GPR baselines do not produce $W(k_y)$ pointwise estimates and only provide flux observables. Best **bold**, second underlined.

Method	$W(k_y)_{\mathrm{WD}}\downarrow$		$Q(k_y)_{\mathrm{WD}}\downarrow$		$W(k_y)_{\mathrm{PC}}\uparrow$		$Q(k_y)_{\mathrm{PC}}\uparrow$	
	ID	**OOD**	**ID**	**OOD**	**ID**	**OOD**	**ID**	**OOD**
Quasilinear	$0.0171_{\pm 0.0152}$	$0.0135_{\pm 0.0117}$	$0.0155_{\pm 0.007}$	$0.0147_{\pm 0.005}$	$0.022_{\pm 0.066}$	$0.097_{\pm 0.171}$	$0.752_{\pm 0.365}$	$0.537_{\pm 0.342}$
GPR	—	—	—	—	—	—	—	—
GyroSwin	$\underline{0.010}_{\pm 0.0076}$	$0.0140_{\pm 0.0061}$	$\mathbf{0.0065}_{\pm 0.0013}$	$\mathbf{0.0120}_{\pm 0.0075}$	$\underline{0.991}_{\pm 0.017}$	$\mathbf{0.888}_{\pm 0.179}$	$0.851_{\pm 0.187}$	$\mathbf{0.932}_{\pm 0.104}$
VAE	$0.0139_{\pm 0.0101}$	$0.0159_{\pm 0.0047}$	$0.0274_{\pm 0.0075}$	$0.0266_{\pm 0.0082}$	$0.988_{\pm 0.025}$	$0.846_{\pm 0.253}$	$0.533_{\pm 0.148}$	$0.664_{\pm 0.209}$
VQ-VAE (rand)	$0.0126_{\pm 0.0092}$	$0.0149_{\pm 0.0047}$	$0.0257_{\pm 0.0079}$	$0.0251_{\pm 0.0071}$	$0.989_{\pm 0.024}$	$0.864_{\pm 0.221}$	$0.621_{\pm 0.136}$	$0.703_{\pm 0.220}$
VQ-VAE (AR)	$0.0105_{\pm 0.0080}$	$\underline{0.0131}_{\pm 0.0057}$	$0.0136_{\pm 0.0052}$	$\underline{0.0133}_{\pm 0.0056}$	$\mathbf{0.992}_{\pm 0.019}$	$\underline{0.883}_{\pm 0.192}$	$\underline{0.891}_{\pm 0.157}$	$0.835_{\pm 0.169}$
GyroFlow (probe)	$\mathbf{0.0093}_{\pm 0.0035}$	$\mathbf{0.0083}_{\pm 0.0032}$	$0.0175_{\pm 0.0065}$	$0.0180_{\pm 0.0080}$	$\mathbf{0.992}_{\pm 0.011}$	$0.876_{\pm 0.223}$	$0.680_{\pm 0.184}$	$0.691_{\pm 0.207}$
GyroFlow (decode)	$0.0123_{\pm 0.0104}$	$0.0149_{\pm 0.0050}$	$\underline{0.0077}_{\pm 0.0021}$	$0.0137_{\pm 0.0041}$	$0.989_{\pm 0.024}$	$0.859_{\pm 0.231}$	$\mathbf{0.938}_{\pm 0.105}$	$\underline{0.874}_{\pm 0.130}$

I.3 Latent depths

The frozen GyroSwin encoder g_ξ of Section 4.5 is a hierarchical Swin5D U-Net composed of three down-up sampling stages with skip connections, a Transformer bottleneck, and an auxiliary decoder that reconstructs the electrostatic potential ϕ. We extract activations at four depths.

- **skip** L=1: the skip connection one level above L=2 (appendix tables only).
- **skip** L=2: the bottleneck-adjacent middle block.
- ϕ **decoder**: the bottleneck-adjacent activation taken from the auxiliary ϕ-reconstruction branch rather than the f-reconstruction branch.
- **flux head**: the activation immediately preceding GyroSwin's scalar flux readout.

Skip L=2, the flux head, and the ϕ decoder constitute the headline of Table 2b.

I.4 SAMPLE-COUNT REGIME

FGyD of Equation (5) is the squared 2-Wasserstein distance between Gaussian fits to two empirical activation clouds. The empirical estimator is biased: with K generated and N_{ref} reference samples and latent dimension D, the bias of $\widehat{\Sigma}$ scales like $D/\min(K, N_{\text{ref}})$ and the bias of the trace term in Equation (5) therefore decays as $\mathcal{O}(1/\min(K, N_{\text{ref}}))$ with a generator-dependent prefactor (Bińkowski et al., 2018; Chong & Forsyth, 2020). Standard image benchmarks resolve this by sampling $N_{\text{ref}} \geq 50{,}000$ (Heusel et al., 2017); in our setting $K = 64$ and $N_{\text{ref}} \sim 80$, so the absolute values of FGyD are biased and serve only as relative rankings at fixed depth, not as estimates of the underlying W_2^2.

I.5 RELIABILITY OF THE RANKING

We recompute FGyD for every combination of the four feature depths, $K \in \{16, 32, 64, 128\}$ generated samples and $N_{\text{ref}} \in \{20, 40, 80\}$ reference samples, with three independent subsample draws each, on the 11 original held-out points. The ranking covers six generative variants: the four methods of Table 2 and two diffusion variants of the latent model (DDPM and EDM). Table 12 reports Kendall's τ between the ranking in each setting and the ranking at the setting of Table 2 (K=64, N_{ref}=80); Table 11 lists the values at that setting for all four depths. The ranking is identical in 47 of 48 settings, the exception being the smallest budget at the flux head. Table 13 reports the $1/K$ extrapolation of Chong & Forsyth (2020), which differs from the K=64 values by at most 4%, well below the gaps between methods. Figure 3 is drawn from the per-point FGyD values of the original evaluation.

Table 11: FGyD at the canonical cell (N_{ref}=80, K=64) at all four feature depths, including two diffusion variants of the latent model (DDPM and EDM).

Method	skip L=1	skip L=2	ϕ decoder	flux head
VAE	6.201e+04	3.646e+04	8.866e+05	416.7
VQ-VAE (rand)	1.03e+04	1.895e+04	1.202e+05	109.1
VQ-VAE (AR)	7992	1.691e+04	7.974e+04	55.86
Latent DDPM	4.801e+04	1.267e+05	6.497e+05	446.6
GyroFlow	3474	1.112e+04	2.574e+05	38.24
Latent EDM	6189	1.021e+04	2.172e+05	42.05

Table 12: Rank stability of FGyD: Kendall's τ between the ranking of the six variants of Table 11 at each (N_{ref}, K) setting and the ranking at the canonical cell, per feature depth (three independent subsample draws each; 11 held-out operating points). The ranking is identical (τ=1) in 47 of 48 settings.

N_{ref}	20				40				80			
K	16	32	64	128	16	32	64	128	16	32	64	128
flux head	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ϕ decoder	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
skip L=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
skip L=2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 13: Finite-sample bias of FGyD: value at $K{=}64$ versus the $1/K$-extrapolated FGyD_∞ (Chong & Forsyth, 2020) at $N_{\mathrm{ref}}{=}80$. The extrapolated values differ from the $K{=}64$ values by at most 4%. Values are means over three subsample draws and therefore differ slightly from the single draw of Table 11.

	skip $L{=}1$		skip $L{=}2$		ϕ decoder		flux head	
Method	$K{=}64$	∞	$K{=}64$	∞	$K{=}64$	∞	$K{=}64$	∞
VAE	6.19e+04	6.19e+04	3.65e+04	3.65e+04	8.89e+05	8.83e+05	420.9	419.0
VQ-VAE (rand)	1.03e+04	1.02e+04	1.90e+04	1.90e+04	1.21e+05	1.21e+05	108.4	108.0
VQ-VAE (AR)	7895	7775	1.68e+04	1.67e+04	7.93e+04	7.90e+04	54.9	54.5
Latent DDPM	4.78e+04	4.78e+04	1.28e+05	1.28e+05	6.49e+05	6.45e+05	451.6	449.3
GyroFlow	3569	3434	1.10e+04	1.10e+04	2.60e+05	2.59e+05	38.8	38.2
Latent EDM	6191	6049	1.03e+04	1.02e+04	2.22e+05	2.21e+05	41.3	40.9