PARADIFFERENTIAL NEURAL OPERATORS: BETTER CALL BONY FOR NON-LINEAR PDES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Operators have emerged as powerful architectures for learning mappings between infinite-dimensional function spaces. While predominantly deployed as PDE surrogates, their broader promise in data-driven AI for Science requires models capable of rigorously resolving complex non-linear dynamics. To address this, we introduce ParaNO, a Fourier-based neural operator grounded in paradifferential calculus. Crucially, ParaNO leverages Bony's decomposition to efficiently learn the cross-scale frequency interactions caused by non-linearities. Alongside our method, we introduce The Non-Linear Six (The NL6), a curated benchmark designed to systematically isolate distinct classes of algebraic non-linearities without geometric confounding. ParaNO improves over SotA baselines by 23% on Navier–Stokes and reduces errors by factors of $2\times$ to $4.6\times$ on the NL6. Our code, benchmark, and network weights will be made available upon publication.

1 INTRODUCTION

Humans strive to understand, model, and interpret the world. For centuries, Partial Differential Equations (PDEs) have served as the foundational language for modelling continuous systems (Cajori, 1928) across most scientific and engineering disciplines. For instance, they describe complex fluid phenomena such as turbulence and weather patterns, while also governing fundamental processes within electrodynamics, solid and quantum mechanics, chemistry, and robotics. Even when the underlying equations are known, most PDEs lack closed-form solutions and must be resolved through computationally expensive numerical approximations, which slow down scientific discovery. Neural Operators (NOs) have recently emerged as an efficient alternative for learning directly from data at arbitrary resolutions, as well as in complex systems where physical laws are unknown and initial conditions are scarce. They learn mappings between functions on continuous domains and generalise standard network layers by applying an integral operator with a learnable continuous kernel across the physical domain (Azizzadenesheli et al., 2024). While several methods operate directly in the physical domain (Li et al., 2021; Tiwari et al., 2025; Curvo et al., 2026; Cheng et al., 2026), popular alternatives carry out the integration in Fourier space via spectral multiplication (Li et al., 2021; Tran et al., 2023; Colagrande et al., 2026; Zhou et al., 2026).

In this work, we propose a novel NO explicitly designed to handle non-linear PDEs in Fourier domain through a theoretical grounding in Bony's *paradifferential calculus* (Bony, 1981; Métivier, 2008). This framework addresses cross-frequency modulation by recognising that not all scale interactions behave identically. It uses Littlewood-Paley decompositions (Bahouri et al., 2011) to split a nonlinear product into *paraproducts* (where low-frequency components interact with high-frequencies) and a remainder (where comparable frequencies interact). By separating these dominant multi-scale modulations from resonances, the framework effectively linearises the equations. The Paradifferential Neural Operator (ParaNO) we propose, translates this principle into deep learning. It employs these decompositions to partition the frequency spectrum into exponentially scaled *dyadic* bands, thereby isolating low-high and high-low frequency modulations from resonant mixing. To ensure computational tractability, each dyadic band is parametrised with a different learnable kernel using a low-rank separable factorisation.

While ParaNO applies to arbitrary PDEs, its structural priors are conceived for systems where spectral energy actively couples, cascades, or propagates across scales (e.g., atmospheric dynamics, turbulent fluid cascades, and acoustic scattering). In light of this, we also introduce the Non Linear

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.

Six (the NL6) benchmark, a dataset of six non-linear PDEs that exhibit these dynamics, numerically simulated on periodic grids.

Empirically, ParaNO establishes a new state-of-the-art on the inherently non-linear Navier–Stokes equations from Geo-PDE. As illustrated in Fig. 1, it reduces the error of the best-performing baseline by 23% while operating with significantly fewer parameters than most contemporary models. Furthermore, on NL6, ParaNO strictly dominates five out of six tasks. Compared to standard operators, it achieves massive error reductions across the most challenging non-linear regimes, demonstrating that our architectural priors effectively translate to empirical gains.

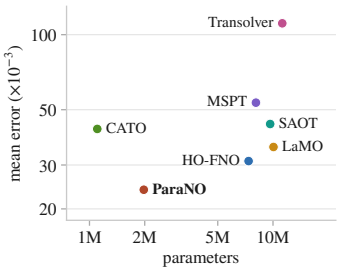


Figure 1: ℓ^2 error against model size on Navier–Stokes.

To summarise, our key contributions are:

1. **A Paradifferential Layer (ParaLayer)**, conceived to mimic physical systems by explicitly separating the evaluation of baseline linear dynamics and non-linear interactions.
2. **A theoretically motivated approach for learning the paradifferential interactions** emerging in non-linear dynamics which is grounded in Bony's decomposition.
3. **A parameter-shared kernel factorisation** that enables highly expressive, band-specific spectral responses while strictly bounding the overall parameter cost.
4. **The Non-Linear Six benchmark**, comprising six PDE-related learning tasks on periodic uniform grids. The suite combines temporal prediction with static operator evaluation, covering distinct sources of algebraic non-linear dependence within a common spatial representation.

2 Background

A common method to assess the algebraic complexity of a PDE is to assess its linearity. By observing how an unknown function and its derivatives appear, we can distinguish four types of PDEs: linear, semi-linear, quasi-linear and fully non-linear. In linear PDEs, an unknown multivariate function $u(x,t)$ (also represented with u for ease of notation) and its derivatives form a linear combination where the coefficients depend only on the independent variables (e.g., $\partial_t u - t\partial_{xx} u = 0$). In semi-linear PDEs, lower-order terms can be non-linear (e.g., $\partial_t u - t\partial_{xx} u = u^2$). In quasi-linear PDEs, the coefficients of the highest order derivatives can be functions of the unknown or its lower-order derivatives (e.g., $\partial_t u + u\partial_{xx} u = 0$). Finally, as the name suggests, the fully non-linear PDEs have non linearities also on the highest-order derivatives (e.g., $(\partial_{xx} u)^2 + \partial_t u = 0$).

To understand the theoretical implications of this classification, we must observe how PDEs behave in Fourier space. Consider, for instance, $\partial_t u - \nu\partial_{xx} u = 0$. By rewriting the unknown in terms of its Fourier series, it can be easily shown that taking the spatial Fourier transform of this linear PDE gives $\partial_t \widehat{u}(k) + \nu k^2 \widehat{u}(k) = 0$. Here, each frequency k evolves independently: the Fourier coefficients $\widehat{u}(k)$ are simply multiplied by a frequency-dependent factor νk^2. More generally, any constant-coefficient linear differential operator becomes a Fourier multiplier, $\widehat{Lu}(k) = m(k)\widehat{u}(k)$, and therefore does not create interactions between different frequencies. On the other hand, non-linear PDEs, like the quasi-linear Burgers' equation $\partial_t u + u\,\partial_x u = 0$, exhibit significant frequency interactions. In fact, just by observing its non-linear part $u\,\partial_x u$ in terms of its Fourier series, we have $u(x,t) = \sum_{k_1} \widehat{u}(k_1) e^{\mathrm{i}k_1 x}$ and $\partial_x u(x,t) = \sum_{k_2} \mathrm{i}k_2\, \widehat{u}(k_2) e^{\mathrm{i}k_2 x}$. Crucially, when multiplying the two terms, the exponentials are combined according to $e^{\mathrm{i}k_1 x} \cdot e^{\mathrm{i}k_2 x} = e^{\mathrm{i}(k_1+k_2)x}$. Through Euler's formula ($e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$), it becomes explicit that the combined exponent directly determines spatial oscillations at frequency $k_1 + k_2$. For a low-high frequency interaction, such as $k_1 = 2$ and $k_2 = 100$, the new spatial frequency of the exponential ($k_1 + k_2 = 102$) remains very close to the original high frequency $k_2 = 100$. The interaction has not created a qualitatively different spatial scale and the low-frequency component ($k_1 = 2$) merely modulates the highly oscillatory one. If we consider two high frequencies, such as $k_1 = 100$ and $k_2 = \pm 100$, their product generates $e^{\mathrm{i}200x}$ and $e^{\mathrm{i}0x} = 1$. The former creates an entirely new and much higher spatial frequency, while the latter resonant interaction produces a constant bias. Neither result can be interpreted as one slowly varying component modulating another.

This observation motivates a more structured treatment of spectral interactions in non-linear PDEs. Rather than treating the full spectral convolution as an unstructured sum of frequency pairs,

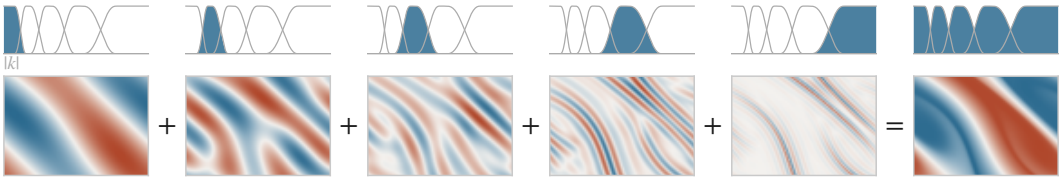


Figure 2: Dyadic Littlewood–Paley decomposition of a 2D scalar field. *Top*: Partition of the frequency domain into dyadic bands. *Bottom*: The corresponding band-pass filtered components $\Delta_j u$ isolating progressively finer spatial scales, which sum to reconstruct the original field (*right*).

paradifferential calculus organises interactions according to their relative scales (Métivier, 2008). In particular, a generic function f is Littlewood-Paley decomposed across frequency octaves as $f = \sum_{n=-\infty}^{+\infty} \Delta_n f$, where $\Delta_n f = \mathcal{F}^{-1}[\Phi_n(k)\hat{f}(k)]$ is a band-pass filtering operation using a smooth bump mask function Φ_n that becomes progressively wider in frequency space and isolates progressively finer scales (Fig. 2). By also introducing a low-pass filter $S_m f = \sum_{n=-\infty}^{m-1} \Delta_n f$ computing the sum of all frequency bands m strictly smaller than n, it is possible to define the paraproduct as $T_f g = \sum_m (S_{m-1} f)(\Delta_m g)$. This captures asymmetric interactions between slowly varying low-frequencies of f that act as a local modulating coefficient on the highly oscillatory frequencies of g. In non-linear PDEs, we can generically represent interactions between the unknown and/or its derivatives as uv. If we index the octaves of u and v with i and j respectively, Bony's decomposition then organises the product into three non-overlapping regions (Fig. 3):

$$uv = \underbrace{T_u v}_{\text{low-high } (i \le j-2)} + \underbrace{T_v u}_{\text{high-low } (j \le i-2)} + \underbrace{R(u,v)}_{\text{comparable } (|i-j| \le 1)} . \tag{1}$$

where $R(u,v) = \sum_{|i-j|\le 1} (\Delta_i u)(\Delta_j v)$ is a remainder capturing resonant interactions at comparable scales and generating entirely new macroscopic frequencies.

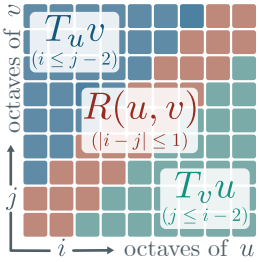


Figure 3: Bony's interacting octaves (Eq. (1)).

Neural Operators act as surrogates for an underlying PDE by predicting how an input function is mapped to its future state. However, current Fourier-based neural operators do not make any distinctions between linear and non-linear PDEs. Standard FNOs treat Fourier modes independently in the spectral layers, relying entirely on pointwise spatial activation functions to implicitly mix frequencies. While this implicit coupling is theoretically universal, it is structurally agnostic to the underlying PDE and the true spectral interactions. Conversely, higher-order models (e.g., Gao et al. (2025); Colagrande et al. (2026)) attempt generic dense coupling between all frequency pairs in the spectral domain, which scales poorly and ignores scale locality. By embedding Bony's paraproduct (Eq. (1)) directly into the NO, we introduce an inductive bias that explicitly models the frequency interactions and resonant behaviours that characterise the vast majority of governing PDEs.

3 Paradifferential Neural Operators (ParaNO)

The Paradifferential Neural Operator (ParaNO) consists of a pointwise lifting, a sequence of paradifferential layers (ParaLayers), and a final projection onto the solution space (Fig. 4). In this section, we first develop the core paradifferential branch, its spectral interaction formulation, kernel parameterisation, and latent factor splitting. Then, we outline their integration into the ParaLayer.

Learning paradifferential interactions. At the core of our Paradifferential layers lies Bony's decomposition introduced in Sec. 2: first we let frequency interactions spontaneously emerge within each of the three regimes introduced by the decomposition in Eq. (1), then we learn how to combine them in the Fourier domain.

Since terms from each component of Eq. (1) are mixed in Fourier space, it is useful to re-write them in the spectral domain. Using the generic band index m established for the paraproduct definition (Sec. 2), we transform the low-high interaction as:

$$\mathcal{F}\big[T_u v\big] = \mathcal{F}\Big[\sum_m (S_{m-1}u)(\Delta_m v)\Big] = \sum_m \mathcal{F}\big[(S_{m-1}u)(\Delta_m v)\big] = \sum_m \widehat{q}_m^{\,\mathrm{LH}}.$$

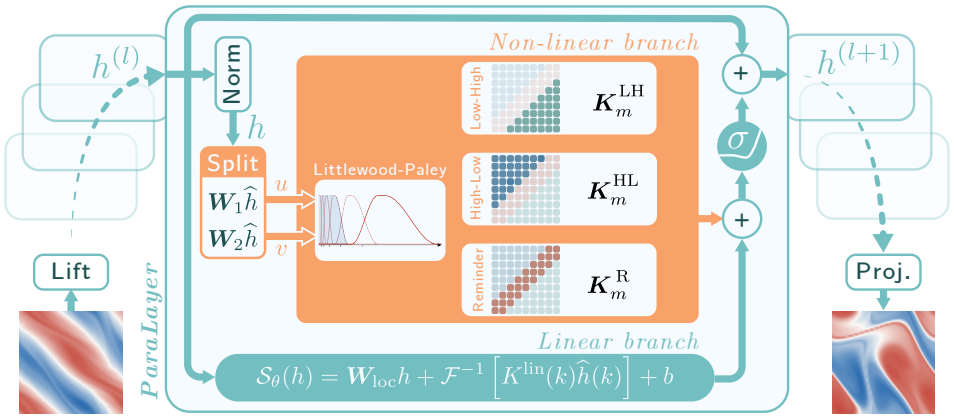


Figure 4: The ParaNO architecture.

Just as this isolates the low-high spectral contribution $\widehat{q}_m^{\mathrm{LH}} = \mathcal{F}[(S_{m-1}u)(\Delta_m v)]$, we similarly obtain the high-low $\widehat{q}_m^{\mathrm{HL}} = \mathcal{F}[(S_{m-1}v)(\Delta_m u)]$ and remainder $\widehat{q}_m^{\mathrm{R}} = \mathcal{F}\big[(\Delta_m u)\sum_{|n-m|\le 1}\Delta_n v\big]$, where we also rearrange the double sum to isolate a single outer band index m, thus ensuring a unified representation across all three terms. It is important to note that, following the product–convolution identity $\mathcal{F}[fg] = \widehat{f} * \widehat{g}$, all the three spectral contributions are effectively convolutions over frequencies of band- and low-pass filtering operations, all applied channel-wise.

By Bony's identity, a direct unweighted sum of these contributions over all dyadic bands m and interaction types $\alpha \in \mathcal{A} = \{\mathrm{LH}, \mathrm{HL}, \mathrm{R}\}$ exactly reconstructs the Fourier transform of the full physical product, $\mathcal{F}[uv]$. Rather than simply rebuilding this rigid product, we define a parametrised paradifferential branch $\mathcal{P}_\theta(u, v)$ that transforms each spectral contribution through a learned kernel before aggregation:

$$\mathcal{P}_\theta(u, v) = \mathcal{F}^{-1}\left[\sum_m \sum_{\alpha\in\mathcal{A}} K_m^\alpha(k)\, \widehat{q}_m^\alpha(k)\right]. \tag{2}$$

This construction deliberately separates the generation of new frequency interactions (handled by the underlying convolutions in $\widehat{q}_m^\alpha(k)$) from their spectral scaling. By applying a learned spectral kernel $K_m^\alpha(k)$ that mixes channels independently at each output frequency k, we grant the network the capacity to scale responses dynamically across the spectrum. Crucially, the kernel's dependence on m and α allows the model to learn distinct, frequency-dependent responses based on both the interaction scale and type, effectively transforming the branch into a highly expressive, learned paradifferential operator.

Kernel parametrisation. While Eq. (2) is entirely defined in continuous space, to introduce our learned spectral kernels $K_m^\alpha(k)$, it is useful to examine the dimensionality of the discretised counterparts. Let $u(x)$ and $v(x)$ represent continuous latent fields with C channels. Upon discretisation onto a 2D spatial grid of size $N_x \times N_y$, these fields are defined as $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{C\times N_x\times N_y}$. Consequently, when transformed to the truncated Fourier domain, the computed spectral contribution $\widehat{\mathbf{Q}}_m^\alpha$ takes the shape of a $C \times M_x \times M_y$ complex tensor. At any single discrete frequency coordinate $k = (k_x, k_y)$, this object evaluates to a simple 1D complex vector in $\mathbb{C}^C$. To mix information across these features, the corresponding kernel $K_m^\alpha(k)$ must evaluate to a dense complex weight matrix $\boldsymbol{K}_m^\alpha[k_x, k_y] \in \mathbb{C}^{C\times C}$. However, instantiating a separate $C \times C$ matrix for every discrete frequency coordinate across all the J frequency bands would impose a prohibitive parameter cost of $\mathcal{O}(JC^2M_xM_y)$. Therefore, to ensure efficiency, we instead parametrise the discrete spectral kernel using a *free modes* separable tensor factorisation (Kolda & Bader, 2009; Kossaifi et al., 2024):

$$\boldsymbol{K}_m^\alpha[k_x, k_y] = \sum_{r=1}^{R} \boldsymbol{X}_m^\alpha[r, k_x]\, \boldsymbol{Y}_m^\alpha[r, k_y]\, \boldsymbol{A}[r], \tag{3}$$

where $\boldsymbol{A} \in \mathbb{C}^{C\times C}$ is a shared channel-mixing matrix for all bands in the layer, while the 2D arrays $\boldsymbol{X}_m^\alpha \in \mathbb{C}^{R\times M_x}$ and $\boldsymbol{Y}_m^\alpha \in \mathbb{C}^{R\times M_y}$ define separate, scale-specific discrete frequency profiles

for each band m. Through this formulation, the scalar weights $\boldsymbol{X}_m^\alpha[r, k_x]$ and $\boldsymbol{Y}_m^\alpha[r, k_y]$ scale the corresponding $C \times C$ matrix slice $\boldsymbol{A}[r]$. By sharing the channel transformations across bands while learning independent spatial frequency filters, this scheme reduces the complex-valued parameter complexity for an interaction family from $\mathcal{O}(JR(C^2 + M_x + M_y))$ to $\mathcal{O}(R(C^2 + J(M_x + M_y)))$. This provides the necessary capacity to learn band-dependent pseudo-differential responses without unnecessary redundancy. Alternative parametrisation schemes exploring different capacity trade-offs are detailed and ablated in Sec. A.2.

Constructing the interacting factors. Bony's paralinearisation theorem establishes that generic PDE non-linearities can be systematically reduced to bilinear paraproducts. Classically, these interactions operate on explicit physical quantities (e.g., $u, v \in \{f, \partial_x f, \nabla f, \dots\}$). Rather than enforcing predefined physical identities on the latent channels, each layer dynamically learns two linear projections that split the incoming latent field h (with C channels) into optimal bilinear operands u and v, allowing successive layers to naturally compose these into higher-order interactions. We construct these factors directly via channel mixing in the Fourier domain:

$$\widehat{u}(k) = \boldsymbol{W}_1 \widehat{h}(k), \qquad \widehat{v}(k) = \boldsymbol{W}_2 \widehat{h}(k), \qquad \text{with } \boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{C}^{C \times C}. \tag{4}$$

Applying this linear complex projection guarantees that $\mathbf{u}$ and $\mathbf{v}$ perfectly inherit the band-limited spectral support of $\mathbf{h}$, ensuring the paradifferential decomposition operates on strictly band-preserved data.

Paradifferential Layer (ParaLayer). Having established all the core components required to ground NOs in paradifferential calculus, we now describe how to assemble them in the core building block of ParaNO, the ParaLayer. While Bony's calculus rigorously isolates the non-linear dynamics into $\mathcal{P}_\theta(u, v)$, physical systems are fundamentally driven by both non-linear interactions and baseline linear dynamics. To mirror this duality, the paradifferential branch operates in parallel with a linear spectral branch $\mathcal{S}_\theta$, defined as:

$$\mathcal{S}_\theta(h) = \boldsymbol{W}_{\text{loc}} h + \mathcal{F}^{-1}\left[K^{\text{lin}}(k)\widehat{h}(k)\right] + b. \tag{5}$$

This operation follows the standard FNO formulation (Li et al., 2021), but factorising the spectral kernel K^{lin} according to Kolda & Bader (2009), $\mathcal{S}_\theta$ closely resembles Kossaifi et al. (2024). Crucially, unlike FNO, we omit any non-linear activation within this branch, as it would inherently render the pathway non-linear, destroying its intended role as a pure linear PDE surrogate. To update the state from layer l to $l + 1$, the incoming representation $h^{(l)}$ is first pre-normalised via $h = \text{Norm}(h^{(l)})$ and then passed to both branches. The linear and non-linear outputs are fused, scaled by learned gating coefficients c_{S} and c_{P}, and passed through a generic activation function σ before updating the unnormalised residual stream:

$$h^{(l+1)} = h^{(l)} + \sigma\left(c_{\text{S}}\mathcal{S}_\theta(h) + c_{\text{P}}\mathcal{P}_\theta(u, v)\right). \tag{6}$$

This layer allows the network to explicitly balance the contribution of linear and non-linear components, seamlessly integrating the spectral pathway with our theoretically-grounded paradifferential interactions.

4 THE NL6, A BENCHMARK FOR NON-LINEAR PDES

In recent years, several benchmarks have been introduced to evaluate neural architectures on PDEs. Unfortunately, none were conceived to isolate and study the effects of algebraic complexity in the governing equations. For instance, Geo-PDE, initiated in Li et al. (2021) and later extended in Li et al. (2023c), became one of the most widely used benchmark for NOs on PDEs, but it evaluates models primarily across different geometries, with only two non-linear PDEs on a regular grid. Similarly, PDEBench (Takamoto et al., 2022) includes just one non-polynomial PDE, which is 1-D and local, while APEBench (Koehler et al., 2024) is restricted to linear and semi-linear PDEs. Comprehensive collections of simulations like The Well (Ohana et al., 2024) couple boundary conditions, multi-field interactions, and shocks, making it difficult to attribute model's gains directly to its handling of non-linearities.

To ensure that NOs' architectural advantages stem solely from a model's capacity to resolve the mathematical complexity of diverse non-linear behaviours, we introduce the Non-Linear Six (The

NL6). This benchmark comprises six learning tasks derived from non-linear PDEs simulated on periodic uniform grids to prevent geometric confounding factors. The NL6 systematically spans the standard classes of PDE non-linearities. It includes two semi-linear evolution systems: surface quasi-geostrophic turbulence (SQG), featuring bilinear transport interactions, and the Kuramoto–Sivashinsky equation (KS), featuring a quadratic gradient non-linearity. Next, it features two quasi-linear systems: shear-thinning flow (STF), governed by non-polynomial fractional gradients, and a static hyperelasticity operator (HYP), involving negative and non-integer powers. Finally, we include two highly non-linear parameter-to-state operator mappings: viscoacoustic wave propagation (VAW) and the Dirichlet–Neumann operator (DNO). While VAW and DNO simulate state evolutions that are technically linear, they represent highly non-linear mappings with respect to their heterogeneous input fields, testing a model's ability to process variable fractional orders and analytic non-polynomials, respectively. Crucially, unlike heavily diffusive PDEs that globally smooth inputs, all six tasks in the NL6 are united by their requirement to actively sustain, couple, and resolve high-frequency spectral information. A detailed description of the NL6 is provided in Sec. A.7, while a visual depiction of arbitrary target functions is depicted in Fig. 5 (*top*).

5 Experiments

Datasets. We conduct experiments across two datasets: Geo-PDE (Li et al., 2021; 2023c) and our newly proposed benchmark, The NL6 (Sec. 4). Although the PDEs in Geo-PDE are not primarily designed to isolate the algebraic non-linearities targeted by our model, we include them due to their broad adoption. In particular, we include the natively gridded Navier–Stokes (non-linear) and Darcy flow (state-linear) data, as well as the plasticity, airfoil, and pipe simulations computed on geometric meshes and subsequently interpolated onto uniform grids. Across the two datasets, tasks are formulated either as autoregressive temporal rollouts (e.g., Navier–Stokes, SQG) or as static parameter-to-state and boundary operator mappings (e.g., Darcy flow, DNO, and hyperelasticity). Detailed specifications for all tasks are provided in Sec. A.5.

Implementation Details. ParaNO's runs on all PDEs and datasets using the same hyperparameters and architecture specifications, which showcases the robustness of our method. We use depth 7 at a width bisected so that the final model has 2M parameters; for ParaNO, that means having width 83 on Navier–Stokes, 82 on Darcy, 80 on Plasticity, 74 on Airfoil, 77 on Pipe, and 83 on every task of the NL6 except Hyperelasticity (82). We train with AdamW (weight decay 10^{-5}, $\beta_2 = 0.999$) at batch size 8 for 500 epochs, and we use a standard `OneCycle` scheduler with a max learning rate of 2×10^{-3}. Both the linear and non-linear branches' kernels use a rank of 8. Normalisation and activation function within each ParaLayer are RMSNorm and GeLU, respectively. We use all modes up to Nyquist for the paraproduct branch (except for in HYP, which is capped at $|k| \leq 32$ per axis to bound memory, and still retains all but 10^{-6} of the energy), and 50% of the full Nyquist modes in each dimension for the linear branch.

Main Results. Following common practice in the NO literature (Tiwari et al., 2025; Colagrande et al., 2026), we evaluate and compare models in terms of the lowest relative ℓ_2 test errors achieved during training. Because published results on Geo-PDE are characterised by single-seed evaluations whose reported gains fall within noise margins, we re-evaluate all baselines across three random seeds, computing the mean and standard deviation. Based on a systematic review of concurrent and SotA NOs, we benchmark against the best performing models on Geo-PDE: Transolver (Wu et al., 2024), LaMO (Tiwari et al., 2025), and SAOT (Zhou et al., 2026), alongside three concurrent models introduced within months of this submission: MSPT (Curvo et al., 2026), CATO (Cheng et al., 2026), and HO-FNO (Colagrande et al., 2026). Baseline experiments on Geo-PDE retain the original codebase, hyperparameters, and parameter counts (Tab. A6) of each model. As CATO has not publicly released an official implementation at the time of writing, we faithfully reimplemented their architecture based on the paper's specification. Our re-evaluation in Tab. 1 successfully reproduces the literature, with most discrepancies strictly bounded within $[0.70, 1.30]\times$ of the originally reported metrics (see Tab. A7 for a detailed reproducibility assessment). The sole exception is Plasticity, as we were unable to reproduce the baseline metrics despite extensive efforts, we report the originally published values for this task. For completeness, Tab. A5 extends these benchmarks to include reported results from other prominent models.

Table 1: Mean and standard deviation ℓ^2 errors ($\times 10^{-3}$) over three seeds on Geo-FNO. While retrained, all baselines retain their original codebase, hyperparameters, and parameter count. Colours indicate first, second and third best results.

	Native Grids		Interpolated Geometries		
	NAVIER–STOKES	DARCY	PLASTICITY*	AIRFOIL	PIPE
Transolver	111.07 ±32.16	5.44 ±0.03	1.20	5.40 ±0.46	4.71 ±0.24
LaMO	35.43 ±3.30	4.14 ±0.19	0.70	4.58 ±0.44	4.49 ±0.33
SAOT	43.81 ±0.80	5.00 ±0.15	0.80	5.00 ±0.17	6.29 ±0.41
MSPT	53.36 ±0.86	11.14 ±0.46	1.00	6.59 ±0.34	**3.98** ±0.32
CATO	41.86 ±1.20	**3.87** ±0.05	0.50	**4.16** ±0.04	4.03 ±0.14
HO-FNO	31.11 ±0.08	5.72 ±0.45	0.60	4.86 ±0.06	5.55 ±0.20
ParaNO	**23.87** ±1.19	4.90 ±0.14	**0.39** ±0.02	5.70 ±0.19	5.45 ±0.16

Across Geo-PDE (Tab. 1), ParaNO excels on Native Grids containing active spectral mixing, improving upon HO-FNO by 23% on Navier–Stokes. However, it underperforms relative to spatial methods (like CATO and MSPT) on Darcy flow and most interpolated geometries. This is theoretically expected: although Darcy flow is a state-linear PDE used in a highly non-linear parameter-to-state mapping task, and thus apparently well posed to be solved with ParaNO, its inverse operator is intensely smoothing, meaning high-frequency interactions are exponentially damped (see Fig. A3). On the other hand, our ParaLayer is strictly engineered for systems where complex cross-frequency interactions are actively preserved, scattered, or cascaded. Similarly, for the interpolated geometries, ParaNO achieves state-of-the-art on Plasticity but struggles on Airfoil and Pipe, falling short of both spatial models and HO-FNO. Forcing these complex domains onto regular grids introduces spectral leakage that disproportionately disrupts ParaNO, as these artifacts corrupt the clean frequency bands required to compute our dyadic cross-modulations.

We next evaluate the tight cluster of top-performing models on Navier–Stokes, the sole non-linear Geo-PDE benchmark without geometric confounders. Specifically, we advance LaMO, CATO, and HO-FNO, alongside FNO$^{(+)}$, to the NL6 (Tab. 2). With FNO$^{(+)}$ we indicate a modernised FNO baseline equipped with pre-norm and residual connections (Colagrande et al., 2026) to ensure stable scaling. For these experiments, all baselines retain their Navier–Stokes architectural and optimization setups, with channel widths calibrated to enforce a uniform budget of 2M parameters. The NL6 (Tab. 2 and Fig. 5) further validates our structural hypothesis. ParaNO achieves massive gains—lowering errors by up to 4.6x over CATO on SQG and 2x over HO-FNO on KS. Bony's decomposition perfectly matches these polynomial non-linearities by neatly separating the baseline linear advection from the cascading dyadic scale interactions. Conversely, ParaNO is narrowly beaten by CATO on Hyperelasticity (HYP), as spatial attention better resolves the system's sharp, localised stress deformations (Upadhyay et al., 2026).

To verify whether ParaNO's gains stem from a structural advantage rather than parameter efficiency, we scale FNO$^{(+)}$ on DNO to 2x, 4x, and 8x capacity. While FNO$^{(+)}$'s error improves from 5.03 to 3.62 at 4M parameters, it strictly plateaus at 3.25 for 8M parameters and 3.28 for 16M parameters. In contrast, ParaNO achieves an error of 1.79 at just 2M parameters. This represents a 45% error reduction using 8x fewer parameters, confirming that ParaNO's paradifferential architecture provides a fundamental inductive bias for resolving complex algebraic non-linearities that brute-force

Table 2: Mean and standard deviation ℓ^2 errors ($\times 10^{-3}$) over three seeds across the NL6. All models are standardised to 2M parameters, but retain the hyperparameters used in Navier–Stokes. Colours indicate first, second and third best results.

	DNO	HYP	KS	SQG	STF	VAW
LaMO	7.95 ±0.23	6.86 ±0.31	57.99 ±2.19	56.54 ±3.41	38.62 ±4.32	12.59 ±0.86
CATO	4.95 ±0.61	**2.84** ±0.30	18.17 ±2.69	19.12 ±4.66	25.40 ±0.62	3.25 ±0.07
HO-FNO	5.22 ±0.25	5.52 ±0.10	15.66 ±0.87	26.89 ±10.40	14.48 ±0.38	1.66 ±0.09
FNO$^{(+)}$	5.03 ±0.12	4.65 ±0.12	17.10 ±0.62	31.31 ±6.50	17.83 ±0.16	1.90 ±0.01
ParaNO	**1.79** ±0.09	3.36 ±0.15	**7.61** ±0.15	**4.13** ±0.09	**8.48** ±0.06	**1.02** ±0.03

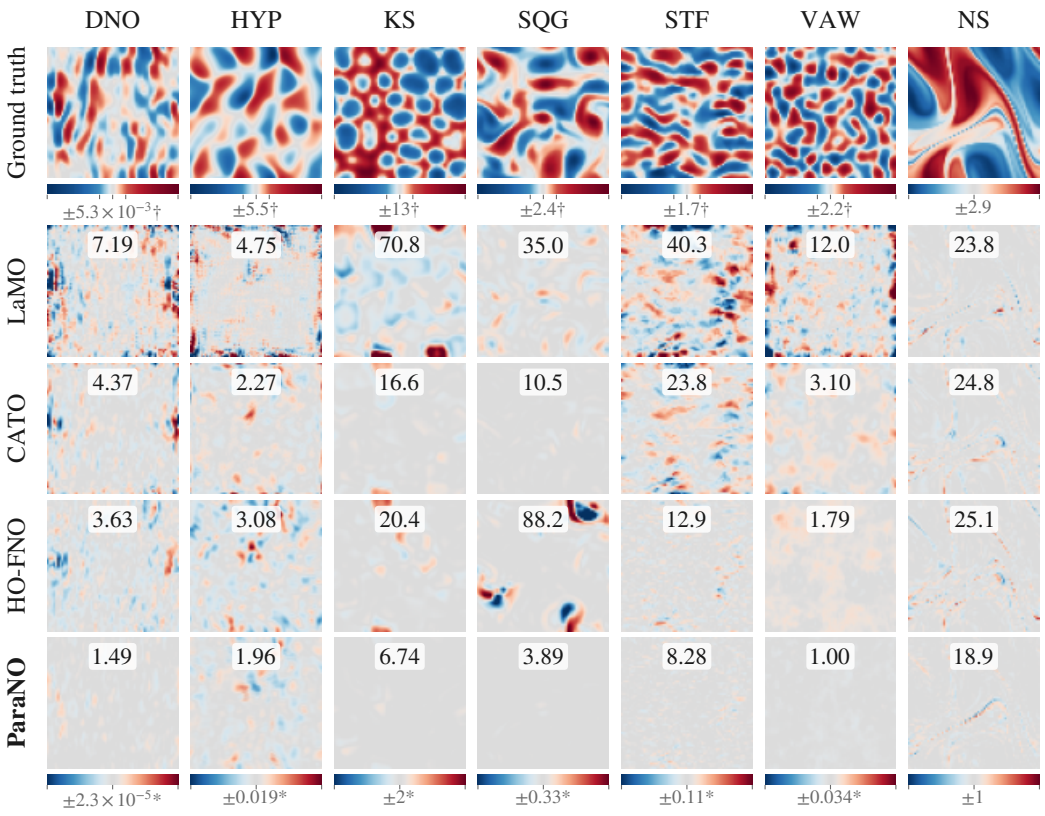


Figure 5: Error maps on the NL6 and Navier–Stokes on ParaNO's median error sample. For each PDE (*columns*), we report the Ground Truth output (*top*) and pointwise relative ℓ^2 errors across baselines (*rows*). Colour bars marked with * are clipped and scaled at 99.5%, while those with † use a symmetric logarithmic scale that is linear within 20% of the maximum absolute value. Overlaid numbers indicate per-sample relative ℓ^2 error.

scaling cannot replicate. Finally, the appendix provides a comprehensive ablation study (Sec. A.2), an experimental evaluation of band dependence (Sec. A.3), confirms ParaNO's zero-shot resolution transfer capabilities (Sec. A.8), and includes detailed spectral error analyses (Sec. A.9) alongside extended qualitative results for all baselines (Figs. A3 and A4).

6 Related work

PDE surrogate solvers have gained popularity due to their inference speed (Kurth et al., 2023; Li et al., 2023d) and their ability to learn unknown PDEs. Convolution based methods, such as TF-Net Wang et al. (2020) and U-Net (Ronneberger et al., 2015), alongside physics-informed neural networks (PINNs) (Lagaris et al., 1998; Raissi et al., 2019; Karniadakis et al., 2021), were the first models used. However, they face clear limitations: PINNs still require knowing the form of the PDE beforehand and are notoriously difficult to optimise (Krishnapriyan et al., 2021; Wang et al., 2021), while CNN surrogates remain tied to a specific grid resolution. NOs (Li et al., 2020a; Kovachki et al., 2023) address both issues: they are PDE-agnostic and resolution invariant (Kovachki et al., 2024; Azizzadenesheli et al., 2024). NOs naturally fall into two classes: models operating in physical space, and models primarily operating in frequency space. The latter extend convolutions by leveraging the Fourier transform, an integral part of the PDE analysis literature (Evans, 2010; Bahouri et al., 2011).

NOs operating in frequency space. The Fourier Neural Operator (FNO) (Li et al., 2021) relies on a spectral convolution that independently scales frequency modes (Kovachki et al., 2023).

This methodological restriction is shared by numerous follow-up works, like U-FNO Wen et al. (2022), U-NO (Rahman et al., 2023), LSM (Wu et al., 2023), and notably TFNO (Kossaifi et al., 2024), which addresses poor parameter via tensor decomposition. SAOT (Zhou et al., 2026) and AFNO (Guibas et al., 2022) apply linear attention and shared MLPs in frequency space, respectively, but their kernels remain diagonally restricted, inducing no cross-frequency coupling. DS-FNO (Gao et al., 2025) tackles this by learning dynamic kernels capable of mixing modes, but they remain unstructured and lack a theoretical underpinning. HO-FNO (Colagrande et al., 2026) instead learns dense frequency convolutions; however, their kernels mix modes uniformly without differentiating the nature of the interaction. Wavelet-based operators (MWT (Gupta et al., 2021), WNO (Tripura & Chakraborty, 2023)) decompose inputs into scale-specific bands, but apply them to linear frequencies rather than mode couplings. ParaNO, in contrast, explicitly models pairwise interactions according to Bony's decomposition. Rather than ignoring cross-frequency mixing (like FNO) or mixing modes uniformly (like HO-FNO), ParaNO groups frequencies into interacting octaves and routes them directly through the distinct algebraic regimes depicted in Fig. 3: the off-diagonal paraproducts ($T_u v$, $T_v u$) and the diagonal resonant band ($R(u, v)$). By factorising the kernel, it affords a distinct, structured response for each dyadic band while maintaining a separate linear branch.

NOs operating in physical space. Early NOs performed kernel integration directly in physical space, via graph-based message passing (Li et al., 2020a). Recently, physical space operators have shifted towards attention-based models, starting with linear-attention variants such as the Galerkin Transformer (Cao, 2021) and GNOT (Hao et al., 2023). Transolver (Wu et al., 2024) implements a kernel based on a sliced version of attention, which tackles the $\mathcal{O}(n^2)$ complexity of regular attention, infeasible for standard PDE grids. MSPT (Curvo et al., 2026) uses a hierarchical, multi-scale attention, which also lowers complexity, and balances the learning of local and non-local interactions. CATO (Cheng et al., 2026) learns a chart that maps the physical coordinates to a learnt coordinate grid, and uses them as inputs to RoPE in axially separated attention, which improves its geometric adaptability. Finally, LaMO (Tiwari et al., 2025) introduces a Mamba inspired kernel Gu & Dao (2024) that is linear in complexity. Interestingly, as highlighted by Colagrande et al. (2026), most of these methods are not resolution equivariant and do not have intrinsic super-resolution capabilities. Because these architectures focus heavily on spatial token routing rather than spectral interaction, their mechanisms are largely orthogonal to our paradifferential framework and could, in principle, be integrated into our spatial lifting and projection layers.

NOs across domains and applications. NOs have expanded beyond learning benchmark PDEs (Azizzadenesheli et al., 2024). Prominent applications include weather and climate forecasting (Bonev et al., 2023; Kurth et al., 2023), clinical MRI reconstruction (Jatyani et al., 2025), infinite-resolution image generation (Bond-Taylor & Willcocks, 2024), and many others. A growing body of research extends NOs to non-Euclidean domains: learning different geometries (Li et al., 2023d;b), extending to graphs (Li et al., 2020b), and tackling topological domains (Bastian et al., 2026). Concurrently, models outside the NO branding, like DiffusionNet (Sharp et al., 2022) and UV3-TeD (Foti et al., 2024), also share their discretisation agnostic and function-to-function properties to process non-uniformly sampled data on manifold surfaces.

7 Discussion & conclusion

We introduced ParaNO, a neural operator grounded in paradifferential calculus that isolates quadratic scale interactions into distinct dyadic bands. By coupling these scale-conditioned nonlinear responses with a dedicated linear pathway, ParaNO explicitly mirrors the algebraic structure of non-linear PDEs. Empirically, ParaNO achieves state-of-the-art results on active spectral cascades, reducing errors by up to $4.6\times$ on the NL6 suite and improving upon HO-FNO by 23% on Navier–Stokes. As theoretically expected, ParaNO yields to spatial methods on intensely smoothing operators like Darcy flow. Our ablations confirm these gains stem from our holistic architectural design rather than band dependence alone. Ultimately, these results validate structured dyadic interactions as a powerful, targeted inductive bias for algebraic non-linearities, rather than a universal replacement for geometry-aware models. Because handling irregular domains and complex boundaries is essential for real-world physical simulations, future work will focus on integrating this paradifferential framework into mesh-native or spatial attention architectures. Adapting these frequency-separation techniques to non-periodic geometries represents a critical next step for extending our algebraic priors to general-purpose, geometry-aware neural operators.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing, but no text was freely generated by AI and there was supervision on every suggested minor edit. They were also used for specific implementation tasks, such as documenting the code, vendoring baseline models, and creating plots, but all code was reviewed and nothing was generated without explicit approval. No generative AI output was used to produce experimental results, and we reviewed, tested and validated all generated code before using it. In particular, every dataset had to pass the solver verification checks described in subsection A.7. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our results, we provide the complete source code, including training scripts, model architectures, and the data generation pipeline for The NL6, in the supplementary material. A comprehensive breakdown of architectural configurations, initialization schemes, and training hyperparameters for ParaNO across all datasets is provided in Tab. A1 and Sec. A.2. For our baseline evaluations, models were re-trained across multiple seeds using their original codebases. Whenever we identified discrepancies between the reported hyperparameters in the literature and those specified within the official codebase, only the best-performing settings were reported. Detailed parameter counts and a strict reproduction fidelity assessment for these baselines are documented in Tab. A6 and Tab. A7. Finally, exhaustive descriptions of dataset generation, physical constants, and numerical validation checks for our proposed benchmark are available in Sec. A.7.

ACKNOWLEDGEMENTS

REFERENCES

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024.

Hajer Bahouri, Jean-Yves Chemin, and Raphaël Danchin. *Fourier analysis and nonlinear partial differential equations*, volume 343 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2011. ISBN 978-3-642-16829-1. doi: 10.1007/978-3-642-16830-7.

Lennart Bastian, Samuel Leventhal, Mustafa Hajij, and Tolga Birdal. Topological neural operators, 2026. URL https://arxiv.org/abs/2606.09806.

Sam Bond-Taylor and Chris G Willcocks. ∞-Diff: Infinite resolution diffusion with subsampled mollified states. In *International Conference on Learning Representations*, pp. 30132–30153, 2024.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2806–2823. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bonev23a.html.

Jean-Michel Bony. Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires. *Annales scientifiques de l'École Normale Supérieure*, 14(2):209–246, 1981. URL https://www.numdam.org/articles/10.24033/asens.1404/.

Florian Cajori. The early history of partial differential equations and of partial differentiation and integration. *The American Mathematical Monthly*, 35(9):459–467, 1928. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2298771.

Shuhao Cao. Choose a transformer: Fourier or Galerkin. *Advances in Neural Information Processing Systems*, 34:24924–24940, 2021.

Chun-Wun Cheng, Sifan Wang, Carola-Bibiane Schönlieb, and Angelica I. Aviles-Rivero. CATO: Charted attention for neural PDE operators, 2026. URL https://arxiv.org/abs/2605.09016.

Young I Cho and Kenneth R Kensey. Effects of the non-Newtonian viscosity of blood on flows in a diseased arterial vessel. part 1: Steady flows. *Biorheology*, 28(3-4):241–262, 1991.

Alex Colagrande, Paul Caillon, Eva Feillet, and Alexandre Allauzen. Higher-order Fourier neural operator: Explicit mode mixer for nonlinear PDEs, 2026. URL https://arxiv.org/abs/2606.28122.

Walter Craig and Catherine Sulem. Numerical simulation of gravity waves. *Journal of Computational Physics*, 108:73–83, 1993. doi: 10.1006/jcph.1993.1164.

Pedro M. P. Curvo, Jan-Willem van de Meent, and Maksim Zhdanov. MSPT: Efficient large-scale physical modeling via parallelized multi-scale attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12924–12933, June 2026.

Douglas G. Dommermuth and Dick K. P. Yue. A high-order spectral method for the study of nonlinear gravity waves. *Journal of Fluid Mechanics*, 184:267–288, 1987. doi: 10.1017/S002211208700288X.

Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2nd edition, 2010. ISBN 978-0-8218-4974-3. doi: 10.1090/gsm/019.

Simone Foti, Stefanos Zafeiriou, and Tolga Birdal. UV-free texture generation with denoising and geodesic heat diffusion. *Advances in Neural Information Processing Systems*, 37:128053–128081, 2024.

Wenhan Gao, Jian Luo, Ruichen Xu, and Yi Liu. Dynamic Schwartz-Fourier neural operator for enhanced expressive power. *Transactions on Machine Learning Research*, 2025.

Albert Gu and Tri Dao. Mamba: linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM 2024)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive Fourier neural operators. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EXHG-A3jlM.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. *Advances in Neural Information Processing Systems*, 34:24048–24062, 2021.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Klaus Hasselmann, T. Barnett, E. Bouws, H. Carlson, D. Cartwright, K Enke, J Ewing, H Gienapp, D. Hasselmann, P. Kruseman, A Meerburg, Peter Muller, Dirk Olbers, K Richter, W. Sell, and H. Walden. Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP). *Deut. Hydrogr. Z.*, pp. 1–95, 01 1973.

Isaac M Held, Raymond T Pierrehumbert, Stephen T Garner, and Kyle L Swanson. Surface quasi-geostrophic dynamics. *Journal of Fluid Mechanics*, 282:1–20, 1995.

Armeet Singh Jatyani, Jiayun Wang, Aditi Chandrashekar, Zihui Wu, Miguel Liu-Schiaffini, Bahareh Tolooshams, and Anima Anandkumar. A unified model for compressed sensing MRI across undersampling patterns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26004–26013, June 2025.

A. Kalogirou, E. E. Keaveny, and D. T. Papageorgiou. An in-depth numerical study of the two-dimensional Kuramoto–Sivashinsky equation. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20140932, 07 2015. ISSN 1364-5021. doi: 10.1098/rspa.2014.0932. URL https://doi.org/10.1098/rspa.2014.0932.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005. doi: 10.1137/S1064827502410633.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. *Advances in Neural Information Processing Systems*, 37:120252–120310, 2024.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500, 2009.

Jean Kossaifi, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, and Anima Anandkumar. Multi-grid tensorized Fourier neural operator for high-resolution PDEs. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=AWiDlO63bH.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Nikola B Kovachki, Samuel Lanthaler, and Andrew M Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*, 25:419–467, 2024.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021.

Yoshiki Kuramoto and Toshio Tsuzuki. Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Progress of Theoretical Physics*, 55(2):356–369, 02 1976. ISSN 0033-068X. doi: 10.1143/PTP.55.356. URL https://doi.org/10.1143/PTP.55.356.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701900. doi: 10.1145/3592979.3593412. URL https://doi.org/10.1145/3592979.3593412.

I.E. Lagaris, A. Likas, and D.I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, 1998. doi: 10.1109/72.712178.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023a. URL https://openreview.net/forum?id=EPPqt3uERT.

Zijie Li, Dule Shu, and Amir Barati Farimani. Scalable transformer for PDE surrogate modeling. *Advances in Neural Information Processing Systems*, 36:28010–28039, 2023b.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020a. URL https://arxiv.org/abs/2003.03485.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew Stuart, Kaushik Bhattacharya, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. *Advances in Neural Information Processing Systems*, 33:6755–6766, 2020b.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023c.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, et al. Geometry-informed neural operator for large-scale 3d PDEs. *Advances in Neural Information Processing Systems*, 36:35836–35854, 2023d.

Xinliang Liu, Bo Xu, Shuhao Cao, and Lei Zhang. Mitigating spectral bias for the multiscale operator learning. *Journal of Computational Physics*, 506:112944, 2024. doi: 10.1016/j.jcp.2024.112944.

Guy Métivier. *Para-differential calculus and applications to the Cauchy problem for nonlinear systems*. Number 5 in Centro di Ricerca Matematica Ennio De Giorgi (CRM) Series. Edizioni della Normale, Pisa, 2008.

David Nicholls and Fernando Reitich. Stability of high-order perturbative methods for the computation of Dirichlet–Neumann operators. *Journal of Computational Physics*, 170:276–298, 06 2001. doi: 10.1006/jcph.2001.6737.

R. W. Ogden. Large Deformation Isotropic Elasticity - On the Correlation of Theory and Experiment for Incompressible Rubberlike Solids. *Proceedings of the Royal Society of London Series A*, 326 (1567):565–584, February 1972. doi: 10.1098/rspa.1972.0026.

Raymond W Ogden. *Non-linear elastic deformations*. Dover Publications, 1997.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B Dalziel, Drummond B Fielding, et al. The Well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 37:44989–45037, 2024.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-NO: U-shaped neural operators. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=j3oQF9coJd.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. DiffusionNet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (ToG)*, 41(3):1–16, 2022.

G.I. Sivashinsky. Nonlinear analysis of hydrodynamic instability in laminar flames—I. derivation of basic equations. *Acta Astronautica*, 4(11–12):1177–1206, 1977. ISSN 0094-5765. doi: 10.1016/0094-5765(77)90096-0. URL https://www.sciencedirect.com/science/article/pii/0094576577900960.

KS Smith, G Boccaletti, CC Henning, I Marinov, CY Tam, IM Held, and GK Vallis. Turbulent diffusion in the geostrophic inverse cascade. *Journal of Fluid Mechanics*, 469:13–48, 2002.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEbench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Karn Tiwari, Niladri Dutta, N M Anoop Krishnan, and Prathosh Ap. Latent Mamba operator for partial differential equations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59639–59667. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tiwari25a.html.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized Fourier neural operators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tmIiMPl4IPa.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

Kshitiz Upadhyay, Jan N Fuhg, Nikolaos Bouklas, and KT Ramesh. Physics-informed data-driven discovery of constitutive models with application to strain-rate-sensitive soft materials. *Computational Mechanics*, 77(2):357–386, 2026.

Rui Wang, Karthik Kashinath, Mustafa Mustafa, Adrian Albert, and Rose Yu. Towards physics-informed deep learning for turbulent flow prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1457–1466, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403198. URL https://doi.org/10.1145/3394486.3403198.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43 (5):A3055–A3081, 2021. doi: 10.1137/20M1318043. URL https://doi.org/10.1137/20M1318043.

Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. U-FNO—an enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022.

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. Solving high-dimensional PDEs with latent spectral models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37417–37438. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wu23f.html.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24r.html.

Zipeng Xiao, Zhongkai Hao, Bokai Lin, Zhijie Deng, and Hang Su. Improved operator learning by orthogonal attention. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller,

Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54288–54299. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xiao24c.html.

K Yasuda, RC Armstrong, and RE Cohen. Shear flow properties of concentrated solutions of linear and star branched polystyrenes. *Rheologica Acta*, 20(2):163–178, 1981.

V. E. Zakharov. Stability of periodic waves of finite amplitude on the surface of a deep fluid. *Journal of Applied Mechanics and Technical Physics*, 9(2):190–194, 1968. doi: 10.1007/BF00913182. English translation of Zhurnal Prikladnoi Mekhaniki i Tekhnicheskoi Fiziki 9(2), 86–94.

Yanzhi Zhang, Xiaofei Zhao, and Shiping Zhou. Fourier pseudospectral methods for the variable-order space fractional wave equations. *Computer Physics Communications*, 318:109876, 2026. doi: 10.1016/j.cpc.2025.109876.

Chenhong Zhou, Jie Chen, and Zaifeng Yang. SAOT: An enhanced locality-aware spectral transformer for solving PDEs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28928–28936, 2026.

Tieyuan Zhu and Jerry M Harris. Modeling acoustic wave propagation in heterogeneous attenuating media using decoupled fractional Laplacians. *Geophysics*, 79(3):T105–T116, 2014.

A Appendix

A.1 Additional implementation details

Initialisation. As previously detailed in Sec. 3, the linear and non-linear branches within each ParaLayer are combined according to $c_{\mathrm{S}}\mathcal{S}_\theta(h)+c_{\mathrm{P}}\mathcal{P}_\theta(u,v)$, with c_L and c_N being learnable scalars and $h = \mathrm{Norm}(h^l)$ being the normalised hidden function later split into the interacting factors u and v. The non-linear branch is sensitive to the input norm and is prone to vanishing or exploding if uncontrolled, so to reduce variance, initialisation is performed by calibrating both branches. One batch is run through ParaNO, and the learned gating coefficient c_P is adjusted so that the norm of the non-linear branch's output at each layer becomes half that of the linear branch. Then, both coefficients c_P and c_S are jointly multiplied by $|\text{layer input}|/|\text{layer output}|$, clamped to $[10^{-3}, 10^3]$. All spectral components (i.e., the paraproduct and linear branch spectral kernels, as well as the factor projections that construct u and v) are initialised from a complex normal distribution. All real parameters (i.e., lift, project, and the pointwise local operation in the linear layer) are initialised from a uniform distribution.

The lift and project components of the model are one layer 1×1 convolutions. Grid coordinates are appended to the input on the four Geo-FNO benchmarks obtained by interpolating PDE initially simulated on the vertices of a mesh. On Plasticity, the 20 output timesteps are queried one at a time through a sinusoidal time embedding, with one optimiser step per query. All relevant hyperparameters are displayed in Tab. A1. Only two details differ between benchmarks: HYP being restricted to half of the Nyquist modes for a more efficient training, since only 10^{-6} of the energy sits above that cutoff, and Plasticity missing branch calibration, since it requires a time input.

Table A1: Hyperparameter and implementation details of ParaNO on all benchmarks.

Configuration	Geo-FNO					The NL6					
	NS	Darcy	Plasticity	Airfoil	Pipe	DNO	HYP	KS	SQG	STF	VAW
Layers	7										
Width	83	82	80	74	77	83	82	83	83	83	83
Parameters (M)	1.98	1.99	1.90	1.99	1.98	1.98	1.97	1.98	1.98	1.98	1.98
Normalisation	RMSNorm										
Activation	GeLU										
Lift / projection	1×1 convolution										
Rank (linear / paraproduct)	8 / 8										
Paraproduct modes	100%					100%	50%	100%			
Linear modes	50%										
Padding / drop path	None										
Time embedding	–	–	Sinusoidal	–	–	–	–	–	–	–	–
Loss	Relative ℓ_2										
Epochs	500										
Batch size	8										
Optimiser	AdamW										
β_1 / β_2 / ϵ	$0.95 \to 0.85 \to 0.95$ / 0.999 / 10^{-8}										
Weight decay	10^{-5}										
Scheduler	`OneCycle`, 30% warm-up										
LR (initial / peak / final)	8×10^{-5} / 2×10^{-3} / 8×10^{-9}										
Gradient clipping	None										
Branch calibration	✓	✓	×	✓	✓			✓			

A.2 Ablation study

This section evaluates the individual impact of four key architectural choices in ParaNO: spectral kernel sharing across dyadic bands, factorisation of the paraproduct kernel, number of retained Fourier modes, and parametrisation of the linear branch. A single architectural axis is modified at a time relative to a baseline configuration. All models are constrained to a fixed budget of 2,000,000 real degrees of freedom (counting complex parameters as two) at a fixed depth of $L = 7$. Performance is reported as relative error normalised to the benchmark baseline, where values below 1.000 represent an improvement.

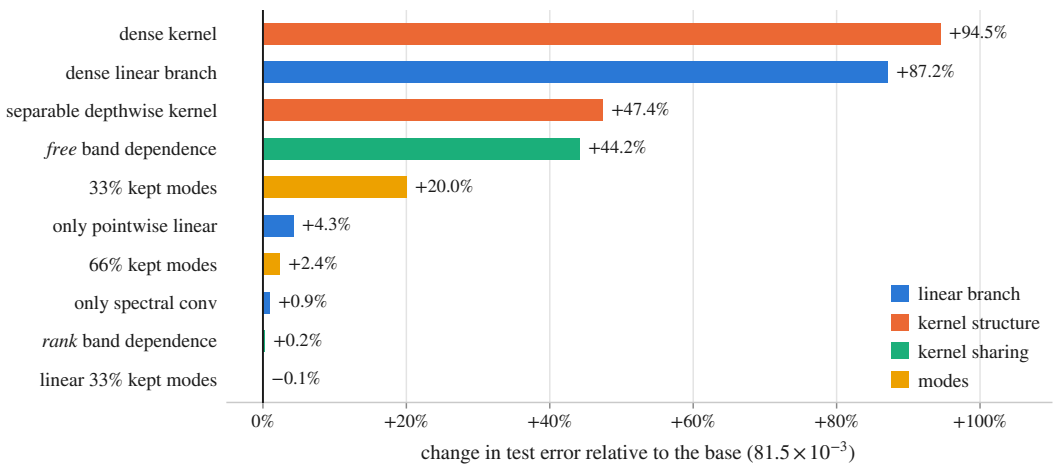


Figure A1: Ablation results as error relative to each benchmark's baseline. Each experiment isolates a single axis by reallocating channel width to maintain exactly 2,000,000 real parameters.

Ablations are conducted using a reduced training budget of 150 epochs on 352 samples, on Navier–Stokes. Validation indicates that this reduced budget preserves relative model rankings and captures intermediate effects within 10.0% of full protocol ratios. Fig. A1 summarises the complete ablation results.

Per-Band Kernel Structure. Three parameter-sharing strategies are evaluated across dyadic frequency bands: *free modes* is the described method in Sec. 3, *free* represents keeping fully independent kernels for each band, and *rank* builds a dictionary of 4 kernels with rank 2, and then each band keeps an independent, learnable coefficient vector that is used to combine the different kernels. *Rank* stays approximately equal in the ablation, while *free* parameter scaling causes it to resolve to width 29 and thus suffers a severe drop in performance.

Kernel Factorisation. Full dense channel kernels are computationally infeasible under the specified parameter constraint, and scale very poorly with increased budget. At the baseline bandwidth, a dense kernel requires 7.58M parameters at width 8. Halving the frequency bands and reducing depth to 4 affords a width of 29, and achieves a $1.9\times$ worse error. Depthwise separable kernels, which are the sum of a depthwise and a separable kernel, are feasible but perform poorly, increasing error by $1.5\times$.

Linear Branch Architecture. Removing the spectral convolution from this branch slightly worsens performance, while removing the pointwise skip stays about equal. However, using a dense kernel for the linear branch forces the width to 21, and thus performs the worst out of all options. Reducing the modes of the branch barely affects performance.

Retained Fourier Modes. The optimal frequency cutoff depends on the physical domain. Truncating retained modes to 66% of the Nyquist frequency very slightly degrades accuracy on Navier–Stokes, while reducing to 33% increases the loss to $1.2\times$ that of the base.

A.3 Experimental evaluation of band dependence

Band dependence is the central architectural contribution of the thesis, so it is ablated on four of NL6's benchmarks (precisely, DNO, HYP, KS, and SQG) under the full training protocol, with three seeds per arm, rather than only at the reduced budget above. To ablate it, the same kernel is used for every band; instead of having independent frequency vectors, those are also matched per band. Because the Littlewood–Paley decomposition is a partition of unity Métivier (2008), this reduces to one spectral convolution applied to all pairwise interactions.

The results are displayed in Tab. A2. Letting the kernel vary across bands reduces the loss on every benchmark, by 8.0% on average.

Table A2: Mean and standard deviation ℓ^2 errors ($\times 10^{-3}$) over three seeds for ParaNO with free opposed to shared modes, at 2M parameters. Bold marks the better result.

	DNO	HYP	KS	SQG
ParaNO (free modes)	**1.79** ± 0.09	**3.36** ± 0.15	**7.61** ± 0.15	**4.13** ± 0.09
ParaNO (shared)	1.92 ± 0.13	3.69 ± 0.18	8.46 ± 0.46	4.40 ± 0.25

The kernels themselves are also compared to analyse whether the model learns distinct representations per band. If the similarity sits significantly below 1, then the model exploits band dependence. Additionally, bands further apart should be less correlated, as their behaviours are more distinct.

Both predictions hold on every model (Fig. A2). Similarity sits far below a shared kernel, all lower than 0.53, yet above its random control on every term (except for on Airfoil), so the bands learn related but distinct kernels rather than one shared kernel or unrelated ones. Restricting the average to adjacent bands raises similarity on 26 of the 33 terms, and only lowers it on the two KS's paraproducts.

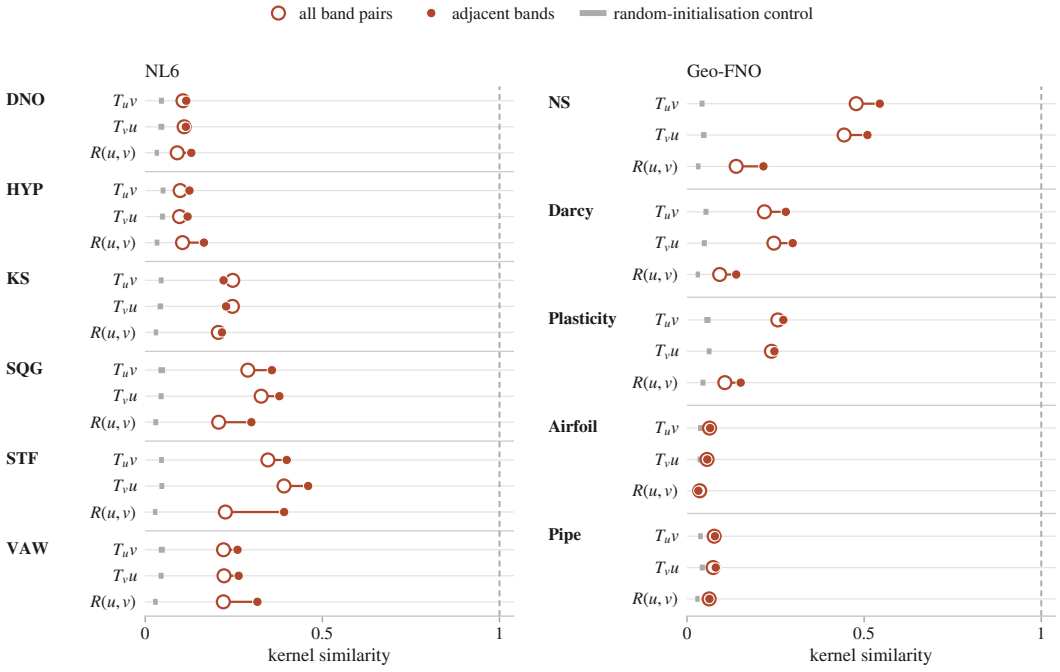


Figure A2: Per-band paraproduct kernels W_j similarity, computed as $|\langle W_j, W_l \rangle| / (\|W_j\| \, \|W_l\|)$, and averaged over band pairs, where 1 means an identical (shared) kernel. It is measured on live frequencies, those holding 99.9% of a band's measured input of trained ParaNO models. Hollow marks average over all band pairs and filled marks over adjacent bands only; grey bars are the control, the similarity after a random initialization.

A.4 Runtime Comparison

Tab. A3 displays the time taken for all baseline runs and ParaNO, on L40 Nvidia GPUs, on the Navier–Stokes benchmark. ParaNO has the longest runtime, around three times higher than all baselines but MSPT and SAOT, and 1.37× higher than those. This higher runtime is mainly driven by the repeated computations across all bands, and the tensor decomposed kernel, which must be reconstructed on every iteration. The peak memory usage is comparable to SAOT, lower than HO-FNO's and higher than the rest. Future work will focus on improving ParaNO's computational efficiency.

Table A3: Training cost on Navier–Stokes, every model on one NVIDIA L40S (46 GB; PyTorch 2.12, CUDA 13.0). NS error: lowest test relative L^2 error during training, mean and sample standard deviation over seeds 42, 1 and 2, as in Tab. A4. Params: real degrees of freedom, complex weights counted twice. h / 500 ep: runtime hours for 500 epochs of training. Peak GiB: largest GPU memory allocated during those steps.

OPERATOR	PARAMS (M)	NS ERROR ($\times 10^{-3}$)	H / 500 EP	PEAK GIB
Transolver	11.23	$111.07^{\dagger}$ ± 32.16	15.6	7.7
LaMO	10.06	35.43 ± 3.30	**12.8**	7.8
SAOT	9.65	43.81 ± 0.80	37.1	13.7
MSPT	8.07	53.36 ± 0.86	37.0	**5.3**
CATO	1.10	41.86 ± 1.20	17.1	11.0
HO-FNO	7.36	31.11 ± 0.08	18.3	23.6
ParaNO	1.98	**23.87** ± 1.19	50.7	14.0

A.5 GEO-FNO BENCHMARK SPECIFICATIONS

All benchmarks formulate a supervised learning problem: given a dataset $\{a_j, u_j\}_{j=1}^{N}$ generated by numerical solvers, the neural operator learns the mapping $\mathcal{G}^{\dagger} : a \mapsto u$. Tab. A4 summarises the inputs, targets, grid discretisations, and dataset partitions. Elasticity is posed on an unstructured point cloud, outside the scope of this project, and is omitted.

Benchmark	**Input** a	**Target** u	**Mesh**	**Train / Test**
Darcy	Permeability field	Pressure field	85×85	1000 / 200
Navier–Stokes	10 vorticity frames	Next 10 frames	64×64	1000 / 200
Plasticity	Die profile	20 steps $\times$ 4 fields	101×31	900 / 80
Airfoil	Mesh coordinates	Mach number field	221×51	1000 / 200
Pipe	Mesh coordinates	Horizontal velocity field	129×129	1000 / 200

Table A4: The six standard benchmark problems. Discretisation sizes reflect the standard published protocol after subsampling.

A.6 GEO-FNO BENCHMARK ADDITIONAL RESULTS

Extended Comparison with Published Baselines. Tab. A5 extends Tab. 1 with thirteen further baselines that were not reproduced, taking their numbers from the compilation in (Cheng et al., 2026, Tab. 1) and Curvo et al. (2026, Tab. 2). No published-only baseline enters the top three on any benchmark.

Table A5: Tab. 1 extended with published-only baselines, test relative L^2 error ($\times 10^{-3}$). The upper block reports the values compiled in Cheng et al. (2026, Tab. 1) and Curvo et al. (2026, Tab. 2) to the precision published there, with — marking a benchmark the paper does not report. Reproduced rows and ParaNO are copied unchanged from Tab. 1. * Plasticity reports published values for every baseline, reproduced or not.

OPERATOR	NAVIER–STOKES	DARCY	PLASTICITY*	AIRFOIL	PIPE
Published, not reproduced					
FNO Li et al. (2021)	155.6	10.8	—	—	—
WMT Gupta et al. (2021)	154.1	8.2	7.6	7.5	7.7
U-FNO Wen et al. (2022)	223.1	18.3	3.9	26.9	5.6
Geo-FNO Li et al. (2023c)	155.6	10.8	7.4	13.8	6.7
U-NO Rahman et al. (2023)	171.3	11.3	3.4	7.8	10.0
F-FNO Tran et al. (2023)	232.2	7.7	4.7	7.8	7.0
LSM Wu et al. (2023)	153.5	6.5	2.5	5.9	5.0
Galerkin Cao (2021)	140.1	8.4	12.0	11.8	9.8
HT-Net Liu et al. (2024)	184.7	7.9	33.3	6.5	5.9
OFormer Li et al. (2023a)	170.5	12.4	1.7	18.3	16.8
GNOT Hao et al. (2023)	138.0	10.5	33.6	7.6	4.7
FactFormer Li et al. (2023b)	121.4	10.9	31.2	7.1	6.0
ONO Xiao et al. (2024)	119.5	7.6	4.8	6.1	5.2
Reproduced (Tab. 1)					
Transolver	111.07 ±32.16	5.44 ±0.03	1.20	5.40 ±0.46	4.71 ±0.24
LaMO	35.43 ±3.30	4.14 ±0.19	0.70	4.58 ±0.44	4.49 ±0.33
SAOT	43.81 ±0.80	5.00 ±0.15	0.80	5.00 ±0.17	6.29 ±0.41
MSPT	53.36 ±0.86	11.14 ±0.46	1.00	6.59 ±0.34	**3.98** ±0.32
CATO	41.86 ±1.20	**3.87** ±0.05	0.50	**4.16** ±0.04	4.03 ±0.14
HO-FNO	31.11 ±0.08	5.72 ±0.45	0.60	4.86 ±0.06	5.55 ±0.20
ParaNO	**23.87** ±1.19	4.90 ±0.14	**0.39** ±0.02	5.70 ±0.19	5.45 ±0.16

Parameter Counts. Tab. A6 lists the real degrees of freedom of every model in Tab. 1, counting each complex parameter as two. Baselines use their authors' default configurations, so budgets are not matched across models.

Table A6: Real degrees of freedom of every model in Tab. 1, with complex parameters counted twice.

OPERATOR	NAVIER–STOKES	DARCY	PLASTICITY	AIRFOIL	PIPE
Transolver	11,232,321	2,826,945	2,844,484	2,810,817	3,073,985
LaMO	10,061,314	1,145,986	3,791,044	1,144,825	4,020,545
SAOT	9,650,433	2,694,017	5,491,204	6,058,689	2,757,313
MSPT	8,071,937	1,042,561	1,060,100	1,026,433	1,158,017
CATO	1,102,723	622,757	1,726,162	1,100,933	622,949
HO-FNO	7,362,049	9,456,897	2,533,636	2,500,353	6,235,905
ParaNO	1,976,600	1,994,081	1,901,714	1,986,947	1,982,702

Reproduction Fidelity. Reproduction is faithful everywhere except for Plasticity; only HO-FNO and Transolver reproduced on this benchmark, and thus the numbers for that column correspond to the published numbers instead. Of the other 24 cells scored against published values, 20 fall within $[0.75, 1.30]\times$ of the report, with a median ratio of $1.05\times$. All six Plasticity reproductions fall outside it, at 1.85–$28\times$ the published error.

Table A7: Ratio of each reproduced median in Tab. 1 to the value published for that cell in the source papers. Above $1.00\times$ means the reproduction is worse than the paper claims. Cells outside $[0.75, 1.3]\times$ are in bold.

OPERATOR	NAVIER–STOKES	DARCY	PLASTICITY	AIRFOIL	PIPE
Transolver	$1.23\times$	$0.95\times$	$\mathbf{1.85}\times$	$1.02\times$	$\mathbf{1.43}\times$
LaMO	$0.77\times$	$1.06\times$	$\mathbf{2.54}\times$	$1.12\times$	$1.18\times$
SAOT	$\mathbf{0.64}\times$	$1.02\times$	$\mathbf{28.24}\times$	$1.04\times$	$1.00\times$
MSPT	$0.84\times$	$\mathbf{1.77}\times$	$\mathbf{3.85}\times$	$1.29\times$	$1.28\times$
CATO	$\mathbf{1.31}\times$	$0.92\times$	$\mathbf{4.24}\times$	$1.02\times$	$1.06\times$
HO-FNO	$0.98\times$	$1.10\times$	$\mathbf{3.28}\times$	$0.95\times$	$1.13\times$

Error Maps for All Baselines. We show the errors maps on the Geo-FNO suite (except for Plasticity, the only benchmark where the baselines were not reproduced) for ParaNO and all baseline models, which can be seen in Fig. A3.

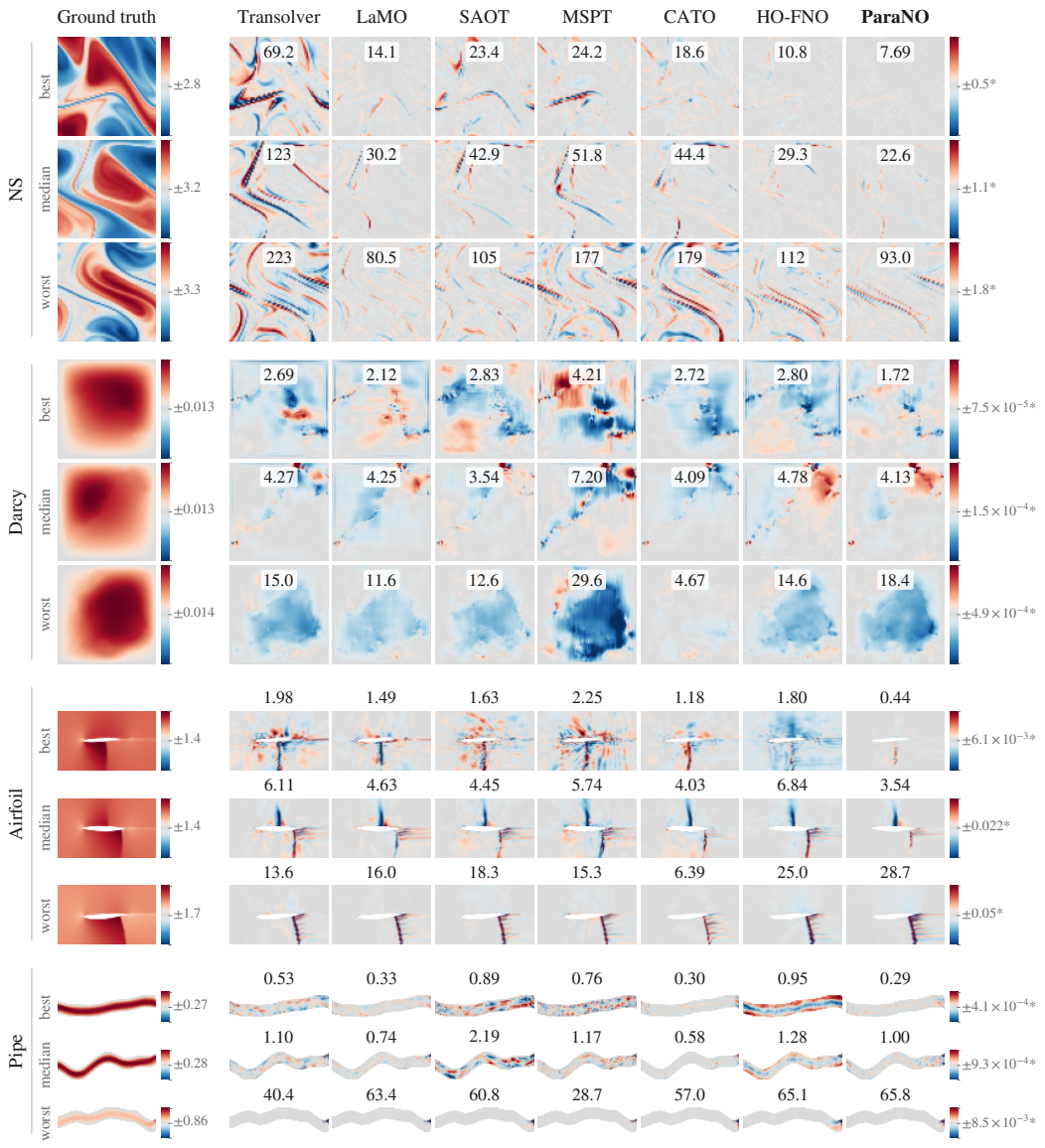


Figure A3: Geo-FNO error maps for all baselines.

A.7 DETAILED DESCRIPTION OF THE NL6

The NL6 comprises six non-linear PDE benchmarks (Tab. A8), partitioned into four autoregressive rollout tasks and two static operator mappings. Each dataset contains 1,200 samples split into 1,000 training and 200 test instances, with rollout trajectories generating 20 temporal frames for each sample: frames 1–10 are the input, and frames 11–20 the prediction target.

Benchmark	Input	Target	Grid
SQG	Surface buoyancy $[\theta]^{10}$	Next 10 frames	64×64
KS	$[u]^{10}$	Next 10 frames	64×64
STF	Velocity field $[u]^{10}$	Next 10 frames	64×64
VAW	Pressure field $[p, \partial_t p]^{10}$, medium $(c_0, \log Q)$	Next 10 frames of $[p, \partial_t p]$	64×64
DNO	Surface state $[\eta, \psi]$	Normal flux $G(\eta)\psi - \lvert D\rvert\psi$	64×64
HYP	Displacement u	Divergence Div $P(F)$	128×128

Table A8: Summary of the NL6 suite. Trajectory rollouts follow the standard Navier–Stokes protocol: 10 input frames stacked along channels, predicting 10 future steps autoregressively with one window per trajectory. Viscoacoustics appends static medium properties $(c_0, \log Q)$ as invariant input channels.

Construction. All benchmarks are formulated on the periodic two-torus $[0, L)^2$ ($L = 2\pi$, except for KS where $L = 16\pi$) discretised on uniform $N \times N$ spatial meshes with modal coordinates $k \in (2\pi/L)\{-N/2, \dots, N/2-1\}^2$. Quadratic and higher-order polynomial products are dealiased using the $2/3$-rule. Each system is solved numerically on a fine grid of at least twice the target spatial resolution ($N \geq 2M$) and downsampled to the stored grid via the spectral projection operator Π_M, which restricts the discrete Fourier transform to modes satisfying $|k_x|, |k_y| < M/2$ and rescales them by $(M/N)^2$.

Initial conditions and heterogeneous media are generated as zero-mean, band-limited Gaussian random fields (GRFs) with uniform random phases:

$$\hat{f}(k) = A(|k|)\big(\xi_1(k) + i\,\xi_2(k)\big), \quad \xi_{1,2} \sim \mathcal{N}(0,1) \text{ i.i.d.}, \quad A(|k|) = e^{-(|k|-k_p)^2/2w^2} \tag{A1}$$

enforcing conjugate symmetry to guarantee real-valued fields in physical space, followed by normalisation to a specified root-mean-square (RMS) amplitude. The DNO benchmark substitutes the Gaussian profile $A(|k|)$ with an empirical directional JONSWAP wave spectrum. Channel data are stored in physical units and normalised during training using the channel-wise mean and standard deviation computed over the training partition.

Checks. Every stored sample is verified against an independent `float64` numerical simulation. Worst-case metric values across each dataset are bounded by the thresholds specified in Tab. A9, while Tab. A10 reports the empirical values obtained for checks C1–C5. Rollout benchmarks include the relative L^2 error between the stored sequence and a fresh re-integration as an additional diagnostic.

benchmark	C1	C2 or C3	C4	C5
SQG	1.6×10^{-6}	2.4×10^{-10}	0.914	0.939
KS	2.4×10^{-5}	7.9×10^{-8}	3.40	1.09
STF	8.4×10^{-5}	2.2×10^{-3}	0.0269	0.443
VAW	2.3×10^{-6}	2.3×10^{-5}	0.196	1.55
DNO	1.4×10^{-6}	9.7×10^{-7}	0.166	—
HYP	2.5×10^{-14}	1.5×10^{-12}	0.344	—

Table A10: Empirical verification metrics representing the worst-case values measured on the released dataset. C1 is the largest error over all refinements and samples, and C2 and C3 the largest over samples; C4 is the mean over the 200 test samples and C5 their median. For static maps (DNO, HYP), C3 substitutes for C2 and C5 does not apply. Every rollout error, C2 included, is the relative L^2 error over the full ten-frame target.

check	definition	threshold
C1	Relative L^2 difference between the stored target and a more precise simulation, refining each truncation parameter	$\leq 10^{-4}$
C2	$\|A - B\|_2/\|A\|_2$, where $A = \Pi_M S(u_{\text{fine}})$, $B = \Pi_M S(\text{Lift}(\Pi_M u_{\text{fine}}))$, S denotes fine-grid forward evolution over frames 11–20 from frame 10, and Lift denotes zero-padding to the fine grid	$\leq 0.1\times$ (mean-symbol baseline error)
C3	Fraction of fine-grid target energy discarded by the projection operator Π_M onto the stored resolution (evaluated for static mappings)	$\leq 10^{-6}$
C4	Mean relative test L^2 error of the mean-symbol baseline, measured against the numerical data floor $\max(\text{C1}, \text{C2})$ for rollouts and $\max(\text{C1}, \sqrt{\text{C3}})$ for static maps	$\geq 10\times$ (numerical data floor)
C5	Test median persistence null error across the target horizon	≥ 0.2
C6	Benchmark-specific conservation laws and operator invariants	specific tolerances

Table A9: Dataset validation checks. C1 and C3 bound the numerical error of the stored data, and C2 ensures that spatial projection retains the information the target depends on. C4 is stated relative to that error floor rather than in absolute terms.

Each benchmark defines two null baselines: projecting the final observed state forward identically for rollouts ($\hat{u}_{11:20} = u_{10}$), or zero for static residual operators ($\hat{y} = 0$); and replacing the true state- or medium-dependent symbol with a trivial multiplier, with free physical constants calibrated on the training split to minimize the loss.

Surface quasi-geostrophic turbulence (SQG). We simulate decaying, unforced two-dimensional surface quasi-geostrophic turbulence on $[0, 2\pi)^2$ Held et al. (1995):

$$\partial_t \theta + u \cdot \nabla \theta = -\nu(-\Delta)^4 \theta, \qquad u = \nabla^\perp \psi = (-\partial_y \psi, \partial_x \psi), \qquad \hat{\psi}(k) = -|k|^{-1} \hat{\theta}(k), \quad \hat{\psi}(0) = 0. \tag{A2}$$

The system is solved at 128^2 and projected to 64^2 via Π_{64}. The hyperviscous parameter $\nu = k_d^{1-2p} = 7.8125 \times 10^{-10}$ balances advective transport and dissipation at the dissipation scale $k_d = 20$ Smith et al. (2002), positioned immediately below the dealiased cutoff ($|k| < 42.7$). Initial conditions are GRF realizations from Eq. (A1) with $k_p = 6$, $w = 3$, band-limited to $|k| \leq 16$ and scaled such that $\theta_{\text{rms}} = u_{\text{rms}} = 1$. Time stepping uses fourth-order integrating-factor Runge–Kutta (IF-RK4), which integrates dissipation exactly, with $\Delta t = 0.005$ (Courant number 1.2). The first $t_{\text{spin}} = 2.0$ is discarded to establish fully developed turbulent cascades, after which 20 frames are recorded at cadence $\tau = 0.5$ over $t \in [2.5, 12.0]$, spanning five active dyadic octaves.

Refinement halves the time step ($\Delta t = 0.0025$) and refines the mesh to 256^2 at the reduced step. The mean-symbol null is pure hyperdiffusion $\partial_t \theta = -\nu(-\Delta)^4 \theta$ with advective terms set to zero; due to the small ν, it closely approximates the persistence baseline. The solver satisfies the dissipation balance $\frac{d}{dt} \frac{1}{2} \|\theta\|_2^2 = -\nu \|(-\Delta)^2 \theta\|_2^2$ across the temporal trajectory (relative error $\leq 10^{-3}$) and conserves the quadratic invariants $\langle \theta^2 \rangle$ and $\langle \theta, |k|^{-1} \theta \rangle$ under inviscid truncation over $t \in [0, 0.5]$ (error $\leq 10^{-6}$).

Shear-thinning flow (STF). We consider decaying, unforced incompressible flow of a non-Newtonian Carreau–Yasuda fluid Yasuda et al. (1981), modifying standard Navier–Stokes dynamics with a generalized strain-rate-dependent viscosity:

$$\partial_t u + (u\cdot\nabla)u = \nabla\cdot\big(2\mu(\dot\gamma)D(u)\big) - \nabla p, \tag{A3}$$

$$\nabla\cdot u = 0, \quad D(u) = \tfrac{1}{2}(\nabla u + \nabla u^\top), \quad \dot\gamma = \sqrt{2D{:}D},$$

$$\mu(\dot\gamma) = s\Big[\mu_\infty + (\mu_0-\mu_\infty)\big(1+(\lambda\dot\gamma)^a\big)^{(n-1)/a}\Big]. \tag{A4}$$

The flow is solved at 256^2 and projected to 64^2 via Π_{64}. Its rheology uses power-law index $n = 0.4$ (matching physiological blood flow approximations, 0.357 Cho & Kensey (1991)), transition parameter $a = 2$, yielding the smooth dependence $(\lambda\dot\gamma)^2 = 2\lambda^2 D{:}D$, zero-shear viscosity $\mu_0 = 1$, infinite-shear ratio $\mu_\infty/\mu_0 = 0.02$, relaxation time $\lambda = 1$ and scaling factor $s = 0.025$. These place the flow at a large-scale Reynolds number $\mathrm{Re} \approx 20$, with viscous-to-advective stress ratio 0.27–0.34 and non-dimensional strain rate $\lambda\dot\gamma$ ranging from 0.41 (p1) to 13.3 (p99), 83% of spatial points satisfying $\lambda\dot\gamma > 2$, and dynamic range $\mu/\mu_0 \in [0.23, 0.96]$ across five active octaves. The initial condition is a divergence-free velocity derived from a stream function GRF (Eq. (A1)) with $k_p = 6$, $w = 2$, $|k| \le 12$, normalized to RMS velocity 1. Time stepping applies the solenoidal Leray projection and IF-RK4 integrating the reference linear diffusion $-\mu_r|k|^2$ with $\mu_r = s\mu_0/2$, leaving the residual viscous stress $2(\mu-\mu_r)D$ explicit ($\mu \le s\mu_0 = 2\mu_r$), with $\Delta t = 0.002$. Twenty frames are recorded at temporal step $\tau = 0.05$ across $t \in (0, 1.0]$; zero spin-up is required, as initial fields are solenoidal and dynamically admissible at $t = 0$.

Refinement halves the time step ($\Delta t = 0.001$) and refines the mesh to 512^2. The mean-symbol null is the equivalent Newtonian flow with globally calibrated viscosity $\bar\mu \approx 0.34, s\mu_0$ (error 0.027). The solver enforces the incompressibility constraint $\|\nabla\cdot u\|_2/\|\nabla u\|_2 \le 10^{-12}$ and satisfies the kinetic energy dissipation identity $\frac{d}{dt}\frac{1}{2}\|u\|_2^2 = -\langle 2\mu D{:}D\rangle$ per recording step (relative error $\le 10^{-6}$).

Viscoacoustic waves (VAW). We solve the decoupled constant-Q viscoacoustic wave formulation of Zhu & Harris (2014), incorporating Kjartansson rheology through spatial fractional Laplacians:

$$\partial_{tt}p = c(x)^2\Big[\eta(x)(-\Delta)^{\gamma(x)+1}p + \tau(x)\,\partial_t(-\Delta)^{\gamma(x)+1/2}p\Big], \tag{A5}$$

$$\gamma = \frac{\arctan(1/Q)}{\pi}, \quad \eta = -c_0^{2\gamma}\omega_0^{-2\gamma}\cos(\pi\gamma),$$
$$\tau = -c_0^{2\gamma-1}\omega_0^{-2\gamma}\sin(\pi\gamma), \quad c = c_0\cos(\pi\gamma/2),$$

where $c_0(x)$ is the acoustic phase velocity at reference frequency ω_0. In the lossless limit $Q\to\infty$, $\gamma\to 0$ and the system recovers the classical acoustic wave equation $\partial_{tt}p = c_0^2\Delta p$. The spatially varying fractional Laplacian is evaluated by expanding around the logarithmic mean attenuation $\bar Q = \sqrt{Q_{\min}Q_{\max}} \approx 31$ Zhang et al. (2026): $|k|^{2s(x)} = |k|^{2\bar s}\sum_{m=0}^{M}\frac{\left(2(s(x)-\bar s)\ln|k|\right)^m}{m!}$, decoupling the operator into $M+1$ invariant Fourier multipliers modulated by local powers of $(s(x)-\bar s)$.

The wavefield is solved at 128^2 and projected to 64^2 via Π_{64}. The medium is $c_0(x) = 1 + 0.25\,g(x)$ with GRF g (Eq. (A1), $k_p = 2$, $w = 3$, $|k| \le 6$, unit peak), and an attenuation profile $\log Q(x)$ generated independently and mapped to $[\log 8, \log 120]$; both profiles are band-limited below the $2/3$-dealiasing cutoff. The expansion uses reference frequency $\omega_0 = 8$ and truncation order $M = 4$, for which the exponent $2 + 2\gamma(x)$ varies smoothly from 2.005 ($Q = 120$) to 2.079 ($Q = 8$) and the fifth-order truncation term at $Q = 8$, $|k| = 16$ is bounded by 9.5×10^{-7}. The source is an initial pressure pulse from Eq. (A1) ($k_p = 6$, $w = 2$, $3 \le |k| \le 10$, RMS amplitude 1) released from rest ($\partial_t p|_{t=0} = 0$), integrated by explicit fourth-order Runge–Kutta (RK4) in first-order state-space form $(p, \partial_t p)$ with $\Delta t = 0.01$ (Courant number 0.25) and dealiased spatial products with the medium coefficients. Twenty frames are recorded at sampling period $\tau = 0.06$ spanning $t \in (0, 1.2]$.

Refinement halves the time step ($\Delta t = 0.005$), raises the expansion order $M : 4 \to 6$, and refines the mesh to 256^2. The mean-symbol null is a homogeneous background medium with $c_0 \equiv \bar c_0$ and $Q \equiv \exp(\overline{\log Q})$ (error 0.196); component ablation yields homogeneous velocity c_0 with true $Q(x)$ (0.195), lossless propagation $Q\to\infty$ (0.053), spatially constant fractional order $\bar\gamma$ (0.023), and spatially uniform attenuation $\bar Q$ (0.020). Setting $Q = 10^8$ recovers the lossless acoustic operator

asymptotically (discrepancy 2.6×10^{-9}), and in homogeneous media the modal invariants $\frac{1}{2}|\widehat{\partial_t p}|^2 + \frac{1}{2}c^2|\eta|\,|k|^{2\gamma+2}|\hat{p}|^2$ decay monotonically, verified across $Q \in \{8, 31, 120\}$.

Dirichlet–Neumann operator (DNO). For free-surface elevation $\eta(x)$ and surface velocity potential $\psi(x) = \phi(x, \eta(x))$ over an infinitely deep fluid layer, the Dirichlet–Neumann operator $G(\eta)\psi$ maps Dirichlet boundary data to the scaled outward normal derivative at the boundary:

$$G(\eta)\psi = (\partial_z\phi - \nabla_x\eta\cdot\nabla_x\phi)|_{z=\eta(x)} = \sqrt{1+|\nabla\eta|^2}\,\partial_n\phi\big|_{z=\eta(x)}. \tag{A6}$$

This represents the non-local mapping of Zakharov's Hamiltonian water-wave equations Zakharov (1968). The target prediction is isolated to the non-linear perturbation away from the flat-surface operator $G(\eta)\psi - |D|\psi$, with $|D| = (-\Delta)^{1/2} = G(0)$, as the flat-surface multiplier $|D|$ dominates the magnitude and is reproduced identically by standard spectral layers.

The forward operator is computed via the High-Order Spectral (HOS) perturbation expansion Dommermuth & Yue (1987), expanding the vertical surface velocity $W(x) = \partial_z\phi|_{z=\eta}$ into a Taylor series about $z = 0$, yielding $G(\eta)\psi = W(1+|\nabla\eta|^2) - \nabla\eta\cdot\nabla\psi$.

Surfaces are drawn from a directional JONSWAP spectrum Hasselmann et al. (1973) with peak enhancement $\gamma = 3.3$, spectral widths $\sigma_a = 0.07$, $\sigma_b = 0.09$, peak wavenumber $k_p = 4$, truncated to $1 \le |k| \le 3k_p$; directional spreading follows $\cos^2(\pi\theta/2\beta)$ for $|\theta| \le \beta = \pi/4$ along the $+x$ axis, wave steepness is $k_p\eta_{\mathrm{rms}} \sim U[0.04, 0.07]$, and the linear surface potential is $\hat{\psi}(k) = -i\,\mathrm{sgn}(k_x)\sqrt{g/|k|}\,\hat{\eta}(k)$ with $g = 1$. Each realization is then integrated for 10 peak wave periods ($T_p = 2\pi/\sqrt{gk_p}$) under Zakharov Hamiltonian dynamics at HOS order 5 on a 128^2 grid, via RK4 with $\Delta t = T_p/126$; non-linear terms are smoothly ramped by $1 - \exp[-(t/5T_p)^4]$ and fields are filtered to $|k| \le 16$ between steps, establishing bound harmonics (skewness shifts from 0.148 at $t = 1.3T_p$ to 0.205 at $t = 10T_p$). The target is finally computed at HOS expansion order 14 on a 256^2 grid from the zero-padded spin-up states and downsampled to 64^2 via Π_{64}. Realizations are rejected and redrawn (up to 12 attempts) if the relative difference against order 16 exceeds 10^{-5}, the truncation tail energy exceeds 10^{-6}, or the self-adjointness defect exceeds 10^{-6}, and steepness is capped at 0.07 to prevent perturbation breakdown Nicholls & Reitich (2001).

Refinement raises the expansion order $14 \to 16$ ($\le 10^{-5}$) and the grid to 512^2 ($\le 10^{-6}$). The mean-symbol null is the first-order Craig–Sulem bilinear correction $G_1(\eta)\psi = -\nabla\cdot(\eta\nabla\psi) - |D|(\eta|D|\psi)$ Craig & Sulem (1993), corresponding to the leading-order non-linear term (residual error 0.13 ($0.11 - 0.17$) below 0.045; 0.20 ($0.16 - 0.25$) above 0.065; mean 0.166); the trivial baseline is zero. The evaluated operator preserves the flat-surface properties $G(\eta)1 = 0$ and $G(0)\psi = |D|\psi$ (errors $\le 10^{-14}$) and remains self-adjoint, $|\langle G(\eta)\psi_1, \psi_2\rangle - \langle\psi_1, G(\eta)\psi_2\rangle| \le 4.8\times10^{-7}$ across band-limited test functions (gated at 10^{-6}), an identity independent of every truncation in the expansion.

Hyperelasticity (HYP). We consider an incompressible thin Ogden rubber sheet under plane-stress deformation, where the principal stretch outside the plane is kinematically strained via incompressibility $\lambda_3 = (\lambda_1\lambda_2)^{-1}$. The static mapping evaluates the internal force divergence:

$$\text{target} = \mathrm{Div}\,P(F), \quad F = I + \nabla u, \quad P = \frac{\partial W}{\partial F}, \tag{A7}$$

$$W(\lambda_1, \lambda_2) = \sum_{p=1}^{3} \frac{\mu_p}{\alpha_p}\Big(\lambda_1^{\alpha_p} + \lambda_2^{\alpha_p} + (\lambda_1\lambda_2)^{-\alpha_p} - 3\Big), \tag{A8}$$

which represents the momentum balance residual computed at each step of an implicit non-linear solver. Material parameters match Ogden's empirical fit to Treloar's experimental data Ogden (1972): $\alpha = (1.3, 5.0, -2.0)$ and $\mu = (1.0, 0.0019048, -0.015873)$ (normalized to $\mu_1 = 1$), yielding an infinitesimal shear modulus $\mu_0 = \frac{1}{2}\sum_p \mu_p\alpha_p = 0.671$.

In-plane singular values $\lambda_1 \ge \lambda_2$ are extracted via an exact 2×2 singular value decomposition $F = U\Sigma V^\top$, yielding the first Piola–Kirchhoff stress tensor $P = U\,\mathrm{diag}(\partial W/\partial\lambda_1, \partial W/\partial\lambda_2)V^\top$ Ogden (1997). Spatial divergence Div is evaluated spectrally, avoiding finite-difference discretization errors. The material satisfies strong ellipticity across $\lambda \in [0.01, 100]$. The formulation assumes the sheet remains planar under tension; out-of-plane compressive buckling is constrained.

The operator is evaluated at 512^2 and projected to 128^2 via Π_{128}; truncation to 64^2 instead yields an energy tail of 3.0×10^{-6}, violating check C3. Displacement components are drawn independently from Eq. (A1) ($k_p = 3$, $w = 1$, $1 \leq |k| \leq 6$), providing three input octaves and generating 3–5 active octaves in the stress divergence target, and are rescaled per sample via root-finding such that the RMS stretch metric $\sqrt{\langle(\lambda_1 - 1)^2 + (\lambda_2 - 1)^2\rangle}$ matches a random draw $U[0.10, 0.25]$. Samples violating $\min_x \lambda_2(x) \geq 0.36$ are rejected and redrawn (up to 200 attempts) to match the valid empirical strain range of Treloar's experimental data, where maximum uniaxial tension reaches $7.6 = 0.36^{-\frac{1}{2}}$.

Refinement doubles the spatial grid to $1{,}024^2$ ($\leq 10^{-13}$). The mean-symbol null is linear isotropic elasticity $L(u) = \mu_L \Delta u + (\lambda_L + \mu_L)\nabla(\nabla \cdot u)$, fitted across the training partition via least squares (relative error 0.22 at strain amplitude 0.10, and 0.50 at amplitude 0.25); the trivial baseline is zero. The stress evaluation is verified against the closed-form neo-Hookean tensor $P = F - J^{-2}F^{-\top}$ under $\alpha = 2$, $\mu = 1$ (2.5×10^{-15}), which shares no decomposition with the solver, and recovers linear elasticity in the infinitesimal limit ($\times 10^{-4}$ amplitude), giving $(\mu_L, \lambda_L + \mu_L) = (0.673, 2.03)$ against the theoretical $(0.671, 2.013)$; the acoustic tensor remains pointwise positive definite (minimum eigenvalue ≥ 0.53).

Kuramoto–Sivashinsky, 2-D (KS). We solve the isotropic, two-dimensional Kuramoto–Sivashinsky equation on $[0, 16\pi)^2$ Kuramoto & Tsuzuki (1976); Sivashinsky (1977):

$$\partial_t u + \Delta u + \Delta^2 u + \tfrac{1}{2}|\nabla u|^2 = 0. \tag{A9}$$

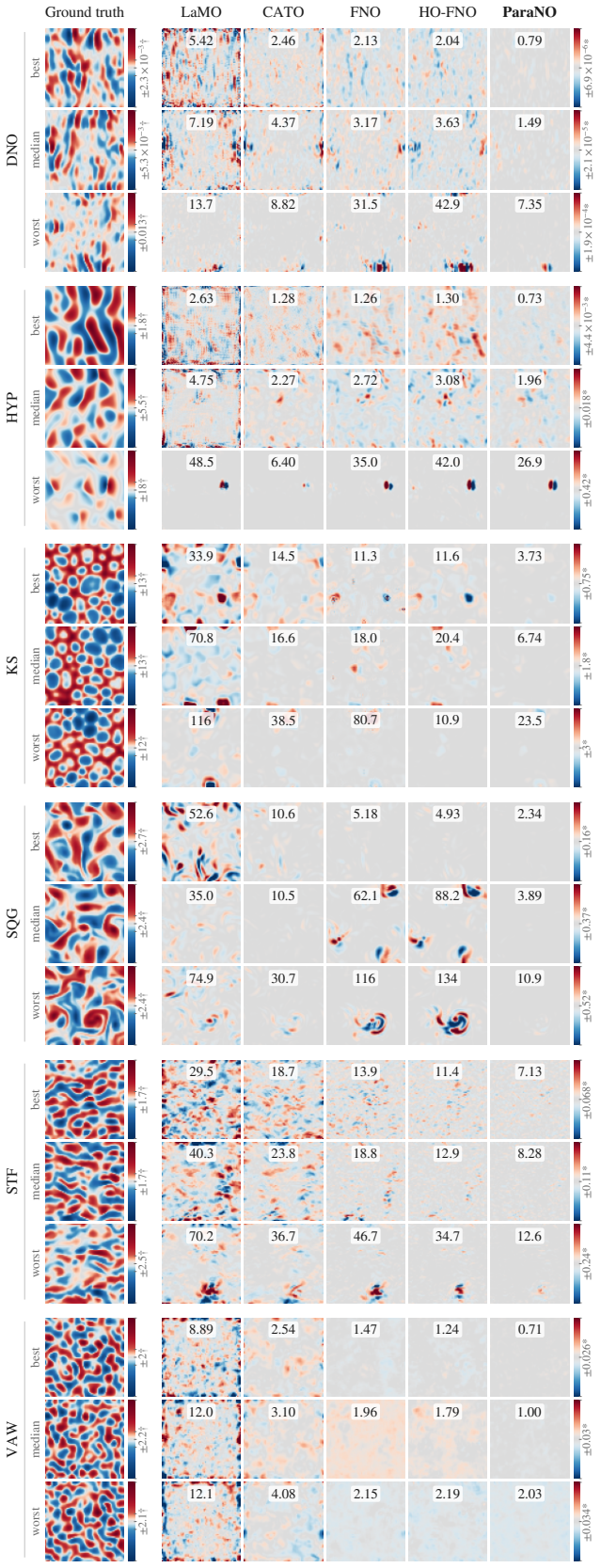
The rotationally invariant biharmonic operator Δ^2 is employed in place of directional derivatives $(\partial_x^4 + \partial_y^4)$ to preserve continuous rotational symmetry. The non-linear advection term drives a spatial mean drift according to $\frac{d}{dt}\bar{u} = -\frac{1}{2}\langle|\nabla u|^2\rangle$, but this uniform shift decouples from the spatial gradients, so the spatial mean is projected out at each time step and stored as zero.

The domain has fundamental wavenumber $k_{\min} = 1/8$, yielding ≈ 200 linearly unstable Fourier modes, and its parameter $(2\pi/L)^2 = 0.016$ lies well within the fully developed chaotic regime (< 0.1 Kalogirou et al. (2015)). The equation is solved at 128^2 and downsampled to 64^2 via Π_{64}, using fourth-order Exponential Time Differencing Runge–Kutta (ETDRK4) Kassam & Trefethen (2005) on the stiff linear operator $|k|^2 - |k|^4$ with 32-point complex contour integration and $\Delta t = 0.0125$, the non-linear gradient products being dealiased via the $2/3$-rule; the field is projected to real space at every step to prevent exponential round-off amplification along the maximum linear growth rate $(1/4)$. Initial conditions are GRF realizations from Eq. (A1) restricted to $|k| \leq 1$ with RMS amplitude 10^{-2}, integrated for $t_{\text{spin}} = 300$ and discarded, attractor dynamics stabilizing by $t \approx 40$. Twenty frames are then recorded at sampling period $\tau = 1.0$ across four active dyadic octaves, with trajectory divergence displaying a Lyapunov e-folding time of 4.5 (2.2 e-folds across the target horizon).

Refinement halves the time step ($\Delta t = 0.00625$) and refines the grid to 256^2. The mean-symbol null is the decoupled linear evolution $\partial_t u = -\Delta u - \Delta^2 u$. Isolated infinitesimal Fourier modes grow at the rates predicted by the linear dispersion $|k|^2 - |k|^4$ (relative error $\leq 10^{-6}$), and attractor stationarity is confirmed by matching the sample medians of $\langle u^2 \rangle$ across subsequent 20-frame windows to within 5%.

Error Maps for All Baselines.. We further show the error maps for all baselines evaluated on NL6, extending the results of Fig. 5 to the worst and best test samples, ranked by ParaNO's error (Fig. A4).

Figure A4: Error maps on NL6 for all baselines



A.8 ZERO-SHOT RESOLUTION TRANSFER

One of the characteristic components of NOs is their function approximation capabilities. Thus, some operators can effectively train at one resolution and evaluate at another without dropping accuracy, and so are resolution equivariant. We show empirically in Fig. A5 that ParaNO's loss, like that of FNO and HO-FNO, stays mostly flat when transfering to higher resolutions.

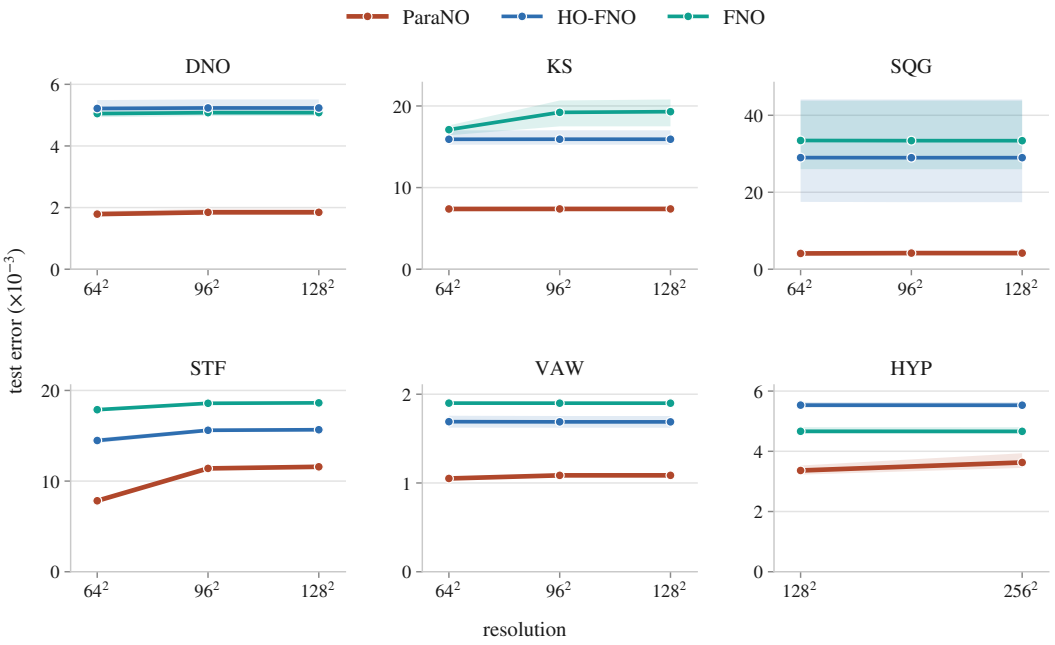


Figure A5: Trained checkpoints at the base resolution (the first x-axis entry), evaluated on higher resolutions without any retraining.

A.9 SPECTRAL ERROR ANALYSIS

We compare ParaNO with every baseline frequency by frequency, using the seed-42 runs of Tab. 2 (LaMO, CATO, FNO and HO-FNO, the baselines that ran on every benchmark) and, for Navier–Stokes. Every figure shares one layout. The top row plots each model's power spectrum against the ground truth's, averaged across every test sample. The bottom row plots each model's relative error at every radial frequency k, $\left(\sum_n \|\widehat{e}_n\|_k^2 / \sum_n \|\widehat{u}_n\|_k^2\right)^{1/2}$, where $\|\cdot\|_k$ is the norm over the Fourier modes with $|\xi| \approx k$, on a logarithmic and on a linear axis. Every panel covers every frequency of the grid, out to the corners of the Fourier box, and the vertical line marks the Nyquist radius. Where the truth holds almost no power, at the zero frequency of a zero-mean field and in the tail of a smooth one, the relative error reaches several orders of magnitude and dominates the linear axis.

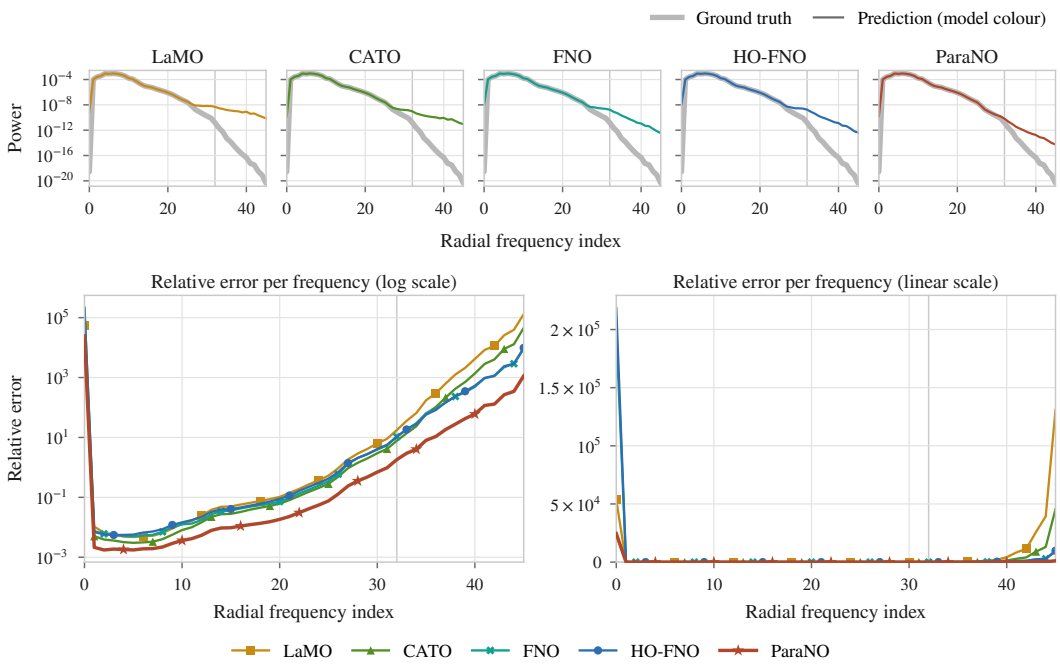


Figure A6: Dirichlet–Neumann spectra. ParaNO follows the true spectrum up to the Nyquist radius, while every baseline deviates from index 27 or 28. ParaNO's error is the lowest at every frequency, and its lead over the best baseline, CATO, grows from $1.8\times$ at the lowest frequencies to $3.7\times$ near the Nyquist radius.

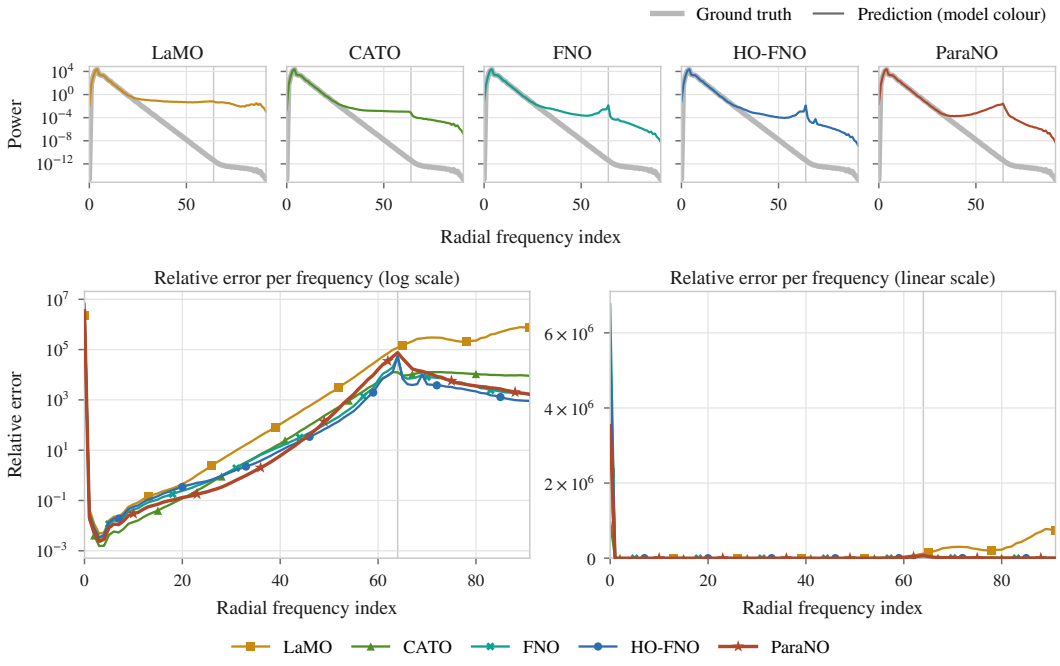


Figure A7: Hyperelasticity spectra. CATO has the lowest error up to index 20, ParaNO from 21 to 44, and HO-FNO at most frequencies above that; ParaNO's energy peaks at the Nyquist radius.

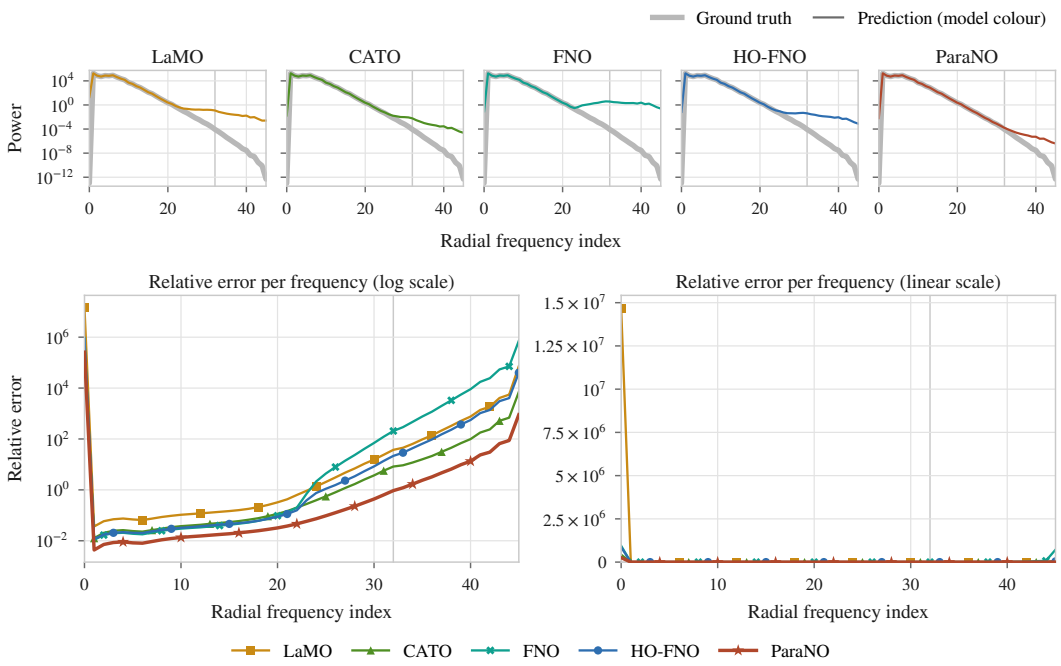


Figure A8: Kuramoto–Sivashinsky spectra. ParaNO's error is the lowest at every frequency, $2.3\times$ below the best baseline's at the lowest frequencies and $4.8\times$ below it near the Nyquist radius.

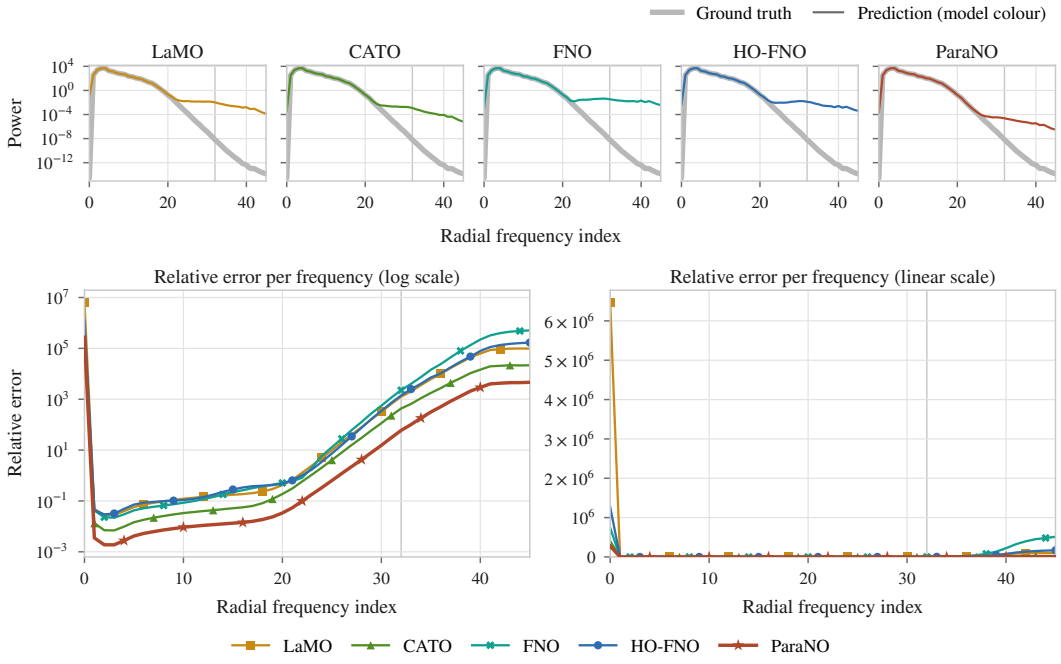


Figure A9: Surface quasi-geostrophic spectra. ParaNO's error is the lowest at every frequency, $3.6\times$ below the best baseline's, CATO's, at the lowest frequencies and $6.2\times$ below it near the Nyquist radius.

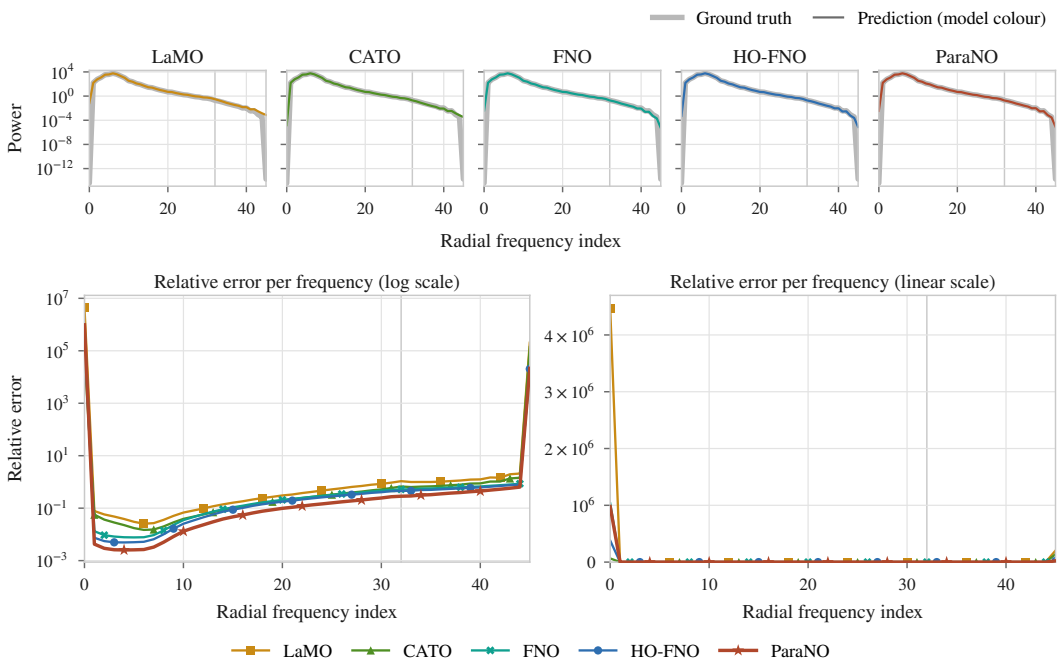


Figure A10: Shear-thinning flow spectra. Every model follows the true spectrum until its sharp drop at the last corner frequencies, where every error jumps by several orders of magnitude.

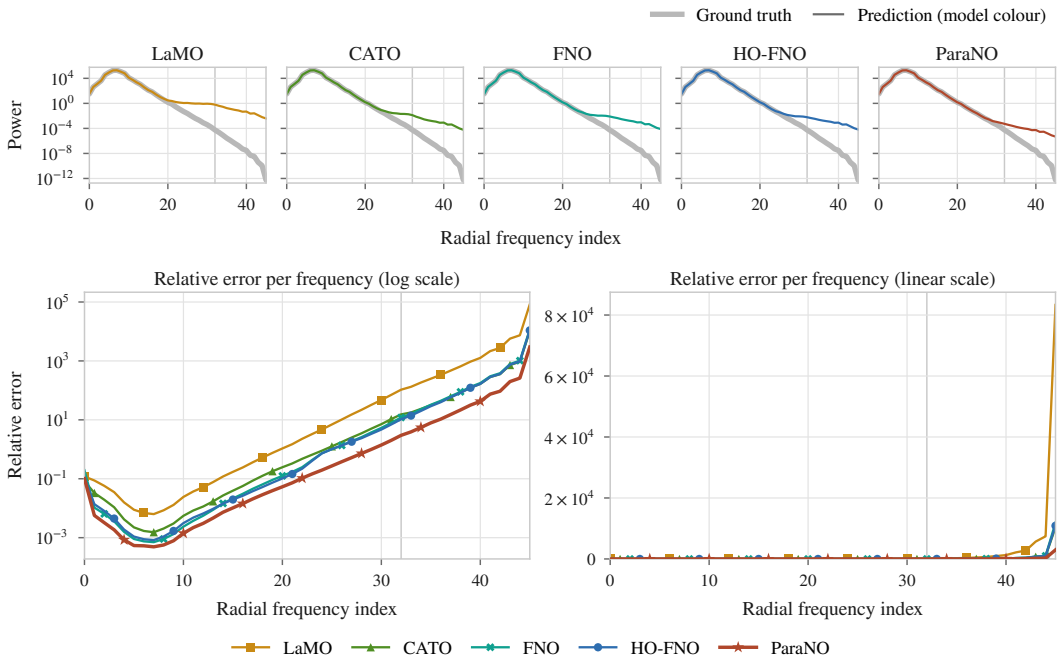


Figure A11: Viscoacoustic spectra. Where the true spectrum decays, the baselines' spectra leave it between indices 20 and 26, ParaNO's only at 30. ParaNO's error is the lowest at every frequency, and the gap widens with frequency.

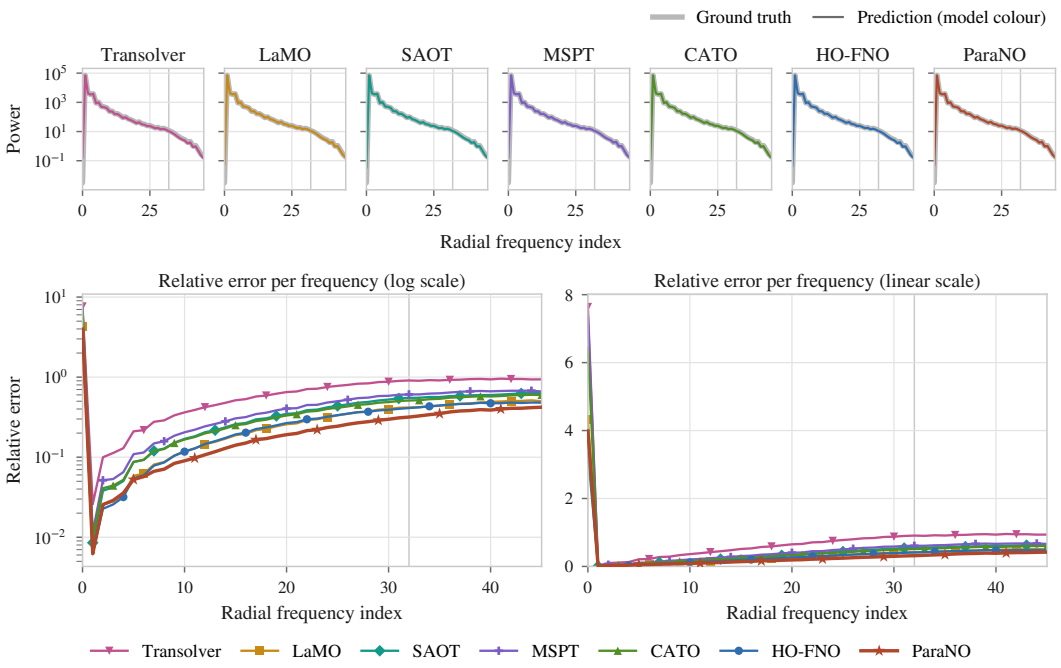


Figure A12: Navier–Stokes spectra. Every model follows the true spectrum over the whole range. ParaNO's error is the lowest from index 5 upwards, and its lead widens by the highest frequencies: 0.24 against LaMO's 0.33 between index 20 and the Nyquist radius.