NEURAL OPERATORS WITH EXACTLY LINEARIZABLE STRUCTURE FOR NEAR-INTEGRABLE PDES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned solution operators accumulate rollout error on systems whose evolution is exactly linear in the right coordinates. For near-integrable PDEs those coordinates exist in closed form: the inverse scattering transform (IST) diagonalizes the integrable flow. The Scattering-Coordinate Neural Operator (SCNO) routes the state through a differentiable nonlinear Fourier transform, applies the exact phase rotation, inverts by machine-precision Darboux dressing, and learns only the integrability-breaking residual. Zero-initialized, the pipeline is the analytic solution operator, verified to round-off; a learned transform cannot satisfy this identity. Soliton perturbation theory predicts that the residual, the learning target, grows as ϵz; we measure that law directly, and matched-capacity ablations show that the coordinates carry the advantage while three residual architectures inside them are interchangeable. We also measure where structure stops paying: across four fiber-physics perturbation channels the loss channel crosses over at $\epsilon^* \in [0.020, 0.030]$ over baseline families, selection rules and budgets, beyond which tuned time-domain baselines win, and the same phase diagram re-emerges on damped Korteweg–de Vries through a different spectral problem and a different inverse. Removing the transmitter-side oracle and re-training from blindly recovered spectra changes the diagram by a median of 0%. Off the reflectionless manifold the error is set by the energy the coordinates exclude: a radiative fraction ρ costs $\approx\sqrt{\rho}$, flat in distance. The closest prior method, an FNO with learned scattering structure, is competent in distribution and diverges at twice its training horizon: learned approximate linearizations compound under extrapolation, exact ones cannot.

1 INTRODUCTION

Learned solution operators train on short horizons and are asked to extrapolate; we measure what that costs on the integrable nonlinear Schrödinger equation (NLSE), with relative L_2 error as the metric throughout. A tuned U-Net, selected from a 60-configuration sweep on its in-distribution error, reaches 0.0533 at the top of its training range and 1.62 at a $10\times$ horizon, where unity is what predicting zero achieves. On the same trajectories a pipeline with no learned parameters, built from the inverse scattering transform (IST) of Zakharov & Shabat (1972), holds 4.6×10^{-4}, a separation of 3.6×10^3. The gap is about where the capacity goes: learned operators spend it on a map that has had a closed form since 1972, and accumulate error while doing so.

Recent architectures borrow scattering *structure* without the scattering *transform*. The inverse-scattering-inspired FNO of Yu (2025) constrains an operator with invertible lifting pairs and exponential spectral layers motivated by IST. Its linearizing map is learned, so it stays approximate at every distance: the best variant reaches 0.104 at $z = 4$, a factor 2.3×10^3 above the exact line. When the exact transform exists, use it as a computational layer.

SCNO runs the state through a differentiable nonlinear Fourier transform, evolves the scattering data by its closed-form law, and inverts by machine-precision Darboux dressing; learning is restricted to the integrability-breaking residual. We study the NLSE under four physical perturbations: loss, third-order dispersion, Raman-type self-frequency shift, and distributed noise. Each breaks integrability in optical fiber and carries a known first-order soliton perturbation rate. In co-rotating scattering coordinates the unperturbed flow is the identity, so the learned vector field represents

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.

the perturbation alone, of magnitude $O(\epsilon)$, and its integrated drift is $O(\epsilon z)$ by soliton perturbation theory (Prop. 2).

The paper's weight rests on an exact backbone: its $10\times$-horizon separation is one no selection rule can relabel. Beside it sit a learning target measured against a first-principles prediction and a mechanism isolated by ablation at matched capacity. The structure transfers to a second family, the phase diagram survives removal of its own oracle, and off the manifold the excluded energy alone sets the error. Our contributions:

The learning target, measured (C1). Prop. 2 bounds the amplitude residual at $O(\epsilon z)$ and the phase at $O(\epsilon z^2)$, and §5.3 measures both: pooled log-log fits over 57 grid points per channel return exponents $1.00 \pm 5.1 \times 10^{-3}$ for third-order dispersion and 0.908 ± 0.0693 for loss against the predicted 1; the local slope is 0.93 up to the crossover and bends to 0.79 only past it. Zero-initialized, the pipeline is the analytic solution operator (Prop. 1), verified to 2.1×10^{-15}, a separation of 3.6×10^3 from the strongest tuned baseline at a $10\times$ horizon.

The coordinates carry the effect (C2). At matched capacity, moving the residual from scattering to time coordinates costs $5.8\times$, removing the exact rotation drives the model to the no-information level, and three parameterizations inside the coordinates span only $1.1\times$ (§5.6). Out of distribution the tuned baseline crosses the no-information line while SCNO degrades smoothly (§5.5).

A phase diagram that transfers (C3). Across four channels, tuned against tuned at three to eight seeds per arm, SCNO leads the noise channel, third-order dispersion and weak loss against the strongest baseline, a tuned U-Net at all 11 points ($19.6\times$, $6.1\times$, $2.33\times$ at the smallest perturbations). The loss channel crosses over at $\epsilon^* \in [0.020, 0.030]$, the low end from tripling both training budgets and the high end from the selection rules; the U-Net's family crossover lands inside the range at 0.022. Raman is a loss against it at both strengths. Against the strongest closed-form pipeline (exact rotation plus first-order drifts wherever the rate is exact) the residual wins 10 of 16 perturbed cells, by up to $5.0\times$ (Table 9). Rebuilt for damped KdV, the residual buys 14.4 to $29.4\times$ over the exact-rotation line (1.29 to $2.57\times$ over the first-order line) and crosses over at $\epsilon^*_{\mathrm{KdV}} = 0.051$, moving to 0.037 against per-point tuned baselines (Appendix C, Fig. 10b).

The protocol survives its own removal (C4). Every rollout result grants exact scattering data at $z = 0$. Re-training with a spectrum recovered blindly from the field by the §5.8 eigensolver changes the phase diagram by a median of 0% over all 11 points, largest Welch $|t| = 1.1$.

How learned linearizations fail (C5). The closest prior method, an FNO whose architecture learns scattering structure, received a four-rate learning-rate sweep and 3 architecture variants at 4 contested grid points. In distribution it is competent, 0.431 at worst at $z = 4$; at $z = 8$ its best point reads 5.0×10^5. A learned exponential evolution compounds under extrapolation in z; the analytic rotation has no learned component to extrapolate, and the inverse, the spectrum estimate and the residual are each measured.

Failure is priced by excluded energy (C6). Darboux dressing inverts the discrete spectrum, so the coordinates carry the soliton energy and nothing else. A radiative fraction ρ of the input costs the pipeline $\approx\sqrt{\rho}$, nearly flat in z, and the trained residual adds the same $\sqrt{\rho}$ in quadrature, fitting none of it. By the trace formula that excluded energy is at most 3.3×10^{-3} where SCNO wins and 0.290 where it loses (§5.9).

2 Related Work

Integrable systems and the nonlinear Fourier transform. The IST linearizes the focusing NLSE through the Zakharov–Shabat spectral problem (Zakharov & Shabat, 1972; Ablowitz et al., 1974). Direct scattering is computed numerically by transfer-matrix methods (Boffetta & Osborne, 1992), with fast implementations in standard tooling (Wahls et al., 2018), and the reflectionless inverse by Darboux dressing (Matveev & Salle, 1991; Faddeev & Takhtajan, 1987). Optical communications engineered the transform into transmission systems (Yousefi & Kschischang, 2014; Turitsyn et al.,

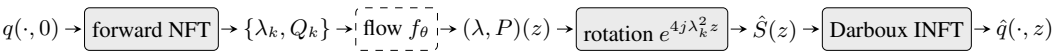


Figure 1: The SCNO pipeline. Shaded blocks carry no learned parameters. The dashed block is learned: f_θ integrates a residual field on the co-rotating coordinates (λ, P) of equation 2, and the analytic rotation is applied afterwards with the drifted $\lambda_k(z)$. It is zero-initialized, so the composite computes exactly the NLSE solution operator (Prop. 1) and the flow of f_θ only ever represents the $O(\epsilon z)$ drift of Prop. 2.

2017) and brought machine learning to their receivers (Zibar et al., 2016); we repurpose the transform as a differentiable computational layer for operator learning. The perturbed-soliton theory that supplies our $O(\epsilon z)$ bound is classical (Kaup, 1976; Karpman & Maslov, 1977; Gordon, 1986).

Neural operators. FNO (Li et al., 2021), DeepONet (Lu et al., 2021) and their successors (Raonić et al., 2023; Colagrande et al., 2026) learn solution operators in the time or space domain. Their spectral parameterizations diagonalize linear constant-coefficient flows, which is what forces ever-richer mixers once the flow is nonlinear. Rollout stabilization by the pushforward trick or iterative refinement (Brandstetter et al., 2022; Lippe et al., 2023) lengthens learned steppers' horizons in general; at the one-step precision these rollouts demand it lowers no autoregressive arm here (Appendix A), and the linearizing map stays approximate, so the C5 argument applies unchanged. In scattering coordinates the nonlinear flow is already diagonal, which removes the need for richer mixers.

Physics-structured learning. IS-FNO (Yu, 2025) is the closest neighbor: it motivates invertible lifting pairs and exponential spectral layers from IST, and it learns the transformation. We treat it as a baseline under both its official protocol and the proposed one, in the integrable limit and in the contested regime. Solver-in-the-loop methods (Um et al., 2020) learn corrections to a numerical coarse solver; the analytic stage here is exact, which shrinks the correction target from $O(1)$ discretization error to $O(\epsilon z)$ physics. Koopman autoencoders (Lusch et al., 2018) learn approximate linearizing embeddings, and SCNO is the case where the Koopman map is known in closed form. Hamiltonian and physics-informed networks (Greydanus et al., 2019; Raissi et al., 2019) inject structure as architecture or as penalty; §5.6 measures what the change of coordinates alone contributes. Zero-initialized residual paths follow ReZero and ControlNet (Bachlechner et al., 2021; Zhang et al., 2023), with one difference: the zero-init point here is an exactly correct physical model.

3 METHOD: THE SCNO PIPELINE

Setup. We study the focusing NLSE in normalized form under a perturbation,

$$j\,\partial_z q + \partial_{\tau\tau} q + 2|q|^2 q \;=\; \epsilon\, R[q], \tag{1}$$

with j the imaginary unit, and four channels taken one at a time: loss $R = -jq$ (ϵ_1), third-order dispersion $R = j\,\partial_{\tau\tau\tau} q$ (ϵ_2), Raman-type self-frequency shift $R = q\,\partial_\tau |q|^2$ (ϵ_3), and distributed noise $\sigma \dot{W}$. The choice is physical: these four mechanisms break integrability in optical fiber, and each has a known first-order soliton perturbation rate, which is what makes the residual bound of Prop. 2 checkable.

Scattering coordinates. At $\epsilon = 0$, equation 1 is integrable. The Zakharov–Shabat problem $\partial_\tau v = \begin{bmatrix} -j\lambda & q \\ -q^* & j\lambda \end{bmatrix} v$ associates to q a discrete spectrum $\{\lambda_k = \zeta_k + j\eta_k\}$ with $\eta_k > 0$ and norming constants Q_k (Zakharov & Shabat, 1972; Ablowitz et al., 1974). Under the flow the eigenvalues are invariant and the norming constants rotate in closed form, $Q_k(z) = Q_k(0) e^{4j\lambda_k^2 z}$; the convention-dependent factor 4 is pinned against the reference solver to 1.9×10^{-6} relative error (G4, Table 4). We then work in *co-rotating* coordinates

$$P_k(z) \;:=\; Q_k(z)\, e^{-4j\lambda_k^2 z}, \tag{2}$$

in which the integrable flow is exactly the identity: $\dot{\lambda}_k = 0$, $\dot{P}_k = 0$. A learned vector field on (λ, P) then represents the perturbation and nothing else.

Pipeline. Fig. 1 shows the four stages. Forward scattering uses a transfer-matrix discretization (Boffetta & Osborne, 1992) with bidirectional Jost integration for the norming constants. The inverse is a differentiable Darboux dressing (Matveev & Salle, 1991), exact to 4.6×10^{-15} against the

closed-form single soliton. Composing the forward transform with the inverse returns the norming constants to 0.0499 relative error on the production grid and 6.7×10^{-4} on a 4096-point validation grid (E0-G2, Table 4); the Newton eigensolver of §5.8 reconstructs fields to 2.9×10^{-6}. Data generation supplies (λ_k, Q_k) by Darboux synthesis, so the rollouts of §5.1 through §5.6 never invoke the forward transform. §5.8 measures what it costs when a receiver has only the field. Between the two transforms sit the analytic rotation, which has zero parameters, and the learned residual flow, a neural ODE (Chen et al., 2018)

$$\frac{d}{dz}\begin{pmatrix}\lambda\\ P\end{pmatrix} \;=\; f_\theta(\lambda, P, z; \boldsymbol{\epsilon})\,, \qquad \text{integrated by RK4 with step } \Delta z = 0.1, \tag{3}$$

where f_θ is permutation-equivariant over the soliton set: a per-soliton encoder, a masked mean pool, and a per-soliton head, so relabeling solitons leaves the prediction fixed. The composite map $q \mapsto \hat{q}$ is discretization-invariant by construction: the two transforms absorb the grid, and the learned flow acts on grid-free spectral data. The state carries $\arg P_k$ unwrapped as an ODE variable, so no branch cut enters the flow; Fig. 4(c) measures that unwrapped drift. Appendix A, Fig. 6 renders one stored trajectory, the analytic rollout, and their residual. The output layer of f_θ is zero-initialized, following ReZero and ControlNet (Bachlechner et al., 2021; Zhang et al., 2023). The zero point carries more weight here than in those works, because it is an exactly correct physical model (Prop. 1). The reduction is exact to floating-point round-off: eigenvalues return unchanged and the field matches the analytic composition to 2.1×10^{-15} relative L_2 out to $z = 40$ (E0-G7, Table 4); the state is carried as $(\ln|P_k|, \arg P_k)$, so the agreement is not bitwise. Every ablation in §5.6 therefore measures what learning contributes.

Domain. Darboux dressing inverts the discrete spectrum, so the pipeline operates on the reflectionless (multi-soliton) class, the class nonlinear-Fourier transmission systems use (Yousefi & Kschischang, 2014); §5.9 measures the response to energy outside it.

Training and budget. The objective is a multi-distance relative L_2 error, $\mathcal{L} = \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} \|\hat{q}(\cdot,z) - q(\cdot,z)\|_2^2 / \|q(\cdot,z)\|_2^2$ with $\mathcal{Z} = \{0.5, 1, 2, 3, 4\}$; a conservation-penalty variant of the baseline loss is reported in Appendix A. Two asymmetries favor the baselines and we left both in place: the z-conditioned direct baselines train on all 16 stored frames with $z > 0$, a strict superset of SCNO's five, and SCNO carries $68\,484$ learned parameters against $2\,122\,562$ for the FNO baselines, about $31\times$ more, counting a complex weight as two real degrees of freedom.

4 Theory

For reflectionless initial data the scattering map is $\mathcal{S} : q \mapsto \{(\lambda_k, Q_k)\}_{k=1}^N$ with $\lambda_k = \zeta_k + j\eta_k$ in the upper half-plane, and we use the co-rotating coordinates of equation 2. Proofs are in Appendix B.

Proposition 1 (Exactness at $\epsilon = 0$, zero learned parameters)**.** *Let $\epsilon = 0$ and let q_0 be reflectionless with distinct eigenvalues $\{\lambda_k\}_{k=1}^N$ in the upper half-plane. Then the map*

$$q_0 \;\longmapsto\; \mathrm{Darboux}\Big(\{\lambda_k,\; Q_k\, e^{4j\lambda_k^2 z}\}\Big)$$

equals the NLSE solution operator $\Phi_z(q_0)$ for every $z \geq 0$. Moreover, the SCNO pipeline with a zero-initialized residual module ($f_\theta \equiv 0$) computes exactly this map.

Measured in §5.1, the pipeline's discrepancy against the split-step reference is 4.5×10^{-5} at $z = 4$, the order of the reference's own truth error, and 4.6×10^{-4} at $z = 40$, growing with that error.

Proposition 2 (Size of the learning target)**.** *Let $q_\epsilon(z)$ solve equation 1 with reflectionless data whose solitons stay mutually separated on $[0, Z]$, for one of the deterministic channels (loss, third-order dispersion, Raman), and let $\lambda_k(z)$, $Q_k(z)$ be its adiabatically continued spectrum and norming constants. Write $\tilde{P}_k(z) = Q_k(z)\exp\big(-4j\int_0^z \lambda_k(s)^2\, ds\big)$ for the path-integrated co-rotating coordinate. Then for $z \in [0, Z]$ with ϵZ bounded, and under the validity of first-order adiabatic perturbation theory on that interval (Assumption A, Appendix B),*

$$\big\|\lambda(z) - \lambda(0)\big\| + \big\|\tilde{P}(z) - \tilde{P}(0)\big\| \;\leq\; C_R\,\epsilon\, z \;+\; O(\epsilon^2 z^2), \tag{4}$$

where C_R depends on the channel and on the soliton parameters through first-order soliton perturbation theory (Kaup, 1976; Karpman & Maslov, 1977): for loss, $d\eta_k/dz = -2\epsilon_1\eta_k$ and $d\zeta_k/dz = 0$; for Raman, the self-frequency shift of Gordon (1986); for third-order dispersion, velocity drift plus $O(\epsilon_2)$ radiation. The stochastic channel is excluded: Gordon–Haus jitter grows as z^3 in variance (Gordon & Haus, 1986), an $O(\sigma z^{3/2})$ root-mean-square drift that satisfies no pathwise bound of this form. The coordinate P_k of equation 2, which uses the instantaneous $\lambda_k(z)$ in place of the path integral, differs from $\tilde{P}_k$ by the deterministic phase

$$\Delta\phi_k(z) \;=\; 4\Big(\lambda_k(z)^2 z - \int_0^z \lambda_k(s)^2 ds\Big) \;=\; O(\epsilon z^2). \tag{5}$$

The target f_θ actually learns is then bounded by $C_R\epsilon z + |\Delta\phi_k| = O(\epsilon z + \epsilon z^2)$, one order larger in z.

The frame term is real and the network absorbs it. The frame term equation 5 is large: for the loss channel it evaluates to $16\epsilon_1\eta_k^2 z$ per unit distance, reaching 18.4 at the strongest loss we study. The path-integrated frame removes the secular term by construction and trains worse, 0.544 against 0.317 at $\epsilon_1 = 0.03$, $z = 8$, because the path integral couples every step to the final phase and lengthens the gradient path. We report the corrected theory and keep the coordinate that trains better (Appendix A).

What this buys. Prop. 2 bounds the *size of the learning problem*, and that is the whole argument for the architecture. The learned object is the whole right-hand side f_θ of equation 3, with $\boldsymbol{\epsilon}$ a conditioning input: on the state $(\zeta, \eta, \ln|P|, \arg P)$ it has magnitude $O(\epsilon)$ and, on the training family, Lipschitz constant $O(\epsilon)$. A time-domain solution operator for the same system carries $O(1)$ content at every ϵ; Barron-type rates (Barron, 1993) scale the budget with the norm of the target, so the scattering-coordinate problem is smaller by a factor ϵ (§5.4). The closed-form rates leave the residual its job: they are exact on one channel and order-of-magnitude on the rest. §5.2 shows the learned residual beating the strongest closed-form pipeline exactly where they stop being exact.

Two consequences fail in different places. That the residual is *small*, §5.3 verifies directly in every channel; that a small residual is *easy to learn* is the stronger claim, inheriting the adiabatic hypothesis of Prop. 2. Once ϵz leaves the first-order regime the backbone itself degrades and the advantage should vanish. §5.2 locates the crossover; §5.3 shows the exponent bending past it.

5 Experiments

Every learned model is trained with 3 seeds, and both arms carry 8 on the loss grid; each caption states its own line and band statistic, and we read every band as spread. All randomness is seeded and logged to machine-readable records (Appendix A). An E0 physics gate of 10 checks runs before any training, and all 10 pass (Table 4). Every baseline gets more parameters than SCNO, a superset of its supervision distances, and its own tuning sweep at every grid point. Every rollout starts from exact scattering data at $z = 0$; §5.8 removes that oracle.

5.1 E1: The integrable limit separates exact from learned

We train every baseline on trajectories over $z \in [0, 4]$ and roll out to $z = 40$, a $10\times$ extrapolation, on a bound-state ensemble ($\zeta_k = 0$), so no soliton leaves the time window.

Baseline configurations are selected on in-distribution accuracy at $z = 4$; each FNO family received a 24-configuration sweep and the U-Net 12 (Appendix A). The winner, U-Net (autoregressive), reaches 0.0533 at $z = 4$ and first passes 0.1 at $z = 8.0$. At $z = 40$ its best run reads 1.62 (median 3.11, the value Fig. 2 draws), while the zero-parameter SCNO line sits at 4.6×10^{-4} and never crosses 0.1 out to > 40. No configuration of the 60 swept reaches below 1.00 at $z = 40$, and both baseline numbers are best-of-4-runs (Appendix A).

Refining the split-step size lowers the SCNO line (Appendix A), so every separation quoted is a lower bound. From the field alone, through the production forward transform, the same pipeline reaches 1.2×10^{-4} at $z = 40$ on 64 fields (Fig. 2, dashed black): the soliton count is exact on every field, eigenvalues are recovered to 4.3×10^{-7} and norming constants to 5.2×10^{-6} (§5.8).

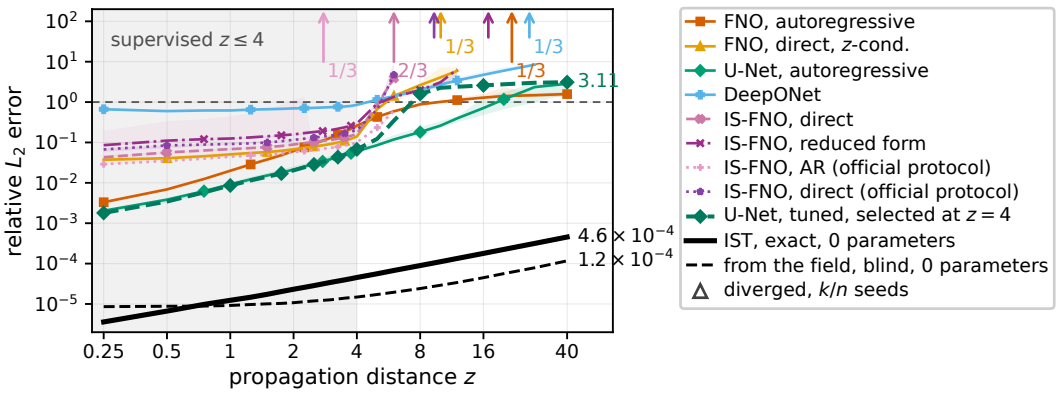


Figure 2: Integrable limit ($\epsilon = 0$): relative L_2 error against distance, both axes logarithmic, training range shaded. Lines are medians over seeds, bands span the seed extremes, arrows mark where k of n seeds pass error 10. Every learned operator ends at or above unity, the score a zero prediction receives, while the zero-parameter IST line stays near 4.6×10^{-4}, with the same pipeline driven from the field alone (dashed black, §5.8) below it. IS-FNO appears in four variants: the direct (dashed) and reduced (dash-dotted) forms of Yu, 2025 under the proposed protocol, and two arms under its official protocol (dotted).

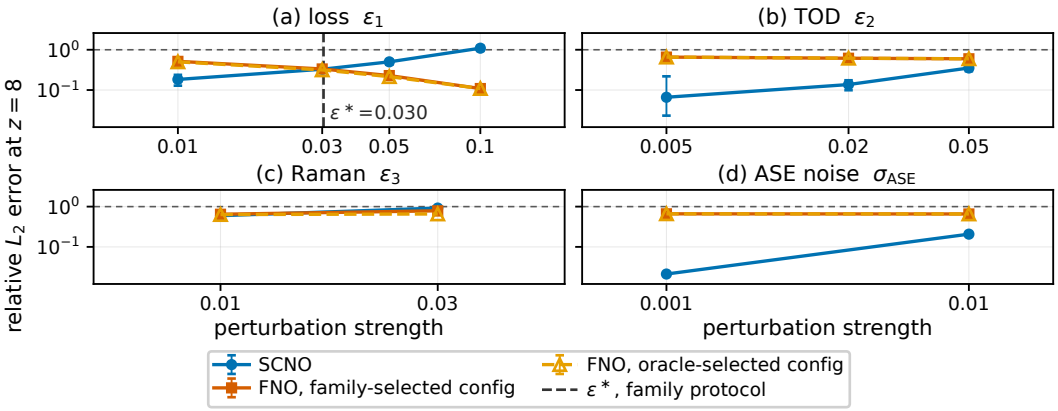


Figure 3: Near-integrable phase diagram: relative L_2 error at $z = 8.0$, twice the training horizon, one panel per channel, logarithmic error axis. Both sides are tuned at every grid point; the baseline appears under the family protocol (filled) and the per-point oracle (open), which coincide almost everywhere. ϵ^* is drawn in panel (a); no other channel crosses in the swept range, and panels (c, d) carry two strengths each, bracketing the regime. The horizontal line is unit error. SCNO leads across noise and third-order dispersion and loses at the right of the loss panel. On Raman it is level with the plotted FNO arms at $\epsilon_3 = 0.01$ and behind them at 0.03. The tuned U-Net of Table 7 is not drawn. Table 8 lists every point.

IS-FNO (Yu, 2025) is the comparison that carries the argument, because its architecture is motivated by the same transform. In the integrable limit it diverges by $z = 40$ under both its official training protocol and the proposed one. The same test was run where nothing is guaranteed, with a four-rate learning-rate sweep and 3 architecture variants (modes, width, evolution depth) at 4 contested grid points of §5.2. IS-FNO is a working operator in distribution (0.431 at worst at $z = 4$, ahead of the direct FNO); by $z = 8$ its best point reads 5.0×10^5. The error grows smoothly through $z = 5$ and then compounds at every rate, every variant and every point (Table 10, Appendix A), which is what a learned exponential evolution does under extrapolation.

5.2 E2: NEAR-INTEGRABLE PHASE DIAGRAM AND THE CROSSOVER ϵ^*

Integrability is broken here and no method is exact: each channel is swept independently, every learner trained separately at every grid point, evaluated at $z = 8.0$, twice the training horizon. DeepONet, the weakest E1 arm (0.855 at $z = 4$ against 0.0533 for the winner), is not carried into the sweep.

Tuning symmetry. Both sides are tuned, because an asymmetric comparison would decide this experiment on protocol. Baselines receive the 24-configuration sweep and SCNO a learning-rate

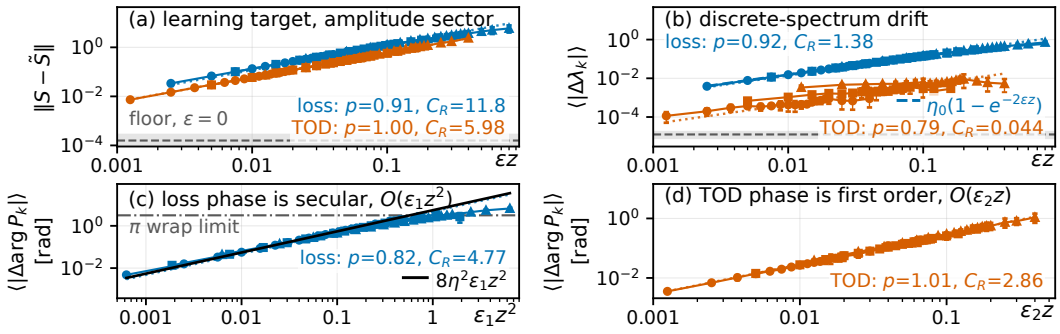


Figure 4: The learning target measured directly, 57 grid points per channel, log–log. *(a)* Distance between measured scattering data $S(z)$ and its analytic rotation $\tilde{S}(z)$, with the pooled fit and the $\epsilon = 0$ floor. *(b)* Eigenvalue drift against the closed-form rate. *(c, d)* The phase sector, secular $O(\epsilon_1 z^2)$ for loss, first order $O(\epsilon_2 z)$ for third-order dispersion. Exponent intervals are 95% cluster bootstraps over the three ϵ values of each channel. Means are over the solitons the Newton continuation tracks (all for $\epsilon_1 \leq 0.03$; 35.4% at $\epsilon_1 = 0.1$, $z = 8$).

selection on a held-out validation split. That sweep selects on the test metric, which is generous to the baselines and unavailable to a practitioner. We therefore also report a *family* protocol: one configuration per baseline chosen by mean performance across all *other* grid points, so no test information about the scored point enters the selection. The crossover is 0.0303 family against 0.0288 oracle, a 5% shift. We quote the family protocol and give both in Appendix A.

Where structure wins. SCNO leads across the whole noise channel (31× at $\sigma = 10^{-3}$, 3.2× at 10^{-2}), the whole third-order-dispersion channel (9.9× down to 1.7× as ϵ_2 grows), and weak loss (2.8×). The advantage widens monotonically as ϵ decreases in every channel.

Where it stops. On the loss channel the ranking inverts at $\epsilon^* = 0.030$, located by log-linear interpolation between the bracketing grid points (distance sweeps in Fig. 9). At $\epsilon_1 = 0.03$ SCNO is level under the family protocol, 0.328 against 0.333 for the strongest FNO, Welch $p = 0.354$ at 8 seeds per arm. By $\epsilon_1 = 0.1$ that arm is 10× better: loss decays every η_k exponentially toward the regime where the discrete spectrum stops organizing the dynamics (§6). Under one held-out validation protocol for every arm, test set untouched, every lead above holds and the crossover reads 0.032 (Table 2).

The E1 winner, a U-Net swept at all 11 points (Table 7), is the strongest baseline on every channel: the $\epsilon_1 = 0.03$ tie becomes a 1.41× deficit and the family crossover moves to 0.022. The leads on weak loss, third-order dispersion and noise hold at 2.33×, 1.17× to 6.1× and 2.17× to 19.6×; on Raman it is ahead at both strengths (§6).

The residual pays wherever first-order theory is incomplete. Table 9 settles what the learning buys. It scores SCNO against the strongest zero-parameter rollout at all 16 perturbed cells of both families: exact rotation, tightened by the drifts of Prop. 2 wherever the rate is exact. The residual wins 10 of 16: 3.0 to 5.0× on third-order dispersion, 1.36 to 1.87× on Raman, 1.29 to 2.57× on KdV, widening with ϵ (Appendix C). The 4 loss cells go to theory, the one rate G5a verifies to 10^{-12}: the integrated drift reads 0.0935 to 0.794, ahead of SCNO by 1.24 to 1.96×. Nothing is left to learn there. On noise the unforced evolution against the realized trajectory is a floor for any deterministic predictor up to the conditional-mean correction, 0.0207 at $\sigma = 10^{-3}$; SCNO sits on it (0.0212) and learns nothing, so that channel measures robustness to unpredictable forcing. One rule covers all 16 cells: the residual pays exactly where first-order theory stops being exact.

Against learned baselines the Raman channel is a loss: level with the FNO at $\epsilon_3 = 0.01$ (1.1×) and behind the U-Net at both strengths, an unsolved case (§6).

5.3 Theory–experiment closure: measuring the residual

Prop. 2 predicts the scattering data drifts from its analytically rotated value as ϵz with a channel constant; we measure that drift directly (Fig. 4), extracting $S(z)$ on a 4× oversampled grid, 57

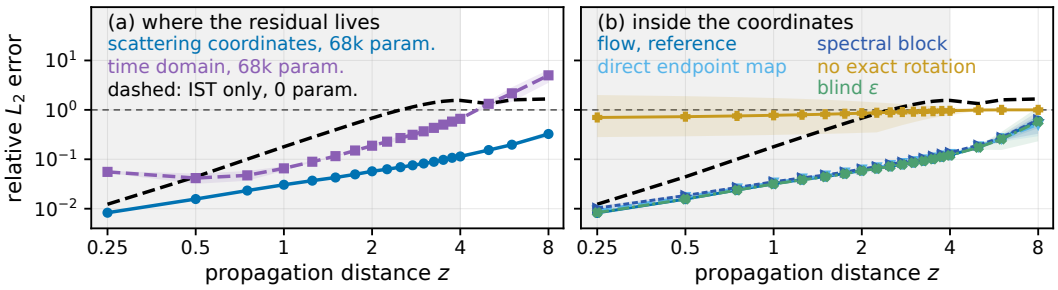


Figure 5: Ablations at $\epsilon_1 = 0.03$: relative L_2 error against distance, log–log, training range shaded, bands 95% over 3 seeds. *(a)* One parameter budget in the co-rotating coordinates and in the time domain, over the same exact IST rollout (dashed; non-monotone: the distance between two soliton trains oscillates with their relative phase). *(b)* Parameterizations inside the coordinates; four of the five arms overlap by design (the $1.1\times$ finding). Read at $z = 4$; at or above 1.00 a curve carries no information.

points per channel. A pooled log-log fit returns $1.00 \pm 5.1 \times 10^{-3}$ on the third-order-dispersion channel against the predicted 1; the loss channel returns 0.908 ± 0.0693, slightly shallower, the signature of first-order saturation of the exponential amplitude decay.

The fit bends past the crossover. On the loss channel the local slope is 0.93 from $\epsilon_1 = 0.01$ to 0.03, the interval holding ϵ^*. From there to 0.1, where SCNO loses by $16.9\times$, it is 0.79. Near ϵ^* the target is still first order; the ranking turns because the baselines' error falls with ϵ_1 while the residual grows at its predicted rate (Appendix A).

5.4 E3: TRAINING-SET SCALING

A secondary check of the §4 target-norm heuristic runs from 64 to 4096 training trajectories at a mid-strength point (Fig. 7, Appendix A). SCNO at $n = 64$ reaches 0.135 at $z = 4$ and the tuned autoregressive FNO 0.150 with all 4096, a thin margin that reverses at $z = 8.0$ from $n = 1024$.

5.5 E4: OUT-OF-DISTRIBUTION GENERALIZATION

Out-of-distribution tests target the two axes the representation makes explicit, against the tuned baseline of §5.2 (Fig. 8, Appendix A). On the energy axis the tuned baseline beats SCNO in distribution by $2.61\times$ (0.0450 against 0.117). One step outside, at $\eta_k \in [0.9, 1.1]$, its error passes unity, the zero-prediction score, and stays there. At $[1.3, 1.5]$ the numbers are 0.321 against 1.47.

The soliton-count axis shows the same as a trend. Training draws one to three solitons; against the strongest baseline the ratio is $1.49 \to 1.91 \to 2.20 \to 2.77$ at $N = 3, 4, 5, 6$, widening monotonically to 0.252 against 0.697 at $N = 6$.

The representation explains both axes: soliton count is a cardinality a permutation-equivariant f_θ extends without retraining, and amplitude enters through η_k, a number the encoder extrapolates along.

5.6 E5: THE REPRESENTATION IS THE CONTRIBUTION, THE ARCHITECTURE INSIDE IT IS INTERCHANGEABLE

The advantage lives in the coordinates: Fig. 5 holds the parameter budget fixed and moves one design element at a time, every variant read at $z = 4$ and $z = 8.0$.

Moving the residual to the time domain is the largest single effect: at the same parameter count it reaches 0.661 at $z = 4$ against 0.114 in scattering coordinates, a factor of 5.8 at $z = 4$, where both arms still carry information. The ablation prices the bound of Prop. 2, which holds only in these coordinates.

Removing the exact rotation removes the advantage entirely: when f_θ must learn the phase evolution too, the $z = 4$ error moves from 0.121 to 1.00, the constant-zero score.

Inside the coordinates the parameterization matters little: a neural-ODE flow (0.121), a direct map (0.129) and a spectral residual (0.136) span $1.1\times$ at $z = 4$, with disjoint seed ranges yet small

beside the $5.8\times$ cost of leaving them. At $z = 8.0$ the three seed ranges overlap, and withholding the perturbation parameters changes nothing (Appendix A).

5.7 Cost against prediction distance

Inference cost: one CPU, batches of 32 fields, quoted per field, cost fitted $\sim z^p$ against distance. SCNO with an active residual integrates, so $p = 1.02$; the autoregressive baseline steps, so $p = 0.95$. At $z = 40$ that leaves SCNO at 87 ms per field against 524 ms ($6.0\times$); the direct z-conditioned FNO is cheaper and flat (2.67 ms) but exceeds unity past the training range. The split-step reference at production accuracy costs 2.39 s per field, $27.5\times$ over SCNO.

5.8 Test-time protocol: removing the transmitter oracle

A deployed receiver must recover the $z = 0$ scattering data from the field, and a hard eigensolver with Newton continuation does, driving the same pipeline (Fig. 11, Appendix A). Blind extraction recovers the exact soliton count on 34 of 35 datasets, and SCNO was re-trained from the cached spectra at all 11 points against matched oracle references. The median change at $z = 8$ is 0%, the range -5% to $+13\%$, the largest Welch $|t| = 1.1$, inside the seed spread everywhere (Fig. 10a, Appendix A). On clean fields the analysis costs 0.238 s per field on CPU (Fig. 2). At $\sigma = 0.003$ the end-to-end error rises to 0.0508 and the §5.1 separation shrinks to $32\times$; the recovered (λ_k, Q_k) lose precision (Appendix A).

5.9 The excluded energy controls the error

Off the manifold the error is set by the energy the coordinates exclude. By the trace formula the energy a reflectionless reconstruction cannot carry, measured at $z = 8$, is at most 3.3×10^{-3} below the crossover and 0.290 at $\epsilon_1 = 0.1$: negligible where SCNO wins, dominant where it loses. Injecting a radiative component of energy fraction ρ at $z = 0$ costs the zero-parameter pipeline $\approx\sqrt{\rho}$: 0.0314 at $\rho = 10^{-3}$, 0.100 at 10^{-2}, 0.380 at 10^{-1}, and the cost is nearly flat in z (0.312 at $z = 4$). The trained residual adds the same amount in quadrature, the excess over $\rho = 0$ matching $\sqrt{\rho}$ within 11% at $z = 4$ and 28% at $z = 8$ ($\epsilon_1 = 0.03$). Natural continuous spectra follow the same law at 1.01 to $1.13 \times \sqrt{\rho}$ on single-soliton families and 1.44 to $1.53\times$ on two-soliton ones (Table 1).

6 Limitations

One perturbation channel is not captured. Under Raman the FNO families are level with SCNO at $\epsilon_3 = 0.01$ ($1.1\times$) and ahead by $1.2\times$ at 0.03; the tuned U-Net leads at both ($1.32\times$, $1.79\times$). Raman drags ζ_k resonantly, a drift the trained f_θ does not capture; an open problem.

Near-integrable means near. Beyond the crossover the ranking inverts: at $\epsilon_1 = 0.1$ the strongest baseline, the tuned U-Net there, is $16.9\times$ better. Strong loss dissolves the soliton hierarchy, the failure Prop. 2's adiabatic hypothesis predicts. The selection rules agree within 5% at 0.030 and a seed bootstrap gives $[0.029, 0.032]$. The training budget moves it to 0.020, the U-Net family value is 0.022, and one held-out validation protocol for every arm reads 0.032. Under that tripled budget the $\epsilon_1 = 0.03$ seed spread of SCNO also grows from 9% to 62%. All results use reflectionless initial data on two integrable families.

7 Conclusion

Exact linearization can be a computational layer. SCNO leaves a residual whose predicted $O(\epsilon z)$ size we measure, whose mechanism the ablations isolate down to the coordinates, and whose structure recurs on a second integrable family; the phase diagram survives removal of its own oracle. The next layer is the reflection coefficient $r(\xi)$: a differentiable Gelfand–Levitan–Marchenko solve on the mixed spectrum extends the backbone to general data, and §5.9 prices what it recovers at $\sqrt{\rho}$.

STATEMENT ON AI USAGE

We disclose the use of large language models in this work, as ICLR policy requires. They were used to polish the writing of the manuscript and to search the related literature.

The authors reviewed every claim, number and proof in this manuscript and take full responsibility for its content.

REPRODUCIBILITY STATEMENT

Appendix A records the environment, the seeds, the data card for NearIntegrableBench-NLSE, the E0 physics-gate table and the hyperparameters as logged. On acceptance we will release publicly the data-generation scripts, every JSONL result log, the figure scripts and the number-generation script. That script writes the macro file behind every number in this paper, and one command regenerates the figures and the numbers from the logs. Each training run appends one record carrying its seed, configuration, per-distance errors, wall-clock time and parameter count, so any number here traces to the run that produced it. The appendix also reports a truth-solver convergence study: refining the split-step size lowers the integrable-limit line, which is why §5.1 reports its separations as lower bounds.

ETHICS STATEMENT

This is foundational methods research on surrogates for integrable and near-integrable evolution equations, and the nearest application is modeling of fiber-optic transmission links. No human subjects, personal data or human-annotated data are involved; every dataset is generated by a numerical solver written for this work. The compute footprint is one consumer GPU for training and CPU for the blind test-time analysis, itemized in Appendix A.

REFERENCES

Mark J. Ablowitz, David J. Kaup, Alan C. Newell, and Harvey Segur. The inverse scattering transform—Fourier analysis for nonlinear problems. *Studies in Applied Mathematics*, 53(4):249–315, 1974. doi: 10.1002/sapm1974534249.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021. arXiv:2003.04887.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Guido Boffetta and Alfred R. Osborne. Computation of the direct scattering transform for the nonlinear Schroedinger equation. *Journal of Computational Physics*, 102(2):252–264, 1992. doi: 10.1016/0021-9991(92)90370-E.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alex Colagrande, Paul Caillon, Eva Feillet, and Alexandre Allauzen. Higher-order Fourier neural operator: Explicit mode mixer for nonlinear PDEs. *arXiv preprint arXiv:2606.28122*, 2026.

Steven M. Cox and Paul C. Matthews. Exponential time differencing for stiff systems. *Journal of Computational Physics*, 176(2):430–455, 2002.

Ludwig D. Faddeev and Leon A. Takhtajan. *Hamiltonian Methods in the Theory of Solitons*. Springer, Berlin, Heidelberg, 1987. doi: 10.1007/978-3-540-69969-9.

James P. Gordon. Theory of the soliton self-frequency shift. *Optics Letters*, 11(10):662–664, 1986. doi: 10.1364/OL.11.000662.

James P. Gordon and Hermann A. Haus. Random walk of coherently amplified solitons in optical fiber transmission. *Optics Letters*, 11(10):665–667, 1986. doi: 10.1364/OL.11.000665.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. arXiv:1906.01563.

Vladimir I. Karpman and Evgenii M. Maslov. Perturbation theory for solitons. *Soviet Physics JETP*, 46(2):281–291, 1977.

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

David J. Kaup. A perturbation expansion for the Zakharov–Shabat inverse scattering transform. *SIAM Journal on Applied Mathematics*, 31(1):121–133, 1976. doi: 10.1137/0131013.

Zongyi Li, Nikola 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 (ICLR)*, 2021. arXiv:2010.08895.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. doi: 10.1002/cpa.21413.

Vladimir B. Matveev and Mikhail A. Salle. *Darboux Transformations and Solitons*. Springer Series in Nonlinear Dynamics. Springer, Berlin, Heidelberg, 1991. doi: 10.1007/978-3-662-00922-2.

Maziar Raissi, Paris Perdikaris, and George Em 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. doi: 10.1016/j.jcp.2018.10.045.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. arXiv:2302.01178.

Sergei K. Turitsyn, Jaroslaw E. Prilepsky, Son Thai Le, Sander Wahls, Leonid L. Frumin, Morteza Kamalian, and Stanislav A. Derevyanko. Nonlinear Fourier transform for optical data processing and transmission: advances and perspectives. *Optica*, 4(3):307–322, 2017. doi: 10.1364/OPTICA.4.000307.

Kiwon Um, Robert Brand, Yun (Raymond) Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-loop: Learning from differentiable physics to interact with iterative PDE-solvers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. arXiv:2007.00016.

Sander Wahls, Shrinivas Chimmalgi, and Peter J. Prins. FNFT: A software library for computing nonlinear Fourier transforms. *Journal of Open Source Software*, 3(23):597, 2018.

Mansoor I. Yousefi and Frank R. Kschischang. Information transmission using the nonlinear Fourier transform, Part I: Mathematical tools. *IEEE Transactions on Information Theory*, 60(7):4312–4328, 2014. doi: 10.1109/TIT.2014.2321143.

Rixin Yu. An inverse scattering inspired Fourier neural operator for time-dependent PDE learning. *arXiv preprint arXiv:2512.19439*, 2025.

Vladimir E. Zakharov and Aleksei B. Shabat. Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Physics JETP*, 34(1):62–69, 1972.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2302.05543.

Darko Zibar, Molly Piels, Rasmus Jones, and Christian G. Schäffer. Machine learning techniques in optical communication. *Journal of Lightwave Technology*, 34(6):1442–1452, 2016.

A Reproducibility details

A.1 Environment

Experiments run under Python 3.11.4 with PyTorch 2.0.1+cu117. Training uses a single NVIDIA RTX 4090 (24 GB), and the blind test-time analysis protocol of §5.8 runs on CPU at $0.238\,\mathrm{s}$ per sample. The E0 physics gate and the Darboux reference constructions use double precision (complex128); training uses complex64, a choice the gate validates (Table 4).

Asymptotic cost. Forward analysis evaluates the transfer-matrix Jost coefficient $a(\lambda)$ in $O(N_\tau)$ per trial eigenvalue. The search runs a 25×20 grid and then at most 17 Newton iterations for each of up to 12 candidates. Darboux dressing costs $O(NN_\tau)$ for N solitons. The residual flow costs $O(N)$ network evaluations per RK4 step over $z/\Delta z$ steps, independent of N_τ. Every stage is batched over fields.

A.2 Seeds and statistics

All learned models are trained with 3 seeds ($\{0, 1, 2\}$), and the loss-grid arms of §5.2 carry five more ($\{3, \ldots, 7\}$: `t12_scno_seeds.jsonl` for SCNO, the `family_seeds` and `seeds` phases of `r1_tune.jsonl` for the baselines); figures report mean $\pm$ 95% CI over seeds. Dataset generation uses fixed, disjoint seeds for train/test splits (train seed 2; test seeds 900/901 for the E1/E2 ensembles). Every run appends one JSON record (seed, config, all per-z errors, wall-clock, parameter count) to a machine-readable log in `results/logs/*.jsonl`; every number in the main text is generated from these logs by `paper/make_numbers.py` followed by `paper/finalize_numbers.py`. The second script also fixes the two selection rules the paper states: E1 baseline configurations are chosen on the $z = 4$ error and E5 ablation ratios are quoted at $z = 4$.

Seed stability of the E1 baseline. The E1 sweep trains one seed per configuration, so we re-ran the selected configuration (U-Net, lr 2×10^{-3}, width 64) on seeds $\{0, 1, 2\}$. That gives 4 runs over 3 seeds, since the sweep already contributed one. Selection was fixed before the re-run, so the extra seeds cannot leak the $z = 40$ metric into the choice. The runs are not interchangeable. The best reaches 1.62 at $z = 40$, the median 3.11, and 1 diverges to 6.5×10^7 while also failing inside the training range. Every number the main text attributes to this baseline is the minimum over the four runs, which is the best case for the baseline and the smallest separation the data support. Aggregating by mean or median would enlarge the separation we report.

Perturbation conditioning is not used. SCNO can take the perturbation parameters $\boldsymbol{\epsilon}$ as an input, and withholding them changes nothing. Training on a mixture of three loss strengths gives 1.04 with the parameters and 1.00 without, indistinguishable under a Welch t-test, and supplying

deliberately wrong values is indistinguishable too. The conditioning channel is decorative at these settings. Mixed-perturbation training is worse than per-point training for every method: the tuned autoregressive FNO trained on the same three-strength mixture at matched budget saturates past unity at every point (1.22 to 1.69 at $z = 8$, 3 seeds, `results/logs/t10_mixed_fno.json`), the same collapse the SCNO arms show. The failure is a property of the mixture and of no single method; it limits the present setup and leaves the coordinates untouched.

Off-manifold inputs (the §5.9 measurement). Injecting a smooth radiative component of energy fraction ρ into the $z = 0$ field (`experiments/t9_radiative_inputs.py`, 32 fields, SSFM truth) costs the zero-parameter blind pipeline a relative error of $\approx\sqrt{\rho}$: 0.0313/0.0314 at $\rho = 10^{-3}$, 0.0989/0.100 at 10^{-2}, and 0.312/0.380 at 10^{-1} (at $z = 4$/$z = 8$), with $\rho = 0$ reducing to the 2.0×10^{-5} floor of Prop. 1. The error tracks the amplitude of the discarded radiation and barely grows with distance, since the dropped component disperses out of the window; the discrete spectrum retains 99.9% of the energy at $\rho = 10^{-3}$ and 91% at 10^{-1}.

The trained residual was measured on the same inputs (`experiments/t16_trained_offmanifold.py`: three seeds at $\epsilon_1 = 0.03$, 64 held-out fields, the hard eigensolver supplying the $z = 0$ spectrum, SSFM truth with the loss active). At $z = 4$ its error is 0.107, 0.113, 0.148 and 0.339 at $\rho = 0$, 10^{-3}, 10^{-2}, 10^{-1}; at $z = 8$ it is 0.307, 0.308, 0.318 and 0.482, with a seed range of 0.432 to 0.520 at 10^{-1}. The excess over $\rho = 0$ in quadrature is 0.0229, 0.0830 and 0.372 at $z = 8$, the $\sqrt{\rho}$ law of the zero-parameter arm again. The rotation-only arm ignores the loss channel and sits near 1.75 throughout, so at $\epsilon_1 > 0$ the residual carries the perturbation and the radiation is paid separately.

Natural continuous spectra. The injected component is then replaced by initial fields whose reflection coefficient is nonzero by construction (`experiments/t18_mixed_spectrum.py`, 32 fields per family, random centre and phase): $A\,\mathrm{sech}\,\tau$, Gaussian pulses $Ae^{-\tau^2/2}$ and chirped solitons $\mathrm{sech}\,\tau\, e^{jC\tau^2/2}$, 12 families in all. The radiative fraction ρ is measured by the analyzer; on the sech family it agrees with the Satsuma–Yajima closed form to four digits. Table 1 lists the zero-parameter error at $\epsilon = 0$ and the trained residual at $\epsilon_1 = 0.03$. Across ρ from 7.0×10^{-3} to 0.202 the single-soliton families sit at 1.01 to $1.13 \times \sqrt{\rho}$ at $z = 8$, the two-soliton families, each carrying a weak soliton, at 1.44 to 1.53×. The trained residual fails on the three two-soliton families, whose eigenvalues (1.5 and 0.5; 1.46 and 0.18; 1.2 and 0.2) fall outside the training range $\eta \in [0.3, 1.2]$: an extrapolation in η, present already at $\rho = 0$ ($A = 2$), and the behaviour §5.5 measures.

Family	A or C	ρ	$\sqrt{\rho}$	zero-param. z=4	z=8	trained z=8
$A\,\mathrm{sech}\,\tau$	1	0	0	0	0	0.162
$A\,\mathrm{sech}\,\tau$	1.15	0.0170	0.130	0.140	0.137	0.207
$A\,\mathrm{sech}\,\tau$	1.3	0.0532	0.231	0.266	0.237	0.306
$A\,\mathrm{sech}\,\tau$	1.45	0.0963	0.310	0.330	0.329	0.443
$A\,\mathrm{sech}\,\tau$	1.7	0.0311	0.176	0.237	0.254	0.370
$A\,\mathrm{sech}\,\tau$	2	0	0	4.4×10^{-4}	1.0×10^{-3}	0.976
$Ae^{-\tau^2/2}$	1	0.0505	0.225	0.241	0.233	0.275
$Ae^{-\tau^2/2}$	1.5	0.0481	0.219	0.221	0.232	0.334
$Ae^{-\tau^2/2}$	2	0.0754	0.275	0.393	0.419	0.732
$\mathrm{sech}\,\tau\, e^{jC\tau^2/2}$	0.1	7.0×10^{-3}	0.0838	0.0911	0.0867	0.168
$\mathrm{sech}\,\tau\, e^{jC\tau^2/2}$	0.3	0.0585	0.242	0.255	0.274	0.251
$\mathrm{sech}\,\tau\, e^{jC\tau^2/2}$	0.6	0.202	0.449	0.502	0.453	0.436

Table 1: Inputs with a natural continuous spectrum: measured radiative fraction, the zero-parameter pipeline at $\epsilon = 0$ (median relative L_2), and the trained residual at $\epsilon_1 = 0.03$ (mean over three seeds of the median). The zero-parameter error tracks $\sqrt{\rho}$; the trained arm holds on the single-soliton families and fails where an eigenvalue lies outside its training range.

Co-rotating frame and the residual rate cap. Two implementation choices were probed at $\epsilon_1 = 0.03$ with three seeds per cell, evaluated at $z = 8$ (`results/logs/r7_frame.jsonl`). The instantaneous frame of equation 2 reaches 0.317 at the tight rate cap and 0.320 at the loose one, a difference inside the seed spread, so the cap is not binding there. The path-integrated frame is

sensitive to it, 0.544 against 0.728, and loses to the instantaneous frame at either setting. Production runs use the instantaneous frame and the loose cap.

An invalidated batch. For one day the `ResidualFlow` constructor carried `frame="integrated"` as its default, and no training driver passed the keyword, so every SCNO run started in that window silently used the path-integrated frame while its references used the instantaneous one. The two differ by $2.2\times$ on the loss channel (the calibration cells `cal_*` in `e2_results.jsonl` measure both frames on the same data). The affected records are `r12_budget.jsonl` and the `_dz`-tagged rows written in that window; they stay in the logs for the audit trail, no macro reads them, and every affected comparison was re-run with the frame passed explicitly and written into each record. The training-budget and integrator-step numbers quoted in this paper come only from the re-run, frame-matched cells. Two claims drawn from the contaminated runs were retracted before any of them entered a submitted draft.

Two staged sweeps, re-run. A U-Net sweep at the contested phase-diagram points and a per-point learning-rate sweep of the KdV baselines were first run in a staging area whose raw logs were deleted before archiving; only the run summaries survived, kept as `t5_unet_e2_summary.json` and `t6_kdv_tune_summary.json`. Both sweeps were then re-run from scratch with every per-seed row kept (`t13_unet.jsonl`, all eleven points; `t14_kdv_tune.jsonl`). The KdV medians reproduce the recovered values exactly and the U-Net medians within seed noise; every number quoted from either sweep comes from the re-run.

One validation protocol for every arm. The sweeps above select the baselines on the test metric (oracle) or on the test sets of the other grid points (family), while SCNO selects its rate on a held-out split. Table 2 re-runs the phase diagram under one rule for every arm (`t17_val_protocol.jsonl`, 723 training runs). The split is the t12 rule, seed-keyed and 90/10. Every configuration of the method's grid trains on the 90% at seed 0 and is ranked by its rollout error at $z = 4$ on the held-out 10%; the winner is re-seeded to three; the test file is read only afterwards. A re-seeded run whose held-out error reaches the no-information level (≥ 0.9) fails the check and the next-ranked configuration is taken. The check fired only for SCNO at learning rate 3×10^{-3}, which diverged on re-seeding at 4 points and at seed 0 at five others, so the accepted SCNO rate is 10^{-3} everywhere; no baseline re-seed failed it. Every lead of §5.2 holds. The U-Net selected on the held-out split is weaker than the test-selected one of Table 7 (0.745 against 0.426 at $\epsilon_1 = 0.01$), and the loss crossover reads 0.032 against 0.030 under the family protocol.

Point	SCNO	FNO (AR)	FNO (direct)	U-Net	best baseline / SCNO
$\epsilon_1 = 0.01$	0.274 [0.188, 0.286]	0.549 [0.513, 0.649]	1.69 [1.55, 1.81]	0.745 [0.728, 1.40]	2.00×
$\epsilon_1 = 0.03$	0.319 [0.298, 0.327]	0.357 [0.309, 0.357]	1.84 [1.60, 1.85]	0.378 [0.266, 2.68]	1.12×
$\epsilon_1 = 0.05$	0.516 [0.462, 0.540]	0.225 [0.221, 0.245]	2.11 [1.37, 2.21]	0.231 [0.142, 0.432]	1/2.29×
$\epsilon_1 = 0.1$	1.07 [0.924, 1.09]	0.112 [0.109, 0.123]	1.39 [1.29, 1.54]	0.0699 [0.0694, 0.0716]	1/15.3×
$\epsilon_2 = 0.005$	0.0506 [0.0366, 0.120]	0.660 [0.657, 0.660]	1.52 [1.40, 2.19]	0.460 [0.438, 0.605]	9.1×
$\epsilon_2 = 0.02$	0.137 [0.134, 0.151]	0.661 [0.620, 71.1]	1.29 [1.18, 1.95]	0.434 [0.427, 0.480]	3.2×
$\epsilon_2 = 0.05$	0.317 [0.310, 0.377]	0.621 [0.612, 0.633]	1.38 [1.19, 1.79]	0.432 [0.408, 0.454]	1.36×
$\epsilon_3 = 0.01$	0.602 [0.568, 0.641]	0.665 [0.626, 0.672]	1.91 [1.47, 2.06]	0.493 [0.481, 0.507]	1/1.22×
$\epsilon_3 = 0.03$	0.917 [0.911, 0.978]	1.13 [0.715, 2.1×10^4]	1.11 [1.09, 1.15]	0.556 [0.526, 0.564]	1/1.65×
$\sigma = 10^{-3}$	0.0209 [0.0208, 0.0210]	0.660 [0.625, 0.665]	1.56 [1.30, 2.48]	0.465 [0.378, 0.519]	22.3×
$\sigma = 10^{-2}$	0.207 [0.207, 0.209]	0.649 [0.633, 2.18]	1.74 [1.67, 1.83]	0.481 [0.448, 0.485]	2.32×

Table 2: The phase diagram under one validation protocol for every arm: relative L_2 error at $z = 8.0$, median over three seeds with the seed range in brackets, every configuration selected on a held-out split at $z = 4$ and the test set untouched. The last column is the strongest validation-selected baseline over SCNO; $1/x$ marks a loss.

A.3 Data card: NearIntegrableBench-NLSE

Table 3 states the grid, the initial-data distribution, the perturbation channels and the splits.

Grid	$N_\tau = 512$, $\tau \in [-25, 25]$ ($h \approx 0.098$)
Initial data	$N \sim \mathrm{Unif}\{1..3\}$ solitons, Darboux-synthesized (reflectionless)
Parameter ranges	$\eta_k \in [0.3, 1.2]$, $\zeta_k \in [-0.8, 0.8]$, $\tau_{0,k} \in [-10, 10]$, $\phi_k \in [0, 2\pi)$
Separation constraints	$\lvert\lambda_i - \lambda_j\rvert \geq 0.25$, $\lvert\tau_{0,i} - \tau_{0,j}\rvert \geq 2$
Trajectories	4096 train; 512 (E1) / 256 (E2 per grid point) test
Stored frames	$z \in \{0, 0.25, \ldots, 4\}$ (17 frames per trajectory)
Truth solver	SSFM, 2000 steps/unit distance (convergence study below)
Perturbation grid	$\epsilon_1 \in \{0, .01, .03, .05, .1\}$, $\epsilon_2 \in \{0, .005, .02, .05\}$
	$\epsilon_3 \in \{0, .01, .03\}$, $\sigma \in \{0, 10^{-3}, 10^{-2}\}$
E1 ensemble	bound states ($\zeta = 0$) to keep $z = 40$ rollouts in-window
E4 sub-ranges	soliton axis trains on $N \leq 3$; energy axis on $\eta_k \in [0.3, 0.8]$

Table 3: Data card (generation scripts and seeds are part of the release).

A.4 E0 PHYSICS GATE (MEASURED)

Table 4 lists the ten checks with their criteria and measured values; every check runs before any training.

Gate	Check	Criterion	Measured	Pass
G1	Darboux 1-soliton vs. closed form (L_∞, complex128)	$< 1.0 \times 10^{-12}$	4.6×10^{-15}	✓
G2a	NFT∘INFT round trip on Q_k (production grid, N_τ=512)	< 0.0600	0.0499	✓
G2b	NFT∘INFT round trip on Q_k (validation grid, N_τ=4096)	$< 1.5 \times 10^{-3}$	6.7×10^{-4}	✓
G3	SSFM vs. Darboux$(\lambda_k, Q_k e^{4j\lambda_k^2 z})$ at z=4 (rel. L_2)	$< 5.0 \times 10^{-4}$	5.3×10^{-5}	✓
G4	Phase-law slope $4\eta^2$ (relative error)	$< 1.0 \times 10^{-3}$	1.9×10^{-6}	✓
G5a	Perturbed solver vs. adiabatic theory: loss $e^{-2\varepsilon_1 z}$	$< 1.0 \times 10^{-6}$	8.1×10^{-13}	✓
G5b	Raman self-frequency shift (order-of-magnitude window)	< 20.0	4.00	✓
G6	Autograd through full Q_k extraction vs. finite differences	$< 1.0 \times 10^{-4}$	1.4×10^{-7}	✓
G7a	Zero-init pipeline vs. analytic composition: λ_k (rel.)	$< 1.0 \times 10^{-15}$	0	✓
G7b	Same on the field q (rel. L_2, worst over $z \leq 40$)	$< 1.0 \times 10^{-12}$	2.1×10^{-15}	✓

Table 4: E0 physics gate, all checks green (read from `results/logs/e0_gate.jsonl`). Fixed implementation facts: Q–b convention $Q_k = jb_k$; $4\times$ oversampled FFT band-limited interpolation for scattering-data extraction; co-rotating state (λ_k, P_k) so that zero-initialized SCNO matches the analytic IST pipeline at the round-off level G7 measures. G5b verifies the Raman rate only to order of magnitude (measured/theory ≈ 4), consistent with Raman being the channel the residual model does not capture; the loss rate G5a is verified to 10^{-12}. G2a exercises the transfer-matrix forward transform on the coarse production grid, where the Q_k round trip is dominated by grid aliasing; G2b on the 4096-point grid is the convergence check, and the production pipeline never invokes this forward pass (§3). The §5.8 blind protocol uses the Newton-continuation eigensolver instead, whose field-level reconstruction reaches 2.9×10^{-6}.

A.5 TRUTH-SOLVER CONVERGENCE

The integrable-limit line of §5.1 measures the reference integrator, so its level has to be checked against the integrator's own step. Halving the split-step size lowers that line by close to the factor expected from the second-order symmetric splitting, which places the line at the truth error of the reference and leaves the SCNO pipeline error below it. Every separation quoted in §5.1 is a lower bound for this reason. The step-halving curve is regenerated by the convergence script in the release.

A.6 HYPERPARAMETERS (AS LOGGED)

Table 5 lists the training configuration of every model in the paper and Table 6 lists the winner of each baseline tuning sweep, both read from the run records. Every model, baselines included, trains with AdamW at weight decay 10^{-4}, cosine annealing of the learning rate over the run, and gradient-norm clipping at 1.0. One qualification belongs here. The E1 sweep of §5.1 trains one seed per configuration, so the numbers quoted for its $z = 4$-selected winner are single-seed measurements; the 3-seed protocol applies to every comparison in §5.2 onward.

Rollout-stabilized training of the autoregressive baselines. The pushforward trick (Brandstetter et al., 2022) was applied to the three autoregressive arms at the 4 contested points, three seeds each (`results/logs/t15_pushforward.jsonl`): per batch an unroll length $m \sim \mathrm{U}\{1..M\}$ is

Model	Params	Epochs	lr	Batch	Notes
SCNO f_θ (flow)	68,484	60	0.0003	128	RK4 Δz=0.1, z-supervision $\{0.5,1,2,3,4\}$
FNO (AR stepper)	2,122,562	200	0.002	–	step 0.25, adjacent-frame pairs
FNO (direct, z-cond.)	2,122,626	200	0.002	–	supervised on all 16 frames $z \in \{0.25, \ldots, 4\}$
U-Net (AR stepper)	577,682	200	0.002	128	E1 protocol
DeepONet	778,400	200	0.002	128	E1 protocol
IS-FNO (direct)	1,461,762	200	0.002	128	E1 protocol

Table 5: Training configurations extracted from `results/logs` (defaults; per-point deviations recorded in the per-record jsonl fields). The supervision sets differ by model by design and the difference favors the baselines (§3).

Baseline	Grid point	lr	Width	Modes	Err. (z=8)
FNO (autoregressive)	eps1_0.01	0.004	64	32	0.499
FNO (autoregressive)	eps1_0.03	0.004	64	32	0.315
FNO (autoregressive)	eps1_0.05	0.004	64	32	0.221
FNO (autoregressive)	eps1_0.1	0.004	96	64	0.118
FNO (autoregressive)	eps2_0.005	0.004	96	64	0.627
FNO (autoregressive)	eps2_0.02	0.004	96	64	0.601
FNO (autoregressive)	eps2_0.05	0.004	96	64	0.602
FNO (autoregressive)	eps3_0.01	0.004	64	64	0.634
FNO (autoregressive)	eps3_0.03	0.002	96	64	0.665
FNO (autoregressive)	sigma_ase_0.001	0.004	96	64	0.640
FNO (autoregressive)	sigma_ase_0.01	0.004	96	64	0.638
FNO (direct)	eps1_0.01	0.0005	64	32	1.18
FNO (direct)	eps1_0.03	0.001	32	64	1.44
FNO (direct)	eps1_0.05	0.001	64	64	1.30
FNO (direct)	eps1_0.1	0.001	64	64	1.19
FNO (direct)	eps2_0.005	0.001	32	64	1.20
FNO (direct)	eps2_0.02	0.004	64	32	1.31
FNO (direct)	eps2_0.05	0.004	96	64	1.40
FNO (direct)	eps3_0.01	0.004	96	64	1.07
FNO (direct)	eps3_0.03	0.001	32	32	1.12
FNO (direct)	sigma_ase_0.001	0.004	64	32	1.28
FNO (direct)	sigma_ase_0.01	0.004	64	32	1.16

Table 6: r1 baseline-tuning winners (search space: lr $\times$ width $\times$ modes, $\geq$24 configs per baseline per point; full sweep in `results/logs/r1_tune.jsonl`).

Point	SCNO	U-Net	config	n
eps1_0.01	0.183	0.426	lr 0.002, ch 32	3
eps1_0.03	0.328	0.232	lr 0.001, ch 64	3
eps1_0.05	0.501	0.125	lr 0.002, ch 64	3
eps1_0.1	1.09	0.0646	lr 0.002, ch 64	3
eps2_0.005	0.0661	0.405	lr 0.002, ch 64	3
eps2_0.02	0.137	0.383	lr 0.002, ch 64	3
eps2_0.05	0.352	0.414	lr 0.002, ch 64	3
eps3_0.01	0.595	0.451	lr 0.002, ch 64	3
eps3_0.03	0.921	0.515	lr 0.002, ch 64	3
sigma_ase_0.001	0.0212	0.416	lr 0.002, ch 64	3
sigma_ase_0.01	0.207	0.448	lr 0.002, ch 64	3

Table 7: The tuned U-Net at all 11 phase-diagram points (`results/logs/t13_unet.jsonl`): per-point winner of a six-configuration sweep (two rates, three widths) re-seeded to three, $z = 8.0$ medians against the matched SCNO reference. It is the strongest baseline on every channel, ahead of SCNO on strong loss and on Raman.

drawn, the first $m - 1$ steps run without gradient, and the loss differentiates through the last step only. Three settings were trained: $M = 4$ from the first epoch, and a curriculum that trains one-step for half the epochs and then ramps the cap to $M = 4$ or $M = 2$. The $M = 1$ arm of the same loop reproduces the standard baselines (FNO 0.326 to 0.651 against 0.333 to 0.662; U-Net 0.221 to 0.425 against 0.232 to 0.426), so the loop is sound. Every stabilized setting lands past unity at $z = 8$ at every point: the FNO reads 1.18 to 1.28 uniform, 1.21 to 1.28 and 1.07 to 1.27 under the curriculum; the U-Net 1.25 to 2.15, 1.24 to 6.11 and 1.91 to 2.81. The training curves show why:

Point	SCNO	FNO (family)	FNO (oracle)	ratio
eps1_0.01	0.183 [0.140, 0.349]	0.515 [0.482, 0.558]	0.500	2.81
eps1_0.03	0.328 [0.298, 0.400]	0.333 [0.298, 0.359]	0.316	1.02
eps1_0.05	0.501 [0.439, 0.686]	0.229 [0.215, 0.239]	0.215	0.46
eps1_0.1	1.09 [0.807, 1.33]	0.109 [0.106, 0.113]	0.109	0.10
eps2_0.005	0.0661 [0.0508, 0.164]	0.657 [0.624, 0.674]	0.657	9.9
eps2_0.02	0.137 [0.130, 0.159]	0.616 [0.605, 0.640]	0.616	4.5
eps2_0.05	0.352 [0.341, 0.386]	0.597 [0.561, 0.655]	0.597	1.69
eps3_0.01	0.595 [0.576, 0.634]	0.647 [0.631, 0.661]	0.634	1.09
eps3_0.03	0.921 [0.918, 0.986]	0.787 [0.713, 1.7×10^5]	0.656	0.85
sigma_ase_0.001	0.0212 [0.0210, 0.0217]	0.662 [0.612, 0.701]	0.662	31
sigma_ase_0.01	0.207	0.657 [0.637, 0.676]	0.657	3.2

Table 8: The §5.2 phase diagram as a table: relative L_2 error at $z = 8.0$, median over seeds with the seed range in brackets. The family configuration carries eight seeds at every point and SCNO eight on the loss grid (three elsewhere); oracle winners were reseeded to three, and to eight on the loss grid. One family seed at $\epsilon_3 = 0.03$ diverged and sits in the range as logged. The last column is the strongest FNO family arm over SCNO.

Cell	reference	ref.	SCNO	gain	winner
eps1_0.01	first-order drift	0.0935	0.183	0.51	closed form
eps1_0.03	first-order drift	0.265	0.328	0.81	closed form
eps1_0.05	first-order drift	0.402	0.501	0.80	closed form
eps1_0.1	first-order drift	0.794	1.09	0.73	closed form
eps2_0.005	exact rotation	0.203	0.0661	3.1	residual
eps2_0.02	exact rotation	0.690	0.137	5.0	residual
eps2_0.05	exact rotation	1.07	0.352	3.0	residual
eps3_0.01	exact rotation	1.11	0.595	1.87	residual
eps3_0.03	exact rotation	1.25	0.921	1.36	residual
sigma_ase_0.001	exact rotation	0.0207	0.0212	0.98	tie
sigma_ase_0.01	exact rotation	0.205	0.207	0.99	tie
KdV $\epsilon = 0.005$	first-order drift	0.0161	0.0125	1.29	residual
KdV $\epsilon = 0.01$	first-order drift	0.0321	0.0249	1.29	residual
KdV $\epsilon = 0.02$	first-order drift	0.0637	0.0291	2.19	residual
KdV $\epsilon = 0.05$	first-order drift	0.156	0.0635	2.45	residual
KdV $\epsilon = 0.1$	first-order drift	0.295	0.115	2.57	residual

Table 9: The zero-parameter accounting behind §5.2: SCNO against the strongest closed-form pipeline at every perturbed cell of both families, from `results/logs/t11_zero_param.json`. The reference is the exact rotation, tightened by the integrated first-order drifts wherever the rate is verified exact (loss via G5a; KdV damping); G5b verifies the third-order dispersion and Raman rates to order of magnitude only, so no drift line exists there. NLSE cells are test-set means at $z = 8.0$, KdV cells seed medians of test-set means at $t = 4$, matching each family's reporting protocol. Gain is reference over SCNO: above one the learned residual improves on everything closed-form theory supplies.

the one-step loss sits near 10^{-4} at the end of the warm-up, and the first unrolled epoch lifts it to a few 10^{-2}, where it stays. The mixed objective buys robustness to its own drift at the price of one-step precision, and at $\Delta z = 0.25$ over 32 steps the precision is what the rollout needs. IS-FNO as a unit stepper diverges at every point under one-step training and pushforward, uniform or curriculum, holds it at 1.17 to 1.28, no information either way.

A.7 Figures for the main-text experiments

Three experiments are reported numerically in the main text and plotted here. Fig. 7 is the sample-efficiency curve of §5.4, Fig. 8 the out-of-distribution axes of §5.5, and Fig. 9 the distance sweep that accompanies the phase diagram of §5.2. Fig. 6 renders one stored trajectory of the loss channel in the time domain.

The two sector exponents below one in Fig. 4. The loss-channel phase sector fits 0.817 ± 0.0886 against the predicted 1. First-order theory itself predicts the shortfall. The proof of Prop. 2 takes the eigenvalue drift as linear in z, while the loss channel decays $\eta_k(z) = \eta_k(0)e^{-2\epsilon_1 z}$ exactly; the accumulated phase then grows sublinearly in ϵ_1 once $2\epsilon_1 z$ approaches one, and the top grid point reaches $2\epsilon_1 z = 0.8$ by $z = 4$. This saturation sits at the top grid point, past ϵ^*. The eigenvalue

Point	lr	z=4	z=5	z=6	z=8
eps1_0.01	0.0005	0.592	4.3×10^2	1.4×10^6	2.2×10^{11}
eps1_0.01	0.001	0.464	17.2	1.5×10^4	7.1×10^8
eps1_0.01	0.0025	0.409	1.116	31.8	5.0×10^5
eps1_0.01	0.005	0.415	0.872	40.8	2.8×10^7
eps1_0.03	0.0005	0.599	2.3×10^2	3.2×10^5	3.5×10^{10}
eps1_0.03	0.001	0.419	5.778	1.8×10^3	1.6×10^7
eps1_0.03	0.0025	0.364	0.947	42.2	3.2×10^6
eps1_0.03	0.005	0.365	0.792	22.1	6.6×10^6
eps2_0.005	0.0005	0.690	3.0×10^2	1.3×10^6	3.2×10^{11}
eps2_0.005	0.001	0.551	54.4	5.9×10^4	8.3×10^9
eps2_0.005	0.0025	0.431	1.089	30.6	1.2×10^6
eps2_0.005	0.005	0.447	0.949	45.7	4.5×10^7
sigma_ase_0.001	0.0005	0.682	3.6×10^2	9.6×10^5	1.7×10^{11}
sigma_ase_0.001	0.001	0.552	39.9	2.9×10^4	2.1×10^9
sigma_ase_0.001	0.0025	0.424	1.075	28.1	5.6×10^5
sigma_ase_0.001	0.005	0.449	0.957	48.0	4.8×10^7

Table 10: IS-FNO in the contested regime (§5.1): median relative L_2 error at four distances, for every learning rate in the sweep at every contested grid point, from `results/logs/t3_isfno_e2.jsonl`. Every rate is competent at $z = 4$, grows through $z = 5$, and has compounded by $z = 8$; the divergence is a property of the learned exponential evolution under extrapolation, of no single run or rate. Widening the sweep to 3 architecture variants at the selected rate (modes 256; width $d_b = 45$; evolution depth 8; `results/logs/t15_isfno_wide.jsonl`) changes nothing: the variants read 0.328 to 0.427 at $z = 4$ and 12 of 12 cells sit past unity at $z = 8$, the best at 5.1×10^5.

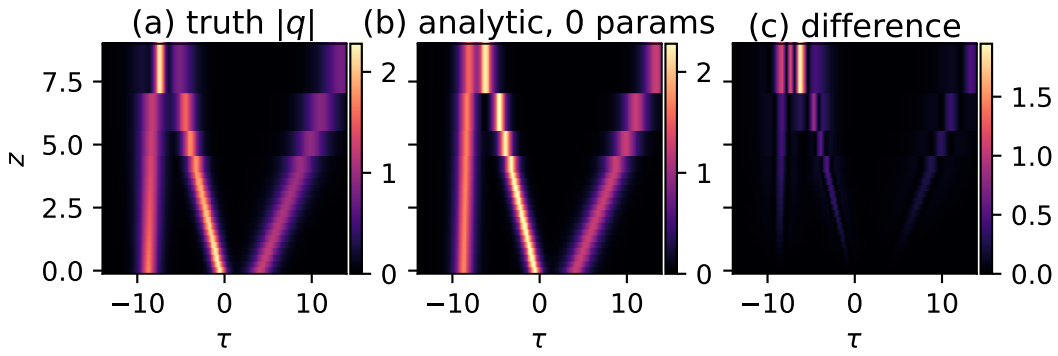


Figure 6: One stored test trajectory at $\epsilon_1 = 0.03$, $|q(\tau, z)|$ over the stored frames (z spacing widens past 4). *(a)* The reference solver. *(b)* The zero-parameter analytic rollout: Darboux applied to the exactly rotated $z = 0$ spectrum. *(c)* Their pointwise difference. A small spectral drift displaces solitons, so the field-level difference grows toward $O(1)$ by $z = 8$; supplying that drift is the job of the residual flow, and no trained model enters any panel.

sector under third-order dispersion fits 0.788 ± 0.330 with $C_R = 0.0440$, the smallest constant of the six sector fits, and its bootstrap interval includes 1. At the weakest strength the median drift is 4.7×10^{-4}, a factor 2.3 above the 2.0×10^{-4} eigenvalue-extraction error the pipeline measures on the unperturbed reference itself, and an additive floor of that size biases a log–log slope downward. The amplitude sectors, two orders larger, carry the closure claim of §5.3.

Physics as a penalty. A conservation-penalty variant of the direct FNO baseline adds to its loss the violation of the first three Kruskal invariants of the Zakharov–Shabat problem, the PINN-A form of the eigenvalue constraint, with weights $0, 0.001, 0.01, 0.1, 1$ at 3 seeds. The invariants are evaluated on the predicted field through their trace formulas; on reflectionless fields they are the power sums $\sum_k \operatorname{Im} \lambda_k^m$, $m = 1, 2, 3$. Every arm, weight zero included, saturates above 1.2 at $z = 8.0$, where a curve carries no information. The logs are in the release with the rest.

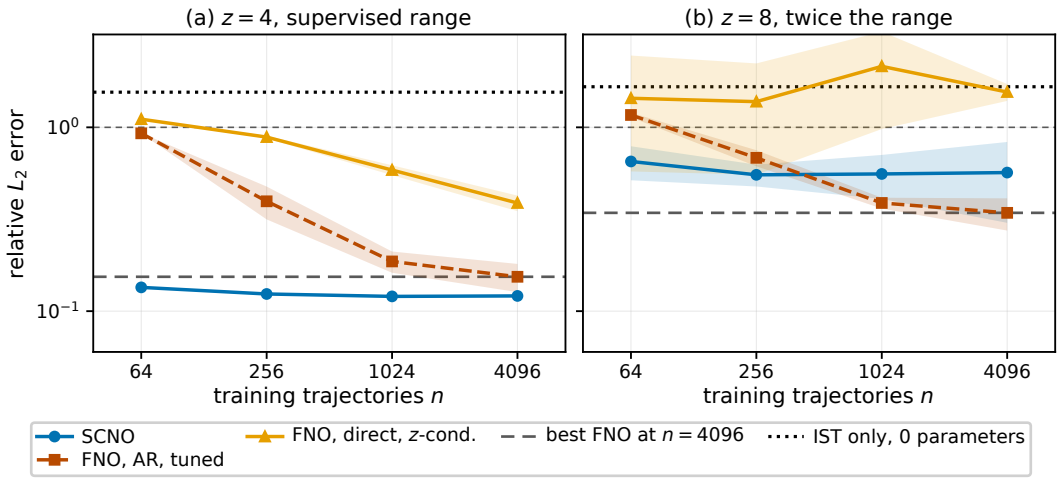


Figure 7: Sample efficiency at $\epsilon_1 = 0.03$, the figure behind §5.4: relative L_2 error against the number of training trajectories n, both axes logarithmic, mean over 3 seeds with a 95% confidence band. *(a)* At $z = 4$, the top of the supervised range. *(b)* At $z = 8.0$, twice it. The dashed line is what the strongest FNO arm reaches with all 4096 trajectories. Read where the SCNO curve first passes under it, at the left edge of panel (a), $n = 64$; in panel (b) the tuned autoregressive arm ends below SCNO, consistent with §5.2's tie at this grid point.

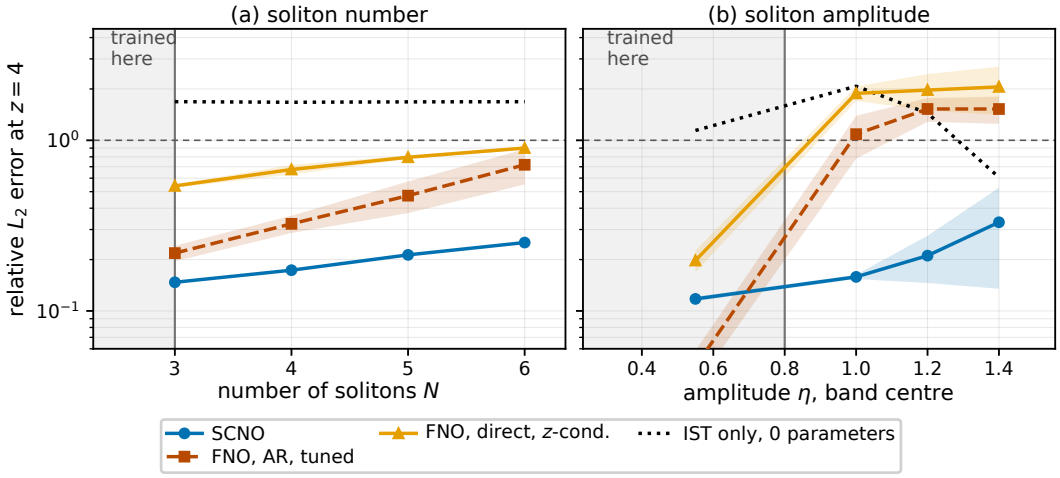


Figure 8: The figure behind §5.5: the two axes the scattering representation makes explicit, both at $z = 4$ and $\epsilon_1 = 0.03$, error axis logarithmic. *(a)* Error against soliton number N, training capped at $N \leq 3$ (shaded) and testing carried to $N = 6$. *(b)* The same against soliton amplitude η_k, trained on $[0.3, 0.8]$ and tested up to $[1.3, 1.5]$. Read the vertical separation outside the shaded region: the baseline degrades toward and past unity on both axes while the SCNO curve rises slowly.

A.8 Blind test-time analysis

The end-to-end numbers of the blind protocol are in §5.8; this section holds what the main text omits. Cardinality detection is exact at every noise level tested for the hard eigensolver. The soft fallback places each eigenvalue at the softmin-weighted centroid of $|a(\lambda)|$ on the coarse grid, with no Newton root finding, and drops to 0.89 in the worst case. The full detection grid, including the soft-detection fallback and the higher noise levels, is recorded in `results/logs/testtime.jsonl` and `results/logs/testtime_z8.jsonl`.

B Proofs

Proof of Prop. 1. IST completeness for the focusing Zakharov–Shabat problem (Zakharov & Shabat, 1972; Faddeev & Takhtajan, 1987) gives the bijection between reflectionless q and its discrete scattering data. The evolution $b_k(z) = b_k(0)e^{4j\lambda_k^2 z}$ with λ_k invariant is the AKNS linearization

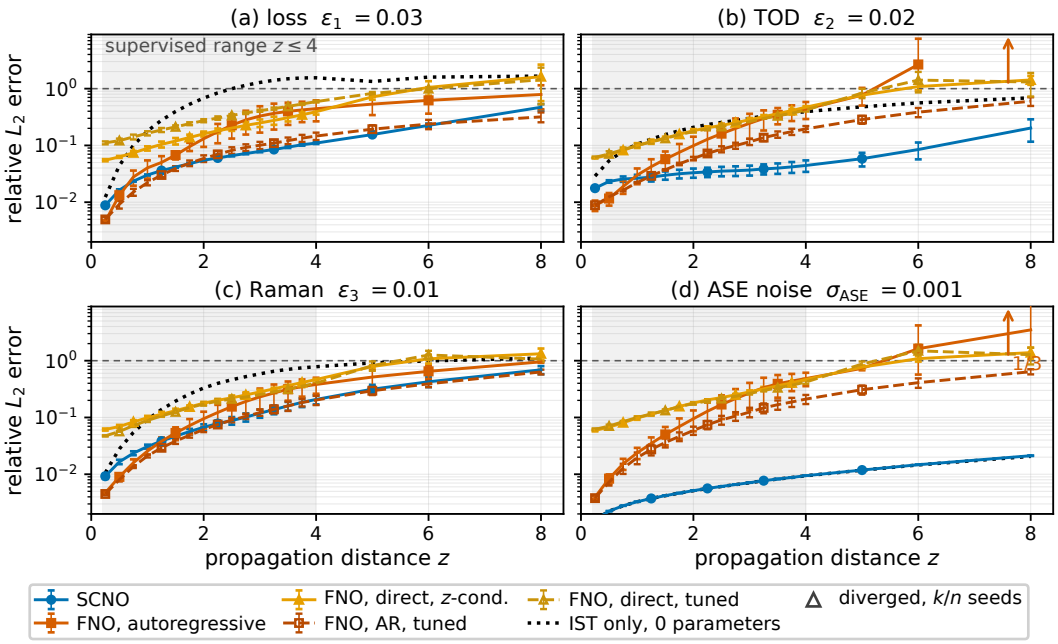


Figure 9: Companion to Fig. 3: relative L_2 error against propagation distance at one mid-range strength per channel, error axis logarithmic, with the supervised range $z \le 4$ shaded and bands giving 95% confidence intervals over 3 seeds. The reading is the slope past the shaded region, where the SCNO curve keeps the ordering it had inside the training range on every channel except loss.

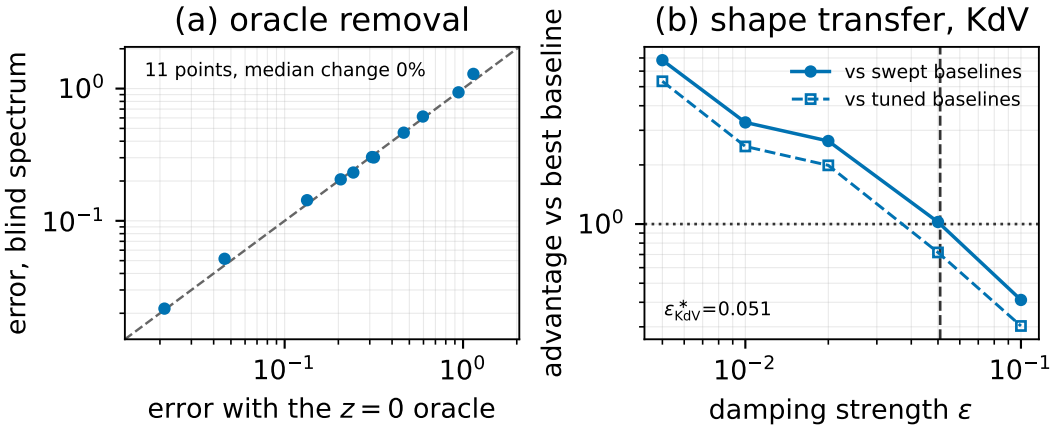


Figure 10: *(a)* The oracle-removal result of §5.8 as a picture: z=8 error re-trained from the blindly recovered spectrum against the matched oracle reference, one marker per phase-diagram point, identity line dashed. *(b)* The KdV phase diagram of Table 11 as a shape: the advantage over the strongest baseline decays monotonically in the damping and crosses unity at $\epsilon^*_{\mathrm{KdV}} = 0.051$ against the swept baselines (solid) and 0.037 against the per-point tuned ones (dashed, re-run per-seed medians of `t14_kdv_tune.jsonl`), the same form the NLSE loss channel takes.

(Ablowitz et al., 1974), and the Darboux construction realizes the inverse map exactly (Matveev & Salle, 1991). Composing the three gives Φ_z. In the SCNO pipeline, zero initialization makes f_θ vanish identically, so the computed map is the composition above. We measured the difference between the two at 0 on the eigenvalues and 2.1×10^{-15} relative L_2 on the field, the round-off floor of the $(\ln|P_k|, \arg P_k)$ parameterization. □

Proof of Prop. 2. Apply first-order adiabatic perturbation theory of the Zakharov–Shabat spectrum through Kaup's squared-eigenfunction expansion (Kaup, 1976), evaluated on the unperturbed soliton, and integrate over z. The channel constant C_R collects the resulting rates: $d\eta_k/dz = -2\epsilon_1\eta_k$ for loss, the self-frequency shift of Gordon (1986) for Raman, and velocity drift with $O(\epsilon_2)$ radiation for third-order dispersion. Assumption A is that the second-order remainder of this expansion

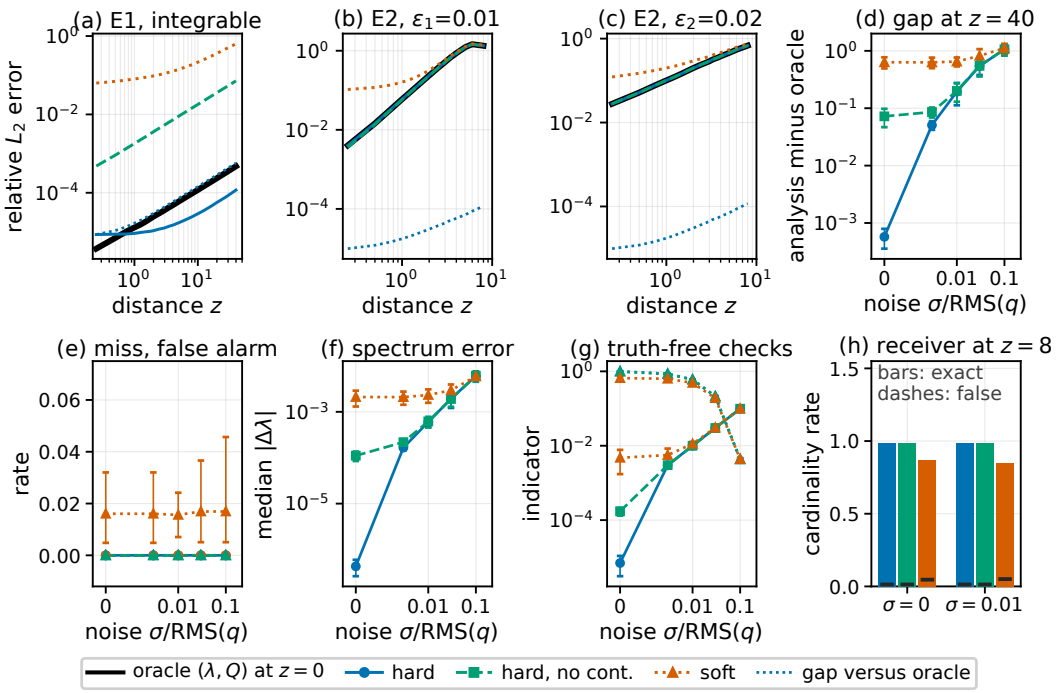


Figure 11: Cost of the forward-transform analysis step, measured on the zero-parameter rollout so that no learned component is involved. *(a, b, c)* Relative L_2 error against distance for the z=0 oracle and for three receiver-side detectors, with the dotted curve giving the truth-independent gap between the hard eigensolver and the oracle. *(d)* That gap against observation noise. *(e)* Miss and false-alarm rates, which stay at zero for the hard eigensolver across the whole grid. *(f)* Median error of the recovered spectrum. *(g)* Confidence metrics computable without ground truth. *(h)* Blind extraction at a receiver at $z = 8$. The reading is panel (d): the gap, and not the end-to-end error, is what isolates the analysis step, because the stored trajectory carries solver drift of the same order at large z.

admits a Grönwall estimate bounding it by $O(\epsilon^2 z^2)$ on bounded ϵZ; we state this as an assumption, since a proof needs decay estimates on the squared-eigenfunction basis, and §5.3 is the empirical check. Dividing out $\exp\big(4j\int_0^z \lambda_k(s)^2\,ds\big)$ removes the $O(1)$ analytic rotation exactly, because that integral is the phase the norming constant accumulates along the perturbed spectral path, so what survives in $\tilde{P}_k$ is the perturbation-induced drift and equation 4 follows.

For equation 5, write $\lambda_k(s) = \lambda_k(0) + \delta_k(s)$ with $\delta_k(s) = O(\epsilon s)$ by the first-order rates above. The $\lambda_k(0)^2$ parts of the two phases cancel, leaving $8\lambda_k(0)\big(z\delta_k(z) - \int_0^z \delta_k\,ds\big) + O(\epsilon^2 z^3)$. A drift that is linear to leading order, $\delta_k(s) \simeq \delta_k' s$, gives $z\delta_k(z) - \int_0^z \delta_k\,ds = \frac{1}{2}\delta_k' z^2$, hence $\Delta\phi_k = 4\lambda_k(0)\delta_k' z^2 + O(\epsilon^2 z^3) = O(\epsilon z^2)$. For the loss channel $\delta_k' = -2j\epsilon_1\eta_k$ and $\lambda_k(0) = j\eta_k$ on a centered soliton, so $\Delta\phi_k = 8\epsilon_1\eta_k^2 z^2$, and differentiating in z gives the rate $16\epsilon_1\eta_k^2 z$ quoted in §4. The term is deterministic given the spectrum, so a residual flow with access to (λ, P) can represent it. □

C Transfer to a second integrable family

Nothing in §3 is specific to the focusing NLSE beyond the choice of spectral problem. For any equation admitting a Lax pair $L_z = [M, L]$, the scattering data of L evolves linearly. Three ingredients are then required: a differentiable forward transform, the closed-form evolution of the scattering data, and a differentiable inverse. For the Korteweg–de Vries equation these are, respectively, the time-independent Schrödinger operator $L = -\partial_{\tau\tau} - u(\tau)$, the norming constant law $c_k(z) = c_k(0)e^{4\kappa_k^3 z}$, and Crum–Darboux/Wronskian dressing. We take KdV in the form $u_z + 6uu_\tau + u_{\tau\tau\tau} = 0$, whose solitons are positive, $u = 2\kappa^2\,\mathrm{sech}^2(\kappa(\tau - 4\kappa^2 z))$; the textbook pairing of $L = -\partial_{\tau\tau} + u$ with $u_z - 6uu_\tau + u_{\tau\tau\tau} = 0$ follows from $u \mapsto -u$, and only the sign convention differs. Verified numerically before use: on this sign of u the operator $-\partial_{\tau\tau} + u$ has no bound state at all. The discrete spectrum $\{-\kappa_k^2\}$ of $L = -\partial_{\tau\tau} - u$ replaces the Zakharov–

Shabat eigenvalues, and Wronskian dressing replaces the matrix Darboux transform used here. The co-rotating construction of equation 2 carries over unchanged: dividing out the analytic exponential makes the integrable flow the identity, so the learned field again represents only the perturbation.

The exact limit transfers, measured. We ran the Prop. 1 analogue on KdV. The check needs no reference solver and no training, because the reflectionless N-soliton is known in closed form through $u = 2\,\partial_{\tau\tau} \log\det A$ with $A_{mn} = \delta_{mn} + c_m c_n (\kappa_m + \kappa_n)^{-1} e^{-(\kappa_m+\kappa_n)\tau}$. We evolve the norming constants by $c_k(z) = c_k(0) e^{4\kappa_k^3 z}$, dress with Crum–Wronskian, and compare against that closed form. Over 3 eigenvalues and 3 evolution times the worst relative L_∞ error is 2.2×10^{-15}. On a two-soliton state the evolved momentum holds its analytic value to 5.5×10^{-14} and drifts by at most 1.0×10^{-14} across the interaction. The zero-parameter pipeline is therefore exact on a second family, with a different spectral problem and a different dressing formula, and the co-rotating construction unchanged.

Second derivatives of $\log\det A$ must be taken through the trace identity $\partial_{\tau\tau} \log\det A = \mathrm{Tr}(A^{-1}A'') - \mathrm{Tr}\big((A^{-1}A')^2\big)$. A spectral derivative of $\log\det A$ fails here: the function carries a linear ramp as $\tau \to -\infty$, so a periodic derivative on a finite window returns Gibbs noise instead of the soliton.

Table 11: Perturbed KdV, $u_z + 6uu_\tau + u_{\tau\tau\tau} = -\epsilon u$. Relative L_2 of the field at $z = 4$, median over 3 seeds, 2048 training and 256 test trajectories per point. The two ratio columns are SCNO against the stronger of the two baselines, swept and per-point tuned. The advantage decays monotonically and crosses over under both protocols, which is the shape the loss channel has on NLSE.

ϵ	zero-param.	SCNO	base (swept)	base (tuned)	ratio	tuned
0	1.8×10^{-8}	1.2×10^{-4}	0.0908	0.0687	729×	552×
0.005	0.270	0.0125	0.0850	0.0664	6.8×	5.3×
0.01	0.509	0.0249	0.0818	0.0617	3.3×	2.48×
0.02	0.858	0.0291	0.0771	0.0580	2.65×	1.99×
0.05	1.31	0.0635	0.0650	0.0455	1.02×	0.72×
0.1	1.64	0.115	0.0471	0.0347	0.41×	0.30×

The residual, trained on the second family. We then built the rest of the stack and trained on damped KdV, $u_z + 6uu_\tau + u_{\tau\tau\tau} = -\epsilon u$. The symmetric split-step is exact only for a phase-rotation nonlinearity, so the reference solver is ETDRK4 (Cox & Matthews, 2002; Kassam & Trefethen, 2005), validated against the closed-form two-soliton through a collision. The co-rotating coordinate is a position: with $\Xi_k = \tau_k - \int_0^z 4\kappa_k(s)^2\, ds$, the unperturbed flow is again exactly the identity, and first-order theory gives $d\kappa_k/dz = -\frac{2}{3}\epsilon\kappa_k$ and $d\Xi_k/dz = -\epsilon/(3\kappa_k)$. Carrying the NLSE coordinate over literally, as $c_k e^{-4\kappa_k^3 z}$, does not work: c_k is exponentially sensitive to position, so an $O(\epsilon)$ drift in κ_k multiplies an $O(z)$ position and the target acquires an unbounded secular term.

Data and hyperparameters. Each damping strength has 2048 training (seed 11) and 256 test (seed 901) trajectories on $N_x = 1024$ points of $x \in [-60, 60)$ in a frame moving at $v = 4$, with $N \in \{1, 2, 3\}$ solitons, $\kappa \in [0.6, 1.4]$ and positions in $[-30, 30]$. Pairs closer than $\Delta\kappa = 0.15$ or $\Delta\xi = 8$ are redrawn. Frames are stored at $t \in \{0, 0.5, 1, 1.5, 2, 3, 4\}$ from ETDRK4 at step 5×10^{-4}. SCNO trains for 60 epochs at learning rate 10^{-3}, batch 64, RK4 step 0.05 on the residual flow. The FNO baselines share epochs and batch, use width and modes 64 over four layers, and take their learning rate from a four-rate sweep (5×10^{-4} to 4×10^{-3}). Optimizer, weight decay, schedule and clipping are those of Appendix A.

Table 11 reports 6 damping strengths, 2048 training and 256 test trajectories each, three seeds per arm. The structure transfers, not only the construction. The advantage decays monotonically with the perturbation and crosses over at $\epsilon^*_{\mathrm{KdV}} = 0.051$, by the same log-linear rule as ϵ^*; the measured advantage at the 0.05 grid point itself is 1.02×, so the crossing is essentially that grid point. Tuning the baselines per point (learning-rate sweep over both families) moves the crossing to $\epsilon^*_{\mathrm{KdV}} = 0.037$, the same inward shift the NLSE loss channel shows under baseline tuning. This is the shape the loss channel has on NLSE, reached through a different spectral problem and a different inverse. Against the analytic line the learned residual buys a steady 14.4 to 29.4× across the sweep. A sharper zero-parameter reference integrates the stated first-order rates in closed form: that adiabatic line reaches 0.0161 at $\epsilon = 0.005$ and 0.295 at 0.1, and the trained residual still beats it by 1.29 to 2.57×,

widening with ϵ. The first-order rates omit the trailing shelf, and that omission is what the residual learns; Table 9 carries the same accounting for every cell of both families.

At $\epsilon = 0$ the learned arm makes the exact solution worse. The zero-parameter pipeline reaches 1.8×10^{-8} there and training degrades it to 1.2×10^{-4}, four orders. With no perturbation there is no physical residual, and the loss is computed against the reference solver, whose own discretization error is of order 10^{-7}. That error is deterministic, therefore learnable, and the residual flow fits it. The noise channel of §5.2 is the complement: there the perturbation is random, the gradient averages to nothing, and the model correctly learns to do nothing. Together they say that a residual model absorbs whatever is deterministic in its target, including a solver's own error, and only randomness protects it. At the exactness floor the right number of parameters is zero, which this pipeline can supply by construction and a learned operator cannot.