Journal of Nonlinear Waves (2026), **2**, e21, 1–22
doi:10.1017/jnw.2026.10048

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Learning effective soliton dynamics from scattering data using weak-form system identification

Seth Minor, Vanja Dukic and David M. Bortz

Department of Applied Mathematics, University of Colorado Boulder College of Engineering and Applied Science, Boulder, CO, USA
Corresponding author: Seth Minor; Email: seth.minor@colorado.edu

Keywords: inverse scattering transform; solitons; weak-form system identification

MSC Codes: *Primary* 37K15; *Secondary* 93B30; 35Q53

(Received 2 July 2026 UTC; revised 2 July 2026 UTC; accepted 20 August 2026 UTC)

Abstract
The *inverse scattering transform* (IST) provides the standard theoretical framework for deriving soliton dynamics. Traditionally, such derivations have been of an analytical, rather than data-driven, nature. In this paper, we combine the conceptual framework of the IST with weak-form system identification methods to discover effective soliton dynamics directly from observed scattering data, without assuming prior knowledge of the scattering equations. Our method avoids parameterizing solitary waves via ad hoc curve-fitting by working in the scattering domain, yielding interpretable low-dimensional models that remain valid in perturbed and near-integrable regimes. We demonstrate the performance of the proposed approach on synthetic and experimental data governed by shallow-water equations of Korteweg–de Vries-type and recover models that are consistent with canonical IST theory.

1. Introduction

A central mathematical problem in the study of nonlinear dispersive waves is the construction of effective, low-dimensional models for coherent solitary wave structures (*solitons*). In such settings, the *inverse scattering transform* (IST) of Ablowitz, Kaup, Newell and Segur [2] has enjoyed a rich and successful history, and is now the standard theoretical framework for deriving reduced-order evolution equations for soliton dynamics. Although these derivations are traditionally of an analytical – rather than data-driven – nature, recent work has employed the IST formalism as a tool for experimental data analysis, using the technique to analyse soliton content from empirical measurements [8, 15, 23]. Moreover, recent approaches using alternative parameterization techniques have demonstrated that the learning of reduced-order, interpretable equations of motion for solitons is tenable in a data-driven setting [6, 25, 26]. Despite the success of this recent work, however, little effort has been devoted to developing a data-driven modelling approach based on the IST itself, most likely due to the fact that the framework is fundamentally problem-specific.

In this paper, we address the question of whether effective soliton dynamics can be inferred directly from observed scattering data (as opposed to being derived or approximated analytically). We show that the conceptual framework of the IST can be combined with modern weak-form equation learning methods to identify interpretable models for the scattering dynamics, without assuming prior knowledge about the form of the equations. Our proposed method avoids parameterizing solitary waves via ad hoc

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

curve-fitting by working in natural spectral domains, yielding interpretable low-dimensional models that remain valid in perturbed, near-integrable regimes. In this paper, we specifically focus on examples related to the canonical Korteweg–de Vries (KdV) equation; however, we do briefly comment on the potential for extension to other integrable equations in Section 5.

The paper is organized as follows. In Section 2, we relate to previous work (§2.1), establish some notational conventions and background assumptions (§2.2), identify the specific class of models being considered (§2.3), and review background material on the IST (§2.4) as well as the paradigm of weak form modelling (§2.5). Later, in Section 3, we outline our approach (§3.1) and numerical implementation (§3.2 and §3.3). In Section 4, we discuss validation metrics (§4.1) before presenting some preliminary numerical results obtained using simulated data (§4.2) and experimental water-wave data (§4.3), pausing briefly to comment on structural identifiability concerns (§4.4). Finally, we conclude with a summary and discussion in Section 5. Supplemental information is given in the Appendix.

2. Background material

2.1. Relation to previous work

It is important to note that there are many ways of parameterizing solitary waves, and different choices can lead to distinct effective models [9, 27]. In the recent literature, Chen, Yang, Zhu and Kevrekidis have explicitly fit parameterized waveforms (centres, widths, amplitudes, phases) to simulated data [26], and have also considered parameterizing the statistical moments of physically relevant quantities [6, 25], before then applying equation learning approaches to discover effective dynamic models governing the corresponding parameters. Here, we instead propose to use *scattering data* (formally defined in §2.4) directly as a parameterization of the soliton field. This choice has at least two advantages:

- it is natural in the context of nonlinear wave equations, inheriting a conceptual framework and set of known analytical results from existing *direct/inverse scattering transform* (DST/IST) theory;
- it avoids fitting ad hoc, problem-specific waveforms to the data.

Unfortunately, there is no free lunch regarding the latter point – a significant caveat of using scattering data is that, while the IST framework of [2] does supply a fairly unified language across many integrable equations of interest, the definitions of both DST and IST are inherently problem-specific.

2.2. Notation and assumptions

Herein, we will let $u = u(x,t)$ denote a scalar-valued field which represents a solution to a perturbed, dispersive partial differential equation (PDE) of the form

$$u_t + N[u] = \epsilon F[u], \quad \text{with} \quad \begin{cases} u(x,0) = u_0(x), \\ (x,t) \in (-\infty,\infty) \times [0,T], \end{cases} \tag{1}$$

where N and F are (potentially nonlinear) differential operators and $0 \leq \epsilon \ll 1$ is a small parameter representing a perturbation to the system. Of particular interest are *integrable systems* (for our purposes, it will suffice to define an 'integrable system' as a system of PDEs that admits a Lax pair representation), which for $\epsilon = 0$ admit a *Lax pair* – linear operators $L = L(u)$ and $M = M(u)$ which depend on u and satisfy

$$\big(L_t(u) + [L,M](u)\big)\phi = \big(u_t + N[u]\big) \cdot \phi = 0, \tag{2}$$

for any test function $\phi = \phi(x,t)$ of appropriate smoothness, where $[\cdot,\cdot]$ is the commutator operator and $L_t := [\partial_t, L]$ denotes the operator-valued time derivative of L. This identity arises as a consistency condition for an associated overdetermined eigenvalue problem (EVP) given by

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

$$\begin{cases} L(u)\phi = \lambda\phi, \\ \phi_t = M(u)\phi. \end{cases} \tag{3}$$

To keep the exposition somewhat concise and self-contained, we have deferred a more detailed review of Lax pairs to Appendix B. Throughout the paper, the following two assumptions are in play: (1) the field decays asymptotically (i.e., $|u| \to 0$ as $x \to \pm\infty$; from an analytical perspective, the rate of decay must be sufficiently fast – a standard assumption is that $u \in L^1$ and $(1 + |x|) \cdot u_0 \in L^1$), and (2) the underlying PDE and Lax pair are available in closed form. Regarding (2), we note that, e.g., the WSINDy [18] and SILO [3] algorithms are respectively capable of identifying the N and L, M operators from nonlinear wave equation data, indicating that an end-to-end data-driven extension of the current work is also tenable.

2.3. *Equations considered*

To illustrate our proposed method on a canonical and well-understood testbed, we focus on examples related to the *Korteweg–de Vries* (KdV) *equation*, a prototypical model of shallow water waves given by

$$u_t + \alpha u u_x + \beta u_{xxx} = 0, \quad \text{where} \quad \begin{cases} L(u) = -\left(\partial_x^2 + \frac{\alpha}{6\beta}u\right), \\ M(u) = -\left(4\beta\partial_x^3 + \alpha u\partial_x + \frac{\alpha}{2}[\partial_x, u]\right). \end{cases} \tag{4}$$

We also consider non-trivial perturbations of the KdV equation, leading to models of the form

$$u_t + \alpha u u_x + \beta u_{xxx} = \epsilon F[u]. \tag{5}$$

In particular, for a fifth-order dispersive forcing term $F[u] := \partial_x^5 u$, the perturbed KdV model given above collapses to the *Kawahara equation*,

$$u_t + \alpha u u_x + \beta u_{xxx} = \epsilon u_{xxxxx}. \tag{6}$$

Our focus on this specific class of governing equations is partly motivated by the recent success of Heinrich et al. [10] in applying data-driven equation-learning methods to experimental wave-tank data, in which a Kawahara-type model was identified. (In [10], the identified model was equivalent to eq. (6) with coefficients $\alpha \approx 1.36$, $\beta \approx 0.62$ and $\epsilon \approx 0.06$; see §4.3 below.) However, a near-identical pipeline extends to other integrable equations that admit a Lax pair representation; we return to this point in Section 5.

2.4. *The inverse scattering transform*

In standard acoustic scattering problems, one is often concerned with identifying the form of a spatially varying parameter $v(x)$ associated with a linear hyperbolic PDE such as the classical wave equation,

$$v(x)^{-2}u_{tt} - u_{xx} = 0,$$

in some inhomogeneous region $x \in \Omega$ where $v(x) \to v_0$ collapses to a background value as $x \to \infty$. Adopting a temporally modulated ansatz of the form $u(x,t) = \phi(x)e^{-i\omega t}$ and heuristically introducing a *scattering potential* $V(x) := k^2(1 - \eta(x)^2)$, where $\eta(x) := v_0/v(x)$ is the *index of refraction* and $k := \omega/v_0$ is an associated wavenumber, one obtains a Schrödinger-type EVP (cf. eq. (3)) given by

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

$$L\phi = k^2\phi, \quad \text{with} \quad L = -\partial_x^2 + V(x). \tag{7}$$

For fixed $V(x)$, the self-adjoint operator L has a point spectrum defined by a finite set of eigenvalues $\lambda_i = k_i^2 < 0$ for $i = 1, \ldots, n$, and a continuous spectrum which consists of all $\lambda = k^2 \in [0, \infty)$. Traditionally, the solutions to eq. (7) are decomposed linearly as $\phi = \phi_{\text{in}} + \phi_{\text{out}}$, where $\phi_{\text{in}}(x; k) = e^{ikx}$ represents an incoming wave and $\phi_{\text{out}}(x; k_i) \sim c_i e^{ik_i x}$ or $\phi_{\text{out}}(x; k) \sim R(k)e^{ikx}$ represents bound or scattered waves corresponding to the point spectrum or the continuous spectrum, respectively.

It is often convenient to write the asymptotic weighting coefficients c_i and $R(k)$, respectively referred to as the *norming constants* and *reflection coefficient*, in terms of the positive *Jost solution*, ϕ_+, of the Schrödinger EVP in eq. (7). For $k \neq 0$, this eigenfunction can be uniquely defined as the solution of the EVP subject to the boundary condition $\phi_+(x; k) = \rho(x; k)e^{ikx} \sim e^{ikx}$ as $x \to \infty$, or equivalently:

$$L\rho = 2ik_i\rho', \quad \text{subject to} \quad \begin{cases} \rho(\infty) = 1, \\ \rho'(\infty) = 0. \end{cases} \tag{8}$$

Following the Marchenko convention and defining $\kappa := -ik$ with $\lambda_i = -\kappa_i^2 < 0$, one typically expresses

$$c(\kappa_i) := \left[\int_{-\infty}^{\infty} \phi_+(x, t; i\kappa_i)^2 \, dx \right]^{-1} \quad \text{and} \quad R(k) := \frac{b(k)}{a(k)}, \tag{9}$$

given that $\phi_+(x; k) \sim a(k)e^{ikx} + b(k)e^{-ikx}$ as $x \to -\infty$. We refer to the set of n eigenvalues $\lambda_i := -\kappa_i^2$ in the point spectrum of L, together with the corresponding constants $c_i := c(\kappa_i)$ and reflection coefficient $R(k)$ for each $\lambda = k^2$ in the continuous spectrum, as the *scattering data*, σ, corresponding to the EVP:

$$\sigma(L) := \{(\kappa_i, c(\kappa_i)) \, : \, i = 1, \ldots, n\} \cup \{(k, R(k)) \, : \, k \in \mathbb{R}\}. \tag{10}$$

In a *forward scattering problem*, one aims to compute the scattering data for a known potential $V(x)$, while in an *inverse scattering problem*, one instead aims to recover $V(x)$ from observed scattering data.

At a high level, the IST formalism of [2] turns the standard scattering paradigm described above on its head by associating a nonlinear PDE of the form of eq. (1) with an equivalent linear EVP of the form of eq. (3), by way of a Lax pair representation $L_t + [L, M] = u_t + N[u]$. For the shallow-wave models introduced in §2.3, this can be interpreted as solving a one-parameter family of Schrödinger EVPs (just as in eq. (7) above) defined by the scattering potentials

$$V(x; t) := \gamma u(x, t), \quad \text{with} \quad \gamma := -\frac{\alpha}{6\beta},$$

where $u = u(x, t)$ denotes the solution to the nonlinear PDE and $L(u) = -\partial_x^2 + V(x; t)$ is now the Lax operator corresponding to the KdV equation. Here, the time derivative of the Lax operator is $L_t = \gamma u_t$.

For a fixed time $t \geq 0$, the *direct scattering transform* $\mathcal{S}_t : u \mapsto \sigma(L(u)|_t)$ is defined as the nonlinear map which uniquely associates a field u to its scattering data,

$$\mathcal{S}_t[u] := \sigma\big(L(u)|_t\big) = \{(\kappa_i(t), c_i(t)) \, : \, i = 1, \ldots, n\} \cup \{(k, R(k, t)) \, : \, k \in \mathbb{R}\}. \tag{11}$$

Conversely, the *inverse scattering transform* $\mathcal{S}_t^{-1} : \sigma(L(u)) \mapsto u(\cdot, t)$ uses scattering data to reconstruct the field. In the context of the KdV-type models, the inverse transform $\mathcal{S}_t^{-1}$ can be implicitly defined for scattering data σ of the form of eq. (10) via

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

$$\mathcal{S}_t^{-1}[\sigma](x) := -\frac{2}{\gamma}\,\partial_x K_\sigma(x,x,t), \quad \text{with} \quad \mathcal{S}_t^{-1}\big[\sigma\big(L(u)\big)\big](x) = u(x,t), \tag{12}$$

where K_σ is the unique interaction kernel satisfying the *Gelfand–Levitan–Marchenko* (GLM) *equation*,

$$K_\sigma(x,y,t) \;+\; G_\sigma(x+y,t) \;+\; \int_x^\infty K_\sigma(x,x',t)\,G_\sigma(x'+y,t)\,dx' = 0, \quad \text{for} \quad y \geq x. \tag{13}$$

Here, the function G_σ denotes the observed reflection response and is defined by

$$G_\sigma(x;t) := \sum_{i=1}^{n} c_i(t)e^{ik_i x} + \frac{1}{2\pi}\int_{-\infty}^{\infty} R(k,t)e^{ikx}dk.$$

The GLM equation reconstructs a hidden 1D medium from its reflected waves, and thus can be thought of as a kind of deconvolution equation.

An important consequence of eq. (3) is that the spectrum of L is invariant in time; for any eigenpair (λ,ϕ) initially satisfying $L\phi = \lambda\phi$ at $t = 0$, one has $\phi_t = M\phi$ and $\dot\lambda = 0$ for all $t \geq 0$ (see Appendix B).

For unperturbed systems (i.e., for $\epsilon = 0$), the scattering data also evolve linearly according to

$$\dot R = \boldsymbol{i}\omega(k)R, \quad \text{with} \quad \begin{cases} \dot\kappa_i = 0, \\ \dot c_i = \omega(\kappa_i)c_i, \end{cases} \quad \text{for each } i = 1,\dots,n, \tag{14}$$

where $\omega(k) := 8\beta k^3$ is the dispersion relation of the KdV equation (4). The upshot is that the original nonlinear PDE can then be reduced to the system of linear ordinary differential equations (ODEs) in eq. (14) and solved via the following recipe:

$$u_0 \xrightarrow{\text{DST}} \mathcal{S}_0[u] \xrightarrow{\text{integrate scattering ODEs}} \mathcal{S}_t[u] \xrightarrow{\text{IST}} u.$$

In the case that $0 < \epsilon \ll 1$, the adiabatic perturbation theory of Karpman and Solov'ev [14] confirms that the equations of motion continuously deform to the tune of

$$\dot R = \boldsymbol{i}\omega(k)R + \epsilon f_R(k) + \mathcal{O}(\epsilon^2), \quad \text{with} \quad \begin{cases} \dot\kappa_i = 0 + \epsilon f_{\kappa_i}(\boldsymbol{\kappa},\mathbf{c},R) + \mathcal{O}(\epsilon^2), \\ \dot c_i = \omega(\kappa_i)c_i + \epsilon f_{c_i}(\boldsymbol{\kappa},\mathbf{c},R) + \mathcal{O}(\epsilon^2). \end{cases} \tag{15}$$

A natural connection between the scattering ODEs given above and the governing PDE is obtained by recognizing that the field u can be decomposed into n localized waves parameterized in the fashion

$$u(x,t) = u_{\text{sol}}\big(x;\mathbf{x}_1(t),\dots,\mathbf{x}_n(t)\big) \;+\; u_{\text{rad}}(x,t), \tag{16}$$

where $\mathbf{x}_i := (\kappa_i, c_i)$ represents the scattering data associated with the ith soliton and u_{rad} represents any excess *radiation*. Importantly, for a pure-soliton solution ($u_{\text{rad}} = 0$), the system is *reflectionless* ($R = 0$).

2.5. Sparse regression with WSINDy

Consider a state vector $\boldsymbol{x}(t) = (x_1,\dots,x_d)(t) \in \mathbb{R}^d$ governed by a system of ODEs (in this paper, we will be concerned with ODEs; however, the SINDy [4] and WSINDy [19] paradigms can also be extended to work with PDEs [18, 20, 21]) of the form

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

$$\dot{\boldsymbol{x}} = \boldsymbol{\Theta}(\boldsymbol{x})\mathbf{w}, \quad \text{with} \quad \begin{cases} \boldsymbol{\Theta}(\boldsymbol{x}) := [f_1(\boldsymbol{x}), \ldots, f_J(\boldsymbol{x})] \in \mathbb{R}^J, \\ \mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{J\times d}, \end{cases} \tag{17}$$

where each $f_j : \mathbb{R}^d \to \mathbb{R}$ represents a scalar-valued function of the state, given component-wise by

$$\dot{x}_i = \boldsymbol{\Theta}(\boldsymbol{x})\mathbf{w}_i, \quad \text{for each} \quad i = 1, \ldots, d.$$

The *weak-form sparse identification of nonlinear dynamics* (WSINDy) algorithm [18] is a data-driven technique which attempts to infer dynamics of the form of eq. (17) above from noisy training data

$$\mathcal{X} = \left\{\boldsymbol{x}(t_m) + \boldsymbol{\varepsilon}(t_m) \,:\, m = 1, \ldots, N_t\right\}, \quad \text{where, e.g.,} \quad \boldsymbol{\varepsilon}(t) \sim \mathcal{N}\big(0, \sigma_{\varepsilon}^2 \mathbf{I}\big).$$

This is accomplished by integrating the governing ODE against a collection of smooth and compactly supported test functions,

$$\varphi_k(t) := \varphi(t_{m_k} - t), \quad \text{for} \quad k = 1, \ldots, K,$$

where $\varphi \in C_c^\infty$ is symmetric about $t = 0$ and the *query points*, $t_{m_k} \in [0, T]$, are uniformly placed within the domain. Using integration by parts, models of the form of eq. (17) are then transformed into their *convolutional weak formulations* (note that the factor of '-1' resulting from integration by parts is eliminated by adopting the sign convention $(\varphi * f)(t_k) = \langle \varphi(t_{m_k} - \cdot), f(\cdot)\rangle$),

$$\big(\dot{\varphi} * \boldsymbol{x}\big)(t_{m_k}) = \big(\varphi * \boldsymbol{\Theta}(\boldsymbol{x})\mathbf{w}\big)(t_{m_k}). \tag{18}$$

Accordingly, we investigate the linear system given by

$$\mathbf{b}(\boldsymbol{x}) = \mathbf{G}(\boldsymbol{x})\mathbf{w}, \quad \text{where} \quad \begin{cases} \mathbf{b}(\boldsymbol{x})_{k,i} := \big(\dot{\varphi} * x_i\big)(t_{m_k}) = \langle \dot{\varphi}_k, x_i\rangle, \\ \mathbf{G}(\boldsymbol{x})_{k,j} := \big(\varphi * f_j(\boldsymbol{x})\big)(t_{m_k}) = \langle \varphi_k, f_j(\boldsymbol{x})\rangle, \end{cases} \tag{19}$$

where $\langle \cdot, \cdot \rangle$ denotes the L^2 inner-product. Observe that in eqs. (18) and (19), no differential operators are applied directly to the state variable $\boldsymbol{x} = \boldsymbol{x}(t)$ – a property which increases the fidelity of the results in noisy regimes (that is, compared to the strong-form of SINDy [4]; see, e.g., Table 6 in [18]).

In contrast to standard parameter estimation problems, here the form of the governing equation is *not* assumed to be known prima facie. Instead, one typically assumes in the context of sparse regression that the *library* of candidate terms $\boldsymbol{\Theta}(\boldsymbol{x})$ is substantially over-specified, and that only a small subset of its parameters are nonzero (i.e., the parameter matrix $\mathbf{w}$ is sparse). In turn, WSINDy targets a regularized least-squares problem of the form

$$\hat{\mathbf{w}} := \underset{\mathbf{w}\in\mathbb{R}^{J\times d}}{\operatorname{argmin}} \, \mathcal{L}_\mu(\mathbf{w}), \quad \text{where} \quad \mathcal{L}_\mu(\mathbf{w}) := \|\mathbf{b}(\boldsymbol{x}) - \mathbf{G}(\boldsymbol{x})\mathbf{w}\|_F^2 + \mu\|\mathbf{w}\|_0. \tag{20}$$

The regularization term $\mu\|\mathbf{w}\|_0$ in the loss function promotes the selection of a parsimonious model by penalizing the number of nonzero coefficients, where $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm. Since the ℓ_0 penalty is non-differentiable, in practice the loss function in eq. (20) is approximately minimized via iterative thresholding schemes, e.g., *modified sequential thresholding least squares* (MSTLS) [4, 18], which progressively restricts the columns of $\boldsymbol{\Theta}(\boldsymbol{x})$ available to the model (see Appendix A).

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

3. Methods

3.1. High-level overview

As briefly mentioned in §2.1, Yang et al. have in recent work [26] explored data-driven soliton dynamics (specifically, in the context of the nonlinear Schrödinger equation) by explicitly fitting parameterized waveforms to data. In particular, these authors developed a pipeline of the form

$$\underbrace{\{u(x,t)\}}_{\text{measure data}} \quad \to \quad \underbrace{\{\boldsymbol{x}_i = (\xi_i, a_i, v_i, \theta_i)\}_{i=1}^{n}}_{\text{extract fit parameters}} \quad \to \quad \underbrace{\dot{\boldsymbol{x}}_i \approx \boldsymbol{\Theta}(\boldsymbol{x})\hat{\mathbf{w}}_i, \;\; \text{for } \; i = 1, \ldots, n,}_{\text{infer effective ODEs with SINDy}}$$

where ξ_i, a_i, v_i and θ_i respectively denote the centres, amplitudes, velocities, and phases of modulated sech-like profiles fit to the ith soliton. For the reasons detailed in §2.1, we propose an alternative pipeline:

$$\underbrace{\{u(x,t)\}}_{\text{measure data}} \quad \to \quad \underbrace{\{\boldsymbol{x}_i = (\kappa_i, c_i)\}_{i=1}^{n}}_{\text{compute scattering data}} \quad \to \quad \underbrace{\mathbf{b}(\boldsymbol{x}_i) \approx \mathbf{G}(\boldsymbol{x})\hat{\mathbf{w}}_i, \;\; \text{for } \; i = 1, \ldots, n\,.}_{\text{infer effective ODEs with WSINDy}}$$

Given our focus on effective models for solitons (and to simplify the core ideas of the paper), in this work we specifically address reflectionless settings with $R = 0$; however, in the regime of nonzero radiation, note that one can simply append additional components of the form $R(\mathbf{k}, t) \in \mathbb{R}^{N_{\mathbf{k}}}$ onto the state vector.

3.2. Numerical implementation

In our numerical implementation, we discretize the state vector $\boldsymbol{x}(t) \in \mathbb{R}^d$ over a uniform temporal grid $\mathbf{t} = [t_1, \ldots, t_{N_t}]^T \subset [0, T]$ with spacing Δt, producing a data matrix $\mathbf{X} := \mathbf{X}^* + \boldsymbol{\epsilon}(\mathbf{t}) \in \mathbb{R}^{N_t \times d}$ given by

$$\mathbf{X} = \begin{bmatrix} x_1(t_1) & \cdots & x_d(t_1) \\ \vdots & & \vdots \\ x_1(t_{N_t}) & \cdots & x_d(t_{N_t}) \end{bmatrix} = \begin{bmatrix} x_1^*(t_1) & \cdots & x_d^*(t_1) \\ \vdots & & \vdots \\ x_1^*(t_{N_t}) & \cdots & x_d^*(t_{N_t}) \end{bmatrix} + \begin{bmatrix} \epsilon_1(t_1) & \cdots & \epsilon_d(t_1) \\ \vdots & & \vdots \\ \epsilon_1(t_{N_t}) & \cdots & \epsilon_d(t_{N_t}) \end{bmatrix},$$

where $\mathbf{X}^* := \boldsymbol{x}^*(\mathbf{t})$ denotes the matrix of uncorrupted (i.e., non-noisy) data. The data matrix $\mathbf{X}$ represents the scattering data computed from observations of a noisy scalar field $\mathbf{u}(t) \in \mathbb{R}^{N_x}$ of the form

$$\mathbf{u}(t) := \mathbf{u}^*(t) + \boldsymbol{\varepsilon}(t), \quad \text{where} \quad \mathbf{u}(t) := \texttt{vec}\big\{u(n\Delta x, t) \;:\; n\Delta x \in \Omega\big\},$$

and a discretized Lax operator $\mathbf{L}(\mathbf{u}) \in \mathbb{R}^{N_x \times N_x}$ defined by

$$\mathbf{L}(\mathbf{u}) := -\mathbf{D}_{xx} + \gamma \mathbf{u}' = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T, \quad \text{with} \quad (\,\cdot\,)' := \texttt{diag}(\,\cdot\,),$$

where $\mathbf{D}_{xx}$ is a second-order centred finite difference operator and $\mathbf{Q}, \boldsymbol{\Lambda}$ give a unitary diagonalization of $\mathbf{L}(\mathbf{u})$. After ordering the entries of $\boldsymbol{\Lambda}$ in ascending order, the first n negative eigenvalues approximate the point spectrum of $L(u)$, and we thus define $\lambda_i = -\kappa_i^2 := \Lambda_{ii} < 0$ and $\phi_i(\mathbf{x}, t) := \mathbf{q}_i(t)$ for $i = 1, \ldots, n$, where $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{N_x}]$; see Figure 1. The discrete scattering data then take the form

$$\boldsymbol{x}(t) := \mathcal{S}_t[\mathbf{u}] = [\kappa_1(t), \ldots, \kappa_n(t), c_1(t), \ldots, c_n(t)] \in \mathbb{R}^d, \quad \text{where} \quad d := 2n,$$

and where the norming constants c_i are computed as per eq. (9) after numerically solving the boundary value problem in eq. (8) for the Jost solution.

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

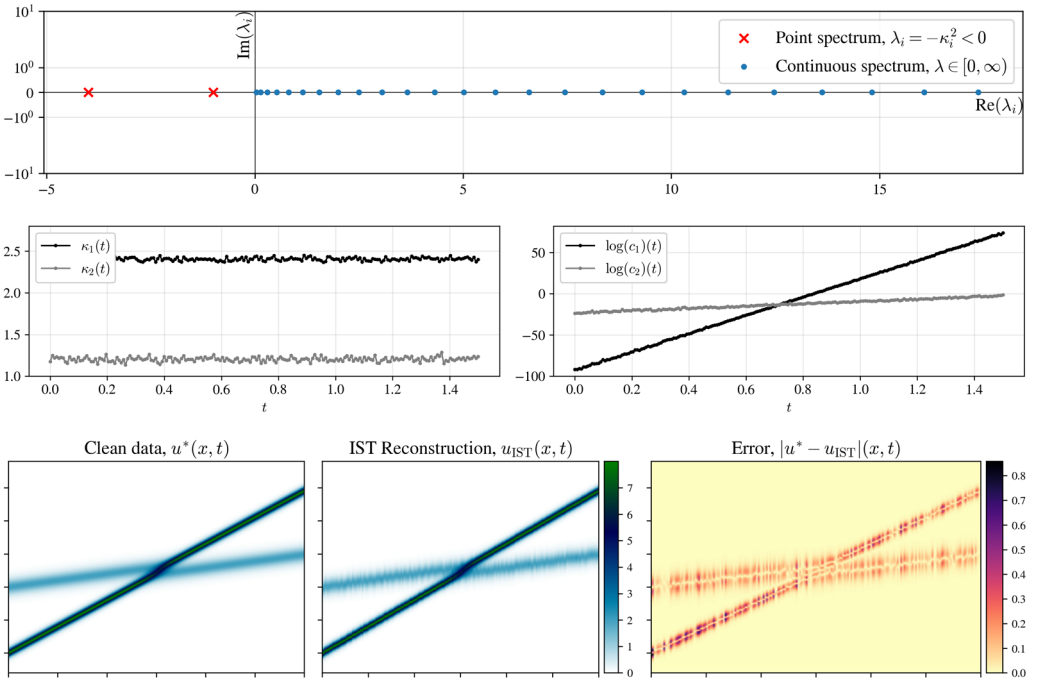


Figure 1. *Illustrating the numerical computation of the scattering data, in this case using the noisy measurements from the two-soliton collision example of §4.2 (see Figure 2). [Top panel] Plotting the first few eigenvalues in the time-invariant spectrum of the discretized Lax operator* $\mathbf{L}(\mathbf{u}^*)$*, for reference; the 'bound state' eigenvalues* $\lambda_i = -\kappa_i^2 < 0$ *in the point spectrum parameterize the soliton dynamics. [Middle panel] Plotting the numerically computed* $\kappa_i(t)$ *and* $\log(c_i)(t)$ *time series; note that these quantities correlate with the soliton widths and locations, respectively. [Bottom panel] Comparing the true state* $\mathbf{u}^*(t)$ *to the reconstructed state* $\mathbf{u}_{IST}(t)$ *obtained by applying the IST to these scattering data.*

It is important to recognize that the DST map $u \mapsto \mathcal{S}_t[u]$ is highly nonlinear and, as such, can be sensitive to noisy perturbations of the form $u = u^* + \varepsilon$. Although this presents limitations when working with experimental data, this sensitivity is partially alleviated by working within the noise-robust weak formulation described in §2.5. Letting $\mathbf{Q}^*, \mathbf{\Lambda}^*$ provide a unitary diagonalization of $\mathbf{L}(\mathbf{u}^*)$, one finds that

$$\mathbf{L}(\mathbf{u}) = \mathbf{L}(\mathbf{u}^*) + \gamma\boldsymbol{\varepsilon}', \quad \text{or equivalently,} \quad \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T = \mathbf{Q}^*\mathbf{\Lambda}^*\mathbf{Q}^{*T} + \gamma\boldsymbol{\varepsilon}'.$$

Applying first-order eigenvalue perturbation theory, one in turn finds that the eigenvalues are perturbed to the tune of

$$\lambda_i = \lambda_i^* + \gamma\|\mathbf{q}_i\|_{\boldsymbol{\varepsilon}'}^2 + \mathcal{O}\big(\|\mathbf{q}_i\|_{\boldsymbol{\varepsilon}'}^4\big),$$

meaning that, to first order, $\lambda_i = -\kappa_i^2$ is approximately normally distributed about $\lambda_i = -(\kappa_i^*)^2$ with a standard deviation that is proportional to the weighted L^2 norm $\|\mathbf{q}_i\|_{\boldsymbol{\varepsilon}'}^2$. Similar results also hold for the norming constants c_i, which implicitly depend on the sensitivity of the Jost solution ϕ_+ to measurement error. Adding noise can also result in the generation of spurious small negative eigenvalues, which in practice we find can be avoided by numerically enforcing a lower bound of the form $\kappa_{\min} \in [0.2, 0.6]$.

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

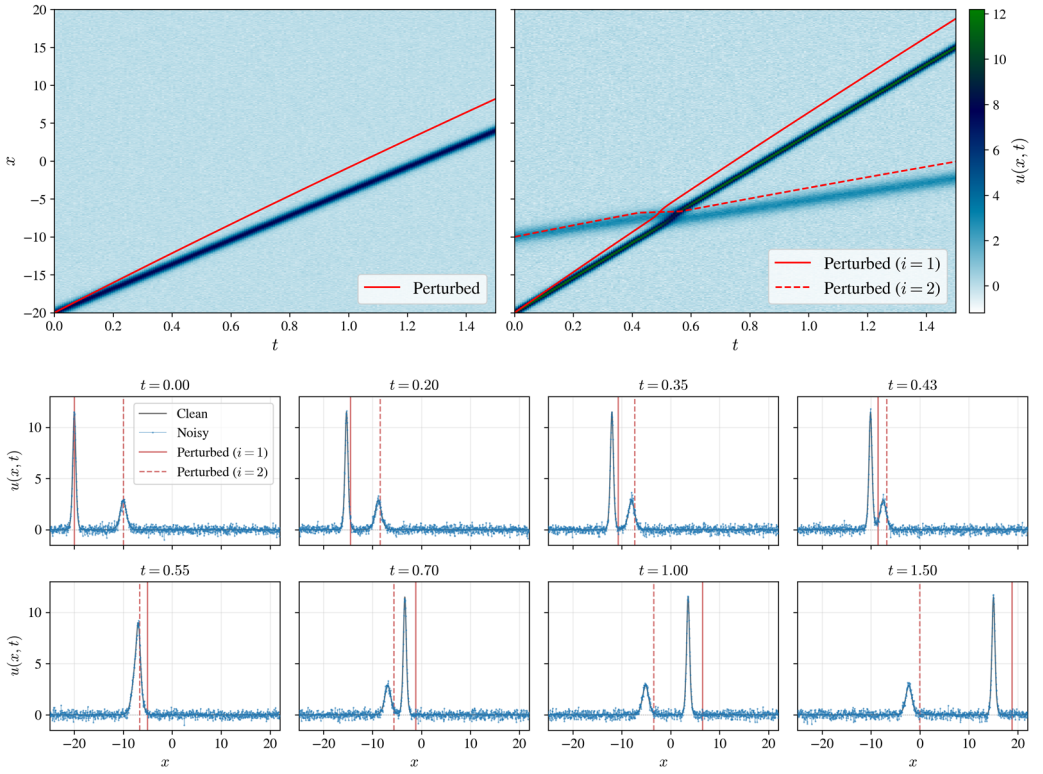


Figure 2. *[Top panel] Visualizing one of three sets of numerical simulations used to generate synthetic data, which each follow the evolution of one or two solitons evolving under the perturbed KdV equation in eq. (5), pictured at* 20% *noise with* $\kappa = 2$ *and* $(\kappa_1, \kappa_2) = (2.4, 1.2)$ *in the left and right panels, respectively. [Bottom panel] Snapshots of the double-soliton simulations evolving in time.*

We discretize the variational WSINDy problem posed in eq. (19) by reshaping the candidate library matrix to the tune of $\boldsymbol{\Theta}(\boldsymbol{x}) \in \mathbb{R}^J \mapsto \boldsymbol{\Theta}(\mathbf{X}) \in \mathbb{R}^{M \times J}$, where

$$\boldsymbol{\Theta}(\mathbf{X}) := \begin{bmatrix} f_1(\boldsymbol{x})(t_1) & \cdots & f_J(\boldsymbol{x})(t_1) \\ \vdots & & \vdots \\ f_1(\boldsymbol{x})(t_{N_t}) & \cdots & f_J(\boldsymbol{x})(t_{N_t}) \end{bmatrix}.$$

In turn, banded convolution matrices $\boldsymbol{\Phi}, \dot{\boldsymbol{\Phi}} \in \mathbb{R}^{K \times N_t}$ can be defined via

$$\boldsymbol{\Phi} := \begin{bmatrix} \varphi_1(t_1) & \cdots & \varphi_1(t_{N_t}) \\ & \ddots & \\ \varphi_K(t_1) & \cdots & \varphi_K(t_{N_t}) \end{bmatrix} \quad \text{and} \quad \dot{\boldsymbol{\Phi}} := \begin{bmatrix} \dot{\varphi}_1(t_1) & \cdots & \dot{\varphi}_1(t_{N_t}) \\ & \ddots & \\ \dot{\varphi}_K(t_1) & \cdots & \dot{\varphi}_K(t_{N_t}) \end{bmatrix}, \quad \text{with} \quad \begin{cases} \mathbf{b}(\mathbf{X}) := \dot{\boldsymbol{\Phi}}\mathbf{X}, \\ \mathbf{G}(\mathbf{X}) := \boldsymbol{\Phi}\boldsymbol{\Theta}(\mathbf{X}). \end{cases}$$

The discretized analogue of the weak convolutional formulation in eq. (18) is then given, up to a quadrature error $\|\mathbf{e}_{\text{int}}\| = \mathcal{O}(\Delta t^{p+1})$ with $2p$ being the polynomial order of φ, by $\mathbf{b}(\mathbf{X}^*) \approx \mathbf{G}(\mathbf{X}^*)\mathbf{w}^*$. Similarly, the discrete analogue of the sparse regression problem in eq. (20) takes the form

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

$$\hat{\mathbf{w}} := \underset{\mathbf{w}\in\mathbb{R}^{J\times d}}{\operatorname{argmin}} \mathcal{L}_\mu(\mathbf{w}), \quad \text{where} \quad \mathcal{L}_\mu(\mathbf{w}) := \|\mathbf{b}(\mathbf{X}) - \mathbf{G}(\mathbf{X})\mathbf{w}\|_F^2 + \mu\|\mathbf{w}\|_0, \tag{21}$$

the solution of which we approximate via the MSTLS algorithm described in Appendix A.

3.3. Test functions and candidate library

Following [18], we use a rescaled Bernstein polynomial for the generating test function φ, defined in a piecewise fashion by

$$\varphi(t; m, p) := \left(1 - \frac{t^2}{m^2\Delta t^2}\right)^p \quad \text{within} \quad t \in \operatorname{supp}(\varphi) := [-m\Delta t,\, m\Delta t],$$

where the degree p is defined for a highest derivative order $\bar{\alpha} := 1$ and support tolerance $\tau_0 := 1\mathrm{e}-10$ via

$$p = \max\left\{\left\lceil \frac{\ln(\tau_0)}{\ln((2m-1)/m^2)} \right\rceil,\ \bar{\alpha}+1\right\} \geq 2. \tag{22}$$

Convolving the scattering data against localized test function kernels $\varphi_k(\cdot; m, p)$ approximately projects these data onto the temporal scales $t \gtrsim m\Delta t$; in particular, as the support radius $m\Delta t \to 0$, the weak (WSINDy) formulation in eq. (18) collapses to the strong (SINDy) formulation in eq. (17). From a distributional perspective, the test functions φ_k converge to Dirac delta distributions $\delta(t_k)$ while the derivatives $\dot{\varphi}_k$ converge to centred finite difference kernels. We provisionally set $m := 20$ throughout. For additional information about our numerical implementation, we refer the reader to Appendix A.

Our choice of library $\boldsymbol{\Theta}(\boldsymbol{x})$ is motivated by the analytic structure of the scattering ODEs in eq. (14). It is convenient to work in terms of $(\kappa_i, \log(c_i))$ rather than, e.g., (κ_i, c_i) or (λ_i, c_i), as the scattering dynamics of the KdV equation become polynomial in these coordinates: $\dot{\kappa}_i = 0$ and $\frac{d}{dt}\log(c_i) = 8\beta\kappa_i^3$. Accordingly, we use a library of low-order (i.e., cubic) monomials in κ_i and in $\log(c_i)$,

$$\boldsymbol{\Theta}(\boldsymbol{x}_i) = \left\{\kappa_i,\ \kappa_i^2,\ \kappa_i^3,\ \log(c_i),\ \log(c_i)^2,\ \log(c_i)^3\right\}. \tag{23}$$

In order to respect the index-invariant physical symmetry of the system – i.e., that each soliton follows the same physics – we enforce index-invariant ODEs of the form $\dot{\kappa}_i = f_{\boldsymbol{\kappa}}(\kappa_i, c_i)$ and $\dot{c}_i = f_{\mathbf{c}}(\kappa_i, c_i)$.

4. Results

4.1. Validation

To illustrate how the performance of our scheme scales with increasingly noisy data $\mathbf{u} = \mathbf{u}^* + \boldsymbol{\varepsilon}$, we run repeated trials on corrupted versions of the datasets and report the resulting metrics (see Figure 3). Following [18], we add distinct realizations of artificial i.i.d. noise $\varepsilon \in \mathcal{N}(0, \sigma_\varepsilon^2)$ in a pointwise fashion to each element of $\mathbf{U} := \mathbf{u}(\mathbf{t})$, which are computed by enforcing a standard deviation $\sigma_\varepsilon = \sigma_{\text{NR}}\|\mathbf{U}\|_F$ so that $\sigma_{\text{NR}} = \|\boldsymbol{\varepsilon}(\mathbf{t})\|_F/\|\mathbf{U}\|_F$. We explore noise ratios in the range $0 \leq \sigma_{\text{NR}} \leq 0.5$, i.e., up to 50% of the magnitude of the data. To give a sense of the symbolic form of the discovered equations, in Table 1 we explicitly compare the ground-truth and identified models at zero noise for the synthetically generated KdV data of §4.2. Metrics given below then track how the identified models deform as σ_{NR} increases.

To help gauge the quality of the weak-form regression, we report the coefficient of determination corresponding to each identified WSINDy model, which is defined by

$$R^2 := 1 - \frac{\|\mathbf{b} - \mathbf{G}\hat{\mathbf{w}}\|_2^2}{\|\mathbf{b} - \bar{\mathbf{b}}\|_2^2}, \quad \text{where} \quad \bar{\mathbf{b}} := \left[\frac{1}{Kd}\sum_{k,i}\mathbf{b}_{k,i}\right]\mathbf{1}.$$

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

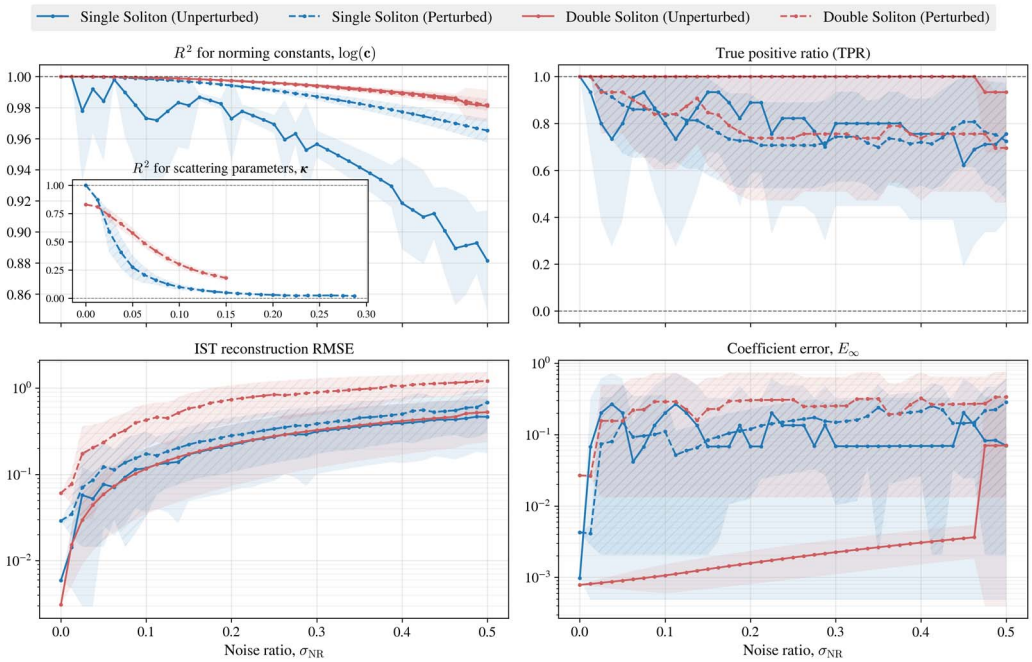


Figure 3. *Illustrating how the validation metrics of §4.1 scale with increasing noise for each of the four configurations of synthetic data. The noise ratio σ_{NR} is swept from* 0 *to* 0.5 *in increments of* $\Delta\sigma_{NR} = 0.0125$ *using* 15 *distinct realizations per nonzero noise level; solid lines denote sample means while the shaded bands denote* ± 1 *standard deviation. [Top-left]* R^2 *values for the identified norming constant models, with an inset corresponding to the scattering parameter models; as per §4.1, we omit* R^2 *values for runs in which the zero model* $\hat{\mathbf{w}} = 0$ *was selected. The TPR [top-right], coefficient error [bottom-right] and IST reconstruction error [bottom-left] are computed via eqs. (25) and (24).*

This metric, which equals the proportion of the variance of $\mathbf{b}$ that is explained by the identified model $\mathbf{G}\hat{\mathbf{w}}$, satisfies $R^2 \leq 1$, with values near one indicating an accurate model (this metric is unstable for models with small or vanishing coefficients, so we do not report a value when the weights vanish; see Figure 3). We additionally report the pointwise RMSE between the true state $\mathbf{U}^*$ and the reconstructed field $\mathbf{U}_{\text{IST}}$ obtained by applying the IST to a forward simulation of the discovered scattering dynamics (see the bottom panel of Figure 1),

$$\text{RMSE}(\mathbf{U}_{\text{IST}}) := \frac{1}{\sqrt{N_x N_t}} \left\| \mathbf{U}^* - \mathbf{U}_{\text{IST}} \right\|_F, \tag{24}$$

which quantifies how faithfully the recovered dynamics reproduce the original field. For reflectionless data, the IST admits a convenient closed-form 'τ-determinant' representation [12] given by

$$u(x,t) = -\frac{2}{\gamma}\frac{d^2}{dx^2}\log|\mathbf{A}(x;\boldsymbol{\kappa}(t),\mathbf{c}(t))|, \quad \text{where} \quad \mathbf{A}_{ij}(x;\boldsymbol{\kappa},\mathbf{c}) := \delta_{ij} + \frac{\sqrt{c_i c_j}\, e^{-(\kappa_i+\kappa_j)x}}{\kappa_i + \kappa_j},$$

which we use to recover $\mathbf{U}_{\text{IST}}$ from the scattering data $(\hat{\kappa}_i, \hat{c}_i)(t)$ forecast with the identified models. The τ-determinant representation is more numerically convenient than the GLM equation of eq. (13), but only applies to reflectionless cases. When forecasting a WSINDy model forward in time, we use an adaptive RK-45 scheme instantiated with the exact initial condition; see Figure 4.

The metrics above are model-agnostic and can be computed in every case, including the experimental setting of §4.3. In cases where the underlying model is available, i.e., the synthetic experiments of §4.2,

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

Table 1. *Results obtained using synthetic, noise-free ($\sigma_{NR} = 0$) data sourced from numerical simulations of the perturbed KdV model in eq. (5) with the canonical parameters $(\alpha, \beta) = (6, 1)$. For each configuration, ground-truth reference dynamics are compared against the model identified by WSINDy. The reference equations are given by the exact scattering dynamics of eq. (14) for the unperturbed runs ($\epsilon = 0$) and by the near-integrable scattering dynamics of eq. (15) for the perturbed runs ($\epsilon = 0.2$). For $\epsilon = 0.2$, note that $\frac{1}{3}\epsilon \approx 0.0667$ and $\frac{2}{3}\epsilon \approx 0.1333$*

Number of solitons (n)	Perturbation (ϵ)	Forcing (F)	Ground truth model	Identified model (0% noise)
$n = 1$	$\epsilon = 0$	N/A	$\begin{cases} \dot{\kappa} = 0, \\ \frac{d}{dt}\log(c) = 8\kappa^3 \end{cases}$	$\begin{cases} \dot{\kappa} = 0, \\ \frac{d}{dt}\log(c) = 7.9981\kappa^3 \end{cases}$
$n = 2$	$\epsilon = 0$	N/A	$\begin{cases} \dot{\kappa}_i = 0, \\ \frac{d}{dt}\log(c_i) = 8\kappa_i^3 \end{cases}$	$\begin{cases} \dot{\kappa}_i = 0, \\ \frac{d}{dt}\log(c_i) = 7.9984\kappa_i^3 \end{cases}$
$n = 1$	$\epsilon = 0.2$	$F[u] = (xu)_x$	$\begin{cases} \dot{\kappa} = \frac{1}{3}\epsilon\kappa + O(\epsilon^2), \\ \frac{d}{dt}\log(c) = 8\kappa^3 - \frac{2}{3}\epsilon\log(c) + O(\epsilon^2) \end{cases}$	$\begin{cases} \dot{\kappa} = 0.0666\kappa, \\ \frac{d}{dt}\log(c) = 8.0072\kappa^3 - 0.1336\log(c) \end{cases}$
$n = 2$	$\epsilon = 0.2$	$F[u] = (xu)_x$	$\begin{cases} \dot{\kappa}_i = \frac{1}{3}\epsilon\kappa_i + O(\epsilon^2), \\ \frac{d}{dt}\log(c_i) = 8\kappa_i^3 - \frac{2}{3}\epsilon\log(c_i) + O(\epsilon^2) \end{cases}$	$\begin{cases} \dot{\kappa}_i = 0.0660\kappa_i, \\ \frac{d}{dt}\log(c_i) = 8.0192\kappa_i^3 - 0.1364\log(c_i) \end{cases}$

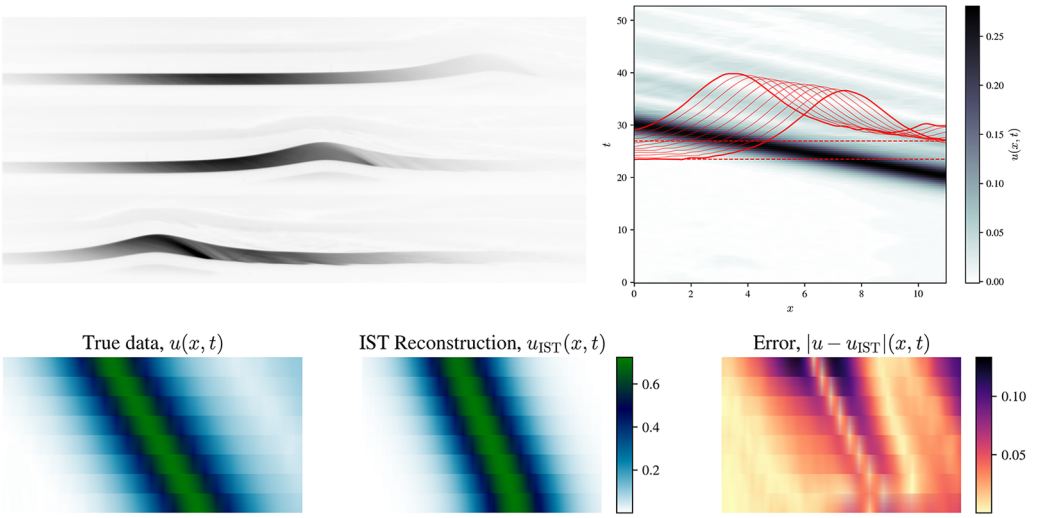


Figure 4. *[Top-left] Example snapshots from the experimental dataset of [10, 11]. [Top-right] The selected temporal window for one trial in which the soliton sits solidly in-frame, showing the extracted profiles (red) overlaid on the measured field $u = u(x,t)$. [Bottom] Visualizing, for one held-out validation trial, the measured field, the reconstruction obtained by forward-integrating the WSINDy-discovered scattering dynamics and applying the IST, and the corresponding pointwise error.*

we also compare the identified coefficients against a reference; for unperturbed synthetic data ($\epsilon = 0$) the reference is the exact scattering ODEs of eq. (14), while for perturbed synthetic data ($\epsilon > 0$) it is the perturbed scattering ODEs of eq. (15). In these cases, we follow [18] in reporting the normalized ℓ^∞ coefficient error and the true positive ratio (TPR), respectively defined by

$$E_\infty := \max_{j=1,\dots,J} \frac{|w_j - w_j^*|}{|w_j^*|}, \quad \text{and} \quad \mathrm{TPR} := \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{25}$$

The E_∞ coefficient error represents maximum element-wise relative error incurred by the discovered model, while the TPR assesses the extent to which the identified models recover the correct terms (here, TP denotes the number of terms that were correctly identified as nonzero, FP denotes the number of terms that were falsely identified as nonzero, and FN denotes the number of terms that were falsely identified as zero.). Note that a TPR of 1 means that the true model has been discovered in its entirety, while a TPR of 0 indicates that none of the correct terms was identified.

4.2. Synthetic data

We consider several examples using synthetic data sourced from numerical simulations of the perturbed KdV equation given in eq. (5). In particular, we explore configurations featuring either one or two solitons (see Figure 2) and investigate both the unperturbed ($\epsilon = 0$) and perturbed cases ($\epsilon = 0.2$) subject to a mass-conserving forcing function, $F[u] = (xu)_x$, for which the reference ODEs in eq. (15) can be explicitly computed (see Table 1). In each instance, we use the canonical KdV parameters $\alpha = 6$ and $\beta = 1$. The numerical simulations are computed on the domain $x \in \Omega = [-25, 25]$ using $N = 2^{10}$ uniformly spaced nodes and adaptively integrated for $t \in [0, 1.5]$ using a pseudo-spectral solver from the open-source `sangkuriang-ideal` [13] Python package with $M = 200$ saved snapshots. (This package does not natively support perturbations, so for the perturbed cases we use a slightly hand-modified version of the original code.)

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

In the unperturbed case, the scattering parameters κ_i are conserved quantities (i.e., $\dot{\kappa}_i = 0$), meaning that any single simulation samples only a constant value for each parameter. Consequently, the candidate monomials κ_i, κ_i^2, κ_i^3 in the WSINDy library of eq. (23) are linearly dependent along the trajectory, and the regression problem in eq. (21) becomes ill-posed. To account for this degeneracy, we stack scattering data $\mathbf{X}_r$ from three distinct simulations $r = 1, 2, 3$ (the minimum number required for well-posedness) and concatenate each respective weak-form linear system from eq. (19) into a single block-regression. This is motivated by the nature of the experimental dataset of [10] investigated in §4.3 below, which includes several (25) distinct experimental runs. Each simulation is initialized using sech-squared profiles $u_0(x) = \sum_{i=1}^{n} 2\kappa_i^2 \mathrm{sech}^2(\kappa_i(x - x_i))$ parameterized by a distinct set of scattering parameters κ_i and initial positions $x_j := 5(j - 5)$. For the single soliton examples in which $n = 1$, we vary $\kappa \in \{2, 2.2, 2.4\}$, while for the two-soliton cases with $n = 2$, we instead use $(\kappa_1, \kappa_2) \in \{(2, 1.8), (2.2, 1.4), (2.4, 1.2)\}$; the $\kappa = 2$ and $(\kappa_1, \kappa_2) = (2.4, 1.2)$ cases are shown in Figure 2. Finally, because the unperturbed dynamics obey a trivial conservation law of the form $\dot{\boldsymbol{\kappa}} = 0$, we extend the standard MSTLS of Algorithm A1 to the zero-admissible formulation given in Algorithm A2, which uses BIC-based hypothesis testing to allow WSINDy to select model coefficients with empty support.

As reported in Table 1, the recovered coefficients agree with both the exact scattering ODEs and the leading-order perturbation theory to within a fraction of a per cent at zero noise; e.g., in the perturbed single soliton example, WSINDy estimates $\dot{\kappa} \approx 0.0667\kappa$ versus the ground-truth of $\dot{\kappa} = \frac{1}{3}\epsilon\kappa \approx 0.0666\kappa$. In turn, Figure 3 tracks how the model identification results deform as the noise ratio σ_{NR} increases, revealing a few consistent trends. In particular, the weak-form regression for the $\log(\mathbf{c})$ models stays robust across the entire noise range – the R^2 for these ODEs remains above ~ 0.96 for three of the four configurations and declines only mildly to ~ 0.88 for the weakest (i.e., unperturbed single-soliton) case at 50% noise. By contrast, the R^2 values decay rapidly for the perturbed $\dot{\boldsymbol{\kappa}}$ equations, which feature small $\mathcal{O}(\epsilon)$ coefficients. Mechanistically, beyond $\sim 15\%$ background noise this perturbation becomes indistinguishable from a conservation law of the form $\dot{\kappa} = 0$, at which point the BIC test in Algorithm A2 selects the zero model $\hat{\mathbf{w}} = 0$ and a corresponding R^2 value is no longer reported.

Although the R^2 eventually becomes unstable for the perturbed cases featuring small coefficients, the correct model terms are nonetheless reliably recovered: the mean TPR stays above ~ 0.7 at all noise levels for each of the four configurations, while the unperturbed two-soliton example in particular retains a TPR of exactly one until $\sim 45\%$ noise. In a similar vein, the E_∞ coefficient error remains $\mathcal{O}(\texttt{1e-3})$ across nearly the entire noise range in the unperturbed two-soliton case, rising only beyond $\sigma_{\mathrm{NR}} \approx 0.45$. By contrast, in the single-soliton and perturbed cases, the E_∞ coefficient error plateaus near $\mathcal{O}(\texttt{1e-1})$, again reflecting the difficulty of resolving the small perturbative coefficients. Moreover, occasional sharp oscillatory peaks are present in several cases – features which we predominantly attribute to sampling error. Lastly, we observe that the reconstruction RMSE, a proxy for the predictive fidelity of the identified models, grows monotonically with σ_{NR} in every case. Unsurprisingly, the most dynamically complex configuration (i.e., the perturbed two-soliton collision) tends to incur the largest reconstruction error.

4.3. Empirical data

In this section, we illustrate the discovery of effective soliton dynamics using real experimental data. Specifically, we consider data from a recent study by Heinrich et al. [10, 11] in which a Kawahara-type PDE was identified from video recordings of solitary shallow-water waves propagating down a flume apparatus (see Figure 4). In this study, the edge-detection algorithm of [5] was used to extract a nondimensionalized (the rescaling was $u^*, x^*, t^* \mapsto hu, hx, \sqrt{h/g}t$, where u^*, x^*, t^* are the measured dimensional quantities and $h \approx 32$ mm is the water depth) surface field $u_k(x, t)$ for each of $k = 1, \ldots, 25$ distinct experimental trials, 18 of which were subsequently concatenated into a single data matrix $\mathbf{U}$ (cf. §4.2 above). The authors then applied two distinct equation-learning methodologies to the data – WSINDy and a novel "Fourier multiplier" method [10] – both of which independently selected a sparse

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

Table 2. *The models identified from the experimental data of [10, 11]. Here, we report the RMSE of eq. (24) normalized to the peak amplitude $A := \max|u|$ and averaged over the seven validation trials*

Identified model	Library	RMSE	R^2
$\begin{cases} \dot{\kappa} = 0, \\ \frac{d}{dt}\log(c) = 3.46\kappa + 2.60\kappa^3 - 0.20\log(c) \end{cases}$	$\begin{cases} \{1\} \cup \{\kappa^p, \log(c)^p\}_{1\le p\le 3}, \\ \{1, \kappa, \kappa^3\} \cup \{\log(c)^p\}_{1\le p\le 3}, \end{cases}$	17.6%	0.803

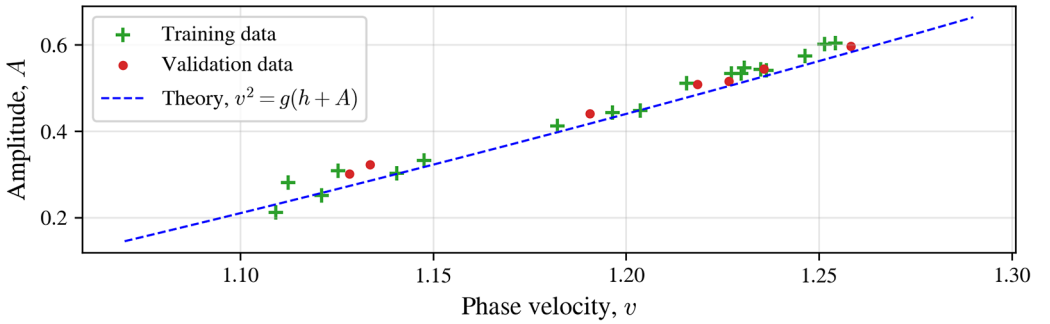


Figure 5. *A reproduction of Figure 9 from Ref. [10], which illustrates an empirical amplitude–phase velocity relation obeyed by the shallow-water waves. Here, the dimensional amplitudes $A^* := \max|hu^*|$ are plotted against the phase-velocity $v^* := |\dot{x}|/\sqrt{gh}$ for the 18 training and 7 validation trials, showing the validity of the empirical relation. Note that the first two terms in the identified $\frac{d}{dt}\log(c)$ model listed in Table 2 are consistent with the first two terms of a Taylor series expansion of the nondimensionalized $v \approx \sqrt{1 + A}$ relationship; see §4.3.*

model of the form

$$u_t + v_0 u_x + \alpha u u_x + \beta u_{xxx} = \epsilon u_{xxxxx}. \tag{26}$$

Moreover, similar coefficient values were estimated by each method (e.g., $\epsilon \approx -0.06$ from both methods). In a moving frame of reference $\xi = x - v_0 t$, this model is equivalent to the Kawahara model in eq. (6),

$$u_t + \alpha u u_\xi + \beta u_{\xi\xi\xi} = \epsilon u_{\xi\xi\xi\xi\xi}.$$

To validate the discovered model, the authors of [10] numerically integrated eq. (26) forward in time and compared the predicted field to the measured soliton surface heights from each of the seven held-out experimental trials, achieving errors of just 4–6% relative to the wave amplitudes.

Note that, for our purposes here, we apply two additional lightweight preprocessing steps to the data: (1) a small number of spike artefacts were repaired via local temporal averaging; (2) we subtracted a small offset from the water level $u \mapsto u - \bar{u}$, where $\bar{u} \sim 0.02$–0.13 mm, so that $u \to 0$ as $x \to \partial\Omega$. Just as in [11], each run is sampled over $N_x = 1200$ points in space ($\Delta x \approx 0.29$ mm) and $N_t = 151$ frames in time ($\Delta t = 0.02$ s). In each case, we isolate a specific temporal window in which the soliton stays solidly in-frame, which we then temporally super-sample via cubic-spline interpolation before computing the scattering data (see Figure 4). We estimate that the experimental noise level is roughly $\sigma_{\mathrm{NR}} \approx 1.4\%$, comfortably within the regime in which WSINDy provides reliable model estimates (cf. §4.2).

For each equation listed in Table 2, we use a candidate library inspired both by the identified PDE model in eq. (26) and by the empirical amplitude–phase-velocity relation $v^2 \approx 1 + A$ reported in [10] (see Figure 9 therein), where $A := \max|u(x,t)|$ and $v := |\dot{x}|$ respectively denote the nondimensionalized amplitude and phase velocity of the travelling water wave (see Figure 5). In particular, a single exact KdV soliton satisfies $\log(c) = 2\kappa\xi + \log(2\kappa)$ in a co-moving frame of reference $\xi(t) = \xi_0 + 4\beta\kappa^2 t$ (cf.

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

eq. (14)); however, a stationary observer (i.e., in the lab frame $x = \xi + v_0 t$) would measure an additional advective force,

$$\tfrac{d}{dt}\log(c) = 2\kappa v = 2v_0\kappa + 8\beta\kappa^3. \tag{27}$$

Inserting an expansion of the empirical relation $v \approx \sqrt{1+A} = 1 + (1/2)A + \mathcal{O}(A^2)$ into eq. (27) and using the exact soliton amplitude $A(\kappa) = -(2/\gamma)\kappa^2 = 12(\beta/\alpha)\kappa^2$ independently yields

$$\tfrac{d}{dt}\log(c) = 2v\kappa \approx 2\kappa\left(1 + \tfrac{6\beta}{\alpha}\kappa^2 + \mathcal{O}(\kappa^4)\right) = 2\kappa + \tfrac{12\beta}{\alpha}\kappa^3 + \mathcal{O}\left(\kappa^5\right). \tag{28}$$

Consistency between the idealized and empirical ODEs in eqs. (27) and (28) requires that both $v_0 = 1$ (empirically, $v_0 \approx 1.2$) and $8\beta = 12(\beta/\alpha)$, which occurs if and only if $\alpha = 3/2$ ($\alpha \approx 1.4$ and $8\beta \approx 5$). Therefore, the observed $v^2 \approx 1 + A$ relationship may be viewed as a restatement of the ODE in eq. (28). Guided by this structure, we use a candidate library based on the odd powers $\{\kappa, \kappa^3\}$, where we interpret the linear term as capturing advection and the cubic term as capturing the soliton's nonlinear evolution, together with low-order powers of $\log(c)$ (see Table 2). Although the identified Kawahara-type PDE is translation-invariant, our hope is that the latter terms allow us to absorb any residual position-dependence introduced by the experimental or numerical setup rather than by the physical dynamics (recall that $\log(c)$ is highly correlated with the soliton position ξ).

Using the setup described above, WSINDy selects the conservation law $\dot{\kappa} = 0$ together with a model for $\frac{d}{dt}\log(c)$ that is consistent with the form of the cubic expansion given in eq. (27), with the exception of an additional small $\log(c)$ term reminiscent of the perturbed scattering ODEs from Table 1. We emphasize, however, that we do not realistically expect the additional $\log(c)$ term to play an active role in the physical dynamics – instead, we attribute its existence to experimental and numerical artefacts, and note that it likely skews the relative magnitudes of the identified coefficients (cf. eq. (27)). Forward integration of the discovered scattering dynamics and subsequent reconstruction of the water-line via the IST yields an amplitude-normalized RMSE of 8–27% (or ~18% on average; see Table 2) across the seven held-out trials (cf. Figure 4). In each case exhibiting large RMSE, the reconstruction errors are predominantly caused by a secondary wave trailing behind the main soliton (i.e., additional radiation), which the reflectionless IST is not capable of representing.

4.4. A note on identifiability

We conclude the results section with a few brief remarks on the identifiability of nonlinear dynamics from solitary wave trajectories. Rudy et al. [20] observed that when SINDy-based methods are applied to measurements of a single soliton evolving along a single trajectory, they tend to recover trivial linear transport equations in place of any underlying nonlinear wave dynamics. (As in §4.2, this problem can be avoided by stacking trajectories – for example, Heinrich et al. [10] use 18 distinct single soliton trajectories.) Interestingly, this behaviour represents an inherent feature of the data, rather than a deficiency of the SINDy architecture; i.e., every lone KdV soliton *exactly* solves a corresponding linear transport equation,

$$u(x,t) = 2\kappa^2\,\mathrm{sech}^2\big(\kappa\big(x - x_0 - 4\kappa^2 t\big)\big) \quad \text{solves both:} \quad \begin{cases} u_t + 4\kappa^2 u_x = 0, \\ u_t + 6uu_x + u_{xxx} = 0. \end{cases} \tag{29}$$

Since both candidate models in eq. (29) reproduce the data exactly, a sparse-regression architecture naturally selects the simpler one – in this case, the transport equation is a more parsimonious model than the KdV equation. Physically speaking, nonlinearity manifests itself in the KdV equation via the dependence of wave-speeds on wave-amplitudes, and no such dependence can be observed from a soliton in isolation. Conversely, nonlinearities can be identified by, e.g., observing a two-soliton interaction

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

or, more generally, a collection of solitons of differing amplitudes [20]. A closely related identifiability issue was subsequently reported by Vasey et al. [24] in the setting of magnetohydrodynamic blast waves.

Working with scattering data $(\kappa_i, c_i)(t)$ instead of field data $u(x,t)$ does not fundamentally remove the degeneracy described above, although it does change the nature of the identifiability problem in an interesting and potentially useful way – in particular, the scattering ODEs in eqs. (14) and (15) are nonlinear in κ_i, regardless of the number of solitons present in the data. However, in analogy with eq. (29), each $\kappa_i(t) = \kappa_i(0)$ is conserved and thus constant in the absence of an external perturbation; consequently, the candidate monomials $\{\kappa_i, \kappa_i^2, \kappa_i^3\}$ in the library of eq. (23) are nearly collinear along a single trajectory, and linear regression cannot uniquely determine which monomial governs the evolution of $\log(c_i)$. It is important to recognize that the collinearity of the κ_i^p is the manifestation, observed at the level of the scattering data, of the very same mechanism responsible for the spurious transport equation observed at the PDE level. That being said, the nature of this degeneracy does become more benign in the scattering coordinates – the recovered model never collapses onto a structurally distinct equation. The conservation laws $\dot{\kappa}_i = 0$ and the linear-in-time growth of $\log(c_i)$ can still be identified, and only the exponents of the $\frac{d}{dt}\log(c_i) = \kappa_i^p$ models remain unknown. As noted above, the identifiability problem dissolves as the data begin to feature either more than one soliton or more than one trajectory, the latter of which is a phenomenon that has been observed in settings as diverse as network dynamics [22].

To concretely illustrate the above points, we observe that applying WSINDy for PDEs [18] to just the first experimental training dataset from [11] yields the transport equation $u_t \approx 1.2u_x$ with $R^2 \approx 0.99$. Interestingly, when using the scale-invariant preconditioning method described in [18], WSINDy instead identifies a nonlinear PDE of the form $u_t \approx -0.0u + 0.1u^2 + 0.9u_x + 0.8(u^2)_x$, where again $R^2 \approx 0.99$, indicating that scaling becomes particularly important when attempting to recover such nonlinearities. (We note that additional radiation content in the experimental data may also have aided in the identification of the nonlinear $(u^2)_x$ term.) In a similar vein, regressing the scattering data against a simple polynomial library of the form $\{\kappa, \kappa^2, \kappa^3\}$ (evaluated over the same experimental trial data) yields an advective law of the form $\frac{d}{dt}\log(c) \approx -2\kappa$, irrespective of whether the library columns are rescaled. In this case, $\kappa(t) \approx 2/3$ is nearly constant along the trajectory, and the identified model is almost indistinguishable from both $-3\kappa^2$ and $-5\kappa^3$.

5. Discussion

In this paper, we have proposed and investigated a data-driven method for modelling effective soliton dynamics that works directly with scattering data. In particular, this technique combines the conceptual framework of the IST [2] with the weak-form equation-learning paradigm of WSINDy [19] to identify interpretable symbolic models without requiring prior knowledge of the form of the scattering equations. Parameterizing a scalar field u via its scattering data (κ_i, c_i) is natural in the context of near-integrable wave equations and inherits a large body of analytical structure from existing theory, e.g., the fact that unperturbed KdV scattering dynamics take a convenient polynomial form in $(\kappa_i, \log(c_i))$ coordinates. We have focused on shallow-water models of KdV-type and demonstrated that for this class of governing equations, the technique reliably recovers both the exact unperturbed scattering ODEs as well as the leading-order corrections predicted by the perturbation theory [14].

Our numerical experiments indicate that the inferred scattering dynamics are robust to low amounts of additive i.i.d. Gaussian measurement noise, both in unperturbed and perturbed regimes. On synthetic data, the recovered coefficients agree with the exact and perturbed scattering ODEs to within a fraction of a percent at 0% noise, and the $\frac{d}{dt}\log(\mathbf{c})$ models remain accurate (i.e., $R^2 \gtrsim 0.96$ in three of four cases) across the entire 0–50% noise range tested. Unsurprisingly, the small $\mathcal{O}(\epsilon)$ corrections to the $\dot{\boldsymbol{\kappa}}$ equations are harder to resolve in the presence of measurement noise; beyond $\sigma_{\text{NR}} \gtrsim 0.15$, these data become statistically indistinguishable from the conservation law $\dot{\boldsymbol{\kappa}} = 0$ and the BIC test of Algorithm A2 defaults to the zero model. When applied to the experimental shallow-water wave dataset of Heinrich et al. [10, 11], the method recovers $\dot{\kappa} = 0$ together with a $\frac{d}{dt}\log(c)$ model consistent with a cubic

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

expansion of the empirical amplitude–phase-velocity relationship reported therein; forward-integration of the recovered dynamics and a subsequent IST reconstruction $(\kappa, \log(c)) \mapsto u$ yields an amplitude-normalized RMSE of $\sim 18\%$, with the largest errors stemming from the existence of excess radiation trailing the main wave.

Notably, numerical experiments that featured a single, unperturbed soliton consistently saw worse performance than their two-soliton counterparts across each validation metric (R^2, E_∞, RMSE, TPR), despite the fact that these cases are significantly simpler from a dynamical perspective. We attribute this somewhat counterintuitive result to a structural identifiability mechanism similar to that observed in both the KdV and magnetohydrodynamic settings of [20] and [24], in which solitary waves measured along a single trajectory were found to obey advection equations that were much simpler than the expected underlying nonlinear dynamics. Although working with scattering data does not fundamentally remove this degeneracy, it does change the nature of the identifiability problem in a convenient way – while lone solitons exactly obey linear dynamics at the PDE level, the ODEs governing their scattering data are, by contrast, inherently nonlinear in κ_i regardless of the number of solitons present. In the scattering formulation, this degeneracy enters only at a more benign numerical conditioning level, with constant κ_i values rendering the candidate monomials κ_i^p collinear. Moreover, this ill-conditioning is avoided in perturbed regimes where the κ_i are non-constant, and in these cases the identified $\mathcal{O}(\epsilon)$ corrections conceivably allow one to back out a set of consistent forms for the forcing term $F[u]$.

Although we have focused here on shallow-water equations of KdV-type, a similar pipeline applies, in principle, to any PDE admitting a Lax-pair representation (e.g., any integrable wave equation from [2]).

Practically speaking, extending the framework discussed here to a new integrable system would require three main ingredients:

(1) the Lax operator corresponding to the PDE in question;
(2) a numerical DST scheme for the associated EVP;
(3) a choice of coordinates in which the scattering dynamics can be reasonably well-approximated by a low-order library of candidate terms.

Applying this pipeline to the focusing nonlinear Schrödinger equation would be a natural next step, subject to the caveat that a treatment of this equation would introduce additional difficulties that haven't been considered herein. In particular, rather than the self-adjoint EVP considered above, its scattering data are defined in terms of a non-self-adjoint EVP (i.e., the 2×2 Zakharov-Shabat system) that has a fundamentally complex-valued point spectrum featuring non-trivial real *and* imaginary components. Additionally, addressing the defocusing regime (in which dark solitons sit atop a nonzero background) would require developing a strategy for relaxing the assumption of decay as $x \to \infty$.

We conclude by considering a few natural extensions of the current work. In addition to addressing other integrable PDEs and more straightforward methodological improvements such as an end-to-end data-driven pipeline utilizing WSINDy [18] and SILO [3], one could consider accommodating nonzero radiation (i.e., $R \neq 0$) by appending reflection coefficients to the state vector as outlined in §3.1. This would generalize the modelling technique to non-solitonic settings and, seeing as how excess radiation was the dominant source of reconstruction error in our experimental results, would likely increase the predictive capacity of the discovered models. From a more theoretical perspective, the inclusion of radiation might also lead one to consider *purely dispersive* regimes in which $\mathbf{c} = 0$ while $R \neq 0$, including connecting the long-time asymptotics of such systems to Painlevé-type equations [1]. Lastly, we expect that it would be interesting and potentially fruitful to explore geophysical modelling applications related to, e.g., internal ocean waves, coastal shoaling waves, and morning-glory waves – settings where coherent nonlinear wave phenomena are ubiquitous and high-resolution field data are becoming increasingly available [7, 15].

Data availability statement. Replication data and code are publicly available: https://github.com/MathBioCU/Scattering.

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

Acknowledgements. The authors would like to thank Mark Ablowitz for his helpful insights, especially concerning future work with Painlevé equations. This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

Author contributions. Conceptualization: S.M.; V.D.; D.M.B. Methodology: S.M; V.D.; D.M.B. Data curation: S.M. Data visualization: S.M. Writing original draft: S.M. All authors approved the final submitted draft.

Funding statement. This work is supported in part by National Science Foundation Grants 2054085 and 2109774, National Institute of Food and Agriculture Grant 2019-67014-29919 and Department of Energy Grant DE-SC0023346.

Competing interests. The authors declare no competing interests.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

[1] **Ablowitz MA and Clarkson PA** (1991) *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, 1st Edn. Cambridge University Press. https://doi.org/10.1017/CBO9780511623998.

[2] **Ablowitz MJ, Kaup DJ, Newell AC and Segur H** (1974) The inverse scattering transform-fourier analysis for nonlinear problems. *Studies in Applied Mathematics* **53**, 249–315.

[3] **Adriazola J, Zhu W, Kevrekidis PG and Aceves A** (2026) Computer assisted discovery of integrability via SILO: sparse identification of lax operators. *SIAM Journal on Applied Dynamical Systems* **25**, 131–159.

[4] **Brunton SL, Proctor JL and Kutz JN** (2016) Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **113**(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113.

[5] **Canny J** (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-8**, 679–698.

[6] **Chen S, Yang S, Kevrekidis PG and Zhu W** (2025) Robust moment identification for nonlinear PDEs via a neural ODE approach *Chaos: An Interdisciplinary Journal of Nonlinear Science* **35**, 123135.

[7] **El G and Hoefer M** (2016) Dispersive shock waves and modulation theory *Physica D: Nonlinear Phenomena* **333**, 11–65.

[8] **Feng Y-J, Sun Z-Y and Yu X** (2023) Nonlinear fourier analysis of matter-wave soliton interferometry. *Frontiers in Physics* **10**, 1092207.

[9] **Guo S, Koh SM, Fritsch AR, Spielman BI and Zwolak JP** (2022) Combining machine learning with physics: A framework for tracking and sorting multiple dark solitons. *Physical Review Research* **4**, 023163.

[10] **Heinrich KS, Seth DS, Ehrnstrom M and Ellingsen SA** (2025) Rediscovering shallow water equations from experimental data.

[11] **Heinrich KS, Seth DS, Ehrnstrom M and Ellingsen SA** (2026) Replication data for: rediscovering shallow-water equations from experimental data.

[12] **Hirota R** (1971) Exact solution of the Korteweg—de Vries equation for multiple collisions of solitons. *Physical Review Letters* **27**, 1192–1194.

[13] **Irawan DE, Herho SHS, Khadami F, Anwar PI, Sujatmiko KA, Handayani AP, Fajary FR and Suwarman R** (2026) An open-source pseudo-spectral solver for idealized Korteweg-de Vries soliton simulations.

[14] **Karpman V and Solov'ev V** (1981) A perturbation theory for soliton systems. *Physica D: Nonlinear Phenomena* **3**, 142–164.

[15] **Lee Y-C, Brühl M, Doong D-J and Wahls S** (2024) Nonlinear fourier classification of 663 rogue waves measured in the Philippine sea. *PLOS ONE* **19**, e0301709.

[16] **Mangan NM, Kutz JN, Brunton SL and Proctor JL** (2017) Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**(2204), 20170009. https://doi.org/10.1098/rspa.2017.0009.

[17] **Messenger D, Dwyer G and Dukic V** (2024) Weak-form inference for hybrid dynamical systems in ecology. *Journal of The Royal Society Interface* **21**, 20240376.

[18] **Messenger DA and Bortz DM** (2021) Weak SINDy for partial differential equations. *Journal of Computational Physics* **443**, 110525.

[19] **Messenger DA and Bortz DM** (2021) Weak SINDy: Galerkin-based data-driven model selection. *Multiscale Modeling & Simulation* **19**, 1474–1497. https://doi.org/10.1137/20M1343166.

[20] **Rudy SH, Brunton SL, Proctor JL and Kutz JN** (2017) Data-driven discovery of partial differential equations. *Science Advances* **3**, e1602614.

[21] **Schaeffer H** (2017) Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**(2197), 20160446. https://doi.org/10.1098/rspa.2016.0446.

[22] **Tian M, Messenger DA, Dukic V, Rodríguez N and Bortz DM** (2026) Learning effective models from network dynamics data with multiple initial conditions using weak form SINDy.
[23] **Tikan A, Bonnefoy F, Ducrozet G, Prabhudesai G, Michel G, Cazaubiel A, Falcon A, Copie F, Randoux S and Suret P** (2022) Nonlinear dispersion relation in integrable turbulence. *Scientific Reports* **12**, 10386.
[24] **Vasey G, Messenger D, Bortz D, Christlieb A and O'Shea B** (2025) Influence of initial conditions on data-driven model identification and information entropy for ideal mhd problems. *Journal of Computational Physics* **524**, 113719.
[25] **Yang S, Chen S, Zhu W and Kevrekidis PG** (2024) Identification of moment equations via data-driven approaches in nonlinear Schrödinger models. *Frontiers in Photonics* **5**, 1444993.
[26] **Yang S, Chen S, Zhu W and Kevrekidis PG** (2025) *Data-Driven Soliton Manifold Approximations for Dark and Bright Waves: Some Prototypical 1d Case Examples*.
[27] **Zhang Z, Liu M, Xing X, Zhang S, Yan Z and Liu W** (2025) Based on purely physical information in deep learning optimizes soliton system parameter identification problem. *Computer Methods in Applied Mechanics and Engineering* **438**, 117852.

Algorithm A1. Modified sequential thresholding least squares (`MSTLS`)

Inputs: response vector $\mathbf{b}$ (size K), library matrix $\mathbf{G}$ (size $K \times J$), thresholding parameter $\lambda \in (0, 1)$.
Outputs: thresholded weights $\mathbf{w}^{\lambda}$ (size J).

1. $n \leftarrow 0$
2. $\mathbf{w}^0 \leftarrow \mathbf{w}_{\text{LS}}$
3. `stopping_criterion` $\leftarrow$ `false`
4. **for** $j = 1, \ldots, J$ **do**:
   - $L_j \leftarrow \max\left(1,\ \|\mathbf{b}\|_2 / \|\mathbf{G}_j\|_2\right)$
   - $U_j \leftarrow \min\left(1,\ \|\mathbf{b}\|_2 / \|\mathbf{G}_j\|_2\right)$
5. **while** `stopping_criterion == false` **do**:
   - $\mathcal{I}_n \leftarrow \left\{1 \le j \le J \ :\ |\mathbf{w}_j^n| \in [\lambda L_j,\ \lambda^{-1} U_j]\right\}$
   - $\mathbf{w}_{n+1} \leftarrow \operatorname{argmin}_{\text{supp}(\mathbf{w}) \subseteq \mathcal{I}_n} \|\mathbf{b} - \mathbf{G}\mathbf{w}\|_2^2$
   - **if** $n \ge 1$ **and** $\mathcal{I}_n = \mathcal{I}_{n-1}$:
     - `stopping_criterion` $\leftarrow$ `true`
     - $\mathbf{w}^{\lambda} \leftarrow \mathbf{w}_{n+1}$
   - $n \leftarrow n + 1$

Appendix A. The MSTLS Algorithm

We use a zero-admissible extension of the MSTLS algorithm developed in [18] to approximately solve the sparse regression problem posed in eq. (20). In MSTLS, a sparse vector of model weights $\hat{\mathbf{w}}_i \in \mathbb{R}^J$ with $\|\hat{\mathbf{w}}_i\|_0 \ll J$ for each $i = 1, \ldots, d$ is obtained by minimizing a normalized version of the loss function given in eq. (20) over a finite set of *thresholding parameters* $\boldsymbol{\lambda} := \{\lambda_l : l = 1, \ldots, N_\lambda\} \subset (0, 1)$. In particular, the model weights are explicitly given by

$$\hat{\mathbf{w}}_i^{\text{MSTLS}} := \texttt{MSTLS}\left(\mathbf{b}_i,\ \mathbf{G},\ \lambda^\star\right), \quad \text{with} \quad \lambda^* := \min\left[\operatorname*{argmin}_{\lambda \in \boldsymbol{\lambda}} \mathcal{L}_{\text{MSTLS}}(\lambda)\right], \tag{A.1}$$

where `MSTLS` denotes the output of Algorithm A1. For a given thresholding parameter $\lambda \in (0, 1)$, the loss function $\mathcal{L}_{\text{MSTLS}}$ is defined as

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

$$\mathcal{L}_{\text{MSTLS}}(\lambda) := \mathcal{L}_{\mu}\left(\mathbf{w}_i^{\lambda}; \frac{\mathbf{b}_i^{\text{LS}}}{\|\mathbf{b}_i^{\text{LS}}\|_2}, \frac{\mathbf{G}}{\|\mathbf{b}_i^{\text{LS}}\|_2}\right), \quad \text{with} \quad \mu := \frac{1}{J},$$

where $\mathbf{b}_i^{\text{LS}} := \mathbf{G}\mathbf{w}_i^{\text{LS}}$ is the projection of the ordinary least-squares estimate given by

$$\mathbf{w}_i^{\text{LS}} := \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{b}_i.$$

Here, the quantity $\mathbf{w}_i^{\lambda} := \texttt{MSTLS}(\mathbf{b}_i, \mathbf{G}, \lambda)$ denotes the vector of λ-thresholded weights, which at every iteration satisfies the following dominant balance relationship:

$$\frac{\|w_{ji}^{\lambda}\mathbf{G}_j\|_2}{\|\mathbf{b}_i\|_2} \in \left[\lambda, \lambda^{-1}\right], \quad \text{for each} \quad j = 1, \ldots, J.$$

We follow [18] in scanning over a set of $N_\lambda = 50$ candidate values $\boldsymbol{\lambda} := \{\lambda_l\}_{l=1}^{50}$ defined by uniformly log-spaced increments $\log_{10}(\lambda_l) \in (-4, 0)$ with the exception of the perturbed $\log(c)$ cases in §4.2 and §4.3, for which we found it helpful to hand-tune the λ^* parameter to `4e-2` and `1e-1`, respectively.

Algorithm A2. MSTLS with BIC testing for trivial dynamics

Inputs: response vector $\mathbf{b}$ (size K), library matrix $\mathbf{G}$ (size $K \times J$), thresholding parameters $\boldsymbol{\lambda} = \{\lambda_l\}_{l=1}^{N_\lambda}$.
Outputs: BIC-tested weights $\hat{\mathbf{w}}$ (size J).

1. $\lambda^* \leftarrow \min\left[\text{argmin}_{\lambda \in \boldsymbol{\lambda}}\ \mathcal{L}_{\text{MSTLS}}(\lambda)\right]$
2. $\hat{\mathbf{w}}^{\text{MSTLS}} \leftarrow \texttt{MSTLS}(\mathbf{b}, \mathbf{G}, \lambda^*)$
3. $\Delta \leftarrow \gamma^2 \log\left(\|\mathbf{b} - \mathbf{G}\hat{\mathbf{w}}^{\text{MSTLS}}\|_2^2 / \|\mathbf{b}\|_2^2\right) + 2\log(\gamma) r(\hat{\mathbf{w}}^{\text{MSTLS}})$
4. **if** $\Delta \geq 0$: $\hat{\mathbf{w}} \leftarrow \mathbf{0}$
   else: $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}}^{\text{MSTLS}}$
5. **return** $\hat{\mathbf{w}}$

It is important to note that the standard MSTLS algorithm does not natively allow for weight vectors with empty support (i.e., with $\text{supp}(\hat{\mathbf{w}}_i) = \varnothing$), and thus does not allow for trivial righthand-side models of the form

$$\dot{x}_i \approx \boldsymbol{\Theta}(\boldsymbol{x})\hat{\mathbf{w}}_i = 0.$$

To resolve this behaviour in the context of the scattering ODEs of eq. (14), where indeed $\dot{\kappa}_i = 0$, we develop a 'zero-admissible' extension of Algorithm A1 inspired by the model selection work of [16] and [17]. In particular, we use a *Bayesian information criterion* (BIC) of the form

$$\mathcal{B}(\hat{\mathbf{w}}_i; \mathbf{b}_i, \mathbf{G}) := \gamma^2 \log\left(\frac{\|\mathbf{b}_i - \mathbf{G}\hat{\mathbf{w}}_i\|_2^2}{\gamma^2}\right) + 2\log(\gamma) r(\hat{\mathbf{w}}_i), \tag{A.2}$$

where $r(\hat{\mathbf{w}}_i) := \text{rank}(\mathbf{G}|_{\text{supp}(\hat{\mathbf{w}}_i)})$ denotes the number of uniquely estimated parameters and $\gamma^2 > 0$ is a 'soft rank' heuristic for the number of statistically independent weak-form equations in eq. (19),

$$\gamma := \frac{\|\dot{\boldsymbol{\Phi}}\|_F^2}{\|\dot{\boldsymbol{\Phi}}\dot{\boldsymbol{\Phi}}^T\|_F} \approx \sqrt{K}\frac{\|\dot{\varphi}(\mathbf{t})\|_2^2}{\|\dot{\varphi}(\mathbf{t}) \star \dot{\varphi}(\mathbf{t})\|_2}, \quad \text{where} \quad \gamma^2 \sim \frac{K}{m}, \ \text{as} \ K \to \infty.$$

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press

For the zero model $\hat{\mathbf{w}}_i = \mathbf{0}$ (representing conserved quantities), the corresponding BIC value is given by

$$\mathcal{B}(\mathbf{0}; \mathbf{b}_i, \mathbf{G}) = \gamma^2 \log\left(\frac{\|\mathbf{b}_i\|_2^2}{\gamma^2}\right),$$

meaning that

$$\Delta_i := \mathcal{B}(\hat{\mathbf{w}}_i; \mathbf{b}_i, \mathbf{G}) - \mathcal{B}(\mathbf{0}; \mathbf{b}_i, \mathbf{G}) = \gamma^2 \log\left(\frac{\|\mathbf{b}_i - \mathbf{G}\hat{\mathbf{w}}_i\|_2^2}{\|\mathbf{b}_i\|_2^2}\right) + 2\log(\gamma) r(\hat{\mathbf{w}}_i).$$

In principle, for $\Delta_i \geq 0$, trivial dynamics are preferred from an information-theoretic point of view; conversely, the nonzero MSTLS estimate is preferred whenever $\Delta_i < 0$. With this in mind, our final model weights $\hat{\mathbf{w}}_i$ are determined using Algorithm A2 given above. Numerically, we guard against degenerate weak dynamics $\|\mathbf{b}_i\|_2 \approx 0$ by defaulting to the zero model whenever $\|\mathbf{b}_i\|_2 \leq \tau_{\mathbf{b}} := \texttt{1e-8}$. Otherwise, we trigger the BIC test of Algorithm A2 whenever the MSTLS fit is poor – in particular, when the relative residual $\|\mathbf{b}_i - \mathbf{G}\hat{\mathbf{w}}_i^{\text{MSTLS}}\|_2 / \|\mathbf{b}_i\|_2$ exceeds a default cutoff of 0.5.

Appendix B. Properties of Lax pairs

Two linear differential operators L and M, acting on test functions $\phi = \phi(x;t)$ and potentially depending on a one-parameter family of functions $u = u(x;t)$, are called a *Lax pair* if they satisfy *Lax's equation*,

$$L_t = [M, L], \quad \text{or equivalently,} \quad L_t + [L, M] = 0,$$

just as in eq. (2) above. In many cases of interest, L is self-adjoint while M is skew-adjoint, in which case the point spectrum of L is invariant in time (i.e., $L(u) \sim L(u_0)$ for every t). This can be explicitly shown by considering the EVP given by $L\phi = \lambda\phi$ with normalized eigenfunctions $\|\phi\|_2 = 1$, so that $\lambda = \langle \phi, L\phi \rangle$. Differentiating with respect to time then yields

$$\dot{\lambda} = \langle \phi, L_t \phi \rangle = \langle \phi, [M, L]\phi \rangle = \lambda \langle \phi, M\phi \rangle - \lambda \langle \phi, M\phi \rangle = 0,$$

where the second equality above follows from Lax's equation and the third uses $\langle \phi, ML\phi \rangle = \lambda \langle \phi, M\phi \rangle$ together with $\langle \phi, LM\phi \rangle = \lambda \langle \phi, M\phi \rangle$. Moreover, for any eigenpair (λ, ϕ) satisfying $L(u_0)\phi_0 = \lambda\phi_0$ at the initial time $t = 0$, the transported eigenfunction $\phi := P(t)\phi_0$ satisfies the linear system of eq. (3), where $P(t)$ is a unitary operator representing a solution to $\dot{P} = M(u)P$ with $P_0 = I$.

Cite this article: Minor S, Dukic V and Bortz DM (2026) Learning effective soliton dynamics from scattering data using weak-form system identification. *Journal of Nonlinear Waves*, 1–22. https://doi.org/10.1017/jnw.2026.10048

https://doi.org/10.1017/jnw.2026.10048 Published online by Cambridge University Press