# Learning physically consistent differential equation models from data using group sparsity

Suryanarayana Maddu,[1,2,3,4] Bevan L. Cheeseman,[1,2,3,*] Christian L. Müller,[5,6,7] and Ivo F. Sbalzarini[1,2,3,4,8,†]

[1]*Technische Universität Dresden, Faculty of Computer Science, 01069 Dresden, Germany*
[2]*Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany*
[3]*Center for Systems Biology Dresden, 01307 Dresden, Germany*
[4]*Center for Scalable Data Analytics and Artificial Intelligence ScaDS.AI, Dresden/Leipzig, Germany*
[5]*Center for Computational Mathematics, Flatiron Institute, New York, New York 10010, USA*
[6]*Department of Statistics, LMU München, 80539 Munich, Germany*
[7]*Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany*
[8]*Cluster of Excellence Physics of Life, TU Dresden, 01307 Dresden, Germany*

(Received 11 December 2020; revised 15 March 2021; accepted 22 March 2021; published 13 April 2021)

We propose a statistical learning framework based on group-sparse regression that can be used to (i) enforce conservation laws, (ii) ensure model equivalence, and (iii) guarantee symmetries when learning or inferring differential-equation models from data. Directly learning interpretable mathematical models from data has emerged as a valuable modeling approach. However, in areas such as biology, high noise levels, sensor-induced correlations, and strong intersystem variability can render data-driven models nonsensical or physically inconsistent without additional constraints on the model structure. Hence, it is important to leverage prior knowledge from physical principles to learn biologically plausible and physically consistent models rather than models that simply fit the data best. We present the group iterative hard thresholding algorithm and use stability selection to infer physically consistent models with minimal parameter tuning. We show several applications from systems biology that demonstrate the benefits of enforcing priors in data-driven modeling.

DOI: 10.1103/PhysRevE.103.042310

## I. INTRODUCTION

Mathematical modeling is fundamental to understanding and predicting natural phenomena. Usually, mathematical models are formulated from first principles, such as symmetries and conservation laws. This classic approach of modeling natural systems has been successful in many domains of science amenable to mathematical treatment. However, in domains like biology, the success of first-principles modeling is limited [1–4]. This is mostly attributed to the "complexity" of biological systems where nonlinearity, stochasticity, multiscale coupling, nonequilibrium behavior, and self-organization can dominate. Formulating mathematical models from first principles is difficult in complex systems, and the resulting models often have many unknown parameters.

Data-driven modeling has thus emerged as a complementary approach to first-principles modeling. Data-driven analysis and forecasting of complex systems were made possible by unprecedented advances in imaging and measurement technology, computing power, and algorithmic innovations. While purely data-driven models, like reservoir computing, can be successful in predicting future behavior [5], such "black-box" models are often difficult to interpret for domain scientists. This raises the question how *interpretable* mathematical models, such as ordinary differential equations (ODEs) or partial differential equations (PDEs), can be learned directly from data.

The idea of algorithmic inference of differential-equation models from data originated in the field of time-series analysis [6,7]. Early works used least-squares fitting to estimate PDE coefficients from spatiotemporal data [8,9]. Many different approaches have since been proposed, e.g., Bayesian networks [10], linear dynamic models [11], recurrent neural networks [12], symbolic regression [13,14], sparse regression [15,16], Gaussian processes [17], and deep learning [18]. Methods based on sparse regression have been particularly successful, owing to their simplicity, computational efficiency, and applicability in the data-scarce regime [19]. They have therefore found applications in many domains ranging from optics [20] to plasma physics [21], fluid mechanics [22], chemical physics [23], aerospace engineering [24], and biology [19,25]. The sparse-regression methodology has also been extended to incorporate control [26], implicit dynamics [25], parametric dependences [27], stochastic dynamics [28], discrepancy models [29], and multiscale physics [30]. Algorithms based on integral terms [31], automatic differentiation [32], and weak formulations [33] have increased regression robustness by avoiding high-order derivatives of noisy data. All of these developments have corroborated the feasibility of data-driven learning of interpretable mathematical models.

[*] Present address: ONI Inc., Oxford OX2 8TA, UK.
[†] sbalzarini@mpi-cbg.de

Given the feasibility of data-driven modeling and the historic success of first-principles modeling, it seems natural to try combine the two. This requires methods to incorporate or enforce first-principle constraints, like conservation laws and symmetries, into the data-driven inference problem. First attempts in this direction used block-diagonal dictionaries with group sparsity to avoid model discrepancy [29,34–36] and to infer PDEs with varying coefficients [27]. However, there are many more *priors* one may want to exploit when modeling complex systems, including information about symmetries in interactions, knowledge of conservation laws, dimensional similarities, or awareness of spatially and temporally varying latent variables. Such prior knowledge can come from first principles or from model assumptions or hypotheses. To date, there is no statistical inference framework available that would allow flexible inclusion of different types of priors as hard constraints in data-driven inference of differential equations models.

Here we present a statistical learning framework based on group sparsity to enforce a wide range of physical or modeling priors in the regression problem for robust inference of ODE and PDE models from modest amounts of noisy data. We present three representative examples from systems biology to demonstrate how information about conservation laws, latent variables, and symmetries can be encoded into grouped features of a sparse-regression formulation. We therefore present numerical experiments using a mass-conserving ODE model of Janus kinase–signal transducer and activator of transcription (JAK-STAT) signaling in cells, a mechanical transport model for membrane proteins, and $\lambda$-$\omega$ reaction-diffusion systems, respectively. We approximately solve the resulting nonconvex optimization problems using the group iterative hard thresholding (GIHT) algorithm presented here, in combination with stability selection for statistically consistent model identification [19]. We show that stability selection in combination with GIHT enables robust model inference from limited noisy data.

## II. PROBLEM FORMULATION

We aim to learn the functional form of a governing ODE or PDE from data about the corresponding dynamics. We consider the following canonical form, where the left-hand side is a first derivative in time and the right-hand side is a nonlinear function $\mathcal{N}$ of space $\boldsymbol{x}$, time $t$, and derivatives:

$$\frac{\partial u_i}{\partial t} = \mathcal{N}\left(\boldsymbol{x}, t, \Xi(\boldsymbol{x}, t), u_i, \frac{\partial u_i}{\partial x_j}, \frac{\partial^2 u_i}{\partial x_i \partial x_j}, \frac{\partial^2 u_i}{\partial x_j^2}, \dots\right). \quad (1)$$

The quantity $\boldsymbol{u} = (u_i)$ is the state variable of interest (e.g., velocity, concentration, or pressure) and $\Xi(\boldsymbol{x}, t)$ is the set of parameters of the equation, such as diffusion constants or viscosity. The dependence of $\Xi$ on $(\boldsymbol{x}, t)$ allows for equations with varying coefficients in both space and time. Without loss of generality, $\mathcal{N}(\cdot)$ can be written as a linear combination of potentially nonlinear terms. Common models like Navier-Stokes, advection, active mechanochemistry, and reaction-diffusion models are represented by this canonical form. Models requiring a different left-hand side (e.g., wave equations) can be expressed using suitably adjusted canonical forms.

The goal of equation inference [15,16] is to find a specific instance of this canonical differential equation from given data. Data are given as measured or simulated values $\hat{\boldsymbol{u}}(\boldsymbol{x}_i, t_j)$ at discrete locations $\boldsymbol{x}_i$ and time points $t_j$. These data points may contain noise, e.g., from measurement uncertainties or numerical errors. The question then is which right-hand side $\mathcal{N}$ makes Eq. (1) describe the dynamics from which the data are sampled, without describing the noise, in a way that is statistically consistent and stable under data perturbation or different realizations of the noise.

We follow the standard approach to equation inference, constructing an overcomplete *dictionary* of possible right-hand-side terms and approximating their values from the data using discrete approximations of the derivatives [15,16] (e.g., finite differences, polynomial differentiation [16,27], or automatic differentiation [32]). For example, for a model with a single scalar state variable $u \in \mathbb{R}$, a dictionary of $p \in \mathbb{N}$ potential terms numerically evaluated over $n \in \mathbb{N}$ data points is a matrix $\boldsymbol{\Theta} \in \mathbb{R}^{n \times p}$. The canonical form of Eq. (1) then becomes

$$\underbrace{\begin{bmatrix} \vdots \\ u_t \\ \vdots \end{bmatrix}}_{\boldsymbol{U}_t \,\in\, \mathbb{R}^{n\times 1}} = \underbrace{\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u & uu_x & \cdots & u^3u_{xx} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{\boldsymbol{\Theta} \,\in\, \mathbb{R}^{n\times p}} \underbrace{\boldsymbol{\xi}}_{\in \mathbb{R}^{p\times 1}}, \quad (2)$$

where subscripts denote derivatives with respect to the subscripted variable. By default, we include in $\boldsymbol{\Theta}$ all differential operators and polynomial nonlinearities up to and including order and degree 3. Each column of $\boldsymbol{\Theta}$ contains the discrete approximations of one such term at all $n$ data points. The vector $\boldsymbol{\xi}$ contains the unknown coefficients $[\xi_0 \quad \xi_1 \quad \xi_2 \quad \xi_3 \quad \cdots \quad \xi_p]^\top$ of the model.

For systems of differential equations, the dictionary $\boldsymbol{\Theta} \in \mathbb{R}^{N\times P}$ becomes block diagonal with $p_b$ blocks $\boldsymbol{\Theta}_b \in \mathbb{R}^{n\times p}$ [see, for example, Fig. 2(a)]. In this case, we distinguish the number $p$ of potential terms in each block and the number $P = p_b p$ of columns in the overall dictionary. Likewise, $N = p_b n$.

In either case, the problem is to find a statistically consistent $\boldsymbol{\xi}^*$ such that the model in Eq. (2) fits the data while being sparse, i.e., $\|\boldsymbol{\xi}^*\|_0 \ll p$. This trade-off between model simplicity and data fitting can be formulated as a regularized optimization problem

$$\hat{\boldsymbol{\xi}}^\lambda = \arg\min_{\boldsymbol{\xi}} [h(\boldsymbol{\xi}) + \lambda r(\boldsymbol{\xi})], \quad (3)$$

where $\hat{\boldsymbol{\xi}}^\lambda$ is the global minimizer, $h(\cdot)$ a smooth convex data-fitting metric (e.g., least-squares or Huber loss), and $r(\cdot)$ a regularization or penalty function with regularization constant $\lambda \in \mathbb{R}^+$ that controls the trade-off between model simplicity and fitting accuracy. The superscript $\lambda$ to the estimated coefficient vector $\hat{\boldsymbol{\xi}}$ indicates the dependence of the result on the regularization parameter. The data-fitting metric measures the distance (in some norm) between the model output for a given $\boldsymbol{\xi}$ and the data. The regularization function measures model complexity.

Here we use the following standard choice for the data-fitting and regularization functions:

$$\hat{\boldsymbol{\xi}}^\lambda = \arg\min_{\boldsymbol{\xi}} \left[\tfrac{1}{2}\|\boldsymbol{U}_t - \boldsymbol{\Theta}\boldsymbol{\xi}\|_2^2 + \lambda\|\boldsymbol{\xi}\|_0\right]. \quad (4)$$

By choosing $r(\boldsymbol{\xi}) = \lambda\|\boldsymbol{\xi}\|_0$, we directly penalize the number of terms on the right-hand side of the model, hence favoring simpler models (Occam's razor) that are easier to interpret. Such sparsity-promoting regularization has been successful in applications of compressive sensing and signal processing. For the data-fitting function, we choose the standard least-squares metric, hence $h(\boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{U}_t - \boldsymbol{\Theta}\boldsymbol{\xi}\|_2^2$, leading to models that fit the data in the least-squares sense.

## III. SOLUTION METHOD

Classic algorithms that efficiently compute locally optimal solutions to Eq. (4) include greedy optimization strategies [37], compressed sampling matching pursuit [38], subspace pursuit [39], and iterative hard thresholding (IHT) [40]. To avoid the problem of nonconvexity in the objective function, a popular approach is to consider the convex relaxation of the problem in Eq. (4) by replacing the $\|\cdot\|_0$ term with $r(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|_1$ [41]. However, while this formulation benefits from the availability of fast convex optimization algorithms, it does not provide good approximations when right-hand-side terms in Eq. (1) are correlated [42] and leads to biased estimates of model coefficients [43], thus reducing model selection performance [19]. Therefore, we directly consider the original nonconvex problem in Eq. (4) and provide an algorithm to approximately solve it while accounting for modeling priors and guaranteeing statistically stable and consistent models.

### A. Group-sparse regression

In addition to statistical stability, we require the learned models to be consistent with prior knowledge about the physics of the process that generated the data. Examples of such priors one may want to impose are conservation laws, symmetries, and knowledge about latent variables. While the sparsity constraint is imposed as a soft constraint, these physical priors will be imposed as hard constraints on the model. They amount to restrictions on the structure of the coefficient vector $\boldsymbol{\xi}$. We show here how the concept of group sparsity [42,44] can be used to impose modeling priors in a sparse-regression framework.

The concept of group sparsity assumes that prior knowledge about the underlying system can be encoded by partitioning model terms into $m$ groups. During the inference process, group sparsity then imposes that coefficients within the same group can only enter or leave the statistical model *jointly*. We additionally leverage the block-diagonal structure of the dictionary matrix to allow for spatially or temporally varying coefficients and for joint sparse regression of multiple state variables.

Formally, given a partitioning of the coefficients $\boldsymbol{\xi}_k$, $k = 1, 2, \ldots, P$, into $m$ groups $g_j$, $j = 1, 2, \ldots, m$, we thus consider the optimization problem

$$\hat{\boldsymbol{\xi}}^{\lambda} = \arg\min_{\boldsymbol{\xi}} \left[\frac{1}{2}\left\|\boldsymbol{U}_t - \sum_{j=1}^{m}\boldsymbol{\Theta}_{g_j}\boldsymbol{\xi}_{g_j}\right\|_2^2 + \lambda\sum_{j=1}^{m}\sqrt{p_{g_j}}\mathbb{1}(\|\boldsymbol{\xi}_{g_j}\|_2 \neq 0)\right], \quad (5)$$

where $\boldsymbol{\Theta}_{g_j} \in \mathbb{R}^{N\times p_{g_j}}$ is the submatrix of $\boldsymbol{\Theta} \in \mathbb{R}^{N\times P}$ formed by all columns corresponding to the coefficients in group $g_j \subseteq \{1, \ldots, P\}$ and $\boldsymbol{\xi}_{g_j} = \{\boldsymbol{\xi}_i : i \in g_j\}$ is the coefficient vector $\boldsymbol{\xi}$ restricted to the index set $g_j$ of size $p_{g_j}$, i.e., $|g_j| = p_{g_j}$. The indicator function $\mathbb{1}(\cdot)$ over the $\|\cdot\|_2$ norm encourages sparsity on the group level [42]. For groups comprising only a single element, this penalty reduces to the $\|\cdot\|_0$-norm. Here we restrict ourselves to *nonoverlapping* groups where $g_i \cap g_j = \emptyset\, \forall\, i \neq j = 1, \ldots, m$ and $\sum_{j=1}^{m} p_{g_j} = P$. Extensions to overlapping groups are possible [45] and discussed in Sec. V. We solve the nonconvex problem in Eq. (5) using the GIHT algorithm, which generalizes the standard IHT algorithm as detailed in Appendix C.

### B. Stability selection

Robust tuning of the regularization parameter $\lambda$ is of fundamental importance for successful model discovery. Wrong choices of $\lambda$ result in incorrect equation models being identified, even if correct model discovery would have been possible in principle given the data [25,46]. Common methods for tuning $\lambda$ include the Akaike information criterion (AIC) [47], the (modified) Bayesian information criterion (BIC) [48], and cross validation. While AIC or BIC model selection is useful for combinatorial best-subset selection methods in low dimensions, they typically deteriorate in high dimensions since they rely on asymptotic considerations. Cross validation tends to include many false-positive coefficients in the data-limited regime [49].

In order to provide a robust model selection method for the data-limited high-dimensional case, we leverage here the statistical principle of stability selection, which tunes $\lambda$ so as to guarantee model stability under perturbation random subsampling of the data [50]. We perform stability selection by generating $B$ random subsamples $I_b^*$, $b = 1, \ldots, B$, of the data, using the GIHT algorithm to find the set $\hat{S}^{\lambda}[I_b^*] \subseteq \{1, \ldots, P\}$ of coefficients (or groups) for every data subsample $I_b^*$ for different values of $\lambda$ ranging over a *regularization path* $\Lambda = [\lambda_{\max}, \lambda_{\min}]$ with $\lambda_{\min} = \epsilon\lambda_{\max}$ and $\lambda_{\max} = \max_{j\in\{1,\ldots,m\}} \frac{1}{2}\|\boldsymbol{\Theta}_{g_j}^{\top}\boldsymbol{U}_t\|_2^2$ as computed for the group least absolute shrinkage and selection operator (LASSO) [27]. Typical values of $\epsilon$ range from 0.1 to 0.01 with the path discretized evenly on a logarithmic scale. The probability that group $j$ overlaps with the coefficients selected for a given $\lambda$ is approximately [50]

$$\hat{\Pi}_{g_j}^{\lambda} = \mathbb{P}[g_j \cap \hat{S}^{\lambda}[I_b^*] \neq \emptyset] \quad (6a)$$

$$\approx \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}(g_j \cap \hat{S}^{\lambda}[I_b^*] \neq \emptyset), \quad g_j \subseteq \{1, \ldots, P\}. \quad (6b)$$

This is the importance measure [50] for group $j$. Plotting this importance measure as a function of $\lambda \in \Lambda$ provides an interpretable way to assess the robustness of the estimation across levels of regularization in a so-called stability plot [50].

To select a final model, stability selection chooses the set of stable coefficients (or groups) $\hat{S}_{\text{stable}} = \{j : \hat{\Pi}_{g_j}^{\lambda_s} > \pi_{\text{th}}\}$. This means that we search for the components (groups) in the dictionary that consistently appear with probability greater than $\pi_{\text{th}}$ when repeatedly solving the sparse-regression problem in

Eq. (5) for different random subsets of the data. The threshold probability $\pi_{\text{th}}$ controls the type I error of false positives according to [51]

$$\pi_{\text{th}} = \frac{1}{2} + \frac{\binom{q}{p_g}^2}{2\binom{P}{p_g}E_{\text{fp}}}, \tag{7}$$

where $E_{\text{fp}}$ is the upper bound on the expected number of false positives and $q$ is the average number of selected variables (i.e., nonzero components of $\boldsymbol{\xi}$) along the regularization path [50]. The group size $p_g = p_{g_j}$, $j = 1, \dots, m$, if all groups are of equal size; otherwise we set $p_g = 1$ [51]. For a fixed value of $\pi_{\text{th}}$, we use this relation to find a $\lambda_s$ for which a given bound on the expected number of false positives, $E_{\text{fp}}$, is achieved. Equation (7) therefore provides an elegant way of determining the regularization constant based on the importance measure and the intuitively defined parameters $E_{\text{fp}}$ and $\pi_{\text{th}}$. Throughout this work, we set $\pi_{\text{th}} = 0.8$ and $E_{\text{fp}} = 1$. For the examples shown in this paper, we empirically find that regularization paths with $\epsilon = 0.1$ are sufficient to find a solution to Eq. (7) for these parameters. Alternatively, one can determine $\pi_{\text{th}}$ and $\epsilon$ by visual inspection of a stability plot, which usually shows a clear separation between two clusters of coefficients of different stability.

Stability selection not only removes the necessity to manually tune $\lambda$, but also ensures robustness against data sampling and noise in the data. All of these properties are required for statistical consistency in the sense that the inferred models are guaranteed to become accurate with high probability for increasing data size [10].

## IV. APPLICATIONS

We present three different modeling examples from systems biology that illustrate the utility of priors in data-driven modeling. Each example highlights a different type of prior knowledge to be enforced. In order to benchmark the accuracy and robustness of model inference, we need to know the ground-truth model. We therefore generate synthetic data by numerically solving known models and see how well we can recover those models again purely from the data. To emulate noisy measurements from real-world experiments, we corrupt the simulated data $\boldsymbol{u}(\boldsymbol{x}, t)$ with additive Gaussian noise $\hat{\boldsymbol{u}} = \boldsymbol{u} + \sigma\boldsymbol{\eta}(0, \theta)$, where $\boldsymbol{\eta}$ is a vector of elementwise independent and identically distributed Gaussian random numbers with mean zero and empirical variance $\theta = \text{Var}\{u_1, \dots, u_N\}$ of the simulated data. The constant $\sigma$ defines the noise level. In line with previous works, we use polynomial differentiation [16,27] to approximate the spatial and temporal derivatives in the dictionary from the noise input data $\hat{\boldsymbol{u}}$.

### A. Enforcing mass conservation in the JAK-STAT reaction pathway for signal transduction

Signal transduction pathways are the engines of chemical information processing in living biological cells. Using methods from biochemistry and systems biology, the constituent molecules of many signaling pathways have been identified. However, identifying the topology of these chemical reaction networks remains challenging. It typically involves building mathematical models of hypothetical reaction networks and comparing their predictions with the data. A popular choice is to use ordinary differential equation models of the stoichiometry and chemical kinetics of the pathway. However, when discrepancies occur between the ODE model and the experimental data, it is difficult to decide whether the model structure is incorrect or whether the parameters of the model have been badly chosen [52]. Here data-driven modeling can help identify the stable structure of minimal ODE models that can explain the measurement data.

In this example, we consider the JAK-STAT pathway, which communicates chemical signals from outside a biological cell to the cell nucleus. It is implicated in a variety of biological processes from immunity to cell division, cell death, and tumor formation. Mathematical models based on biochemical knowledge of the JAK-STAT pathway have identified nucleocytoplasmic cycling as an essential component of the JAK-STAT mechanism, which has been experimentally verified [52,53]. We therefore consider the simplest ODE model with irreversible reactions that account for nucleocytoplasmic cycling in order to model information transfer from the cell membrane to the nucleus as previously described [52]:

$$\dot{x}_1(t) = -k_1^- x_1(t)c(t) + 2k_4^+ x_4(t), \tag{8a}$$

$$\dot{x}_2(t) = +k_1^+ x_1(t)c(t) - k_2^- x_2^2(t), \tag{8b}$$

$$\dot{x}_3(t) = -k_3^- x_3(t) + \tfrac{1}{2}k_2^+ x_2^2(t), \tag{8c}$$

$$\dot{x}_4(t) = +k_3^+ x_3(t) - k_4^- x_4(t). \tag{8d}$$

A schematic of the JAK-STAT pathway is shown in Fig. 1, illustrating the reaction cascade from outside the cell membrane to inside the cell nucleus. The functions $x_1(t)$, $x_2(t)$, $x_3(t)$, and $x_4(t)$ in the above ODE model are the time courses of the concentrations of monomeric STAT-5, phosphorylated STAT-5, cytoplasmic dimeric STAT-5, and STAT-5 in the nucleus, respectively. The scalar constants $k_1^{\pm}$, $k_2^{\pm}$, $k_3^{\pm}$, and $k_4^{\pm}$ are the kinetic reaction rates of phosphorylation, dimerization, nuclear transport, and nuclear export, respectively. While of course $k_1^- = k_1^+$, $k_2^- = k_2^+$, $k_3^- = k_3^+$, and $k_4^- = k_4^+$, we distinguish different occurrences of the same rate constant by sign superscripts in order to make clear that they are independently learned from data by our regression algorithm.

For sparse-regression model learning, a dictionary matrix $\boldsymbol{\Theta}_b$ of all possible interactions between the molecules is generated [see Eq. (2)]. The left-hand side $\boldsymbol{U}_t$ is the time derivative of each concentration, i.e., $\dot{x}_1$, $\dot{x}_2$, $\dot{x}_3$, and $\dot{x}_4$ as approximated from the data. For this example, $\boldsymbol{\Theta}_b$ contains $p = 19$ polynomial nonlinearities (e.g., $x_1, x_2, x_1^2, x_1x_2, x_1x_2x_3, \dots$), corresponding to chemical kinetics of different orders. The same $\boldsymbol{\Theta}_i = \boldsymbol{\Theta}_b$ is used for each component $x_i$, $i = 1, 2, 3, 4$, leading to the block-diagonal overall dictionary structure with $p_b = 4$ [shown in Fig. 2(a)]. For model inference, we use the simulated concentration time courses shown in Fig. 2(b). They are obtained by numerically solving the model (8) with $k_1^- = k_1^+ = 0.021$, $k_2^- = k_2^+ = 2.46$, $k_3^- = k_3^+ = 0.2066$, and $k_4^- = k_4^+ = 0.106\,58$, as found by fitting experimental data [52,53] (see Fig. 1 inset). The simulated data are corrupted with 10% additive Gaussian noise before inference. The noisy time-series data for the concentration of the activated EpoR receptor $c(t)$ are taken directly from experimental measurements

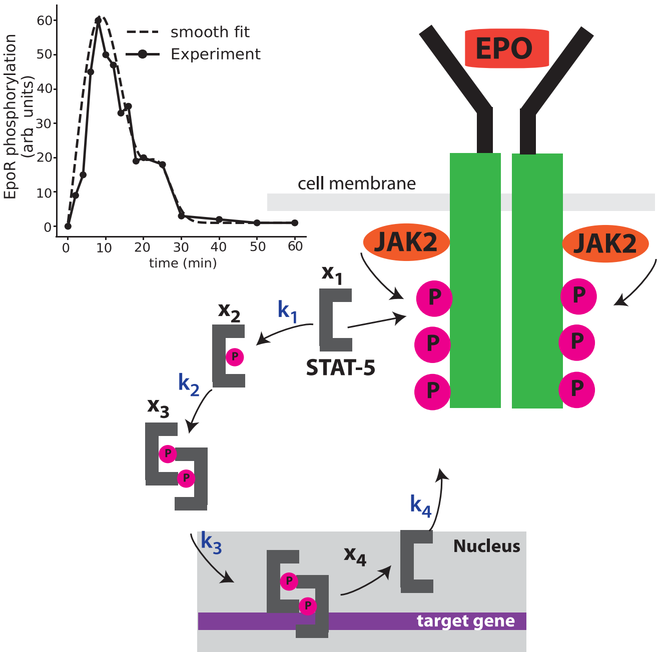

FIG. 1. Core module of the JAK-STAT signaling pathway. The hormone EPO binding to the EpoR receptor results in activation of the receptor [activated form with concentration $c(t)$] by transphosphorylation of JAK2 and subsequent tyrosine phosphorylation (P) of JAK2 and the EpoR cytoplasmic domain. Phosphotyrosine residues 343 and 401 in EpoR mediate recruitment of monomeric STAT-5 (concentration $x_1$). Upon receptor recruitment, monomeric STAT-5 is tyrosine phosphorylated ($x_2$), dimerizes ($x_3$), and translocates to the nucleus ($x_4$), where it binds to the promoters of target genes, is dephosphorylated, and is exported again to the cytoplasm [52]. The inset plot shows an experimentally measured time course of EpoR activation (data from [53]).

[53], both when generating the simulation data and for model inference. All units are relative to the experimental data.

From the simulated data for $x_i(t)$, $i = 1, 2, 3, 4$, and the experimentally measured $c(t)$, we aim to infer back the model equations. The JAK-STAT pathway conserves mass, as evident from the ODE model (8). This can be used as a prior when inferring a model from data. We therefore perform group-sparse regression (see Sec. III A) using the groups

$$g_1 = \{i : \text{column index of } x_1 \text{ in } \mathbf{\Theta}_1, \mathbf{\Theta}_2\}, \tag{9a}$$

$$g_2 = \{i : \text{column index of } x_2^2 \text{ in } \mathbf{\Theta}_2, \mathbf{\Theta}_3\}, \tag{9b}$$

$$g_3 = \{i : \text{column index of } x_3 \text{ in } \mathbf{\Theta}_3, \mathbf{\Theta}_4\}, \tag{9c}$$

$$g_4 = \{i : \text{column index of } x_4 \text{ in } \mathbf{\Theta}_1, \mathbf{\Theta}_4\}. \tag{9d}$$

This is graphically represented by the vertical lines in Fig. 2(a), with each group corresponding to one type of biochemical process in the model, as given in the legend ($g_1$, phosphorylation; $g_2$, dimerization; $g_3$, nuclear transport; and $g_4$, nuclear export). We solve the resulting group-sparse-regression problem using the present GIHT algorithm. This leads to a conservative model *structure*, but the fitted *values* of the rate constants may differ for different signs, i.e., it can be that $k_1^- \neq k_1^+$, etc. Enforcing symmetry also in the coefficient values, and not only in the model structure, would require solving a *constrained* group-sparse-regression problem, which we do not consider here.

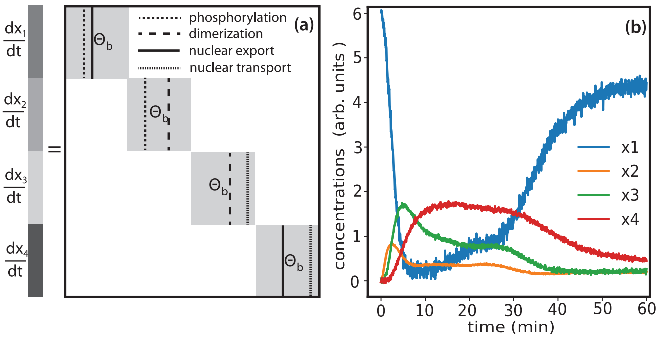

FIG. 2. Dictionary design and simulated data. (a) Dictionary construction and coefficient grouping. The identical dictionaries $\mathbf{\Theta}_i = \mathbf{\Theta}_b$ for each $x_i$, $i = 1, 2, 3, 4$, are stacked in a block-diagonal matrix for joint learning. Vertical lines indicate the coefficient groups $g_1, \dots, g_4$, corresponding to the four biochemical processes named in the legend. (b) Time-series data for different concentrations in the JAK-STAT pathway obtained by numerically integrating the deterministic ODE model in Eqs. (8) using the ode45 MATLAB solver and adding 10% Gaussian noise ($\sigma = 0.1$).

The results are shown in Fig. 3(a). In this benchmark setting, group sparsity helps identify the correct model terms (red curves) out of all terms in the dictionary. There exists a range of $\lambda$ values (shaded gray) where stability selection with threshold $\pi_{\text{th}} = 0.8$ (green dashed line) can identify the correct model, even at the 10% noise level considered here.

Without coefficient grouping, i.e., without imposing the mass-conservation prior, there is no value of $\lambda$ for which the correct model is recovered, as shown in Fig. 3(b). To show consistency of the group-sparsity method, we also provide achievability plots in Figs. 3(c) and 3(d). They show that enforcing the mass conservation prior leads to consistent model selection over a wide range of data sample sizes $N$.

Using group sparsity in combination with stability selection, the correct model can be identified in 100% of cases (over 20 independent repetitions) when more than 200 data points per component are used (i.e., success probability 1), regardless of the noise level in the data (color, see the legend), as shown in Fig. 3(c). Sparse regression without priors suffers from inconsistency, at all noise levels and for all data sizes [Fig. 3(d)]. The learned coefficients at different noise levels are shown in Fig. 8 in Appendix A.

### B. Enforcing model equivalence in advection diffusion with spatially varying velocity

The development of organisms from their zygotic state involves a myriad of biochemical interactions coupled with the mechanical forces that shape the resulting tissue. In past decades, the role of mechanics, including forces and flows, has increasingly been investigated in developmental biology and morphogenesis. On the cell and tissue scale, many developmental processes involve both patterning and flows. Examples include polarity establishment [54], tissue

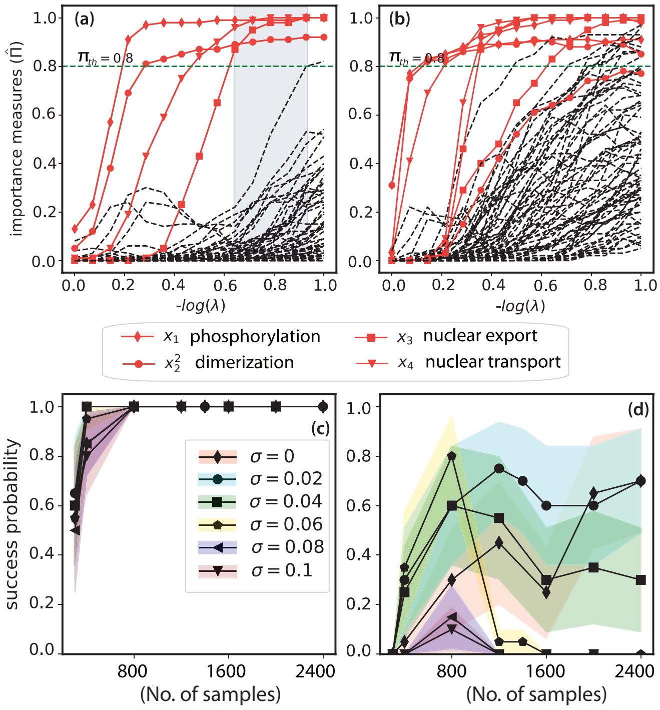

FIG. 3. Inferring JAK-STAT signaling models from noisy data. (a) Stability plot using grouping based on mass conservation. In the gray shaded range of λ values, stability selection with $\pi_{\text{th}} = 0.8$ identifies the correct model. The red solid lines show the behavior of the true components of the ODE model; the black dashed lines are all other $P = 76$ dictionary terms. (b) Stability plot without grouping. There is no value of λ for which the true model is found. In both (a) and (b) Gaussian noise with $\sigma = 0.1$ is added to the simulated data before inference and $n = 200$ time points for each component $x_i$, $i = 1, 2, 3, 4$, are used. (c) Achievability plot for model selection with mass conservation prior. (d) Achievability plot for model selection without mass conservation prior. In (c) and (d) the success probabilities of inferring the correct model over 20 independent trials with different noise realizations and different random data subsampling are shown as a function of the number of data points used in each trial. Colored bands are Bernoulli standard deviations for different amounts of noise added to the simulated data prior to inference (see the legend).

folding [55], and cell sorting [56,57]. The spatiotemporal concentration fields of labeled proteins can be recorded in all of these processes using fluorescence microscopy [54,58]. This has led to quantitative measurements and predictive models of active mechanochemical self-organization in, e.g., cytoplasmic flow [59], endocytosis [60], and tissue patterning [61].

In this example, we consider the simplest case of transport by advection and diffusion of signaling molecules. In order to allow for latent processes, we consider spatially varying model coefficients. We thus construct groups that allow the advection velocity to be a function of space. In addition, we impose a prior that enforces model equivalence, i.e., learning structurally equivalent models for the different chemical species, albeit with different diffusion constants. For the concentration fields $u(x,t)$ and $v(x,t)$ of two chemicals, this

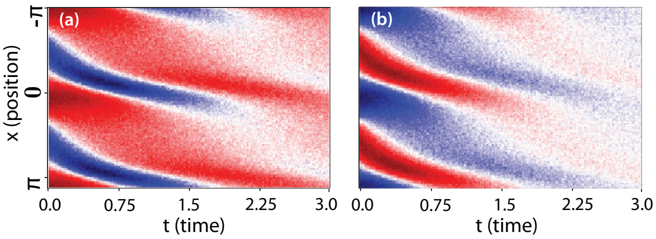

FIG. 4. Simulated data used to learn spatiotemporal models of one-dimensional advection-diffusion dynamics. Visualization of the data is shown for (a) $u(x,t)$ and (b) $v(x,t)$ with 15% additive Gaussian noise ($\sigma = 0.15$). Spatial and temporal discretization use 256 and 200 regularly spaced grid points, respectively. The solution is obtained via spectral differentiation and fourth-order Runge-Kutta time integration. The diffusion constants of the species are $D_u = 0.25$ and $D_v = 0.50$ in nondimensional units. The equations are solved with periodic boundary conditions in the domain $x \in [-\pi, \pi)$ of length $L = 2\pi$ over the time horizon $t \in [0, 3]$ with initial conditions $u(x, t = 0) = \cos(\frac{2\pi x}{L})$ and $v(x, t = 0) = -\cos(\frac{2\pi x}{L})$ for species $u$ and $v$, respectively.

amounts to the model

$$\frac{\partial u}{\partial t} + c(x)\frac{\partial u}{\partial x} + u\frac{\partial c(x)}{\partial x} = D_u\frac{\partial^2 u}{\partial x^2}, \tag{10a}$$

$$\frac{\partial v}{\partial t} + c(x)\frac{\partial v}{\partial x} + v\frac{\partial c(x)}{\partial x} = D_v\frac{\partial^2 v}{\partial x^2}. \tag{10b}$$

Here $D_u = 0.25$ and $D_v = 0.50$ are the ground-truth diffusion constants and the function $c(x) = -\frac{3}{2} + \cos(\frac{2\pi x}{L})$ is the spatially varying advection velocity field in the domain of length $L = 2\pi$. With added chemical reactions, this form of model has previously been successfully used to explain early patterning in the single-cell *C. elegans* zygote [54,58].

We use data from numerical simulations of the above model equations with 15% additive Gaussian noise (see Fig. 4) to show that both priors, model equivalence and spatial variability, are necessary to recover the ground-truth equations including the spatially varying velocity field from the data. We construct two block-diagonal dictionaries, for $u$ and $v$, where each block represents the dictionary constructed at one spatial location. We use $p_b = 10$ blocks, corresponding to five randomly selected spatial data points for each $u$ and $v$. Each of the diagonal blocks $\Theta_b(n, p)$ uses $n = 75$ randomly chosen time points and $p = 15$ potential operators.

We use grouping to enforce that the structure of the model learned from the data must be the same for all spatial locations and that the models learned for $u$ and $v$ must be equivalent. Each group therefore ties a column in a block dictionary to all corresponding columns in the other blocks. This construction results in the following groupings to encode spatial variability:

$$g_l = \{\{l + kp\}\, \forall\, k \in \{0, \ldots, p_b - 1\}\}. \tag{11}$$

Here the set $g_l$ is the group $l$ and $p$ is the number of columns each dictionary block. The groups $g_l^u$ and $g_l^v$, independently constructed for species $u$ and $v$ using Eq. (11), are further grouped to enforce model equivalence between species with the grouping $g_l = g_l^u \cup g_l^v$.

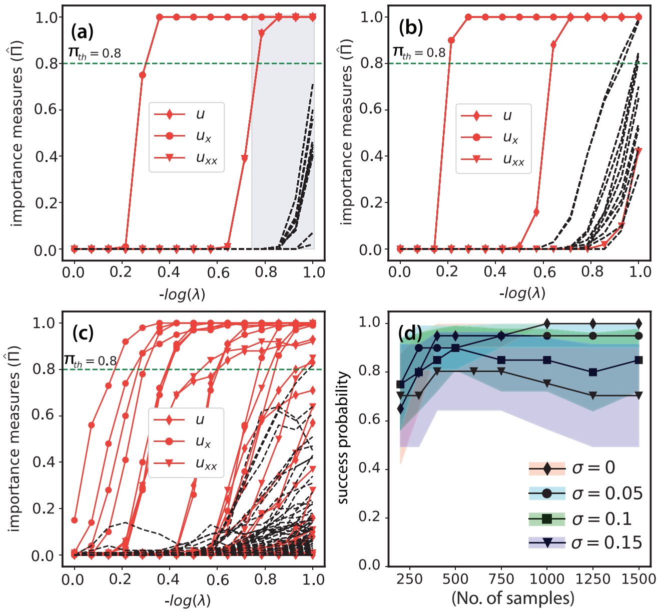

FIG. 5. Inferring advection-diffusion dynamics with unknown spatially varying velocity field. (a) Stability plot with groups to encode both spatially varying coefficients and model equivalence between the species. The gray shaded region is the range of $\lambda$ for which model selection with $\pi_{\rm th} = 0.8$ identifies the correct model. (b) Stability plot with groups only to encode spatially varying coefficients, but no grouping for model equivalence. (c) Stability plot with no groupings at all. In (a)–(c) the red solid lines correspond to the true components of the PDE for the field $u$, with symbols referring to the differential operators as given in the legends. The dictionary block size is $n = 75$ and $p = 15$ with $p_b = 10$ blocks and 15% Gaussian noise ($\sigma = 0.15$) added to the simulated data. (d) Achievability plot for model selection using both priors for different levels of noise in the data. Each point is averaged over 20 independent trials. The colored bands correspond to the Bernoulli standard deviation.

The resulting stability and achievability plots are shown in Fig. 5 when using the noisy data from Fig. 4 for inference. Comparing Figs. 5(a) and 5(b), we see that the prior for model equivalence is necessary to recover the true model. The algorithm is unable to identify the diffusion process of species $u$ when only using the grouping for the spatially varying coefficient [Fig. 5(b)]. Inference without any priors fails to recover the true model even for noise-free data [Fig. 5(c)]. The achievability plot in Fig. 5(d) demonstrates the consistency of our model selection algorithm with grouping over 20 independent realizations of the noise process and of the random subsampling of the data. We observe consistent model recovery with high success probability even at high noise levels, albeit with decreasing fidelity as seen in Fig. 5(d). In contrast, previous studies on advection-diffusion model recovery with unknown velocity field were limited to 1% noise ($\sigma = 0.01$) [27].

The estimated latent velocity fields and their gradients are shown in Fig. 9 and compared with ground truth for different noise levels. In Appendix B we show how these estimates can be further improved by postprocessing with additional smoothness priors.

### C. Enforcing symmetry in reaction-diffusion kinetics

Reaction-diffusion models are widely used in systems biology to describe the dynamics of chemical reaction networks in a continuous space. Their popularity goes back to a seminal paper by Turing [62], proposing that reaction-diffusion mechanisms could be responsible for pattern formation in developing tissues. Since then, reaction-diffusion equations have been successful in modeling nonequilibrium pattern formation [63], dynamics of ecological [64] and biological systems [65], cell polarity [54,63], phase transitions [66], and chemical waves [67].

In this example, we consider the $\lambda$-$\omega$ reaction-diffusion system as a prototypical model of chemical waves [68], showing how it can be inferred from data when including symmetry priors. The model equations for the scalar concentration fields $u(x, y, t)$ and $v(x, y, t)$ of two chemical species in two dimensions are

$$\frac{\partial u}{\partial t} = D_u\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \lambda(r)u - \omega(r)v, \qquad (12a)$$

$$\frac{\partial v}{\partial t} = D_v\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) + \omega(r)u + \lambda(r)v. \qquad (12b)$$

Here we choose $r = \sqrt{u^2 + v^2}$, $\omega(r) = -r^2$, $\lambda(r) = 1 - r^2$, and $D_u = D_v = 0.1$. This system is symmetric in the two species, i.e., swapping $u \leftrightarrow \pm v$ leaves the structure of the model unchanged. Such symmetries are common in biology and can be found in predator-prey models [69], models of fish scale patterning [70], and models of antagonistic protein interactions [54].

If known beforehand, such symmetries can be used as priors. Here we impose the symmetry prior by grouping each column of the dictionary of one species with the corresponding column for the other species, where "corresponding" means pertaining to the same operator upon the swap, i.e., $uv^2 \leftrightarrow u^2v$, $u^2 \leftrightarrow v^2$, $u_{xx} \leftrightarrow v_{xx}$, etc. The dictionary blocks

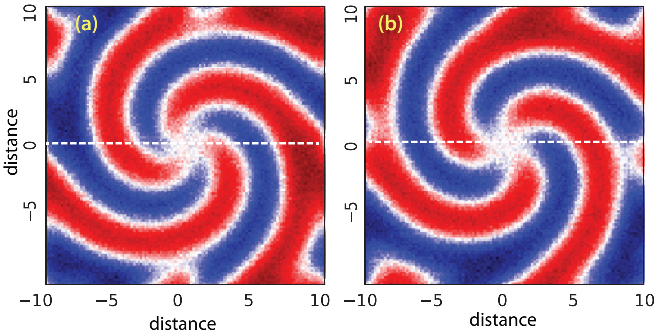

FIG. 6. Simulated data used to learn reaction-diffusion dynamics. Visualization of the two-dimensional concentration fields (a) $u(x, y)$ and (b) $v(x, y)$ is shown at time $t = 7.5$ from numerical solution of the model in Eqs. (12) with $D_u = D_v = 0.1$ and 10% additive Gaussian noise ($\sigma = 0.1$). The solution is obtained via spectral differentiation in the domain $(x, y) \in [-10, 10]^2$ and fourth-order Runge-Kutta time integration with time step size 0.05 on a Cartesian grid of 128 × 128 points with initial conditions $u(x, y, 0) = \tanh\{\sqrt{x^2 + y^2}\cos[3\angle(x + iy) - \sqrt{x^2 + y^2}]\}$ and $v(x, y, 0) = \tanh\{\sqrt{x^2 + y^2}\sin[3\angle(x + iy) - \sqrt{x^2 + y^2}]\}$.

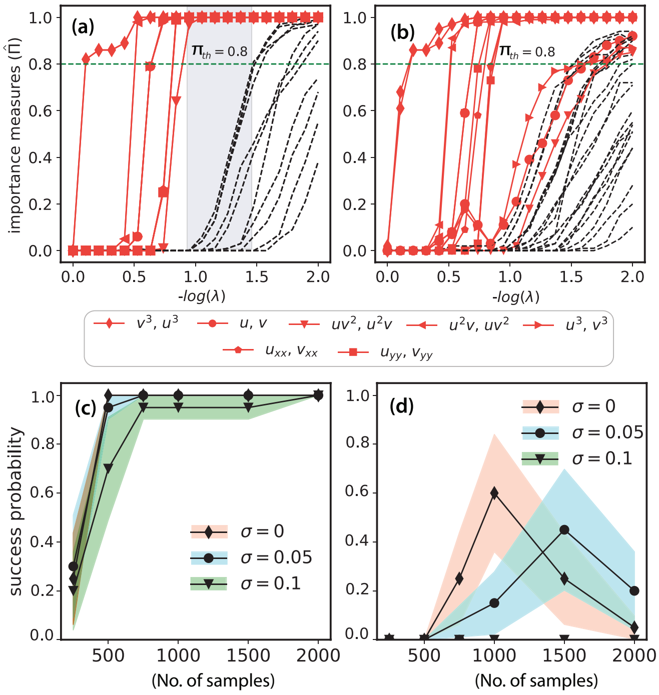

FIG. 7. Inferring reaction-diffusion models from noisy spatiotemporal data. Stability plots (a) with and (b) without symmetry priors for noise level $\sigma = 0.1$. Achievability plots (c) with and (d) without symmetry priors for different noise levels in the data (legends). We show the stability plots for $\lambda_{\min} = 0.01\lambda_{\max}$ to illustrate the difference between (a) and (b), but still run our algorithm with $\lambda_{\min} = 0.1\lambda_{\max}$. Each point is averaged over 20 independent trials. The colored bands correspond to the Bernoulli standard deviation.

for each of the $p_b = 2$ species contain all polynomial nonlinearities up to degree 3 and all spatial derivatives up to second order, resulting in $p = 18$.

We use data obtained by numerically simulating the above model with 10% pointwise Gaussian noise added to the data (see Fig. 6). The stability and achievability plots when using these data are shown in Fig. 7. Comparing Figs. 7(a) and 7(b), we observe that model inference without the symmetry prior fails, whereas it works robustly when the prior is included via group sparsity. This fact is substantiated by the achievability plots in Figs. 7(c) and 7(d) for model inference with and without the prior, respectively, for different noise levels $\sigma$ in the data. Our group-sparse-regression formulation provides remarkable consistency for model recovery over a wide range of $\lambda$ values even at high noise levels of 10%.

### D. Computational cost

Given the block diagonal dictionary matrix $\mathbf{\Theta} \in \mathbb{R}^{N\times P}$ and the vector $U_t \in \mathbb{R}^{N\times 1}$ from data, the computational complexity of Algorithm 1 is $O(NP)$ in each GIHT iteration without the debiasing step. This is the same complexity as matrix-vector multiplication. We also include in our algorithm a debiasing step, which also has a complexity of $O(NP)$. However, debiasing has been reported to lead to faster convergence amortizing its additional computational cost [71]. As an example, the advection-diffusion problem considered here with $n = 75$, $p = 15$, and $p_b = 10$ required less than 1.5 s of runtime to compute a regularization path with 15 different values of $\lambda$ for one data subsample when implemented in PYTHON on a single 2.3-GHz x86_64 processor core. The total time for stability selection over 100 data subsamples was under 150 s. This can be further accelerated using multithreaded programming, since the different data subsamples can be evaluated independently in parallel.

## V. CONCLUSION AND DISCUSSION

We have introduced a flexible and robust inference framework to learn physically consistent differential-equation models from modest amounts of noisy data. We used the concept of group sparsity to provide a flexible way of including modeling priors as hard constraints that render inference more robust. We combined this with the concept of stability selection for principled deduction of regularization parameters in cases where the true model is not known. To approximately solve the resulting nonconvex regression problem, we introduce the group iterative hard thresholding algorithm in Appendix C.

We have benchmarked and demonstrated the use of this algorithm in examples of common mathematical models in biological physics. The examples covered ordinary differential equations and partial differential equations in one- and two-dimensional domains. They demonstrated how different types of priors can be imposed using the concept of group sparsity: Conservation laws, model equivalence, spatially varying latent variables, and symmetries. The results have shown that including such priors enables correct model inference from data containing 10% or even 15% additive Gaussian noise. Without the priors, the correct model could not be recovered in any of the presented cases. The achievability plots furthermore confirmed that relatively little data (here a few hundred space-time points) is sufficient to reliably and reproducibly learn the correct model when group-sparsity priors are included. Without the priors, model inference was inconsistent in all cases.

Importantly, stability selection converts the problem of tuning the regularization parameter $\lambda$ to the easier problem of thresholding the importance measure $\hat{\Pi}$. We argue that this is easier to do, as it relates to an upper bound on the number of false positives one is willing to tolerate [50,51], providing interpretability. Further refining such results in the group-sparse case would be useful for applications that require reliability guarantees.

The concepts introduced here are independent of how the elements of the dictionary are constructed. Exploring more advanced dictionary constructions, such as integral formulations [31] or weak formulations [33], in conjunction with group sparsity and stability selection likely provides a promising future research direction.

In its current form, however, our framework has a number of limitations. First, we only considered nonoverlapping groups, restricting each column of the dictionary to be part of at most one group. This is a limiting assumption, as it is not uncommon in physics or biology to simultaneously use multiple overlapping priors. The more advanced concept of structured sparsity [72] could provide a way to include overlapping priors in future work. Second, we only showed how to include priors about the structure of a model. If additionally one wants to impose priors about coefficient values (e.g., values of diffusion constants and reaction rates), the framework would need to be extended to constrained group-sparse regression [73]. Third, although we have demonstrated robust data-driven inference of the model structure, estimates for the coefficient values can considerable deviate from the ground truth (see Appendix A). Fourth, although we have demonstrated robust data-driven inference of the model structure, estimates for the coefficient values can considerable deviate from the ground truth (see Appendix A). Refitting the coefficient estimates using additional smoothness regularization in a postprocessing step could help, as we hint at in Appendix B.

Especially at high noise levels, coefficient estimation errors likely stem from inaccurate spatial derivative approximations, since the polynomial differentiation schemes used here are known to amplify noise. This issue can possibly be addressed in the future by combining our framework with physics-informed neural networks [18] or with Gaussian processes [17] to more robustly estimate the coefficients of the recovered model once the model structure is fixed. Such hybrid methods, combining the reconstruction abilities of physics-constrained neural networks with the robustness and consistency of sparse inference methods, may be particularly powerful for recovering spatiotemporal latent variables, such as pressure or stresses in continuum mechanics models, that cannot be directly measured in experiments.

## ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) under Germany's Excellence Strategy, Grant No. EXC-2068-390729961, Cluster of Excellence "Physics of Life" of TU Dresden, and by the Center for Scalable Data Analytics and Artificial Intelligence ScaDS.AI Dresden/Leipzig, funded by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung).

## APPENDIX A: REGRESSION ESTIMATES OF THE COEFFICIENTS

The coefficients estimated by the GIHT algorithm from the noisy simulation data in the three application cases are shown in Figs. 8 (for the JAK-STAT example), 9 (for the advection velocity), and 10 (for the reaction-diffusion system). In all cases, the results are compared with ground-truth values for different noise levels.

## APPENDIX B: USING SMOOTHNESS PRIORS TO IMPROVE COEFFICIENT REGRESSION

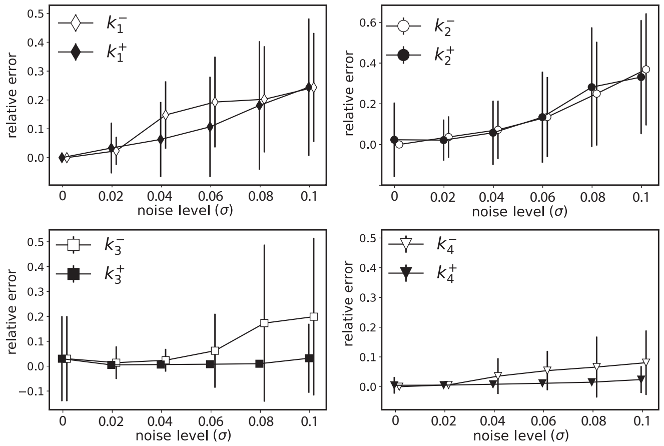

FIG. 8. Relative errors in the coefficients inferred for the JAK-STAT pathway reactions. The plots show the relative errors $\frac{|\xi-\xi^*|}{\xi^*}$ (vs ground truth $\xi^*$) in the reaction rate estimates of the JAK-STAT pathway as inferred by the GIHT algorithm for different noise levels $\sigma$. Symbols show estimated means with bars indicating estimation standard deviations over 20 independent trials. The ground-truth values are $k_1^{\pm} = 0.021, k_2^{\pm} = 2.46, k_3^{\pm} = 0.2066$, and $k_4^{\pm} = 0.106\,58$. The closed and open symbols correspond to the independently estimated rate constants of different signs, which should be identical.

In the results presented in Appendix A, the estimated coefficients are the direct outcome of the GIHT algorithm, which jointly infers the equation structure and the values of the coefficients over the so-determined support. It is possible to further improve the estimation of the coefficient values by imposing an additional smoothness prior on the values of the coefficients. However, this additional prior introduces an additional regularization parameter, which also needs to be determined using stability selection, introducing an additional dimension into the stability plots. Moreover, smoothness priors based on discrete total variation and trend filtering based on discrete higher-order derivatives impose constraints on how the data

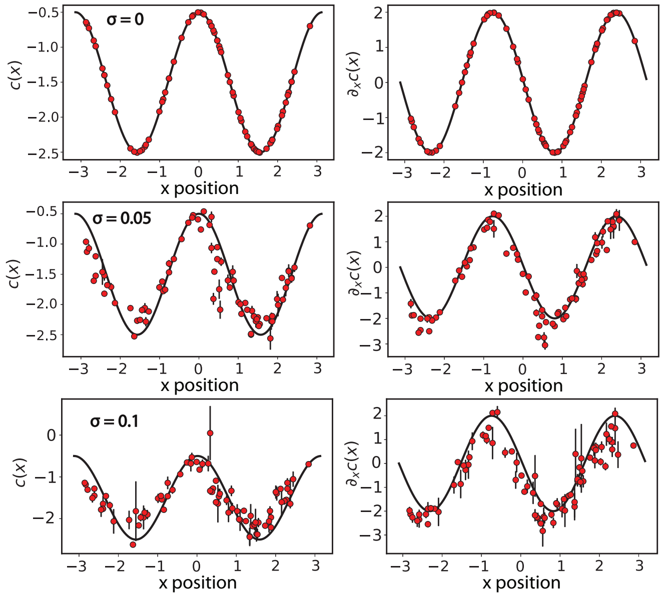

FIG. 9. Spatially varying velocity field and its gradient for the advection-diffusion example. The plots show the estimates for the latent spatially varying velocity $c(x)$ (left column) and its gradient $\partial_x c(x)$ (right column) from the GIHT algorithm. The rows correspond to the inference from data with different noise levels $\sigma$ (shown also in the legend). Symbols show estimated means with bars indicating estimation standard deviations over 20 independent trials. Black solid lines are the ground truth.

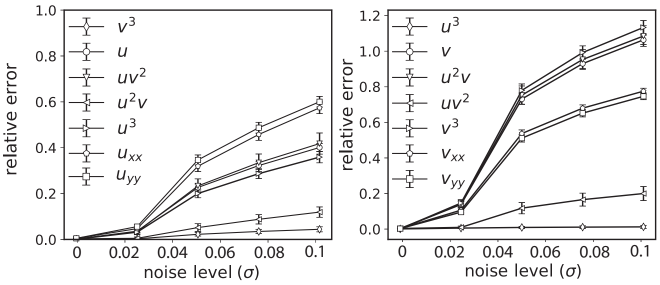

FIG. 10. Relative errors in the coefficient estimation for the $\lambda$-$\omega$ reaction-diffusion system. The plots show the relative errors $\frac{|\xi-\xi^*|}{\xi^*}$ (vs ground truth $\xi^*$) in the GIHT estimates of reaction coefficients and diffusion constants for the species $u$ (left) and $v$ (right) as a function of the noise level $\sigma$ in the data. The ground-truth coefficients for the species $u$ and $v$ are as given in Eq. (12).

points have to be sampled in space and time. This hampers the application of stability selection, which relies on uniformly random subsamples of the data.

We propose to reconcile these two seemingly conflicting requirements of smoothness priors by first identifying the groups using GIHT and then solving the trend-filtering problem as a postprocessing step for each individual group in order to impose the smoothness priors. This uses GIHT with stability selection only to infer the structure of the model (i.e., the support of $\boldsymbol{\xi}$), followed by a separate smoothness-constrained regression to determine the values of the nonzero coefficients by solving

$$\hat{\boldsymbol{\xi}}_s = \arg\min_{\boldsymbol{\xi}} \left[ \frac{1}{2} \left\| U_t - \sum_{j=1}^{K} \boldsymbol{\theta}_{g_j} \boldsymbol{\xi}_{g_j} \right\|_2^2 + \lambda_{\mathrm{f}} \sum_{j=1}^{K} \left\| \Delta_j^{(k+1)} \boldsymbol{\xi}_{g_j} \right\|_1 \right]. \tag{B1}$$

This yields $\hat{\boldsymbol{\xi}}_s$, the smoothed estimates recovered from the trend-filtering problem. Here $K$ is the number of groups identified by stability selection using GIHT, $\Delta_j^{(k+1)} \in \mathbb{R}^{(p_{g_j}-k-1)\times(p_{g_j}-k)}$ is a discrete smoothing filter based on the $(k+1)$th derivative, and $p_{g_j} = |g_j|$ is the size of the respective group. The $\|\cdot\|_1$ norm in the smoothness prior penalizes outliers and favors smooth reconstruction of coefficients. For $k = 0$, this formulation reduces to the classic total variation prior. The regularization constant $\lambda_{\mathrm{f}}$ controls the degree of smoothness imposed on the coefficients by the filter.

As an example, one could regularize the Laplacian (i.e., the curvature) of the coefficients. The discrete filter with $k = 1$ is then given by

$$\Delta^{(2)} = \begin{bmatrix} 1 & -2 & 1 & & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}$$

for all $j$. We demonstrate this in the advection-diffusion example by regularizing smoothness in all recovered groups including the velocity and its gradient field. We solve the optimization problem in Eq. (B1) using the alternating direction method of multipliers algorithm [74] with exhaustive grid search to identify the smoothness regularization $\lambda_{\mathrm{f}}$ that leads to the lowest mean-square error estimate. The reconstructed velocity field and its gradient are shown in Fig. 11 for different levels of noise on the input data. Comparing with the profiles recovered by GIHT directly (Fig. 9), we observe that imposing smoothness priors in a separate postprocessing step significantly improves the reconstruction of the latent fields in this example.

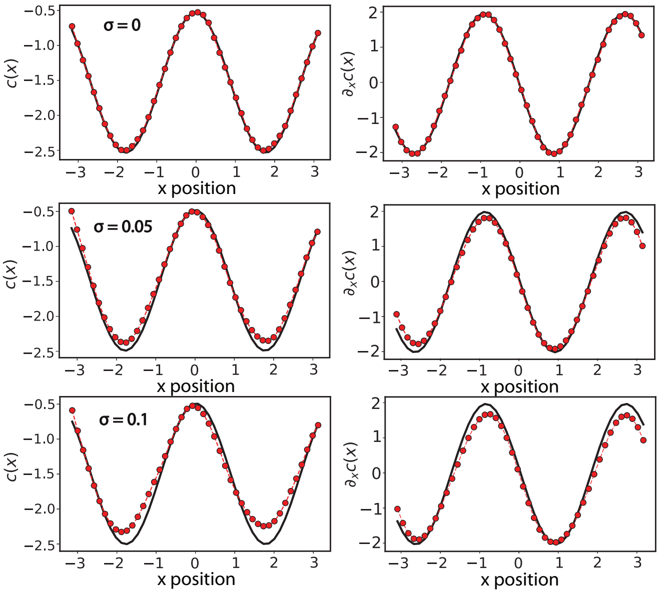

FIG. 11. Reconstructed spatially varying velocity field and its gradient for the advection-diffusion example with additional second-order smoothness prior. The plots show the estimates for the latent spatially varying velocity $c(x)$ (left column) and its gradient $\partial_x c(x)$ (right column) when imposing an additional smoothness prior over each group recovered by stability selection using GIHT. The rows correspond to data with different noise levels $\sigma$ (shown in the legend). For $\sigma = 0$ (top row), we use a smoothness regularization $\lambda_{\mathrm{f}} = 1$, and for $\sigma = 0.05$ and 0.1 (middle and bottom rows), we use $\lambda_{\mathrm{f}} = 20$. Black solid lines are the ground truth.

## APPENDIX C: ALGORITHM FOR GROUP-SPARSE REGRESSION

Given data $\hat{\boldsymbol{u}}(\boldsymbol{x}_i, t_j)$ at discrete locations $\boldsymbol{x}_i$ and time points $t_j$, we use polynomial differentiation [16,27] to approximate the derivatives required to construct the dictionary $\boldsymbol{\Theta}$ and the vector $\boldsymbol{U}_t$. To approximately solve the optimization problem in Eq. (5), we derive the GIHT algorithm. This algorithm is based on an approximate proximal operator for nonoverlapping group sparsity, i.e., for cases where the groups $\{g_l : l \in \mathbb{N}_m\}$ form a partition of the index set $\mathbb{N}_P$. In this case, the approximate proximal operator can be applied to each group separately, and the results summed [75].

### 1. Proximal view of the iterative hard thresholding algorithm

We start from the well-known iterative hard thresholding (IHT) algorithm for $\|\cdot\|_0$-regularized sparse regression [40]. We formulate this algorithm from the perspective of projection and proximal operators. For solving the composite optimization problem in Eq. (3), we use linearization and solve the surrogate problem to generate a sequence $\{\boldsymbol{\xi}^k\}$ as

$$\boldsymbol{\xi}^{k+1} = \arg\min_{\boldsymbol{\xi}} \bigg[ h(\boldsymbol{\xi}^k) + \langle \nabla h(\boldsymbol{\xi}^k), \boldsymbol{\xi} - \boldsymbol{\xi}^k \rangle + \frac{t^k}{2} \|\boldsymbol{\xi} - \boldsymbol{\xi}^k\|^2 + \lambda r(\boldsymbol{\xi}) \bigg]. \tag{C1}$$

This linearization works under the assumption that the data-fitting function $h(\boldsymbol{\xi})$ is continuously differentiable with Lipschitz continuous gradient, i.e., that there exists a positive constant $L > 0$ such that $\|\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y})\|_2 \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|_2\,\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. The problem in Eq. (C1) is equivalent to the proximal operator

$$\text{prox}_r(\boldsymbol{v}^k) = \boldsymbol{\xi}^{k+1} = \arg\min_{\boldsymbol{\xi}} [l(\boldsymbol{\xi}) \equiv \tfrac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{v}^k\|^2 + \lambda r(\boldsymbol{\xi})], \tag{C2}$$

where $\boldsymbol{v}^k = \boldsymbol{\xi}^k - \nabla h(\boldsymbol{\xi}^k)/t^k$ is the gradient-descent iteration with step size $1/t^k$. Thus, we perform gradient descent along $-\nabla h(\boldsymbol{\xi}^k)$ and then apply the proximal operator. In the IHT algorithm with nonconvex penalty function $\lambda\|\boldsymbol{\xi}\|_0$, the proximal operator $\text{prox}_r(\boldsymbol{v})$ is approximated by hard thresholding [40].

### 2. Approximate proximal operator for the nonoverlapping group-sparsity problem

We note that the above alternating gradient-proximal step is similar to the forward-backward splitting algorithm [76]. We therefore propose to use approximate thresholding also for the nonconvex group-sparsity problem.

The proximal operators for proper lower semicontinuous functions $r(\cdot)$ are well defined with the set $\text{prox}_r^\lambda$ being nonempty and compact [77]. By extension of the idea of using thresholding as an approximation to the proximal step, we decompose the separable optimization problem in Eq. (C2) into a sum of subproblems [75] and apply the approximate proximal operator (i.e., thresholding) to each subproblem separately. For nonoverlapping groups, we can decompose the function $l(\boldsymbol{\xi})$ defined in Eq. (C2) into two parts

$$l(\boldsymbol{\xi}) = \left[\frac{1}{2}\|\boldsymbol{\xi}_{g_i} - \boldsymbol{v}_{g_i}\|_2^2 + \lambda\sqrt{p_{g_i}}\mathbb{1}(\|\boldsymbol{\xi}_{g_i}\|_2 \neq 0)\right] + \left(\frac{1}{2}\|\boldsymbol{\xi}_{\bar{g}_i} - \boldsymbol{v}_{\bar{g}_i}\|_2^2 + \lambda\sum_{j\neq i}\sqrt{p_{\bar{g}_j}}\mathbb{1}(\|\boldsymbol{\xi}_{\bar{g}_j}\|_2 \neq 0)\right), \tag{C3}$$

where $\bar{g}_i = \{1, 2, \dots, P\} \setminus g_i$ is the complementary set of the group $g_i$. For a fixed $\boldsymbol{\xi}_{\bar{g}_i} = \boldsymbol{\xi}^*_{\bar{g}_i}$, it can be verified that $\|\boldsymbol{\xi}^*_{g_i}\|_2 = 0$ minimizes both terms in Eq. (C3) if $\|\boldsymbol{v}_{g_i}\| \leqslant \sqrt{\lambda\sqrt{p_{g_i}}}$. For more details, we refer to Lemma 2 in [78] for the zero groups (i.e., for $\boldsymbol{\xi}^*_{g_i} = 0$) in the group LASSO problem [78]. Similar arguments can be made for separable forms other than that shown in Eq. (C3), based on which we can formulate the thresholding rule to minimize the function $l(\boldsymbol{\xi})$:

$$H^\lambda_{\text{group}}(\boldsymbol{v}_g) = \begin{cases} 0 & \text{if } \|\boldsymbol{v}_g\|_2 < \sqrt{\lambda\sqrt{p_{g_i}}} \\ \boldsymbol{v}_g & \text{if } \|\boldsymbol{v}_g\|_2 \geqslant \sqrt{\lambda\sqrt{p_{g_i}}}. \end{cases} \tag{C4}$$

For group size $p_{g_i} = 1\,\forall\, i$, this thresholding rule reduces to the popular hard thresholding algorithm, and the sequence $\{\boldsymbol{\xi}^k\}$ then are iterates of the iterative hard thresholding algorithm [19,40]. Based on the generalized thresholding rule in Eq. (C4), we propose the group iterative hard thresholding algorithm (Algorithm 1) with an additional debiasing step [71,79].

**Algorithm 1** Group iterative hard thresholding with debiasing.

```
Input: $\boldsymbol{\Theta}, \boldsymbol{U}_t, \lambda, \mathcal{G} = \{g_1, g_2, \dots, g_m\}$, maxiter = 10 000
Output: $\hat{\boldsymbol{\xi}}$
Initialization: $\boldsymbol{\xi}^1 = 0$
for $k = 1$ to maxiter do
    $\boldsymbol{v} = \mathbf{H}^\lambda_{\text{group}}(\boldsymbol{\xi}^k - \nabla h(\boldsymbol{\xi}^k)/t^k)$
    $S^k = \text{supp}(\boldsymbol{v}) = \{i \in \{1, \dots, P\} : v_i \neq 0\}$
    $\boldsymbol{\xi}^{k+1} = \arg\min_{\mathbf{z}}\{\|\boldsymbol{U}_t - \boldsymbol{\Theta}z\|_2^2 : \text{supp}(z) \subseteq S^k\}$
end for
```

[1] J. Prost, F. Jülicher, and J.-F. Joanny, Active gel physics, Nat. Phys. **11**, 111 (2015).

[2] X. Trepat and E. Sahai, Mesoscale physical principles of collective cell organization, Nat. Phys. **14**, 671 (2018).

[3] G. Popkin, The physics of life, Nat. News **529**, 16 (2016).

[4] I. F. Sbalzarini, Modeling and simulation of biological systems from image data, Bioessays **35**, 482 (2013).

[5] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach, Phys. Rev. Lett. **120**, 024102 (2018).

[6] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Geometry from a Time Series, Phys. Rev. Lett. **45**, 712 (1980).

[7] J. P. Crutchfield and B. S. McNamara, Equations of motion from a data series, Complex Syst. **1**, 417 (1987).

[8] D. P. Vallette, G. Jacobs, and J. P. Gollub, Oscillations and spatiotemporal chaos of one-dimensional fluid fronts, Phys. Rev. E **55**, 4274 (1997).

[9] M. Bär, R. Hegger, and H. Kantz, Fitting partial differential equations to space-time dynamics, Phys. Rev. E **59**, 337 (1999).

[10] B. C. Daniels and I. Nemenman, Efficient inference of parsimonious phenomenological models of cellular dynamics using S-systems and alternating regression, PLoS One **10**, e0119821 (2015).

[11] K. J. Friston, L. Harrison, and W. Penny, Dynamic causal modelling, Neuroimage **19**, 1273 (2003).

[12] D. Sussillo and L. F. Abbott, Generating coherent patterns of activity from chaotic neural networks, Neuron **63**, 544 (2009).

[13] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data, Science **324**, 81 (2009).

[14] M. D. Schmidt, R. R. Vallabhajosyula, J. W. Jenkins, J. E. Hood, A. S. Soni, J. P. Wikswo, and H. Lipson, Automated refinement and inference of analytical models for metabolic networks, Phys. Biol. **8**, 055011 (2011).

[15] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA **113**, 3932 (2016).

[16] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Data-driven discovery of partial differential equations, Sci. Adv. **3**, e1602614 (2017).

[17] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Machine learning of linear differential equations using Gaussian processes, J. Comput. Phys. **348**, 683 (2017).

[18] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. **378**, 686 (2019).

[19] S. Maddu, B. L. Cheeseman, I. F. Sbalzarini, and C. L. Müller, Stability selection enables robust learning of partial differential equations from limited noisy data, arXiv:1907.07810.

[20] M. Sorokina, S. Sygletos, and S. Turitsyn, Sparse identification for nonlinear optical communication systems: SINO method, Opt. Express **24**, 30433 (2016).

[21] M. Dam, M. Brøns, J. Juul Rasmussen, V. Naulin, and J. S. Hesthaven, Sparse identification of a predator-prey system from simulation data of a convection model, Phys. Plasmas **24**, 022310 (2017).

[22] J.-C. Loiseau, B. R. Noack, and S. L. Brunton, Sparse reduced-order modeling: Sensor-based dynamics to full-state estimation, J. Fluid Mech. **844**, 459 (2018).

[23] M. Hoffmann, C. Fröhner, and F. Noé, Reactive SINDy: Discovering governing reactions from concentration data, J. Chem. Phys. **150**, 025101 (2019).

[24] Y. El Sayed M, R. Semaan, and R. Radespiel, *Proceedings of the 2018 AIAA Aerospace Sciences Meeting, Kissimmee, 2018* (AIAA, Reston, 2018), p. 1054.

[25] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Inferring biological networks by sparse identification of nonlinear dynamics, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 52 (2016).

[26] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Sparse identification of nonlinear dynamics with control (SINDYc), IFAC-PapersOnLine **49**, 710 (2016).

[27] S. Rudy, A. Alla, S. L. Brunton, and J. N. Kutz, Data-driven identification of parametric partial differential equations, SIAM J. Appl. Dyn. Syst. **18**, 643 (2019).

[28] L. Boninsegna, F. Nüske, and C. Clementi, Sparse learning of stochastic dynamical equations, J. Chem. Phys. **148**, 241723 (2018).

[29] B. M. de Silva, D. M. Higdon, S. L. Brunton, and J. N. Kutz, Discovery of physics from data: Universal laws and discrepancies, arXiv:1906.07906.

[30] K. P. Champion, S. L. Brunton, and J. N. Kutz, Discovery of nonlinear multiscale systems: Sampling strategies and embeddings, SIAM J. Appl. Dyn. Syst. **18**, 312 (2019).

[31] H. Schaeffer and S. G. McCalla, Sparse model selection via integral terms, Phys. Rev. E **96**, 023302 (2017).

[32] G.-J. Both, S. Choudhury, P. Sens, and R. Kusters, DeepMoD: Deep learning for model discovery in noisy data, J. Comput. Phys. **428**, 109985 (2021).

[33] P. A. K. Reinbold, D. R. Gurevich, and R. O. Grigoriev, Using noisy or incomplete data to discover models of spatiotemporal dynamics, Phys. Rev. E **101**, 010203(R) (2020).

[34] H. Schaeffer, G. Tran, and R. Ward, Learning dynamical systems and bifurcation via group sparsity, arXiv:1709.01558.

[35] H. Schaeffer, G. Tran, and R. Ward, Extracting sparse high-dimensional dynamics from limited data, SIAM J. Appl. Math. **78**, 3279 (2018).

[36] H. Schaeffer, G. Tran, R. Ward, and L. Zhang, Extracting structured dynamical systems using sparse optimization with very few samples, Multiscale Model. Simul. **18**, 1435 (2020).

[37] J. A. Tropp, Greed is good: Algorithmic results for sparse approximation, IEEE Trans. Inf. Theory **50**, 2231 (2004).

[38] D. Needell and J. A. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. **26**, 301 (2009).

[39] W. Dai and O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Trans. Inf. Theory **55**, 2230 (2009).

[40] T. Blumensath and M. E. Davies, Iterative hard thresholding for compressed sensing, Appl. Comput. Harmon. Anal. **27**, 265 (2009).

[41] R. Tishbirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B Met. **58**, 267 (1996).

[42] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. B **68**, 49 (2006).

[43] M. Kowalski, *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (IEEE, Piscataway, 2014), pp. 4151–4155.

[44] J. Huang and T. Zhang, The benefit of group sparsity, Ann. Stat. **38**, 1978 (2010).

[45] P. Jain, N. Rao, and I. Dhillon, Structured sparse regression via greedy hard thresholding, in *Advances in Neural Information Processing Systems*, edited by D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, and I. Guyon (Curran, Red Hook, 2016), pp. 1516–1524.

[46] Y. Zhang, R. Li, and C.-L. Tsai, Regularization parameter selections via generalized information criterion, J. Am. Stat. Assoc. **105**, 312 (2010).

[47] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Selected Papers of Hirotugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa, Springer Series in Statistics (Springer, New York, 1998), pp. 199–213.

[48] G. Schwarz, Estimating the dimension of a model, Ann. Stat. **6**, 461 (1978).

[49] C. Lim and B. Yu, Estimation stability with cross-validation (ESCV), J. Comput. Graph. Stat. **25**, 464 (2016).

[50] N. Meinshausen and P. Bühlmann, Stability selection, J. R. Stat. Soc. B **72**, 417 (2010).

[51] P. Bühlmann, M. Kalisch, and L. Meier, High dimensional statistics with a view toward applications in biology, Annu. Rev. Stat. Appl. **1**, 255 (2014).

[52] J. Timmer, T. Müller, I. Swameye, O. Sandra, and U. Klingmüller, Modeling the nonlinear dynamics of cellular signal transduction, Int. J. Bifurcat. Chaos **14**, 2069 (2004).

[53] I. Swameye, T. Müller, J. Timmer, O. Sandra, and U. Klingmüller, Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling, Proc. Natl. Acad. Sci. USA **100**, 1028 (2003).

[54] N. W. Goehring, P. K. Trong, J. S. Bois, D. Chowdhury, E. M. Nicola, A. A. Hyman, and S. W. Grill, Polarization of PAR proteins by advective triggering of a pattern-forming system, Science **334**, 1137 (2011).

[55] T. Nishimura, H. Honda, and M. Takeichi, Planar cell polarity links axes of spatial dynamics in neural-tube closure, Cell **149**, 1084 (2012).

[56] M. Mayer, M. Depken, J. S. Bois, F. Jülicher, and S. W. Grill, Anisotropies in cortical tension reveal the physical basis of polarizing cortical flows, Nature (London) **467**, 617 (2010).

[57] T. Mammoto and D. E. Ingber, Mechanical control of tissue and organ development, Development **137**, 1407 (2010).

[58] P. Gross, K. V. Kumar, N. W. Goehring, J. S. Bois, C. Hoege, F. Jülicher, and S. W. Grill, Guiding self-organized pattern formation in cell polarity establishment, Nat. Phys. **15**, 293 (2019).

[59] E. Nazockdast, A. Rahimian, D. Needleman, and M. Shelley, Cytoplasmic flows as signatures for the mechanics of mitotic positioning, Mol. Biol. Cell **28**, 3261 (2017).

[60] C. Collinet, M. Stöter, C. R. Bradshaw, N. Samusik, J. C. Rink, D. Kenski, B. Habermann, F. Buchholz, R. Henschel, M. S. Mueller *et al.*, Systems survey of endocytosis by multiparametric image analysis, Nature (London) **464**, 243 (2010).

[61] S. Eaton and F. Jülicher, Cell flow and tissue polarity patterns, Curr. Opin. Genet. Dev. **21**, 747 (2011).

[62] A. M. Turing, The chemical basis of morphogenesis, Phil. Trans. R. Soc. Lond. B **237**, 37 (1952).

[63] M. C. Cross and P. C. Hohenberg, Pattern formation outside of equilibrium, Rev. Mod. Phys. **65**, 851 (1993).

[64] A. B. Medvinsky, S. V. Petrovskii, I. A. Tikhonova, H. Malchow, and B.-L. Li, Spatiotemporal complexity of plankton and fish dynamics, SIAM Rev. **44**, 311 (2002).

[65] J. D. Murray, *Mathematical Biology: I. An Introduction*, edited by S. S. Antman, J. E. Marsden, L. Sirovich, and S. Wiggins, Interdisciplinary Applied Mathematics Vol. 17 (Springer, New York, 2007).

[66] K.-H. Hoffmann and Q. Tang, *Ginzburg-Landau Phase Transition Theory and Superconductivity* (Birkhäuser, Basel, 2012).

[67] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence* (Courier, Red Hook, 2003).

[68] N. Kopell and L. N. Howard, Plane wave solutions to reaction-diffusion equations, Stud. Appl. Math. **52**, 291 (1973).

[69] H. I. Freedman, *Deterministic Mathematical Models in Population Ecology* (Dekker, New York, 1980), Vol. 57.

[70] L. Yang, M. Dolnik, A. M. Zhabotinsky, and I. R. Epstein, Spatial Resonances and Superposition Patterns in a Reaction-Diffusion Model with Interacting Turing Modes, Phys. Rev. Lett. **88**, 208303 (2002).

[71] S. Foucart, Hard thresholding pursuit: An algorithm for compressive sensing, SIAM J. Numer. Anal. **49**, 2543 (2011).

[72] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, Optimization with sparsity-inducing penalties, Found. Trends Mach. Learn. **4**, 1 (2012).

[73] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).

[74] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* (Now Publishers, Delft, 2011).

[75] A. Argyriou, C. A. Micchelli, M. Pontil, L. Shen, and Y. Xu, Efficient first order methods for linear composite regularizers, arXiv:1104.1436.

[76] P. L. Combettes and J.-C. Pesquet, Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, edited by H. H. Bauschke, R. S. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Springer Optimization and its Applications Vol. 49 (Springer, New York, 2011), pp. 185–212.

[77] S. Zhang, H. Qian, and X. Gong, *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, 2016), Vol. 30, pp. 2330–2336.

[78] L. Yuan, J. Liu, and J. Ye, Efficient methods for overlapping group Lasso, IEEE Trans. Pattern Anal. Mach. Intel. **35**, 2104 (2013).

[79] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, IEEE J. Sel. Top. Signal Process. **1**, 586 (2007).