SINDy for delay-differential equations: application to model bacterial zinc response

Antoine Sandoz[*]   Verena Ducret[†]   Georg A. Gottwald[‡]   Gilles Vilmart[§]
Karl Perron [¶]

December 11, 2022

Abstract

We extend the data-driven method of Sparse Identification of Nonlinear Dynamics (SINDy) developed by *Brunton et al, Proc. Natl. Acad. Sci USA 113 (2016)* to the case of delay differential equations (DDEs). This is achieved in a bilevel optimization procedure by first applying SINDy for fixed delay and then subsequently optimizing the error of the reconstructed SINDy model over delay times. We test the SINDy-delay method on a noisy short data set from a toy delay differential equation and show excellent agreement. We then apply the method to experimental data of gene expressions in the bacterium *Pseudomonas aeruginosa* subject to the influence of zinc. The derived SINDy model suggests that the increase of zinc concentration mainly affects the time delay and not the strengths of the interactions between the different agents controlling the zinc export mechanism.

keywords: data-driven modelling; SINDy; delay-differential equations; *Pseudomonas aeruginosa*, zinc homeostasis

1   Introduction

Our ability to understand, forecast and control dynamical systems depends crucially on our knowledge of its underlying equations. Recently data-driven methods to uncover underlying equations have been proposed to uncover unknown dynamics and to increase our forecast capabilities [6, 9, 39, 24, 46, 2]. A particularly attractive and easy-to-implement method, proposed by Brunton et al. [7], is *Sparse Identification of Nonlinear Dynamics*, or SINDy for short. The problem which is addressed by SINDy is the following: Given observations $\mathbf{x}_n \in \mathbb{R}^d$

---

[*]Section of Mathematics and Microbiology Unit, Department of Plant Sciences, University of Geneva, Switzerland; Antoine.Sandoz@unige.ch

[†]Microbiology Unit, Department of Plant Sciences, University of Geneva, Switzerland; Verena.Ducret@unige.ch

[‡]School of Mathematics and Statistics, University of Sydney, Sydney, Australia; georg.gottwald@sydney.edu.au

[§]Section of Mathematics, University of Geneva, Switzerland; Gilles.Vilmart@unige.ch

[¶]Microbiology Unit, Department of Plant Sciences, and Section of Pharmaceutical Sciences, University of Geneva, Switzerland; Karl.Perron@unige.ch

sampled at (not necessarily equidistant) times t_n which were generated by a dynamical system of the form

$$\dot{\mathbf{x}} = \mathcal{F}(\mathbf{x}),$$

where the dot signifies the time derivative, find an approximation to this dynamical system using only the data. SINDy approaches this question by assuming that the vector field $\mathcal{F}(\mathbf{x})$ lies in the span of a given (potentially very large) library of functions such as simple polynomials. This reduces the problem to linear regression on a library of nonlinear functions. In line with the parsimony principle, SINDy imposes a sparsity constraint leading to a sparse approximation of $\mathcal{F}(\mathbf{x})$ in terms of the members of the library functions. If the system is only partially observed, the method of SINDy can be extended to the reconstructed phase-space using Takens' embedding theorem and it is then known as the Hankel alternative view of Koopman (HAVOK) analysis [5]. SINDy has been successfully applied to systems appearing in a wide range of scientific disciplines, including fluid dynamics, plasma physics and nonlinear optics [30, 12, 40].

Many dynamical systems involve a delayed feedback response and are modelled by delay-differential equations (DDEs). Examples range from the natural world to engineering with applications in, for example, population dynamics [11, 23], biological regulatory systems [19], cardiac dynamics [22, 21], climate dynamics [43, 26], mechanical vibration [49] and in optical systems [45], to name just a few. In this paper, we extend the framework of SINDy to dynamical systems which are described by DDEs and where in addition to the sparse subset of the library of nonlinear functions and their associated coefficients, the delay has to be determined as a parameter. We achieve this by employing a bilevel optimization in which, for a fixed specified delay time, the error in reproducing the observations made by each approximate SINDy model is minimized. Particular emphasis will be given to deal with noisy data. We shall first test the proposed SINDy-delay methodology to a one-dimensional toy model with artificially noisy data before considering a challenging problem with biological data of gene expressions in the bacterium *Pseudomonas aeruginosa* subject to the influence of zinc.

P. aeruginosa is an opportunistic pathogen capable of causing acute infections in hospitals, in particular in immunocompromised patients, in cystic fibrosis patients and in severe burn victims [28, 25]. Therefore, it belongs to the Priority 1 category for research into antibiotic resistance as determined by the world health organization [44]. *P. aeruginosa* has a large genome, coding for 5570 open reading frames, of which 72 are involved in predicting so called two component systems (TCS) [42]. TCSs are crucial biological building blocks. They are composed of a sensor protein which in response to a stimulus, activates a cognate transcriptional regulator by phosphorylation, allowing for a rapid adaptation to environmental changes. *P. aeruginosa* has one of the highest numbers of putative TCSs among bacteria, contributing to the ubiquity of this micro-organism [1]. For instance, the CzcRS TCS promotes resistance to high concentrations as well as to large fluctuations of the concentration of trace metals such as zinc. This is advantageous for the bacteria in an infectious context [36, 13] since to counter the multiplication of bacteria, the host uses nutritional immunity strategies via scavenging essential nutrients including zinc, iron and manganese [27, 8, 31]. Conversely, during phagocytosis, macrophages deliver a toxic amount of zinc and copper into the phagolysosome, leading to the death of the invader organism [14, 41, 18]. Thus, the success of an infection depends largely on the capacity of a pathogen to survive in zinc deficient as well as zinc excess environments and to switch from one to the other of these extreme conditions. *P. aeruginosa* has a whole arsenal of the most effective systems for regulating the entry and exit of the metal. Moreover, zinc was shown to exacerbate the bacterium pathogenicity, enhancing the virulence factor production and rendering this micro-organism

more resistant to antibiotics, especially those belonging to the carbapenem family, a last resort anti-pseudomonas class of compounds [36, 13]. In order to better understand the *P. aeruginosa* zinc homeostasis, we derive a mathematical delay differential equation model focusing on the dynamic of two main zinc export machineries. To address this challenging question, we use the proposed SINDy-delay methodology applied to experimental data.

The paper is organised as follows. In Section 2 we introduce an extension of SINDy to find parsimonious models for delay-differential equations. In Section 3 we illustrate the effectiveness of our method in the context of a known toy model DDE; in particular, we show that the DDE is recovered well even for short noise-contaminated observations. In Section 4 we then apply our method to experimental data of gene expression data of the bacterium *P. aeruginosa* under various concentrations of zinc. We conclude in Section 5 and discuss biological implications of the discovered DDE describing the bacterium's zinc regulation system.

2 Sparse Identification of Nonlinear Dynamics with delay for noisy data (SINDy-delay)

Consider a d-dimensional dynamical system with delay time τ,

$$\dot{\mathbf{x}} = \mathcal{F}(\mathbf{x}(t), \mathbf{x}(t - \tau)), \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^d$, which is probed at times t_n, $n = \ldots, -1, 0, 1, 2, 3 \ldots$, by observations

$$\boldsymbol{\chi}_n = \mathbf{x}_n + \boldsymbol{\Gamma}\boldsymbol{\eta}_n,$$

with measurement error covariance matrix $\Gamma^2 \in \mathbb{R}^{d \times d}$ and independent normally distributed noise $\boldsymbol{\eta}_n \sim \mathcal{N}(0, \mathrm{Id})$. For simplicity we assume here throughout $\boldsymbol{\Gamma} = \gamma \, \mathrm{Id}$. The aim is to find a parsimonious approximation of the vector field $\mathcal{F}(\mathbf{x}(t), \mathbf{x}(t - \tau))$ as a linear combination of nonlinear functions selected from a library $\mathcal{R}$ of cardinality $N_\mathcal{R}$. In particular, the kth component is expressed as a linear combination of all the library functions $\theta_j \in \mathcal{R}$, $j = 1, \ldots, N_\mathcal{R}$, of the form

$$\mathcal{F}_k(\mathbf{x}(t), \mathbf{x}(t - \tau)) = \sum_{j=1}^{N_\mathcal{R}} \xi_k^j \theta_j(\mathbf{x}(t), \mathbf{x}(t - \tau)) + \epsilon_k(\mathbf{x}(t), \mathbf{x}(t - \tau)), \tag{2}$$

for all components $k = 1, \ldots, d$. Simple nonlinear regression would amount to determining the coefficients ξ_k^j using the method of least-squares to minimize the mismatch ϵ_k. In SINDy rather, a sparsity constraint is invoked, seeking a parsimonious model with as many of the coefficients ξ_k^j being zero while still ensuring fidelity of the approximation (2) with respect to the data.

To describe how SINDy finds such an approximation, let us assume for the moment that observations are taken at equidistant times $t_n = n\Delta t$ with constant sampling time Δt. To account for the delay we form the observation vector

$$\hat{\boldsymbol{\chi}}_n^{(s)} = \begin{pmatrix} \boldsymbol{\chi}_n \\ \boldsymbol{\chi}_{n-s} \end{pmatrix} \in \mathbb{R}^{2d},$$

for $n = 1, \ldots, N$, where the positive integer s is related to a delay time $\tau = s\Delta t$. Following the exposition in Brunton et al. [5] and Brunton and Kutz [6], we collect the observation vectors $\hat{\boldsymbol{\chi}}_n^{(s)}$ in a data matrix

$$\boldsymbol{X}^T = \begin{pmatrix} \hat{\boldsymbol{\chi}}_1^{(s)} & \hat{\boldsymbol{\chi}}_2^{(s)} & \ldots & \hat{\boldsymbol{\chi}}_N^{(s)} \end{pmatrix} \in \mathbb{R}^{2d \times N}.$$

3

Similarly we define the matrix consisting of derivatives of the observations χ_n at the observation times

$$\dot{\boldsymbol{X}}^T = \begin{pmatrix} \dot{\chi}_1 & \dot{\chi}_2 & \dots & \dot{\chi}_N \end{pmatrix} \in \mathbb{R}^{d \times N}. \tag{3}$$

Note that for the derivatives we only consider the time derivatives for χ_n and not for χ_{n-s} which would be redundant information. Typically one does not have access to the actual derivatives $\dot{\chi}_n$ but only to the variables χ_n themselves. For noise-free finely sampled observations with $\Delta t \ll 1$ finite differencing can be employed to approximate the derivatives. For noisy observations, however, estimating derivatives via finite-differencing leads to an amplification of the noise. Denoising methods such as the total-variation regularized method are required [10]. Here we propose to use simple polynomial regression for denoising as discussed in the following remark.

Remark 2.1. *(Denoising procedure for computing the derivative matrix (3)) For computing each $\dot{\chi}_n$, $n = 1 \dots, N$ in (3), we use polynomial regression. We define for the corresponding observation time t_n a temporal window $[t_n - \delta, t_n + \delta]$ with $\delta = r\Delta t$ and fit a 3rd order polynomial through the $2r - 1$ observations $\chi_{n-r}, \chi_{n-r+1}, \dots, \chi_{n+r}$ lying within this time window. Choosing a sufficiently large temporal window containing more data points compared to the regression polynomial degree allows for noise reduction.*

The derivatives $\dot{\chi}_n$ can then be analytically determined from the fitted polynomials at each time t_n. This denoising procedure can easily be adapted to handle non-equidistantly sampled observations which may be the situation in experimental data (including the case of determining delayed data at $t_n - \tau$ which may not have been directly observed). Our denoising procedure is closely related to the Savitsky-Golay filter [38] and denoising by splines [48]; for a comparison of various denoising strategies see [47].

At the heart of SINDy lies the choice of a suitably large library $\mathcal{R}$. A natural choice is the set of monomials in $x_k(t), x_k(t - \tau), k = 1, \dots, d$ up to a fixed degree M, $\mathcal{R} = \{1, x_1(t), x_2(t), x_1(t - \tau), x_2(t - \tau) \dots\}$ with cardinality $N_{\mathcal{R}} = \binom{2d+M}{M}$. Given a library $\mathcal{R}$, the associated library matrix $\boldsymbol{\Theta}(\boldsymbol{X}) \in \mathbb{R}^{N \times N_{\mathcal{R}}}$ is constructed from the data by evaluating all functions $\theta_j(\mathbf{x}(t), \mathbf{x}(t - \tau))$ of the library $\mathcal{R}$ at the observation times $t = t_1, \dots, t_N$. When considering the library consisting of monomials of up to order M, the library matrix becomes

$$\boldsymbol{\Theta}(\boldsymbol{X}) = \begin{pmatrix} \mathbf{1} & \boldsymbol{X} & \boldsymbol{X}^2 & \boldsymbol{X}^3 & \dots & \boldsymbol{X}^M \end{pmatrix},$$

where the matrices $\boldsymbol{X}^m \in \mathbb{R}^{N \times \binom{2d+m-1}{m}}$ consist of rows whose coefficients include all possible monomials of degree m between the d-dimensional variables χ_n and χ_{n-s}. For simplicity, we will later in the numerical experiments exclude any products between χ_n and χ_{n-s}. This reduces the number of columns of each $\boldsymbol{X}^m$ to $2\binom{d+m-1}{m}$ and the overall number of columns of $\boldsymbol{\Theta}(\boldsymbol{X})$ to $N_{\mathcal{R}} = 2\binom{d+M}{M}$.

In SINDy the minimization of the error ϵ_k made by the approximation (2) is achieved by an ℓ_1-regularized regression problem. Defining first the ℓ_2-cost function

$$C(\Xi) = \sum_{k=1}^{d} \|\dot{X}_k - \boldsymbol{\Theta}(\boldsymbol{X})\xi_k\|_2^2, \tag{4}$$

where $\dot{X}_k \in \mathbb{R}^N$ denotes the kth column of $\dot{\boldsymbol{X}}$ and $\Xi = \{\xi_k\}_{k=1,\dots,d}$ is the coefficient matrix consisting of column vectors $\xi_k \in \mathbb{R}^{N_{\mathcal{R}}}$ which denote the coefficients associated with the library functions for the kth component of the state variable (cf. (2)). To promote sparsity of the coefficients the cost function is minimized under an ℓ_1-sparsity constraint according to

$$\boldsymbol{\xi_k} = \arg\min_{\xi_k \in \mathbb{R}^{N_{\mathcal{R}}}} \|\dot{\boldsymbol{X}}_k - \boldsymbol{\Theta}(\boldsymbol{X})\xi_k\|_2 + \lambda\|\xi_k\|_1, \tag{5}$$

4

where the regularization parameter λ controls the sparsity. Rather than using a sequential thresholded least-squares algorithm to approximate the solution of the optimisation problem (5), as suggested in [7], we promote here sparsity by the following sequential procedure. Define ξ_k^q to be the coefficient of the qth library function θ_q which is associated with the kth component of the vector field. For each $q = 1, \ldots, N_{\mathcal{R}}$ calculate the least square solution $\xi_k^q \in \mathbb{R}^{N_{\mathcal{R}}}$ corresponding to the minimization of the cost function $C(\Xi)$ with the hard sparsity constraint

$$\xi_k^q = 0.$$

For each of the $N_{\mathcal{R}}$ solutions ξ_k^q record the associated minimized cost $C(\Xi)$, and select the value

$$q^* = \underset{q=1,\ldots,N_{\mathcal{R}}}{\arg\min} \ \underset{\xi_k}{\min} \{ C(\Xi) \ ; \ \xi_k^q = 0 \} \tag{6}$$

corresponding to the hard sparsity constraint $\xi_k^{q^*} = 0$ which leads to the smallest increase in the minimum of the cost $C(\Xi)$. We then set $\xi_k^{q^*} = 0$, i.e. excluding $\theta_{q^\star}$ from the library $\mathcal{R}$ for the kth state variable. Algorithmically this amounts to deleting the qth column of $\Theta(\boldsymbol{X})$ when seeking solutions of (5). This process of eliminating coefficients ξ_k^q is then repeated for the remaining library functions in $\mathcal{R}$ (and the corresponding columns of $\Theta(\boldsymbol{X})$) until a significantly large change of the cost $C(\Xi)$ has been accrued, suggesting that removing any of the remaining functions will lead to a strong increase of the cost function, thereby deteriorating the accuracy of the SINDy model.

Remark 2.2. *(Promoting sparsity to approximate solutions of (5)) Promoting sparsity by envoking (6) avoids having to set a cutoff value p such that coefficients with $|\xi_k^j| \leq p$ are removed as proposed in Brunton et al. [7]. Instead the degree of sparsity is visually determined by plotting the cost function for an increasing number of removals. This does not require the data X to be normalized in a pre-processing step and can be applied to situations in which variables may exhibit widely varying ranges. We shall encounter such a situation for the experimental data in Section 4.*

The above procedure is applied to each of the components $k = 1, \ldots, d$ with each component having their separate subset of eigenfunctions selected. Collecting the typically sparse output vectors ξ_k^*, $k = 1, \ldots d$ in the matrix $\Xi^* = (\xi_1^*, \ldots \xi_d^*)$, the approximate SINDy delay differential equation model for arbitrary fixed delay time $\tau_s = s\Delta t$ is given by

$$\dot{x}_k(t; \tau_s) = \sum_{j=1}^{N_{\mathcal{R}}} (\xi_k^j)^* \theta_j(\mathbf{x}(t), \mathbf{x}(t - \tau_s)), \quad k = 1, \ldots, d. \tag{7}$$

Up to here this is standard SINDy, as described in Brunton et al. [7], except for the proposed alternative method of denoising with local polynomial regressions of the data points described in Remark 2.1, and for the modified algorithm to approximate solutions to the optimization problem (5) described in Remark 2.2.

To account for a delay we extended the nonlinear library $\{\theta_k\}_{k=1,\ldots,N_{\mathcal{R}}}$ to include delay terms $\mathbf{x}(t - \tau)$, which fits in the standard SINDy methodology for fixed delay time parameter $\tau_s = s\Delta t$. To estimate the delay time $\tau = s\Delta t$ of the dynamical system (7) which best matches the data $\boldsymbol{\chi}_n$ an additional optimization procedure is employed: Consider a range of delay times $\tau_s = s\Delta t$ with integer sequence $s \in \{0, 1, 2, \ldots, \}$. For each τ_s we perform the above procedure to obtain the SINDy model (7). We then compute the reconstruction error $\mathcal{E}(\tau_s)$ for fixed delay time parameter τ_s as the ℓ_2-error between the solution of the SINDy

Input: Observational data χ_n, $n = \ldots, -1, 0, 1, 2, \ldots$, and nonlinear function library set $\mathcal{R} = \{\theta_j \; ; \; j = 1, \ldots, N_\mathcal{R}\}$.

Output: Delay time τ^* and coefficient matrix Ξ^* determining the delay differential equation model (10).

Compute the derivative matrix $\dot{\boldsymbol{X}}$ in (3), with denoising procedure (see Remark 2.1);

for *all delay time $\tau_s = s\Delta t$, $s \in \{0, 1, 2, \ldots\}$* **do**

    Compute the data matrix $\boldsymbol{X}$ and the associated library matrix $\Theta(\boldsymbol{X})$.

    for $k \in \{1, \ldots, d\}$ **do**

        Set the list $Q \subset \{1, \ldots, N_\mathcal{R}\}$ of indices of vanishing coefficients $\xi_k^j = 0$ to achieve the sparsity constraint of the SINDy methodology to $Q = \emptyset$;

        while $C = \min_{\xi_k \in \mathbb{R}^{N_\mathcal{R}}} \{\|\dot{X}_k - \boldsymbol{\Theta}(\boldsymbol{X})\xi_k\|_2 \; ; \; \xi_k^j = 0 \text{ for all } j \in Q\}$ *does not increase significantly (see Remark 2.2)* **do**

            Compute
$$(q^*, \xi_k^*) = \underset{q \in \{1, \ldots, N_\mathcal{R}\} \backslash Q, \; \xi_k \in \mathbb{R}^{N_\mathcal{R}}}{\arg\min} \{\|\dot{X}_k - \boldsymbol{\Theta}(\boldsymbol{X})\xi_k\|_2 \; ; \; \xi_k^j = 0 \text{ for all } j \in Q \cup \{q\}\}.$$

            Set $Q = Q \cup \{q^*\}$.

        end

    end

    Keep $\Xi^* = \{\xi_1^*, \ldots, \xi_d^*\}$ and the corresponding error $\mathcal{E}(\tau_s)$ in (8).

end

Save the optimal delay time τ^* given in (9), and the corresponding coefficient matrix Ξ^*.

Algorithm 1: SINDy algorithm for dynamical systems with temporal delay.

model (7) and observations χ_n,

$$\mathcal{E}(\tau_s) = \frac{1}{Z} \sum_{n=1}^{N} \|\mathbf{x}(t_n; \tau_s) - \chi_n\|^2. \tag{8}$$

We set the normalization constant to $Z = \sum_{j=0}^{N} \|\chi_j\|^2$. Note that we define the error $\mathcal{E}(\tau_s)$ here in terms of the trajectories $\mathbf{x}(t)$ rather than via the derivatives as in the cost function (4) used in the SINDy core. This proves to be advantageous as trajectories are less affected by the noise than their derivatives. The optimal delay time is estimated as the solution of

$$\tau^\star = \underset{\tau_s}{\arg\min} \, \mathcal{E}(\tau_s). \tag{9}$$

This finally yields the SINDy-delay differential equation model,

$$\dot{\mathbf{x}}(t)^T = \boldsymbol{\Theta}(\mathbf{x}(t)^T, \mathbf{x}(t - \tau^*)^T)\Xi^*. \tag{10}$$

We remark that the devising strategy proposed in Remark 2.1 allows for non-uniformly sampled data, provided the sampling times are not too far apart. The required values of unobserved $x(t - \tau)$ can be evaluated by the proposed polynomial regression. We summarize the extension of the SINDy methodology to systems involving temporal delays in Algorithm 1. In the next Section we show how the SINDy-delay method performs for artificial data obtained from a simple one-dimensional DDE.

	N	Δt	noise γ	observations	$\tau^\star$	$x(t)$	$x(t-\tau)$	$x^3(t)$	error $\mathcal{E}(\tau^{\tau^\star})$
	–	–	–	–	7	1	−0.75	−1	–
(a)	4000	0.025	0	$x, \dot{x}$	7.00	1.00	−0.75	−1.00	$1.36e-03$
(b)	4000	0.025	0	x	7.00	0.99	−0.75	−0.99	$1.80e-03$
(c)	200	0.25	0.02	x	7.00	0.84	−0.71	−0.88	$2.56e-02$
(d)	4000	0.025	0.02	x	7.025	0.92	−0.74	−0.94	$1.94e-02$

Table 1: Results of the SINDy-delay method (Algorithm 1) applied to data obtained from the toy model (11). The first row denotes the true delay time and coefficients of the DDE (11) used to generate the observations. Rows (a)–(d) present results for the different scenarios described in the main text. We present results the estimated delay times τ, the coefficients as well as the associated reconstruction error $\mathcal{E}(\tau)$ (cf (8)) between the data and the SINDy differential equation model (10) for varying data length N, sampling times Δt, noise levels γ. The columns for the monomial terms $x(t), x(t-\tau), x(t)^3$ display the estimated corresponding coefficients ξ_1^j (coefficients for the remaining monomials were estimated to be 0 in all experiments). Experiments in which only x is observed required polynomial regression as described in Remark 2.1.

3 Application to a toy model

To illustrate how the SINDy-delay method is able to determine an underlying DDE together with the delay time parameter from noisy observations, we consider the following one-dimensional DDE,

$$\dot{x}(t) = x(t) - x(t)^3 - \alpha x(t-\tau). \tag{11}$$

This DDE was introduced as a toy model in the context of climate science to describe, for example, the El Niño – Southern Oscillation (ENSO) phenomenon where $x(t)$ denotes a sea-surface temperature anomaly at time t [43]. We choose in the following as parameter value $\alpha = 0.75$ and a delay time of $\tau = 7$. The initial solution on the time interval $[-\tau, 0]$ is chosen to be the stable periodic solution to (11) and with $x(0) = 1$. For the set of nonlinear library functions $\mathcal{R}$, we consider all monomials up to cubic degree, $\mathcal{R} = \{1, x(t), x(t)^2, x(t)^3, x(t-\tau), x(t-\tau)^2, x(t-\tau)^3\}$, excluding products of $x(t)$ and $x(t-\tau)$. We simulate the DDE (11) using the Matlab dde23 integrator with absolute and relative tolerances of 10^{-8} to produce time series of N observations sampled at equidistant times with sampling time Δt [32].

We present results for several scenarios with increasing difficulty. In particular, we investigate how the accuracy of the method depends on the amount of data available as well as on the level of noise. In the following we restrict the delays τ_s to $\tau_s = s\Delta t$ for $s = 1, \ldots, 8.5/\Delta t$ so that we sample from the interval $[0, 8.5]$.

We begin with the ideal situation of noiseless observations of both the state $\mathbf{x}$ and the derivative $\dot{\mathbf{x}}$. We consider observations with $N = 4\,000$ sampled at equidistant times with sampling interval $\Delta t = 2.5 \cdot 10^{-2}$. Figure 1(a) shows the increase of the normalised cost function $C(\Xi)$ upon removal of members of the library $\mathbf{\Theta}(\boldsymbol{X})$ for fixed delay time τ_s. The normalization is with respect to $C(0)$, the value when all library functions are removed, i.e. the error encountered for a rough model with a constant solution $x(t) = x(0)$. The member of the library $\mathcal{R}$ to be removed at each iteration is chosen as the one leading to the least increase of the normalized cost function. This iterative process terminates with the remaining terms as output, just before the normalized cost function increases by more than 10%. We present results for the delay time $\tau_s = 7$ (blue curve with open circles), corresponding to the true delay time, and for a non-optimal delay time $\tau_s = 6$ (red curve
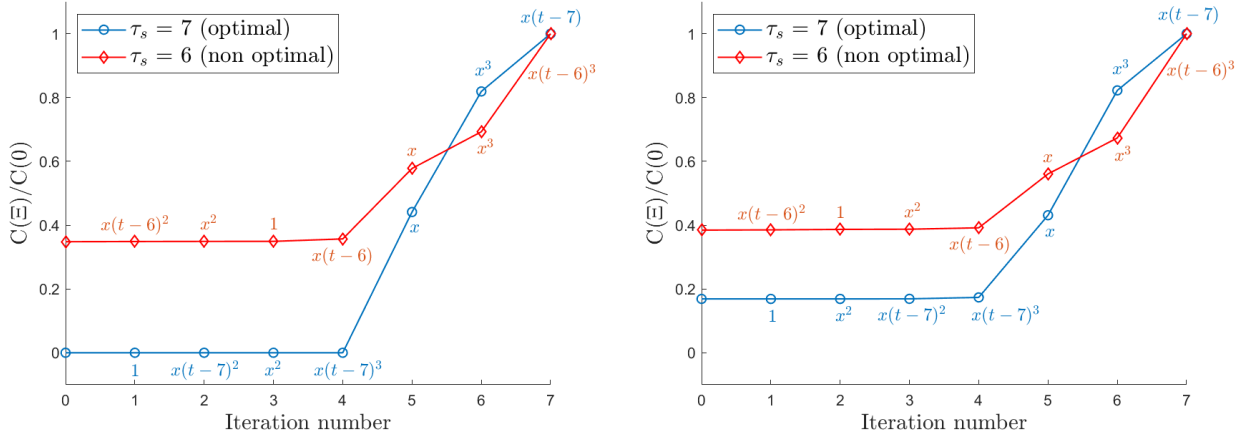
(a) noiseless observations of x and $\dot{x}$ ($N = 4\,000$). (b) noisy observations of x ($N = 200$, $\gamma = 0.02$).

Figure 1: Cost function $C(\Xi)$ (nomalized by $C(0)$) against removed monomials for fixed delay. Results are shown for the true delay time $\tau_s = 7$ (open circles, online blue) and for the non-optimal delay time $\tau_s = 6$ (diamonds, online red). The error increases after iteration 4 of the SINDy algorithm, as soon as the terms $x, x^3, x(t - \tau)$ actually present in the underlying DDE model start to be removed from the library.

with diamonds). We indicate for both delay times the library functions which are removed at each iteration. For the correct delay time $\tau_s = 7$, as expected, we observe a jump in the cost function when one of the terms is being removed which appears in the DDE (11) (i.e. $x(t), x(t)^3, x(t - \tau)$). For the non-optimal delay time $\tau_s = 6$ we also see, as expected, a jump but the selected terms $x(t), x(t)^3, x(t - \tau)^3$ do not correspond to the actual terms appearing in (11). We also observe a significantly lower value of the cost function for the (optimal) delay time $\tau_s = 7$ compared to the non-optimal delay time $\tau_s = 6$ at iteration numbers for which none of the selected monomials have been removed. In Figure 2 (open circles, online blue), we show how the optimal delay time $\tau^\star$ is determined by inspecting the reconstruction error $\mathcal{E}(\tau_s)$ (cf. (8)). The reconstruction error has a clear minimum at $\tau^\star = 7$. In the more challenging case when only ($N = 4\,000$) noise-less observations $x(t_n)$ are available and the derivative matrix $\dot{\boldsymbol{X}}$ has to be estimated in a post-processing step using the polynomial smoothing described in Remark 2.1. We perform the polynomial regression with $r = 25$ with $\delta = r\Delta t = 0.625$. The SINDy algorithm recovers the coefficients and delay times close to the true values as seen in row (b) of Table 1.

We now test the method in the difficult case of short noise-contaminated data with $N = 200$. The variable $x(t)$ is sampled at observation intervals of $\Delta t = 0.25$ and are contaminated with observational noise with $\gamma = 0.02$. To estimate the derivatives from the data polynomial regression is employed with $r = 5$ and $\delta = r\Delta t = 1.25$. Note the smaller value of r compared to the noiseless case considered above accounting for the ten-times larger sampling time used here. Figures 1(b) and 2 (diamonds, online red) show that, remarkably, SINDy identifies the correct members of the library and provides an excellent estimation of the delay time with $\tau^\star = 7$. The estimated parameters for the SINDy model (7) are reported in row (c) of Table 1, and, unsurprisingly, more strongly deviate from the true values with reconstruction error $\mathcal{E}(\tau^\star = 7)$ of 2.56%. If a longer time series with $N = 4\,000$ is used to train the SINDy model, the error is reduced to 1.94% (Table 1, row (d)), indicating that the limiting factor is the noise rather than the length of the time series.
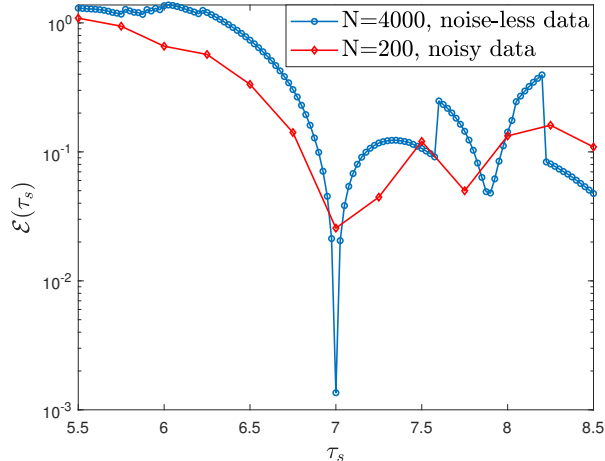
Figure 2: Reconstruction error $\mathcal{E}(\tau_s)$ as a function of the delay time τ_s, showing a clear minimum at the true delay time $\tau = 7$ for the ideal case with $N = 4\,000$ noiseless observations of x and $\dot{x}$ (open circles; online blue) and for the case of $N = 200$ noisy observations of x with noise level $\gamma = 0.02$ (diamonds; online red).

Figure 1(b) shows again the normalized cost function $C(\Xi)$ upon removal of members of the library for fixed delay time τ_s. We show results for the true delay time $\tau_s = 7$ and a for a non-optimal delay time $\tau_s = 6$. The increase of the cost function upon removal of terms appearing in the true model is less pronounced than in the ideal case of noiseless observations (cf. Figure 1(a)). We remark that the value of the cost function for iteration numbers before the removal of the monomials of the DDE (11) is significantly larger than for the ideal noiseless case. Figure 2 shows the reconstruction error $\mathcal{E}(\tau_m)$ as a function of the delay time τ_s. As in the noiseless case a clear minimum is observed at $\tau_s = 7$ corresponding to the delay of the true model. Near the minimum the reconstruction error $\mathcal{E}(\tau_s)$ is continuous. For delay times far away from the minimum the reconstruction error experiences discrete jumps, which are caused by the discrete removal of library terms for those values. The perfect accuracy of the estimation of the delay with $\tau^\star = 7$ is due to the coarse sampling time $\Delta t = 0.25$ implying that the next closest values of delays used for the optimization are $\tau_s = 6.75$ and $\tau_s = 7.25$, which both lead to a significantly larger value of the reconstruction error $\mathcal{E}(\tau_s)$. We remark that one may use interpolation to provide a more accurate estimate for the delay time at which the minimum of the reconstruction error is attained. The minimum will then not be attained necessarily at a multiple of the sampling time.

In Figure 3 we display the trajectories obtained from simulating the estimated SINDy model (7) and compare them to the trajectory of the (noiseless) true model (11), both initialized with the same initial condition as the true solution of (11). Note that the initial value $x(0) = 1$ was not part of the (noisy) observations used for training. Remarkably, the SINDy algorithm permits to recover the true solution for the observed time window even for the noise-contaminated case with trajectories which are hardly discernible with the bare eye. For longer times we will, however, observe increasing phase errors for the noisy case which does not recover the true coefficients exactly (not shown for brevity). The same SINDy model run with a close but non-optimal delay time of $\tau = 6$, however, leads to strong phase and amplitude errors of the SINDy-DDE model as seen in Figures 3(b) and 3(d).

The results presented for the simple one-dimensional toy model (11) suggest that the

SINDy-delay Algorithm 1 described in Section 2 is able to recover the dynamics of a DDE for relatively short data contaminated by moderate measurement noise at least on the time-scale covered by the observations. Indeed, the reconstruction error is only 2.56 % (see Table 1) with only $N = 200$ data measurements, compared to standard applications of SINDy with for instance with $N = 10^5$ data measurements for the Lorenz attractor example in [7, Appendix 4.2]. In the next Section we show how to use the SINDy-delay method to uncover the dynamics from a set of biological experiments where the underlying dynamical system is not known.

# 4 Application to experimental data of gene expressions in *Pseudomonas aeruginosa*

Zinc is an essential element in most living organisms and its proper dosage is vital for their survival. In bacteria, zinc is typically bound to proteins and is responsible for both structural and functional roles of those proteins [3]. Too small zinc concentrations impede on the biological functioning of these proteins. Equally, if zinc is present in excess, it becomes toxic, mainly by nonspecific bindings compromising the cellular integrity [17]. Therefore, intracellular zinc concentration must be tightly regulated. This balance of cellular concentration (also called homeostasis) is finely controlled by zinc import and export systems and their regulators.

Several strategies have evolved in *P. aeruginosa* to mitigate against strong fluctuations of environmental zinc concentrations. In particular, numerous systems composed of transmembrane complexes act to maintain zinc homeostasis [35, 15]. Like all Gram negative bacteria, *P.aeruginosa* possesses a double membrane separated by a particular compartment, the periplasm as illustrated in Figure 4. Several complexes are involved in the uptake of zinc in two stages: the first one allows transport of zinc from the outside into the periplasm, the second allows for transport from the periplasm into the cytoplasm. In presence of zinc excess, the associated import transporters are repressed, giving way to a reversed export systems. The most effective transporter is the efflux pump CzcCBA, which expels metal from the periplasm or the cytoplasm directly out of the cell (cf. Figure 4) [33, 20]. The P-type ATPase CadA on the other hand expels zinc from the cytoplasm to the periplasm [29]. (We follow here the standard convention that proteins have names starting with a capital letter whereas their associated genes have names all in small caps and are written in italics). Other export systems have been described in this bacterium, such as CzcD or YiiP, but do not appear to play a major role in zinc resistance [37, 16].

The expression of the proteins CadA is regulated by CadR that belongs to a family of transcriptional regulators known to be constitutively located on the promoter sequences of their target genes [4]. This configuration provides a fast response as follows: when the cytoplasmic zinc concentration reaches a critical value, CadR binds the metal and immediately induces *cadA* transcription [16]. The threshold of zinc concentration for the activation of this system depends directly on the zinc affinity of CadR. This threshold has not yet been determined in *P. aeruginosa* in the literature, but it may be estimated about $3 \cdot 10^{-12}$ M, as observed in other bacteria for ZntR, a CadR homolog [34]. Conversely, the efflux pump CzcCBA is activated by the CzcRS TCS, where in presence of high periplasmic concentration of zinc, the CzcS sensor activates the CzcR regulator which in turn binds the DNA, promoting the activation of its own transcription and the *czcCBA* efflux pump, but also represses *oprD* porin transcription [16]. OprD is the entry route for carbapenem antibiotics. Therefore, in presence of zinc, CzcR render the bacterium resistant to both metal and antibiotics. Interestingly, the CadA P-type ATPase appeared to be a key component for a full and timely induction of Czc-CBA, suggesting a hierarchical expression in zinc export systems [16] as shown schematically
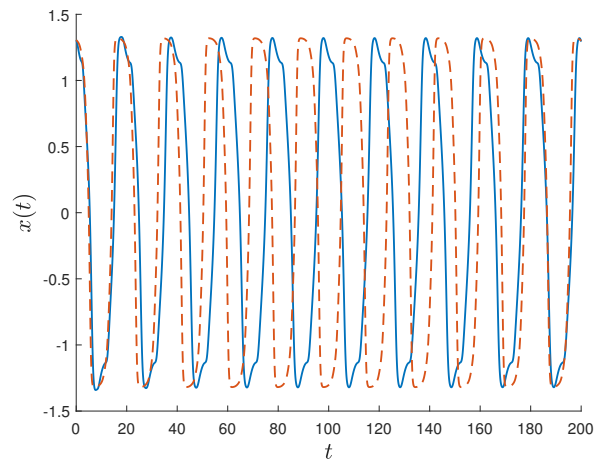
10
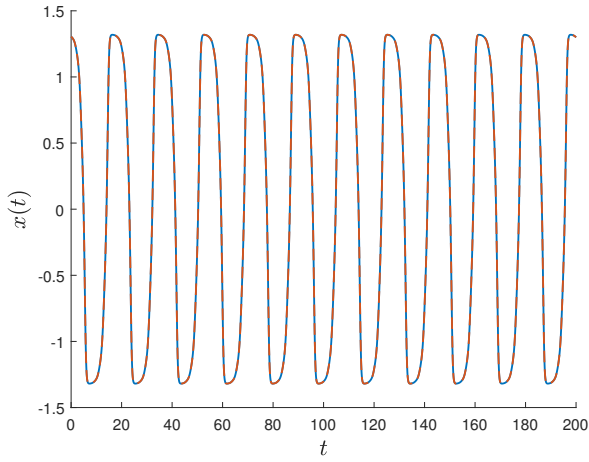
(a) noiseless observations of x and $\dot{x}$
$N = 4\,000$, optimal $\tau = 7$.
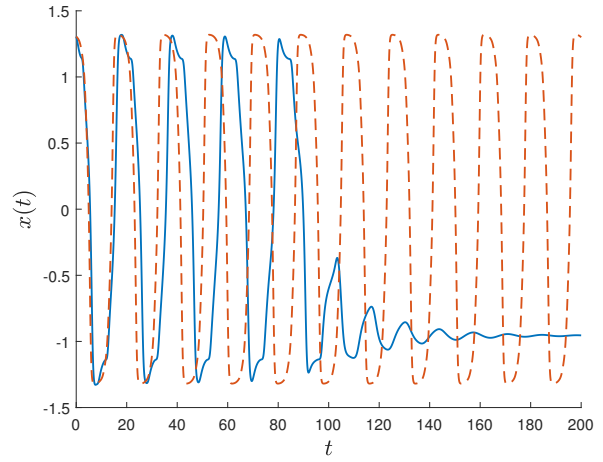
(b) noiseless observations of x and $\dot{x}$
$N = 4\,000$, non-optimal $\tau = 6$.
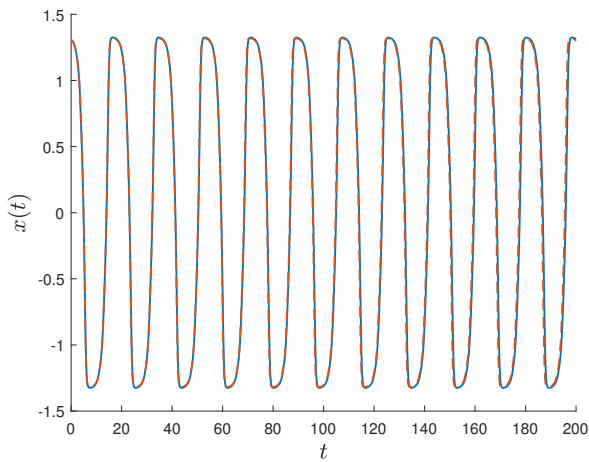
(c) noisy observations of x
$N = 200$, $\gamma = 0.02$, optimal $\tau = 7$.

(d) noisy observations of x
$N = 200$, $\gamma = 0.02$, non-optimal $\tau = 6$.

Figure 3: Comparison of the trajectories obtained from the SINDy model (7) (continuous curves; online blue) and from the true model (11) (dashed curves; online red) for optimal and non-optimal delay times, and for noiseless and noisy data points, respectively.
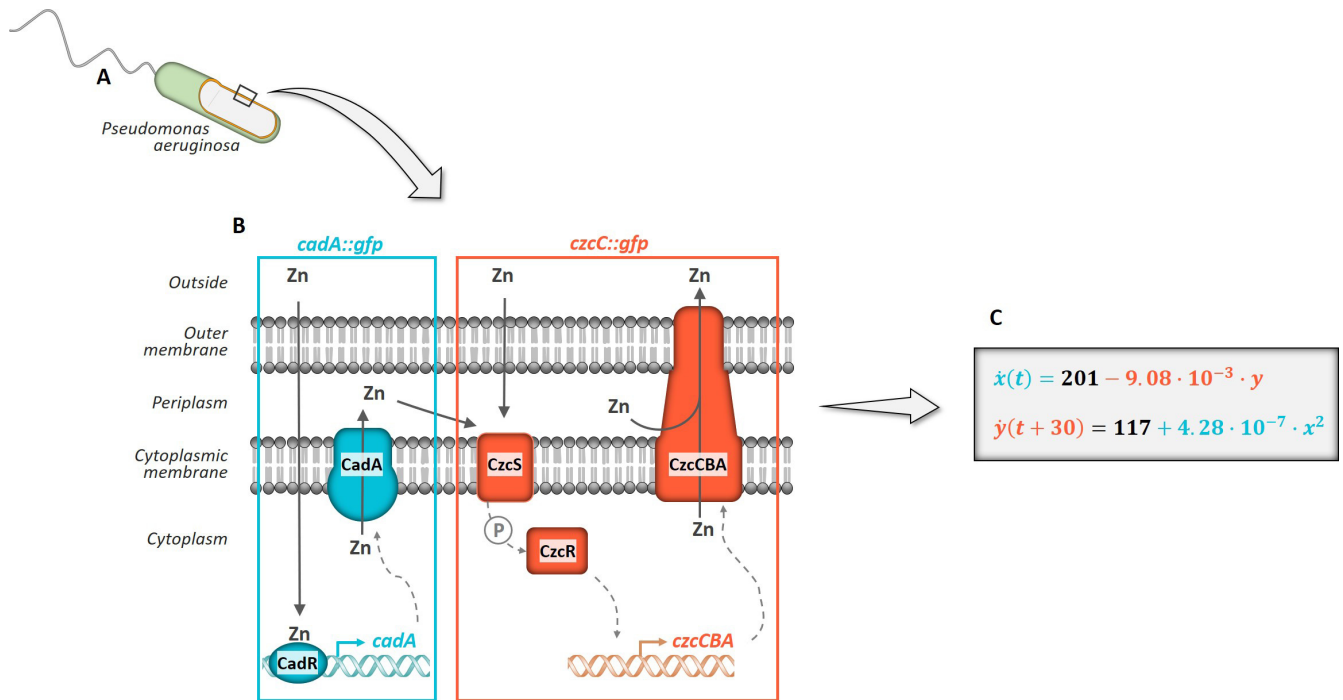
Figure 4: Sketch of the biological model of the two compartment system (TCS) for the regulation of zinc in *Pseudomonas aeruginosa*. **A)** Representation of the bacterium *Pseudomonas aeruginosa*. The square represents the location of the two membranes in which the transport systems visible in **B** are integrated in. **B)** Schematic representation of the two-steps dynamical response of the proteins CadA (blue) and CzcCBA (red) after zinc induction, adapted from [16]. As soon as the metal enters the cell, CadA is rapidly expressed by CadR, leading in a second phase to the induction of CzcCBA via the CzcRS TCS. **C)** The delay differential equation describing the dynamics of the Cad (blue color) and Czc (red color) systems after addition of 2 mM Zn obtained by the SINDy-delay method.

in Figure 4. In a zinc deficient medium, all import systems are expressed. Consequently, zinc accumulates rapidly in the cytoplasm during a metal boost. This results in the closure of the uptake machineries and at the same time the fast induction of CadA, which begins to expel zinc from the cytoplasm to the periplasm, leading subsequently to the activation of the CzcRS TCS and therefore of CzcCBA. This subsequently promotes a strong expulsion of zinc, which in turn decreases CadR activity and hence CadA expression. To better characterize and model this regulatory system, we seek a simplified two-dimensional differential equation system, describing the dynamical induction of the two agents CadA and CzcCBA. The following subsection describes the experimental set-up employed to obtain measurements for CadA and CzcCBA.

4.1 Experimental design and results

We used the transcriptional fusions *cadA::gfp* and *czcCBA::gfp* described in [16]. This method has the advantage of closely reflecting the expression of the gene of interest and naturally yields time series of experimental data. To do so the green fluorescent protein (GFP) were fused with the regulatory sequences of *cadA* or *czcC* genes, respectively. To investigate the interaction between the proteins CadA and CzcCBA, we consider a wild type (wt) strain of *P. aeruginosa* as well as mutants in which either CadA is not expressed ($\Delta cadA$) or CzcCBA is not expressed ($\Delta czcA$). Strains were independently grown in a zinc deficient M-LB medium as described in [16], for 2 hours 30 minutes before the addition of different concentrations of zinc (in the form of $ZnCl_2$). We perform experiments for various zinc concentrations with 0.5, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5 mM, a range of zinc concentrations for which the considered systems are fully induced. The fluorescence of *cadA::gfp* was measured for the wt and the $\Delta czcA$ strains. Similarly, the fluorescence of *CzcCBA::gfp* was measured in the wt and the $\Delta cadA$ strains. Fluorescent values were monitored every 5 minutes for 160 minutes and normalized by the optical density at 600nm (OD600, a standard methodology which permits to estimate the bacterial concentrations). This amounts to a short time series of 33 measurements per experiment. Each experiment is conducted three times and we report on the averages over those three experiments. In the following time $t = 0$ corresponds to the moment of the metal addition. For ease of exposition, fluorescence measurement are shifted to start with a value of 0 at time $t = 0$.

Figure 5 shows the fusion measurements for the wild type and two mutants after adding 2mM of $ZnCl_2$. In agreement with previous work [16], in the wt strain (see Figure 5a), the CadA induction drops when CzcCBA begins to be expressed, i.e. several minutes after the addition of zinc. However in the $\Delta czcA$ mutant we observe a continuous induction of CadA during the time of the experiments (see Figure 5b). The fusion results also reveal a later induction of CzcCBA in the $\Delta cadA$ strain compared to the wt strain (see Figure 5c).

4.2 SINDy-delay method to uncover CadA and CzcCBA system dynamics

The dynamics and induction intensity of CadA and CzcCBA systems depend on several factors, including intracellular (periplasmic and/or cytoplasmic) concentration of zinc, as well as on the response velocity and the metal sensitivity of their respective regulators. Moreover, experimental data obtained from transcriptional fusions are only proxies depending on GFP synthesis and its stability. For simplicity, we ignore these complex interactions and instead consider only two "boxes", one signifying all the variables involved in CadA expression (blue box in Figure 4) and one signifying those responsible of CzcCBA expression (red box in Figure 4). This simplification implies a mathematical model with only CadA and CzcCBA
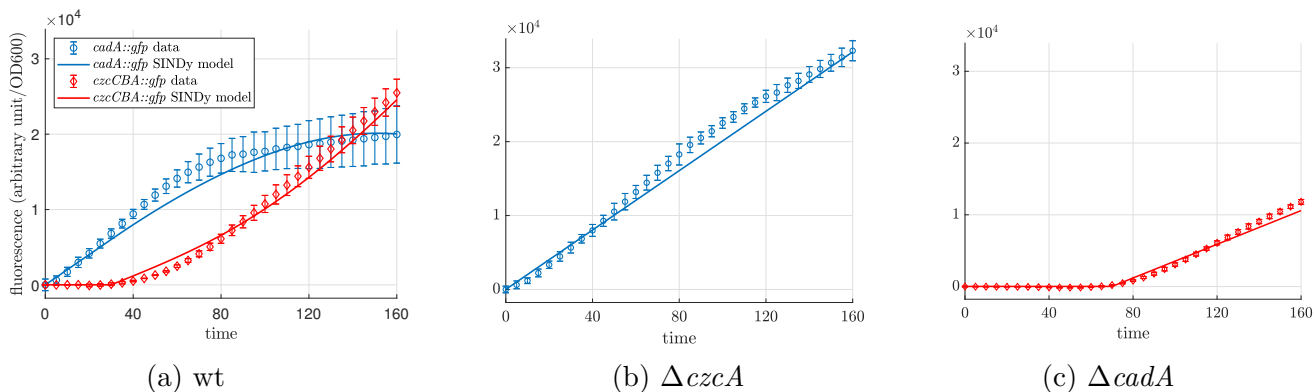
| (a) wt | (b) $\Delta czcA$ | (c) $\Delta cadA$ |

Figure 5: Fluorescence measurements after addition of 2 mM ZnCl$_2$ compared to the corresponding mathematical delay differential equation (DDE) model (solid lines). The fluorescence intensity over time is shown for the wt, $\Delta cadA$ and $\Delta czcA$ strains, containing the *cadA::gfp* (open circles; online blue) or *czcA::gfp* (diamonds; online red) fusions. The values are normalized by the optical density (OD600). Standard deviations of three independent measurements are shown. The mathematical solutions, according to the SINDy selected model, are shown in solid lines.

expressions as dependent variables. We assume that the zinc concentration remains constant during the induction experiment.

For the wild type bacteria (wt) we seek a model of the form

$$\dot{x}_{wt}(t) = f(x_{wt}, y_{wt}), \qquad \dot{y}_{wt}(t + \tau_{wt}) = g(x_{wt}, y_{wt}), \tag{12}$$

where $x(t)$ represents the fluorescence from *cadA::gfp* while $y(t)$ represents *czcA::gfp*. This form is motivated by the experimental data shown in Figure 5 where *cadA::gfp* experiences significant changes within the first minutes whereas *czcA::gfp* remains nearly constant for a significant time suggesting a delayed dynamics. For the mutant $\Delta czcA$, which lacks expression of *czcA* the dynamics is obtained by setting $y = 0$ in the above model for the wild type. We obtain

$$\dot{x}_{\Delta cz}(t) = f(x_{\Delta cz}, 0). \tag{13}$$

Similarly, for the mutant $\Delta cadA$, which lacks expression of *cadA* the dynamics is obtained by setting $x = 0$ in the above model for the wild type, and we obtain

$$\dot{y}_{\Delta ca}(t + \tau_{\Delta ca}) = g(0, y_{\Delta ca}). \tag{14}$$

We allowed here for a delay time $\tau_{\Delta ca} \neq \tau_{wt}$ accounting for the possibility that the delay may depend on the presence of the various agents present in the regulatory process. We also assume that $x(t) = y(t) = 0$ for $t < 0$ which corresponds to the natural assumption that neither CadA (x) nor CzcCBA (y) are produced when no zinc has been added yet into the growing medium, which could activate their expression (see Figure 4).

To determine the model (12)-(14), we apply the SINDy-delay method presented in Sections 2 for the fluorescence measurements of the expression kinetics experiments described in Section 4.1. To estimate the functions f and g in (12) as well as the delay times from the experimental data, we consider a library consisting of all monomials up to cubic order to approximate (12)-(14). To search for a parsimonious model only few terms selected from the library, we apply the sparsity constraints as detailed in Section 2 and search for the delay times $\tau_{\Delta ca}$ and τ_{wt} in the set $[0, 5, \ldots, 160]$ in units of minutes: we remove the less meaningful

14

Zn	Delays		Coefficients for function $x(t)$ ($cadA$::gfp)							Coefficients for function $y(t)$ ($czcA$::gfp)					Error
[mM]	τ_{wt}	$\tau_{\Delta ca}$	1	x	y	x^2	xy	y^2	xy^2	1	x	y	x^2	xy	$\mathcal{E}(\tau_{wt}, \tau_{\Delta ca})$
0.5	20	25	370	$-3.51{\cdot}10^{-2}$	$-3.70{\cdot}10^{-2}$	$1.14{\cdot}10^{-6}$	0	$1.69{\cdot}10^{-6}$	0	118	$1.61{\cdot}10^{-2}$	0	0	$-1.53{\cdot}10^{-6}$	0.0328
1	25	35	317	0	0	$-2.16{\cdot}10^{-7}$	$-3.55{\cdot}10^{-6}$	0	$1.14{\cdot}10^{-10}$	180	$5.30{\cdot}10^{-3}$	$-6.51{\cdot}10^{-3}$	0	0	0.0711
1.25	30	45	316	0	0	$-1.74{\cdot}10^{-7}$	$-2.56{\cdot}10^{-6}$	0	$7.90{\cdot}10^{-11}$	162	0	$-5.18{\cdot}10^{-3}$	$3.36{\cdot}10^{-7}$	0	0.0666
1.5	25	55	216	0	$-1.13{\cdot}10^{-2}$	0	0	0	0	122	0	0	$2.35{\cdot}10^{-7}$	0	0.11
1.75	30	65	209	0	$-1.04{\cdot}10^{-2}$	0	0	0	0	112	0	0	$3.46{\cdot}10^{-7}$	0	0.0865
2	30	70	201	0	$-9.08{\cdot}10^{-3}$	0	0	0	0	117	0	0	$4.28{\cdot}10^{-7}$	0	0.0785
2.25	35	85	200	0	$-8.48{\cdot}10^{-3}$	0	0	0	0	134	0	0	$4.10{\cdot}10^{-7}$	0	0.0730
2.5	45	95	200	0	$-6.54{\cdot}10^{-3}$	0	0	0	0	129	$1.04{\cdot}10^{-2}$	0	0	0	0.0613

Table 2: Results of the SINDy-delay method for the various zinc concentrations. Terms from the library function which were not selected for any zinc concentration are not represented.

terms of the library and stop the process when the minimum of the normalized cost function $C(\Xi)/C(0)$, which refers to equation (4), increases by more than 10 percents. Optimal delays time are found by minimizing the function $\mathcal{E}(\tau_{wt}, \tau_{\Delta ca})$ reconstruction error corresponding to (8).

This process is applied for all experiments with the various zinc concentrations. In Figure 6 we show results for the 2 mM induction of zinc. Figure 6(a) shows the increase of the normalized cost function upon removal of both x and y components. Figure 6(b) shows the reconstruction error $\mathcal{E}(\tau_s)$ with a minimum error of 7.8% for $\tau_{wt} = 30$ and $\tau_{\Delta ca} = 70$ minutes. In particular, we obtain from the SINDy-delay methodology the following delay differential equation model for a concentration of 2 mM of zinc,

$$\dot{x}_{wt}(t) = 201 - 9.08 \cdot 10^{-3}\, y_{wt}, \tag{15}$$

$$\dot{y}_{wt}(t + \tau_{wt}) = 117 + 4.28 \cdot 10^{-7}\, x_{wt}^2 \tag{16}$$

and

$$\dot{x}_{\Delta cz} = 201,$$
$$\dot{y}_{\Delta ca}(t + \tau_{\Delta ca}) = 117 \tag{17}$$

with $\tau_{wt} = 30, \tau_{\Delta ca} = 70$. In Figure 5, solutions of the DDE model (15)-(16) and of (17) are plotted and compared with experimental data for 2 mM ZnCl$_2$, which shows a high degree of similarity with a reconstruction error of 7.85%. The complete results for all zinc concentrations tested (from 0.5 to 2.5 mM) are shown in Table 2. We remark that the coefficient of the linear and quadratic terms in (15),(16), are of the order of 10^{-2} and 10^{-7}, respectively; although their coefficients are small, their presence is crucial. Such small coefficients are hard to detect when employing standard thresholding procedures. This illustrates the advantage of our method based to promote sparsity outlined in Remark 2.2.

DDE model accuracy and consistency We observe in Table 2 that for all zinc concentrations the SINDy-delay method yields DDE models with reconstruction errors smaller than 11%. The SINDy model matches the experimental data very well and is biologically consistent for all ZnCl$_2$ concentrations. This is notable given the very short length of the experimental time series with $N = 33$ data. Remarkably, for moderate ZnCl$_2$ concentrations between 1.5 mM and 2.25 mM (emphasized in dashed lines in Table 2), a unified SINDy DDE model arises which benefits from the sparsity feature, with only the terms 1,y and x^2 selected, allowing for a biologically consistent interpretation of the terms. Importantly, the signs of the associated coefficients are consistent with the biological model: the coefficient associated
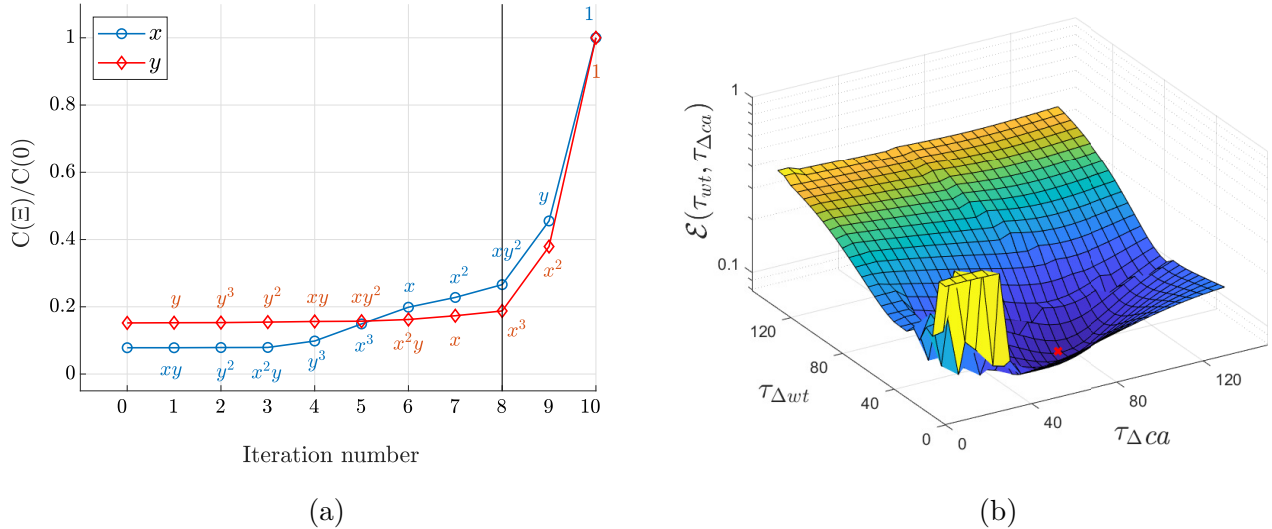
Figure 6: (a) Cost function $C(\Xi)$ (normalized by $C(0)$) against removed monomials for fixed optimal delays, $\tau_{wt} = 30$ (open circles; online blue) and $\tau_{\Delta ca} = 70$ (diamonds; online red). The vertical black line indicates the iteration number where the process is stopped. At iteration 8 the cost function has increased more than 10% for both components. (b) Reconstruction error $\mathcal{E}(\tau_{wt}, \tau_{\Delta ca})$ showing a minimum value equal to $7.8 \cdot 10^{-2}$ for $\tau_{wt} = 30$ and $\tau_{\Delta ca} = 70$, indicated by a red cross.

with the linear term in y in (15), which describes the influence of y (CzcCBA) on x (CadA), is negative in agreement with the biological model where CzcCBA represses CadA. Similarly, the coefficient associated with the x^2 term in (16), which describes the influence of x (CadA) on y (CzcCBA), is positive in agreement with the biological model where CadA accelerates the expression of CzcCBA (12).

For low $ZnCl_2$ concentrations smaller than 1.25 mM the SINDy models are not as sparse, involving more terms than for the moderate concentrations (for instance, up to five functions $1, x, y, x^2, y^2$ for x (CadA) at 0.5 mM of zinc), while for the highest considered $ZnCl_2$ concentration, the different term x is selected in place of x^2. We also remark that the SINDy model is likely to model the response of *P. aeruginosa* to a boost in zinc only for the time duration of the experiment. Indeed, the SINDy models in Table 2 exhibit unphysical negative CadA and CzcA concentrations for all considered $ZnCl_2$ concentrations if a time larger than 800 minutes would be considered (not shown here) in place of the time of 160 minutes considered in the experiments. This suggests that the simplified two box model may be insufficient to capture the impact of the induced stress for longer times, and additional components or mechanisms need to be included in the modelling.

CadA is essential for maintaining a rapid expression of CzcCBA Consider the range of zinc concentrations from 1.5 to 2.25 mM as emphasised with dashed lines in Table 2. A remarkable observation is that the coefficients computed from the SINDy-delay method are only weakly sensitive to the applied zinc concentration, with the exception of the delay time $\tau_{\Delta ca}$, which increases linearly with the zinc concentration, as shown in Figure 7. This linear increase of the delay time $\tau_{\Delta ca}$ in the absence of CadA suggests that the protein CadA is particularly necessary for a rapid zinc response and suggests that the positive effect of CadA on the efflux pump is all the more important as the zinc concentration is high. The OD600
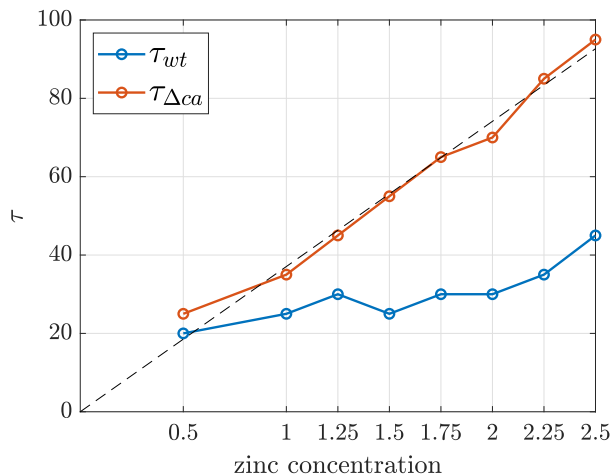
16

Figure 7: Estimated delay times $\tau_{wt}, \tau_{\Delta ca}$ (in minutes) as a function of the zinc concentration (in mM). Remarkably, we observe a linear increase as a function of the zinc concentration of the delay time $\tau_{\Delta ca} = \alpha \cdot [\text{ZnCl}_2]$ with $\alpha = 37.1$ min mM^{-1} (linear regression in dashed line with slope 37.1).

measurement allows the counting of cells independently of whether they are alive or dead. Thanks to colony counting and quantification of cell viability at concentrations of 1.25 mM and 2 mM ZnCl$_2$ after 160 min of incubation (not displayed here for brevity), we observed that the same number of living cells are detected, and hence this difference in delay under different zinc concentrations cannot be attributed to a differential mortality between the $\Delta cadA$ and the wt strains. Biologically, this could reflect a reasonable mechanism whereby the bacterium wants to react as quickly as possible to a stress regardless of its intensity.

5 Conclusion

In this paper, we extended the SINDy methodology introduced in [7] to the case of delay differential equations with a focus on short and noise-contaminated data. To construct the temporal derivatives from noisy measurements we employed a simple denoising procedure based on polynomial regression (Remark 2.1). We further introduced a stopping criterion to promote sparsity which avoids having to introduce sensitive threshold parameters (Remark 2.2). To estimate the temporal delay we applied a bilevel optimization whereby first standard SINDy method is applied for a range of fixed delay times, and then subsequently the optimal delay time is determined by the delay time yielding the minimal reconstruction error. We showed that our method is able to reliably uncover the DDE from noisy data obtained from a known toy model.

Applying the SINDy-delay methodology to model the dynamics of the *Pseudomonas aeruginosa* zinc response from a limited amount of measurement highlighted the subtle interactions between the Cad and Czc regulatory systems. In particular, the SINDy DDE model revealed the importance of CadA on CzcCBA induction for minimizing the time required for the bacterium to respond effectively to a sudden zinc excess. The compatibility between the results of the SINDy DDE models and the biological data supports the hypothesis that the dynamical mechanism of resistance to moderate boosts of zinc can be explained by the interaction of only two systems, namely CadA and CzcCBA. Our results motivate further investigations of this dynamics. The present work was performed over 160 minutes after the

17

metal induction and illustrates only the initial establishment of resistance. Additional experimental data on longer times, which require continuous cultures in a chemostat and a more sensitive method to monitor the *cadA* and *czcCBA* transcriptional expressions would make it possible to compare these mathematical predictions with the biological situation.

References

[1] E. Alm, K. Huang, and A. Arkin. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput. Biol.*, 2(11):1–14, 11 2006.

[2] L. Biggio, T. Bendinelli, A. Neitz, A. Lucchi, and G. Parascandolo. Neural symbolic regression that scales. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 936–945. PMLR, 2021.

[3] D. Blencowe and A. Morby. Zn(II) metabolism in prokaryotes. *FEMS Microbiol Rev.*, 27(2-3):291–311, 2003.

[4] N. Brown, J. Stoyanov, S. Kidd, and J. Hobman. The MerR family of transcriptional regulators. *FEMS Microbiol Rev.*, 27(2-3):145–163, 2003.

[5] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz. Chaos as an intermittently forced linear system. *Nat. Commun.*, 8(1):19, 2017.

[6] S. L. Brunton and J. N. Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control.* Cambridge University Press, 2019.

[7] 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(15):3932–3937, 2016.

[8] D. Capdevila, J. Wang, and D. Giedroc. Bacterial strategies to maintain zinc metallostasis at the host-pathogen interface. *J. Biol. Chem.*, 291(40):20858–20868, 11 2016.

[9] 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(1):312–333, 2019.

[10] R. Chartrand. Numerical differentiation of noisy, nonsmooth data. *ISRN Appl. Math.*, 2011, 5 2011.

[11] J. Cushing. Periodic solutions of two species interaction models with lags. *Math. Biosci.*, 31(1):143–156, 1976.

[12] 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(2):022310, 2017.

[13] G. Dieppois, V. Ducret, O. Caille, and K. Perron. The transcriptional regulator CzcR modulates antibiotic resistance and quorum sensing in *Pseudomonas aeruginosa*. *PLoS One*, 7(5):e38148., 2012.

[14] K. Y. Djoko, C. L. Ong, M. J. Walker, and A. G. McEwan. The role of copper and zinc toxicity in innate immune defense against bacterial pathogens. *J. Biol. Chem.*, 290(31):18954–18961, 2015.

[15] V. Ducret, M. Abdou, C. Goncalves Milho, S. Leoni, O. Martin-Pelaud, A. Sandoz, I. Segovia Campos, M. L. Tercier-Waeber, M. Valentini, and K. Perron. Global analysis of the zinc homeostasis network in *Pseudomonas aeruginosa* and its gene expression dynamics. *Front Microbiol.*, 12:739988, 2015.

[16] V. Ducret, M. R. Gonzalez, S. Leoni, M. Valentini, and K. Perron. The CzcCBA efflux system requires the CadA p-type ATPase for timely expression upon zinc excess in *Pseudomonas aeruginosa. Front. Microbiol.*, 11:911, 2020.

[17] A. Foster, D. Osman, and N. Robinson. Metal preferences and metallation. *J Biol Chem.*, 289(41):28095–28103, 2014.

[18] H. Gao, W. Dai, L. Zhao, J. Min, and F. Wang. The role of zinc and zinc homeostasis in macrophage function. *J Immunol Res.*, 2018:6872621, 2018.

[19] D. S. Glass, X. Jin, and I. H. Riedel-Kruse. Nonlinear delay differential equations and their application to modeling biological network motifs. *Nat. Commun.*, 12(1):1788, 2021.

[20] M. Goldberg, T. Pribyl, S. Juhnke, and D. H. Nies. Energetics and topology of CzcA, a cation/proton antiporter of the resistance-nodulation-cell division protein family. *J Biol Chem.*, 274(37):26065–26070, 1999.

[21] G. A. Gottwald. Bifurcation analysis of a normal form for excitable media: Are stable dynamical alternans on a ring possible? *Chaos*, 18(1):013129, 2008.

[22] G. A. Gottwald and L. Kramer. A normal form for excitable media. *Chaos*, 16(1):013122, 2006.

[23] S. A. Gourley. Travelling fronts in the diffusive Nicholson's blowflies equation with distributed delays. *Math. Comput. Model.*, 32(7):843–853, 2000.

[24] R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F. A. Massucci, M. Miranda, J. Pallarès, and M. Sales-Pardo. A bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci. Adv.*, 6(5):eaav6971, Jan 2020.

[25] I. Jurado-Martín, M. Sainz-Mejías, and S. McClean. *Pseudomonas aeruginosa*: An audacious pathogen with an adaptable arsenal of virulence factors. *Int. J. Mol. Sci.*, 22(6):3128, 2021.

[26] A. Keane, B. Krauskopf, and C. M. Postlethwaite. Climate models with delay differential equations. *Chaos*, 27(11):114309, 2017.

[27] T. E. Kehl-Fie and E. P. Skaar. Nutritional immunity beyond iron: a role for manganese and zinc. *Curr. Opin. Chem. Biol.*, 14(2):218–224, 2010.

[28] K. G. Kerr and A. M. Snelling. *Pseudomonas aeruginosa*: a formidable and ever-present adversary. *J. Hosp. Infect.*, 73(4):338–344, 2009.

[29] S. Lee, E. Glickmann, and D. Cooksey. Chromosomal locus for cadmium resistance in Pseudomonas putida consisting of a cadmium-transporting ATPase and a MerR family response regulator. *Appl Environ Microbiol.*, 67(4):1437–1444, 2001.

[30] J.-C. Loiseau and S. L. Brunton. Constrained sparse Galerkin regression. *J. Fluid Mech.*, 838:42–67, 2018.

[31] Z. Lonergan and E. Skaar. Nutrient zinc at the host-pathogen interface. *Trends Biochem Sci.*, 44(12):1041–1056, 2019.

[32] MATLAB. *version 9.10.0 (R2019b)*. The MathWorks Inc., Natick, Massachusetts, 2019.

[33] D. H. Nies, A. Nies, L. Chu, and S. Silver. Expression and nucleotide sequence of a plasmid-determined divalent cation efflux system from alcaligenes eutrophus. *Proc. Natl. Acad. Sci. U.S.A.*, 86(19):7351–7355, 1989.

[34] D. Osman, C. Piergentili, J. Chen, B. Chakrabarti, A. Foster, E. Lurie-Luke, T. Huggins, and N. Robinson. Generating a metal-responsive transcriptional regulator to test what confers metal sensing in cells. *J Biol Chem.*, 290(32):19806–22, 2015.

[35] V. G. Pederick, B. A. Eijkelkamp, S. L. Begg, M. P. Ween, L. J. McAllister, J. C. Paton, and C. A. McDevitt. ZnuA and zinc homeostasis in *Pseudomonas aeruginosa*. *Sci. Rep.*, 5:13139, 2015.

[36] K. Perron, O. Caille, C. Rossier, C. Van Delden, J. L. Dumas, and T. Köhler. CzcR-CzcS, a two-component system involved in heavy metal and carbapenem resistance in *Pseudomonas aeruginosa*. *J. Biol. Chem.*, 279(10):8761–8768, 2004.

[37] A. Salusso and D. Raimunda. Defining the roles of the cation diffusion facilitators in fe2+/zn2+ homeostasis and establishment of their participation in virulence in *Pseudomonas aeruginosa*. *Front Cell Infect Microbiol.*, 7:84, 2017.

[38] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[39] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[40] M. Sorokina, S. Sygletos, and S. Turitsyn. Sparse identification for nonlinear optical communication systems: SINO method. *Opt. Express*, 24(26):30433–30443, Dec 2016.

[41] S. L. Stafford, N. J. Bokil, M. E. Achard, R. Kapetanovic, M. Schembri, A. G. McEwan, and M. Sweet. Metal ions in macrophage antimicrobial pathways: emerging roles for zinc and copper. *Biosci Rep.*, 33(4):e00049, 2013.

[42] C. K. Stover, X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature*, 406(6799):959–964, 2000.

[43] M. J. Suarez and P. S. Schopf. A delayed action oscillator for ENSO. *J. Atmos. Sci.*, 45:3283–7, 1988.

[44] E. Tacconelli, E. Carrara, A. Savoldi, S. Harbarth, M. Mendelson, D. Monnet, C. Pulcini, G. Kahlmeter, J. Kluytmans, Y. Carmeli, M. Ouellette, K. Outterson, J. Patel, M. Cavaleri, E. Cox, C. Houchens, M. Grayson, P. Hansen, N. Singh, U. Theuretzbacher, N. Magrini, and WHO Pathogens Priority List Working Group. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis.*, 18(3):318–327, 2018.

[45] S. Terrien, V. A. Pammi, B. Krauskopf, N. G. R. Broderick, and S. Barbay. Pulse-timing symmetry breaking in an excitable optical system with delay. *Phys. Rev. E*, 103:012210, Jan 2021.

[46] S.-M. Udrescu and M. Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.*, 6(16):eaay2631, 2020.

[47] F. Van Breugel, J. N. Kutz, and B. W. Brunton. Numerical differentiation of noisy data: A unifying multi-objective optimization framework. *IEEE Access*, 8:196865–196877, 2020.

[48] G. Wahba. Smoothing noisy data with spline functions. *Numer. Math.*, 24(5):383–393, 1975.

[49] F.-B. Wang, S. Gourley, and Y. Xiao. An integro-differential equation with variable delay arising in machine tool vibration. *SIAM J. Appl. Math.*, 79(1):75–94, 2019.