Rheo-SINDy: Finding a Constitutive Model from Rheological Data for Complex Fluids Using Sparse Identification for Nonlinear Dynamics

Takeshi Sato[1,*,†], Souta Miyamoto[2,†], and Shota Kato[3]

[1]Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan
[2]Department of Chemical Engineering, Graduate School of Engineering, Kyoto University, Kyoto 615-8510, Japan
[3]Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

E-mail: takeshis@scl.kyoto-u.ac.jp

*Author to whom any correspondence should be addressed.
†These authors contributed equally to this work.
Keywords: sparse identification, rheology, constitutive model

Abstract. Rheology plays a pivotal role in understanding and predicting material behavior by discovering governing equations that relate deformation and stress, known as constitutive equations. Despite the critical importance of constitutive equations in predicting dynamics of complex fluids, a systematic methodology for deriving these equations from available data has remained a significant challenge in the field. To overcome the problem, we propose a novel method named *Rheo*-SINDy, which employs the sparse identification of nonlinear dynamics (SINDy) for discovering constitutive models from rheological data. *Rheo*-SINDy was applied to five distinct scenarios, including four cases with well-established constitutive equations and one without predefined equations. Our results demonstrate that *Rheo*-SINDy successfully identifies accurate models for the known constitutive equations and derives physically plausible approximate models for the scenario with the unknown one. These findings validate the robustness of *Rheo*-SINDy in handling real-world data complexities and underscore its efficacy as a powerful tool for advancing the development of data-driven approaches in rheology.

1. Introduction

Mathematical models grounded in physical laws are indispensable across science and engineering, offering profound insights into the behavior of complex systems. These models clarify the underlying mechanisms governing system dynamics and empower predictions and innovations in technology and natural science. Traditionally, model derivation has leaned heavily on theoretical and empirical knowledge, often requiring expert knowledge and intuition. Data-driven methods have become capable of assisting in developing mathematical models and constructing models that provide advanced predictions [1]. These data-driven methods involve the sparse identification

methods [2–5], symbolic regression methods [6–11], and physics-informed machine learning methods [12–15]. These methods have emerged as powerful tools for deriving governing equations directly from data, overcoming the limitations of conventional expert-dependent approaches.

Rheology is one of the scientific fields that address the properties of flowing matter, which plays a crucial role in many industries, such as designs of chemical processes, by providing insights into flow behaviors of complex fluids. One of the roles of rheology is to discover or derive governing equations that relate deformation and stress, referred to as *constitutive equations* [16]. From an engineering perspective, accurate constitutive equations are necessary to predict the flows of complex fluids under complex boundary conditions. Nevertheless, it is generally difficult to theoretically obtain constitutive equations for complex fluids. Instead, mesoscopic coarse-grained models, which are based on molecular theories, have been explored in the field of rheology. For example, for polymeric liquids, standard molecular theories have been proposed [17,18], on which refined mesoscopic models have been constructed [19–21]. In these models, the motion of individual (coarse-grained) molecules is numerically tracked. Although these models require significantly more computational time compared to constitutive equations, they can reproduce (nearly) accurate rheological data. Despite these advancements, a clear methodology for obtaining constitutive equations from available data remains elusive.

Data-driven methods are powerful approaches for addressing the aforementioned challenges in rheology. Indeed, such methods have enhanced rheological studies such as constitutive modeling, flow predictions of complex fluids, and model selection [22, 23]. Some applications have successfully identified constitutive relations of complex fluids or governing equations to predict the dynamics of fluids with knowledge of rheology. These studies have employed neural networks (NN), including deep NN [24], graph NN [25], recurrent NN [26], physics-informed NN [27–29], multi-fidelity NN [30], and tensor basis NN [31]. Gaussian processes (GP) have also been employed, for example, for strain-rate dependent viscosity [32] or for viscoelastic properties [33–36].

Despite the success of NNs and GPs, their black-box nature often obscures the underlying physics, making symbolic regression techniques more appealing for transparency and interoperability. These methods, such as the sparse identification of nonlinear dynamics (SINDy) [2], have been frequently utilized to track (reduced order) dynamics in the field of fluid mechanics [37]. Inspired by these successes, symbolic regression methods have recently started to be used in the field of rheology as well. For example, Mohammadamin and coworkers [38] relied on SINDy for flexibly identifying the constitutive equations of an elasto-visco-plastic fluid. Nevertheless, although there are several attempts along this line, a comprehensive study to test SINDy for rheological data has not yet been conducted.

In this study, we employ SINDy to find constitutive models from rheological data, which we call as *Rheo*-SINDy. After verifying the performance of *Rheo*-SINDy when the constitutive equations are known, we apply *Rheo*-SINDy to problems where the constitutive equations are unknown. The details are shown below.

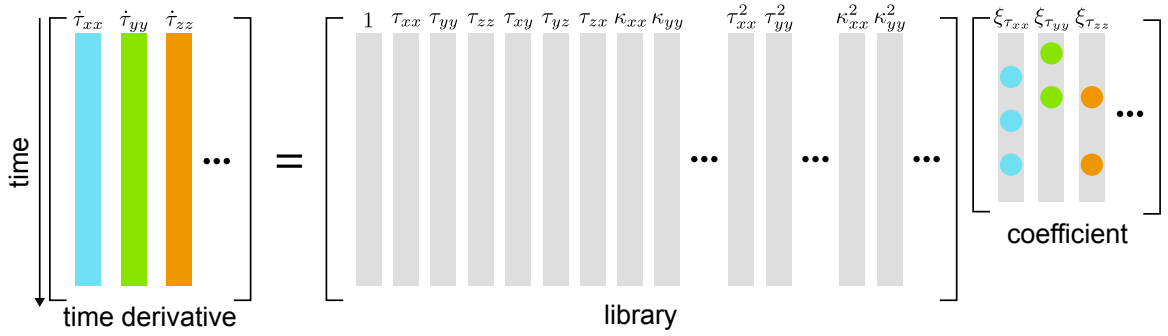


Figure 1. Schematic illustration of *Rheo*-SINDy.

2. Methods

We use a data-driven method known as a sparse identification of nonlinear dynamics (SINDy), which was originally developed by Brunton and coworkers [2]. In this study, we attempt to obtain constitutive equations of complex fluids using SINDy. Here, we briefly explain the basic concepts of SINDy.

We consider dynamical systems generally expressed by the following differential equation:

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{f}[\boldsymbol{x}(t)], \tag{1}$$

where the vector $\boldsymbol{x}(t)$ represents the state of a system at time t and the function $\boldsymbol{f}[\boldsymbol{x}(t)]$ determines the dynamics of the state $\boldsymbol{x}(t)$. The basic idea of SINDy is to find dominant terms for describing the dynamics out of numerous candidates using a sparse identification method. One can determine the (sparse) representation of $\boldsymbol{f}$ by a dataset including a collection of $\boldsymbol{x}(t)$ and $\dot{\boldsymbol{x}}(t)$ (the time derivative of $\boldsymbol{x}(t)$). The regression to points of $\boldsymbol{x}(t)$ and $\dot{\boldsymbol{x}}(t)$ is computed with sparsity-promoting techniques, such as ℓ_1-regularization.

In the rheological community, it is of great importance to determine a relation between stress and strain rate. This relation is a so-called constitutive model or constitutive equation. Most constitutive equations are differential equations that depend on the (extra) stress tensor $\boldsymbol{\tau}$ and velocity gradient tensor $\boldsymbol{\kappa}$. In this study, we prefer to use the so-called extra stress tensor $\boldsymbol{\tau}$ as the stress tensor because this stress tensor is $\boldsymbol{\tau} = \boldsymbol{0}$ at equilibrium, which is convenient for SINDy regression. The total stress tensor $\boldsymbol{\sigma}$ can be obtained by the relation $\boldsymbol{\sigma} = \boldsymbol{\tau} + G\boldsymbol{I}$, where G is the modulus and $\boldsymbol{I}$ is the unit tensor. A general form for such constitutive equations can be written as

$$\frac{\mathrm{d}\boldsymbol{\tau}(t)}{\mathrm{d}t} = \dot{\boldsymbol{\tau}}(t) = \boldsymbol{f}[\boldsymbol{\tau}(t), \boldsymbol{\kappa}(t)]. \tag{2}$$

Here, $\boldsymbol{\kappa}(t)$ is a control variable during rheological measurements. We use SINDy algorithm to find constitutive equations for complex fluids and we refer to this technique as *Rheo*-SINDy.

The training data needed to *Rheo*-SINDy are transient stress data $\boldsymbol{T}$ and those time derivatives $\dot{\boldsymbol{T}}$, which can be summarized as the following two matrices:

$$\boldsymbol{T} = \left[\begin{array}{cccc} \boldsymbol{t}_{xx} & \boldsymbol{t}_{yy} & \cdots & \boldsymbol{t}_{zx} \end{array} \right] = \left[\begin{array}{cccc} \tau_{xx}(t_1) & \tau_{yy}(t_1) & \cdots & \tau_{zx}(t_1) \\ \tau_{xx}(t_2) & \tau_{yy}(t_2) & \cdots & \tau_{zx}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{xx}(t_n) & \tau_{yy}(t_n) & \cdots & \tau_{zx}(t_n) \end{array} \right] \tag{3}$$

and

$$\dot{\boldsymbol{T}} = \left[\begin{array}{cccc} \dot{\boldsymbol{t}}_{xx} & \dot{\boldsymbol{t}}_{yy} & \cdots & \dot{\boldsymbol{t}}_{zx} \end{array} \right] = \left[\begin{array}{cccc} \dot{\tau}_{xx}(t_1) & \dot{\tau}_{yy}(t_1) & \cdots & \dot{\tau}_{zx}(t_1) \\ \dot{\tau}_{xx}(t_2) & \dot{\tau}_{yy}(t_2) & \cdots & \dot{\tau}_{zx}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\tau}_{xx}(t_n) & \dot{\tau}_{yy}(t_n) & \cdots & \dot{\tau}_{zx}(t_n) \end{array} \right], \tag{4}$$

where $\boldsymbol{t}_{\mu\nu}$ ($\mu\nu \in \{xx, yy, zz, xy, yz, zx\}$) is the column of $\boldsymbol{T}$, and we take the stress data for n sequential times. The time derivatives of the stress data $\dot{\boldsymbol{T}}$ are computed by a numerical differentiation method. In this study, we apply the velocity gradient $\boldsymbol{\kappa}$ to systems of prescribed constitutive equations or mesoscopic models for viscoelastic fluids to take the stress data determined by the states of the systems. The data of the velocity gradient tensor $\boldsymbol{K}$ are summarized as

$$\boldsymbol{K} = \left[\begin{array}{cccc} \boldsymbol{k}_{xx} & \boldsymbol{k}_{yy} & \cdots & \boldsymbol{k}_{zx} \end{array} \right] = \left[\begin{array}{cccc} \kappa_{xx}(t_1) & \kappa_{yy}(t_1) & \cdots & \kappa_{zx}(t_1) \\ \kappa_{xx}(t_2) & \kappa_{yy}(t_2) & \cdots & \kappa_{zx}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{xx}(t_n) & \kappa_{yy}(t_n) & \cdots & \kappa_{zx}(t_n) \end{array} \right], \tag{5}$$

where $\boldsymbol{k}_{\mu\nu}$ ($\mu, \nu \in \{x, y, z\}$) is the column of $\boldsymbol{K}$.

In *Rheo*-SINDy, we construct a library matrix of functions, denoted as $\boldsymbol{\Theta}$, which can include various nonlinear functions. $\boldsymbol{\Theta}$ is expressed as

$$\boldsymbol{\Theta} = \left[\begin{array}{ccccccc} \boldsymbol{1} & \boldsymbol{T} & \boldsymbol{K} & (\boldsymbol{T} \otimes \boldsymbol{T}) & (\boldsymbol{T} \otimes \boldsymbol{K}) & (\boldsymbol{K} \otimes \boldsymbol{K}) & \cdots \end{array} \right], \tag{6}$$

where $\boldsymbol{T} \otimes \boldsymbol{K}$, for example, denotes all possible combinations of the products of the row components in $\boldsymbol{T}$ and $\boldsymbol{K}$ for each time t_i ($1 \leq i \leq n$). We note that $\boldsymbol{\Theta}$ can incorporate not only polynomials but also other functions, such as sinusoidal functions. Using these expressions, we can substitute Eq. (2) as

$$\dot{\boldsymbol{T}} = \boldsymbol{\Theta}\boldsymbol{\Xi}, \tag{7}$$

where $\boldsymbol{\Xi}$ is the coefficient matrix written as

$$\boldsymbol{\Xi} = \left[\begin{array}{cccc} \boldsymbol{\xi}_{xx} & \boldsymbol{\xi}_{yy} & \cdots & \boldsymbol{\xi}_{zx} \end{array} \right] = \left[\begin{array}{cccc} \xi_{xx,1} & \xi_{yy,1} & \cdots & \xi_{zx,1} \\ \xi_{xx,2} & \xi_{yy,2} & \cdots & \xi_{zx,2} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{xx,N_{\boldsymbol{\Theta}}} & \xi_{yy,N_{\boldsymbol{\Theta}}} & \cdots & \xi_{zx,N_{\boldsymbol{\Theta}}} \end{array} \right]. \tag{8}$$

where $N_{\boldsymbol{\Theta}}$ is the total number of library functions.

To determine the coefficient matrix $\boldsymbol{\Xi}$, we solve the following optimization problem:

$$\hat{\boldsymbol{\xi}}_{\mu\nu} = \underset{\boldsymbol{\xi}_{\mu\nu}}{\operatorname{argmin}} \|\dot{\boldsymbol{t}}_{\mu\nu} - \boldsymbol{\Theta}\boldsymbol{\xi}_{\mu\nu}\|_2^2 + R(\boldsymbol{\xi}_{\mu\nu}), \tag{9}$$

where $\hat{\boldsymbol{\xi}}_{\mu\nu}$ is the optimized sparse vector, $||\cdots||_2$ is the ℓ_2 norm defined as

$$||\boldsymbol{x}||_2 = \left(\sum_i x_i^2\right)^{1/2}, \tag{10}$$

and $R(\boldsymbol{\xi}_{\mu\nu})$ is the regularization term. To obtain a sparse solution of $\boldsymbol{\Xi}$, we apply the following five methods [37]: (i) the sequentially thresholded least square algorithm (STLSQ), (ii) sequentially thresholded Ridge regression (STRidge), (iii) least absolute shrinkage and selection operator (Lasso), (iv) Elastic-Net (E-Net), and (v) adaptive-Lasso (a-Lasso). These methods employ different regularization terms to obtain sparse solutions (see Sec. S1 in the supporting information for detail). Each method has a hyperparameter α to penalize the solution complexity, which is to be tuned for obtaining good predictive yet parsimonious representations. For this purpose, we test various α values and pick an appropriate value of α that gives a small loss value and the (nearly) correct number of terms for known constitutive equations.

In this study, we limit ourselves to *shear* rheological measurements that give fundamental rheological properties. Under shear flow, among the components of $\boldsymbol{\kappa}$, only κ_{xy} has non-zero values. Here, x is the velocity direction, and y is the velocity gradient direction. Since the major stress components are τ_{xx}, τ_{yy}, τ_{zz}, and τ_{xy} under shear flow, we only use these components to conduct *Rheo*-SINDy.

3. Case Studies

For case studies, we first test whether *Rheo*-SINDy can find appropriate constitutive equations from training data generated by phenomenological constitutive equations, namely the Upper Convected Maxwell (UCM) model and the Giesekus model (the details are summarized in Sec. S2 in the supporting information). Subsequently, we apply *Rheo*-SINDy to constitutive models for the Dumbbell model, which is the most basic mesoscale model of viscoelastic fluids. This section provides a brief overview of the models used in this study and the conditions for creating the datasets.

3.1. Fundamental Equations of Dumbbell Models

The dumbbell-based models have been widely utilized in numerous previous studies for the computation of viscoelastic fluids and are considered a standard mesoscopic model for viscoelastic fluids [39]. As illustrated in Fig. 2, a dumbbell consists of two beads (indexed as 1 or 2) and a spring that connects them. The Langevin equations for the positions of the two beads $\boldsymbol{r}_{1/2}(t)$ can be written as

$$\zeta\left[\frac{\mathrm{d}\boldsymbol{r}_i(t)}{\mathrm{d}t} - \boldsymbol{\kappa}\cdot\boldsymbol{r}_i(t)\right] = -h(t)\left\{\boldsymbol{r}_i(t) - \boldsymbol{r}_j(t)\right\} + \boldsymbol{F}_{\mathrm{B}i}(t), \tag{11}$$

with $(i, j) = (1, 2)$ or $(2, 1)$. Here, ζ is the friction coefficient, $h(t)$ is the spring strength, and $\boldsymbol{F}_{\mathrm{B}i}(t)$ is the Brownian force acting on the bead i. The time evolution equation for

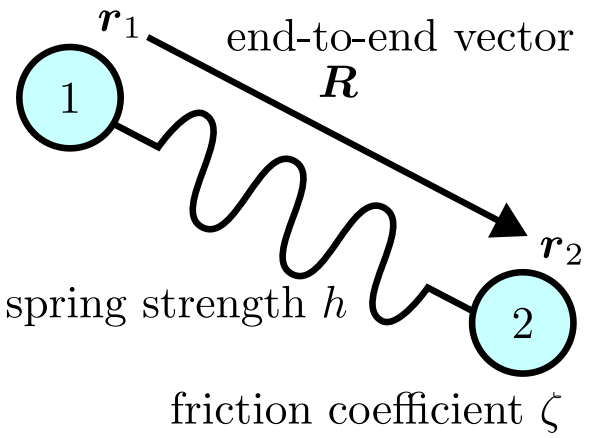


Figure 2. Schematic illustration of the dumbbell model.

the end-to-end vector $\boldsymbol{R}(t)$ $(= \boldsymbol{r}_2(t) - \boldsymbol{r}_1(t))$ of the beads is thus obtained as

$$\zeta \left[\frac{\mathrm{d}\boldsymbol{R}(t)}{\mathrm{d}t} - \boldsymbol{\kappa} \cdot \boldsymbol{R}(t) \right] = -2h(t)\boldsymbol{R}(t) + \{\boldsymbol{F}_{\mathrm{B}2}(t) - \boldsymbol{F}_{\mathrm{B}1}(t)\}. \tag{12}$$

The Brownian force is characterized by the first and second-moment averages as

$$\langle \boldsymbol{F}_{\mathrm{B}i}(t) \rangle = \boldsymbol{0}, \tag{13}$$

and

$$\langle \boldsymbol{F}_{\mathrm{B}i}(t) \boldsymbol{F}_{\mathrm{B}j}(t') \rangle = 2\zeta k_{\mathrm{B}} T \delta_{ij} \delta(t - t') \boldsymbol{I}, \tag{14}$$

where k_{B} is the Boltzmann constant and T is the temperature. From the end-to-end vector $\boldsymbol{R}(t)$, the stress tensor can be expressed as

$$\boldsymbol{\tau}(t) = \rho \langle h(t) \boldsymbol{R}(t) \boldsymbol{R}(t) \rangle - \rho k_{\mathrm{B}} T \boldsymbol{I}, \tag{15}$$

where ρ is the density of dumbbells.

There are several expressions for the spring strength $h(t)$. The most basic one is the Hookean spring, defined as

$$h(t) = h_{\mathrm{eq}} = \frac{3k_{\mathrm{B}}T}{n_{\mathrm{K}} b_{\mathrm{K}}^2}, \tag{16}$$

where n_{K} is the number of Kuhn segments per spring and b_{K} is the Kuhn length. To reproduce several properties of polymers, such as shear thinning under shear flow, it is essential to address finite extensible nonlinear elastic (FENE) effects. Although the exact expression for FENE springs is given by the inverse Langevin function, the following empirical expression is widely used [39]:

$$h(t) = h_{\mathrm{eq}} \frac{1 - \langle R_{\mathrm{eq}}^2 \rangle / R_{\mathrm{max}}^2}{1 - \boldsymbol{R}^2(t) / R_{\mathrm{max}}^2}, \tag{17}$$

where $\langle R_{\mathrm{eq}}^2 \rangle^{1/2} = (n_{\mathrm{K}})^{1/2} b_{\mathrm{K}}$ is the equilibrium length of the springs, and $R_{\mathrm{max}} = n_{\mathrm{K}} b_{\mathrm{K}}$ is the maximum length of the springs. As shown later in Sec. 3.4, a constitutive equation cannot be analytically obtained for the FENE dumbbell model. To address the FENE

spring more analytically, the following approximate expression of the FENE spring has been proposed [39]:

$$h(t) = h_{\mathrm{eq}} \frac{1 - \langle R_{\mathrm{eq}}^2 \rangle / R_{\mathrm{max}}^2}{1 - \langle \boldsymbol{R}^2(t) \rangle / R_{\mathrm{max}}^2} = h_{\mathrm{eq}} f_{\mathrm{FENE}}(t). \tag{18}$$

This spring is referred to as the FENE-P spring. Here, "P" means Peterlin, who proposed the approximate form of the FENE spring law. The average appearing in Eq. (18) makes it possible to obtain the analytical constitutive equation.

We use $\lambda = \zeta / 4h_{\mathrm{eq}}$ as the unit time and $G = \rho k_{\mathrm{B}} T$ as the unit stress for the dumbbell models. To simplify the expressions, in what follows, we omit the tilde representing dimensionless quantities.

3.2. Hookean dumbbell model

The most basic dumbbell model is the Hookean dumbbell model, where Hookean springs are employed (cf. Eq. (16)). From Eqs. (12), (15), and (16), the Hookean dumbbell model reduces to the constitutive equation of the UCM model (cf. Eq. (S5) in the supporting information) in the limit of $N_{\mathrm{p}} \to \infty$ with N_{p} being the number of dumbbells.

For the Hookean dumbbell model, we generate training data by Brownian dynamics (BD) simulations with the finite numbers of dumbbells ($N_{\mathrm{p}} \in \{10^3, 10^4, 10^5\}$). We apply the oscillatory shear flow, $\gamma(t) = \gamma_0 \sin(\omega t)$, with $\gamma_0 = 2$ and $\omega = 0.5$, over a period from $t = 0$ to $t = 100$. The simulations are run with $\Delta t = 1 \times 10^{-3}$ for $0 \leq t \leq 100$ and data are collected at the interval of $\Delta t_{\mathrm{train}} = 1 \times 10^{-2}$. Each simulation is conducted with five different random seeds, and their average data is used for training. Due to the characteristics of the BD simulation, the training data inherently include noise originating from the finite N_{p}. We here test whether *Rheo*-SINDy can find from the noisy data the following equations for the UCM model under shear flow (cf. Eqs. (S6)–(S8) in the supporting information):

$$\dot{\tau}_{xx} = -\tau_{xx} + 2\tau_{xy}\kappa_{xy}, \tag{19}$$

$$\dot{\tau}_{yy/zz} = \tau_{yy/zz} = 0, \tag{20}$$

$$\dot{\tau}_{xy} = -\tau_{xy} + \kappa_{xy} + \tau_{yy}\kappa_{xy} = -\tau_{xy} + \kappa_{xy}. \tag{21}$$

3.3. FENE-P dumbbell model

We next address the so-called FENE-P dumbbell model, where Eq. (18) is utilized as the spring strength. As shown below, the FENE-P dumbbell model has an analytical solution and is utilized for various flow problems, such as turbulent flows [40].

Due to the assumption shown in Eq. (18), a simple representation of the time evolution for the conformation tensor $\boldsymbol{C} = \langle \boldsymbol{R}(t)\boldsymbol{R}(t) \rangle$ can be obtained as

$$\frac{\mathrm{d}\boldsymbol{C}}{\mathrm{d}t} - \boldsymbol{C} \cdot \boldsymbol{\kappa}^{+} - \boldsymbol{\kappa} \cdot \boldsymbol{C} = -f_{\mathrm{FENE}}(t)\boldsymbol{C} + \frac{n_{\mathrm{K}}}{3}\boldsymbol{I}, \tag{22}$$

where $\boldsymbol{\kappa}^+$ is the transposed $\boldsymbol{\kappa}$. The stress tensor is thus obtained by

$$\boldsymbol{\tau}(t) = \rho h(t)\boldsymbol{C}(t) - \rho k_{\mathrm{B}}T\boldsymbol{I}. \tag{23}$$

Under shear flow, Eq. (22) reduces to the following expressions:

$$\dot{C}_{xx} = -f_{\mathrm{FENE}}C_{xx} + 2C_{xy}\kappa_{xy} + \frac{n_{\mathrm{K}}}{3}, \tag{24}$$

$$\dot{C}_{yy/zz} = -f_{\mathrm{FENE}}C_{yy/zz} + \frac{n_{\mathrm{K}}}{3}, \tag{25}$$

$$\dot{C}_{xy} = -f_{\mathrm{FENE}}C_{xy} + C_{yy}\kappa_{xy}. \tag{26}$$

Using *Rheo*-SINDy, we test whether or not Eqs. (24)–(26) can be discovered from the data.

Although it has not been as widely recognized due to its complexity, the FENE-P dumbbell model can also be expressed in the form of the constitutive equation (i.e., the stress expression) [41]. From the textbook of Bird and coworkers [39], the constitutive equation for the FENE-P model is

$$\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}t} - \boldsymbol{\tau}\cdot\boldsymbol{\kappa}^+ - \boldsymbol{\kappa}\cdot\boldsymbol{\tau} = -f_{\mathrm{FENE}}(t)\boldsymbol{\tau} + 2\boldsymbol{D} + \frac{\mathrm{D}\ln Z}{\mathrm{D}t}(\boldsymbol{\tau}+\boldsymbol{I}), \tag{27}$$

where $\mathrm{D}(\cdots)/\mathrm{D}t$ is the substantial derivative and Z is the function expressed as

$$Z = \frac{1}{1-\langle \boldsymbol{R}^2(t)/R_{\max}^2\rangle} = 1 + \frac{1}{3n_{\mathrm{K}}Z_{\mathrm{eq}}^{-1}}(\mathrm{tr}\boldsymbol{\tau}+3). \tag{28}$$

Here, Z_{eq} indicates Z at equilibrium. From Eq. (28), we can see that $\mathrm{tr}\boldsymbol{\tau}$ is tightly related to the (squared) length of dumbbells. Since we do not address the spatial gradient in rheological calculations, $\mathrm{D}(\cdots)/\mathrm{D}t$ simply reduces to $\mathrm{d}(\cdots)/\mathrm{d}t$. Using Eqs. (22), (27), and (28), the constitutive equations for the FENE-P dumbbell model under shear flow can be expressed as

$$\begin{aligned}\dot{\tau}_{xx} = &-\left\{1+\frac{1}{3(n_{\mathrm{K}}-1)}\right\}\tau_{xx} - \frac{1}{3(n_{\mathrm{K}}-1)}(\tau_{yy}+\tau_{zz})\\ &-\frac{1}{9n_{\mathrm{K}}(n_{\mathrm{K}}-1)}(\mathrm{tr}\,\boldsymbol{\tau})^2 - \frac{1}{3n_{\mathrm{K}}}\left(2+\frac{1}{n_{\mathrm{K}}-1}\right)\mathrm{tr}\,\boldsymbol{\tau}\tau_{xx}\\ &+2\left\{1+\frac{1}{3(n_{\mathrm{K}}-1)}\right\}\tau_{xy}\kappa_{xy} - \frac{1}{9n_{\mathrm{K}}(n_{\mathrm{K}}-1)}(\mathrm{tr}\,\boldsymbol{\tau})^2\tau_{xx}\\ &+\frac{2}{3(n_{\mathrm{K}}-1)}\tau_{xx}\tau_{xy}\kappa_{xy},\end{aligned} \tag{29}$$

$$\begin{aligned}\dot{\tau}_{yy/zz} = &-\left\{1+\frac{1}{3(n_{\mathrm{K}}-1)}\right\}\tau_{yy/zz} - \frac{1}{3(n_{\mathrm{K}}-1)}(\tau_{xx}+\tau_{zz/yy})\\ &-\frac{1}{9n_{\mathrm{K}}(n_{\mathrm{K}}-1)}(\mathrm{tr}\,\boldsymbol{\tau})^2 - \frac{1}{3n_{\mathrm{K}}}\left(2+\frac{1}{n_{\mathrm{K}}-1}\right)\mathrm{tr}\,\boldsymbol{\tau}\tau_{yy/zz}\\ &+\frac{2}{3(n_{\mathrm{K}}-1)}\tau_{xy}\kappa_{xy} - \frac{1}{9n_{\mathrm{K}}(n_{\mathrm{K}}-1)}(\mathrm{tr}\,\boldsymbol{\tau})^2\tau_{yy/zz}\\ &+\frac{2}{3(n_{\mathrm{K}}-1)}\tau_{yy/zz}\tau_{xy}\kappa_{xy},\end{aligned} \tag{30}$$

$$\dot{\tau}_{xy} = -\tau_{xy} + \kappa_{xy} + \tau_{yy}\kappa_{xy} - \frac{1}{3n_{\mathrm{K}}}\left(2 + \frac{1}{n_{\mathrm{K}}-1}\right)\mathrm{tr}\,\boldsymbol{\tau}\tau_{xy} - \frac{1}{9n_{\mathrm{K}}(n_{\mathrm{K}}-1)}(\mathrm{tr}\,\boldsymbol{\tau})^2\tau_{xy} + \frac{2}{3(n_{\mathrm{K}}-1)}\tau_{xy}^2\kappa_{xy}. \tag{31}$$

For the derivation, please refer to Sec. S5 in the supporting information. From Eqs. (29)–(31), the constitutive equation for the FENE-P model can be expressed by a polynomial of up to third degree in $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$. Here, we note that Eqs. (29)–(31) become equivalent to the UCM model shown in Eqs. (S6)–(S8) in the supporting information in the limit of $n_{\mathrm{K}} \to \infty$.

To generate noise-free training data, we solve Eqs. (23)–(26) with $n_{\mathrm{K}} = 10$ and $\Delta t = 1 \times 10^{-4}$ for $0 \le t \le 100$. n We apply the oscillatory shear flow with $\gamma_0 = 2$ and various ω values ($\omega \in \{0.1, 0.2, \ldots, 1\}$). From the computed stress data, we collect data at the interval of $\Delta t_{\mathrm{train}} = 1 \times 10^{-2}$.

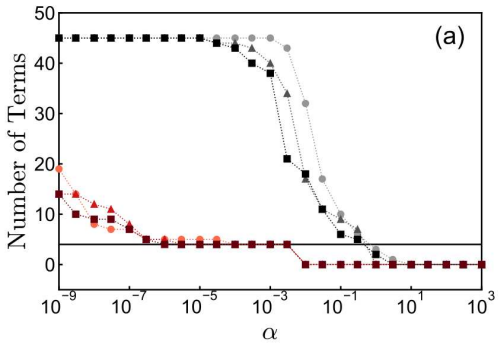
3.4. FENE dumbbell model

We finally address the FENE dumbbell model, where the spring strength is represented by Eq. (17). Since the FENE dumbbell model does not use any simplification for the spring strength (e.g., Peterlin approximation shown in Eq. (18)), its analytical constitutive equation has not been obtained. We apply *Rheo*-SINDy to this case to see if an *approximate* constitutive equation can be obtained. The obtained equations are validated by comparing the data obtained by numerically solving them with the data obtained by BD simulations.

The training data are generated by the BD simulations using Eqs. (12)–(15) and (17) with $n_{\mathrm{K}} = 10$, $N_{\mathrm{p}} = 10^4$, and $\Delta t = 1 \times 10^{-4}$ for $0 \le t \le 100$. We apply the oscillatory shear flows with the same parameters as those in the FENE-P dumbbell model. The BD simulation results with five different random seeds are averaged for each condition. Since we do not use any approximation for the spring strength, the values of $h(t)$ differ for each individual dumbbell. From the computed stress data, we collected data at the interval of $\Delta t_{\mathrm{train}} = 1 \times 10^{-2}$.

4. Results and Discussions

In this section, we present the results of the case studies for the dumbbell models. From the case studies on phenomenological constitutive equations shown in Sec. S3 in the supporting information, we have made the following two findings: (i) taking shear rheological data by an oscillatory shear test is more appropriate than by a simple (constant) shear test, and (ii) among the five optimization methods shown in Sec. 2, the STRidge or a-Lasso is superior to the other three methods. Thus, in what follows, we generate the training data using the oscillatory shear test, and employ the STRidge and a-Lasso as optimization methods.



(b)

method	α	N_p	equations
STRidge	3×10^{-1}	10^3	$\dot{\tau}_{xx} = -0.994\tau_{xx} + 1.999\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -0.992\tau_{xy} + 0.995\kappa_{xy} + 0.559\tau_{yy}\tau_{xy} + 1.007\tau_{yy}\kappa_{xy}$
STRidge	3×10^{-1}	10^5	$\dot{\tau}_{xx} = -1.001\tau_{xx} + 2.002\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -1.001\tau_{xy} + 1.000\kappa_{xy} + 0.387\tau_{yy}\kappa_{xy}$
a-Lasso	3×10^{-4}	10^3	$\dot{\tau}_{xx} = -0.992\tau_{xx} + 1.995\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -0.987\tau_{xy} + 0.983\kappa_{xy}$
a-Lasso	1×10^{-6}	10^5	$\dot{\tau}_{xx} = -1.001\tau_{xx} + 2.002\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -1.001\tau_{xy} + 1.000\kappa_{xy}$

Figure 3. (a) The number of total terms obtained by the STRidge (black) and a-Lasso (red) for the training data generated by the Hookean dumbbell model, and (b) the obtained constitutive equations. The exact equations are given in Eqs. (19)–(21). The horizontal line in (a) indicates the correct number of terms. In (a), circle, triangle, and square symbols indicate the total numbers of terms obtained by the data for $N_\mathrm{p} = 10^3$, 10^4, and 10^5, respectively.

4.1. Hookean Dumbbell Model

We first explain the results for the Hookean dumbbell model. We here used the polynomial library that includes up to 2nd order terms of τ_{xx}, τ_{yy}, τ_{xy}, and κ_{xy}. Thus, the total number of candidate terms is $N_{\boldsymbol{\Theta}} = 15$ for each component.

Figure 3 shows the *Rheo*-SINDy results for the Hookean dumbbell model with the different numbers of dumbbells. We note that the standard deviation of $\boldsymbol{\tau}$ in the training data decreases proportionally with $N_\mathrm{p}^{-1/2}$. From Fig. 3(a), as the value of N_p increases, sparser solutions are obtained especially for *Rheo*-SINDy with the STRidge. Unlike the case of the UCM model (cf. Fig. S1 in the supporting information), which can be considered as the "noise-free" case of the Hookean dumbbell model, the STRidge provides the correct number of terms only within a narrow range of α values. Nevertheless, if we choose the appropriate α value, the (nearly) correct constitutive equations can be found by the STRidge, as shown in the upper part of Fig. 3(b). We note that the terms containing τ_{yy} appear in the time evolution equation for τ_{xy} obtained by the STRidge. Although these terms do not affect the predictions because $\tau_{yy} = 0$, these terms do not appear in the correct equation. We speculate that the appearance of these terms is due to the correlation effects of the noise in R_x and R_y on the stress (cf. Eq. (15)). When comparing the STRidge and a-Lasso, it is evident that the a-Lasso provides stable and sparse solutions across a broader range of α values, regardless of the N_p value. Furthermore, we confirm that the correct equations can be obtained using the a-Lasso, as shown in the lower part of Fig. 3(b). This partially suggests the effectiveness of the a-Lasso in discovering essential terms from noisy data.

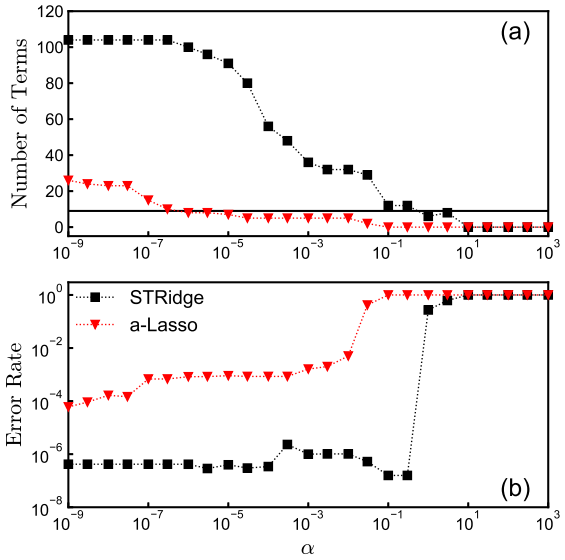


Figure 4. (a) The total number of terms and (b) the error rate for the conformation tensor $\boldsymbol{C}$ of the FENE-P dumbbell model obtained by *Rheo*-SINDy with the STRidge (black squares) and a-Lasso (red reverse triangles). The horizontal line in (a) indicates the correct number of terms. The training data were generated by Eqs. (24)–(26) with $n_{\mathrm{K}} = 10$.

4.2. FENE-P dumbbell Model

We next examine whether *Rheo*-SINDy can find more complex differential equations (i.e., the FENE-P dumbbell model) than the UCM model and the Giesekus model. For the *Rheo*-SINDy regressions of the differential equations for the conformation tensor $\boldsymbol{C}$ of the FENE-P dumbbell model explained in Sec. 3.3, we prepare the following custom library:

$$\boldsymbol{\Theta} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\Omega}(t_1) & \boldsymbol{\Omega}^2(t_1) & f_{\mathrm{FENE}}(t_1)\boldsymbol{\Omega}(t_1) \\ \mathbf{1} & \boldsymbol{\Omega}(t_2) & \boldsymbol{\Omega}^2(t_2) & f_{\mathrm{FENE}}(t_2)\boldsymbol{\Omega}(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \boldsymbol{\Omega}(t_n) & \boldsymbol{\Omega}^2(t_n) & f_{\mathrm{FENE}}(t_n)\boldsymbol{\Omega}(t_n) \end{bmatrix}, \tag{32}$$

where $\boldsymbol{\Omega}$ includes non-zero components of $\boldsymbol{C}$ under shear flow (C_{xx}, C_{yy}, C_{zz}, and C_{xy}) and κ_{xy}, and $\boldsymbol{\Omega}^2$ is the vector composed of all the multiplied combinations of the $\boldsymbol{\Omega}$ components. The total number of library functions is thus $N_{\boldsymbol{\Theta}} = 26$.

Figure 4 indicates (a) the total number of predicted terms and (b) the error rate as a function of the hyperparameter α for the STRidge and the a-Lasso. The error rate is defined as the sum of the mean squared errors (MSEs) of $\dot{\boldsymbol{t}}_{\mu\nu} - \boldsymbol{\Theta}\hat{\boldsymbol{\xi}}_{\mu\nu}$. The MSEs were scaled so that the maximum value of each method was 1. Similar to the results for the phenomenological constitutive equations shown in Sec. S3 in the supporting information, the a-Lasso provides sparser solutions than the STRidge, and the STRidge gives lower error rates than the a-Lasso. Figure 5 presents the differential equations obtained by the STRidge and a-Lasso for two α values that yield the nearly correct

method	α	equations
STRidge	1×10^{-3}	$\dot{C}_{xx} = +0.410 + 0.078C_{xx} + 0.495(C_{yy} + C_{zz}) - 0.011C_{xx}(C_{yy} + C_{zz}) - 0.092C_{yy}^2 - 0.092C_{zz}^2$ $- 0.092C_{yy}C_{zz} + 2.000C_{xy}\kappa_{xy} - 1.038f_{\mathrm{FENE}}C_{xx} + 0.418f_{\mathrm{FENE}}(C_{yy} + C_{zz})$ $\dot{C}_{yy} = +0.970 + 0.573(C_{yy} + C_{zz}) - 0.003C_{xx}C_{yy} - 0.002C_{xx}C_{zz} - 0.080C_{yy}^2 - 0.080C_{zz}^2$ $- 0.080C_{yy}C_{zz} - 0.312f_{\mathrm{FENE}}(C_{yy} + C_{zz})$ $\dot{C}_{zz} = \dot{C}_{yy}$ $\dot{C}_{xy} = +0.500(C_{yy} + C_{zz})\kappa_{xy} - 1.000f_{\mathrm{FENE}}C_{xy}$
STRidge	1×10^{-1}	$\dot{C}_{xx} = +3.333 + 2.000C_{xy}\kappa_{xy} - 1.000f_{\mathrm{FENE}}C_{xx}$ $\dot{C}_{yy} = +3.326 - 0.499f_{\mathrm{FENE}}(C_{yy} + C_{zz})$ $\dot{C}_{zz} = \dot{C}_{yy}$ $\dot{C}_{xy} = +0.500(C_{yy} + C_{zz})\kappa_{xy} - 1.000f_{\mathrm{FENE}}C_{xy}$
a-Lasso	1×10^{-7}	$\dot{C}_{xx} = -0.208C_{xx} + 0.436C_{yy} - 0.042C_{xx}C_{yy} + 1.946C_{xy}\kappa_{xy} + 0.072\kappa_{xy}^2$ $- 0.683f_{\mathrm{FENE}}C_{xx} + 0.683f_{\mathrm{FENE}}C_{yy}$ $\dot{C}_{yy} = 0.000$ $\dot{C}_{zz} = \dot{C}_{yy}$ $\dot{C}_{xy} = +1.101\kappa_{xy} - 0.017C_{xx}C_{xy} - 0.086C_{yy}C_{xy} + 0.314C_{yy}\kappa_{xy} - 0.107C_{zz}C_{xy} + 0.210C_{zz}\kappa_{xy}$ $- 0.289f_{\mathrm{FENE}}C_{xy} + 0.424f_{\mathrm{FENE}}\kappa_{xy}$
a-Lasso	1×10^{-4}	$\dot{C}_{xx} = +1.977C_{xy}\kappa_{xy} - 0.998f_{\mathrm{FENE}}C_{xx} + 1.007f_{\mathrm{FENE}}C_{yy}$ $\dot{C}_{yy} = 0.000$ $\dot{C}_{zz} = \dot{C}_{yy}$ $\dot{C}_{xy} = +3.219\kappa_{xy} - 1.005f_{\mathrm{FENE}}C_{xy}$

Figure 5. The differential equations for the conformation tensor $\boldsymbol{C}$ of the FENE-P dumbbell model found by *Rheo*-SINDy with the STRidge and the a-Lasso. The exact equations are given in Eqs. (24)–(26).

number of terms with a small error rate. From the lower part of Fig. 5, while the a-Lasso can provide sparser solutions, it does not guarantee that these are correct (cf. Eqs. (24)–(26)). Specifically, in all cases for τ_{xx}, τ_{yy}, and τ_{zz}, the a-Lasso has failed to identify the constant term in Eqs. (24) and (25), which is a possible source of larger errors compared to the STRidge. In the case of the STRidge, we confirmed that by choosing the appropriate α ($\alpha = 1 \times 10^{-1}$), nearly correct differential equations can be obtained, as shown in the upper part of Fig. 5. Since the yy-component and zz-component of the stress are equivalent, the exact equations can be recovered by setting $C_{yy} = C_{zz}$. Thus, we found that the correct differential equations for the FENE-P dumbbell model can be obtained if we can prepare the proper library functions and choose the appropriate value of the hyperparameter. Figure 6 shows the test simulation results using the identified differential equations for $\boldsymbol{C}$ in Fig. 5 and the dimensionless form of Eq. (23). Here, the oscillatory shear flow with $\gamma_0 = 4$ and $\omega = 1$, which is outside of the training data, was considered. From Fig. 6, the equations obtained by the STRidge can reproduce the exact solutions even when the equations are not exactly correct ($\alpha = 1 \times 10^{-3}$). In contrast, the test simulations with the differential equations obtained by the a-Lasso show the deviations from the test data, especially for τ_{xx}. These results emphasize the need to choose an appropriate optimization method to obtain reasonable solutions.

We then examine whether the stress expression of the constitutive equation for the FENE-P dumbbell model (cf. Eqs. (29)–(31)) can be found by *Rheo*-SINDy. For such

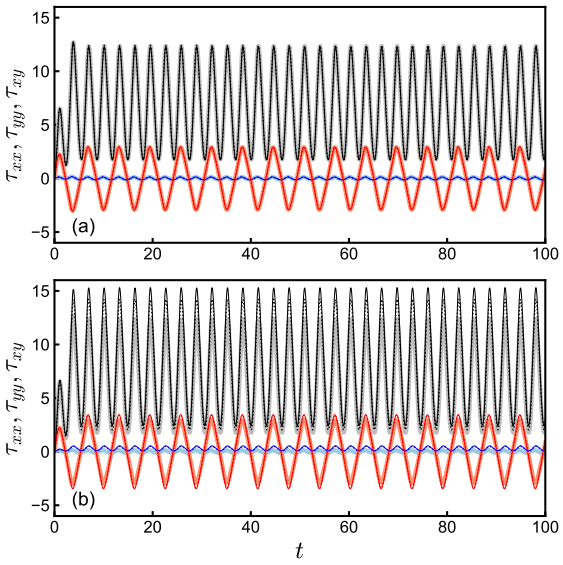


Figure 6. The test simulation results using the equations obtained by (a) the STRidge and (b) a-Lasso. Here, the test simulations were conducted with $\gamma_0 = 4$ and $\omega = 1$. The black, blue, and red lines show τ_{xx}, τ_{yy}, and τ_{xy}. The bold, thin dotted, and thin solid lines indicate the exact solutions, predictions with smaller α values ($\alpha = 1 \times 10^{-3}$ for the STRidge and $\alpha = 1 \times 10^{-7}$ for the a-Lasso), and predictions with larger α values ($\alpha = 1 \times 10^{-1}$ for the STRidge and $\alpha = 1 \times 10^{-4}$ for the a-Lasso).

a purpose, we prepared the following custom library:

$$\boldsymbol{\Theta} = \begin{bmatrix} 1 & \{\mathrm{tr}\,\boldsymbol{\tau}(t_1)\}^p \boldsymbol{T}_\mathrm{s}(t_1) & \{\mathrm{tr}\,\boldsymbol{\tau}(t_1)\}^2 & \{\boldsymbol{T}_\mathrm{s}(t_1)\}^p \kappa_{xy}(t_1) \\ 1 & \{\mathrm{tr}\,\boldsymbol{\tau}(t_2)\}^p \boldsymbol{T}_\mathrm{s}(t_2) & \{\mathrm{tr}\,\boldsymbol{\tau}(t_2)\}^2 & \{\boldsymbol{T}_\mathrm{s}(t_2)\}^p \kappa_{xy}(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \{\mathrm{tr}\,\boldsymbol{\tau}(t_n)\}^p \boldsymbol{T}_\mathrm{s}(t_n) & \{\mathrm{tr}\,\boldsymbol{\tau}(t_n)\}^2 & \{\boldsymbol{T}_\mathrm{s}(t_n)\}^p \kappa_{xy}(t_n) \end{bmatrix}, \tag{33}$$

where $\boldsymbol{T}_\mathrm{s}$ includes $\{\tau_{xx}, \tau_{yy}, \tau_{zz}, \tau_{xy}\}$ and p $(= 0, 1, 2)$ is the polynomial order. Thus, the total number of library functions is $N_{\boldsymbol{\Theta}} = 29$. We prepared the library that includes at least the terms present in Eqs. (29)–(31). Furthermore, we excluded terms that could potentially become large, such as higher-order terms involving κ_{xy}. When such terms are included in the solutions, the differential equations may be unstable, and in worse cases, they may also diverge.

Figure 7 shows (a) the total number of terms and (b) the error rate obtained by *Rheo*-SINDy with the STRidge and a-Lasso. Similar to what we noted previously, the a-Lasso can yield sparser solutions than the STRidge. Based on the number of terms shown in Fig. 7(a) and the error rates shown in Fig. 7(b), we chose several α values with a small number of terms and a low error rate. Figure 8 presents the equations obtained using the chosen α. From Fig. 8, the equations predicted by the STRidge with $\alpha = 1$ and the a-Lasso with $\alpha = 1 \times 10^{-4}$ are almost the same; conversely, the solutions for small α values significantly differ between the two methods. For the STRidge with $\alpha = 1 \times 10^{-2}$, the identified equations are close to the correct equations (cf. Eqs. (29)–(31)). Furthermore, the coefficient values for the correctly obtained terms are close

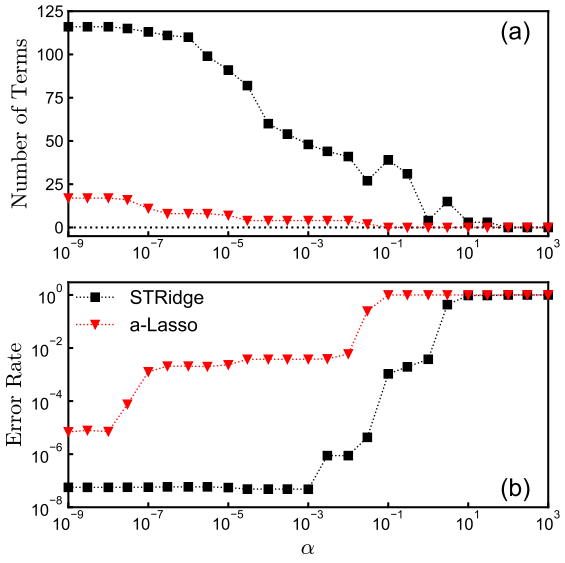


Figure 7. (a) The total number of terms and (b) the error rate for the constitutive equation of the FENE-P dumbbell model obtained by the STRidge (black squares) and the a-Lasso (red reverse triangles). The horizontal short-dashed line in (a) indicates that the number of terms is zero.

method	α	equations
STRidge	1×10^{-2}	$\dot{\tau}_{xx} = -1.033\tau_{xx} - 0.032(\tau_{yy} + \tau_{zz}) - 0.076\mathrm{tr}\boldsymbol{\tau}\tau_{xx} + 0.039\mathrm{tr}\boldsymbol{\tau}(\tau_{yy} + \tau_{zz}) + 2.071\tau_{xy}\kappa_{xy}$ $- 0.031(\mathrm{tr}\boldsymbol{\tau})^2(\tau_{yy} + \tau_{zz}) + 0.076\tau_{xx}\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = -0.037\tau_{xx} - 0.533(\tau_{yy} + \tau_{zz}) + 0.019\mathrm{tr}\boldsymbol{\tau}\tau_{xx} - 0.020\mathrm{tr}\boldsymbol{\tau}(\tau_{yy} + \tau_{zz}) + 0.074\tau_{xy}\kappa_{xy}$ $- 0.020(\mathrm{tr}\boldsymbol{\tau})^2 + 0.037(\tau_{yy} + \tau_{zz})\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{zz} = \dot{\tau}_{yy}$ $\dot{\tau}_{xy} = -0.997\tau_{xy} + 0.999\kappa_{xy} - 0.075\mathrm{tr}\boldsymbol{\tau}\tau_{xy} + 0.506(\tau_{yy} + \tau_{zz})\kappa_{xy}$ $- 0.019\tau_{xx}(\tau_{yy} + \tau_{zz})\kappa_{xy} + 0.048(\tau_{yy}^2 + +\tau_{yy}\tau_{zz} + \tau_{zz}^2)\kappa_{xy} + 0.076\tau_{xy}^2\kappa_{xy}$
STRidge	1	$\dot{\tau}_{xx} = -1.173\tau_{xx} + 2.185\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{zz} = \dot{\tau}_{yy}$ $\dot{\tau}_{xy} = -1.071\tau_{xy} + 1.033\kappa_{xy}$
a-Lasso	3×10^{-8}	$\dot{\tau}_{xx} = -1.024\tau_{xx} + 0.424\mathrm{tr}\boldsymbol{\tau}\tau_{xx} + 1.133\mathrm{tr}\boldsymbol{\tau}\tau_{yy} + 2.067\tau_{xy}\kappa_{xy} - 0.504(\mathrm{tr}\boldsymbol{\tau})^2 + 0.075\tau_{xx}\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = -0.038\tau_{xx} - 1.155\tau_{yy} + 0.073\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{zz} = \dot{\tau}_{yy}$ $\dot{\tau}_{xy} = -1.043\tau_{xy} + 1.047\kappa_{xy} - 0.046\mathrm{tr}\boldsymbol{\tau}\tau_{xy} + 2.489\tau_{yy}\kappa_{xy}$
a-Lasso	1×10^{-4}	$\dot{\tau}_{xx} = -1.173\tau_{xx} + 2.185\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{zz} = \dot{\tau}_{yy}$ $\dot{\tau}_{xy} = -1.070\tau_{xy} + 1.032\kappa_{xy}$

Figure 8. The constitutive equations for the FENE-P dumbbell model obtained by the STRidge and a-Lasso. The exact equations are given in Eqs. (29)–(31).

to the correct values. For the a-Lasso with $\alpha = 3 \times 10^{-8}$, several coefficients for the correctly obtained terms, such as τ_{xx}, τ_{xy}, and $\tau_{xx}\tau_{xy}\kappa_{xy}$ in the equation for $\dot{\tau}_{xx}$, are close to the exact values, but for other several terms, such as $\mathrm{tr}\,\boldsymbol{\tau}\tau_{xx}$ in the equation for $\dot{\tau}_{xx}$, the correct coefficient values are not obtained. Nevertheless, from Fig. 9, which shows

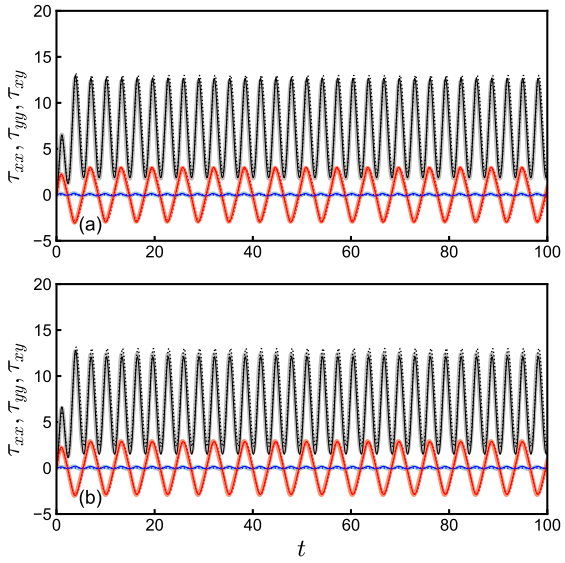


Figure 9. Test simulation results under the oscillatory shear flow with $\gamma_0 = 4$ and $\omega = 1$ for the constitutive equations of the FENE-P dumbbell model obtained by (a) the STRidge and (b) a-Lasso. The black, blue, and red lines represent the xx-, yy-, and xy-components of the stress tensor, respectively. The bold lines show the exact solutions. The thin solid and short-dashed lines indicate the results with smaller α values ($\alpha = 1 \times 10^{-2}$ for the STRidge and $\alpha = 3 \times 10^{-8}$ for the a-Lasso) and with larger α values ($\alpha = 1$ for the STRidge and $\alpha = 1 \times 10^{-4}$ for the a-Lasso).

Table 1. The mean squared error (MSE) between predicted and exact solutions for the FENE-P dumbbell model.

method	α	MSE (τ_{xx})	MSE (τ_{yy})	MSE (τ_{xy})
STRidge	1×10^{-2}	1.1×10^{-1}	5.0×10^{-5}	5.1×10^{-3}
STRidge	1	2.3	8.2×10^{-3}	8.9×10^{-2}
a-Lasso	3×10^{-8}	3.4×10^{-1}	3.0×10^{-3}	3.8×10^{-2}
a-Lasso	1×10^{-4}	2.3	8.2×10^{-3}	9.0×10^{-2}

the test simulation results, the equations obtained by the STRidge with $\alpha = 1 \times 10^{-2}$ and the a-Lasso with $\alpha = 3 \times 10^{-8}$ can well reproduce the exact solutions including the small oscillation of τ_{yy}. Although the equations obtained by the STRidge and a-Lasso demonstrate the similar performance in the test simulations shown in Fig. 9, the difference in predictions is quantified by their MSEs shown in Table 1. When α is small, the error in τ_{xx} is of the same order for both methods, but for predictions of τ_{yy} and τ_{xy}, the STRidge outperforms the a-Lasso. The STRidge, however, provides a sparse solution within a narrow range of α values, requiring careful selection of α.

4.3. FENE dumbbell Model

Finally, we address the FENE dumbbell model. As explained in Sec. 3.4, the FENE dumbbell model does not have an analytical expression of the constitutive equation.

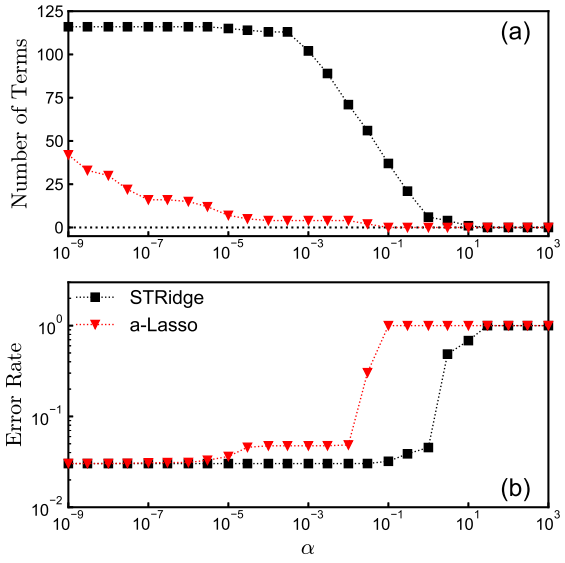


Figure 10. (a) The total number of terms and (b) the error rate for the FENE dumbbell model predicted by the STRidge (black squares) and the a-Lasso (red reverse triangles). The horizontal short-dashed line in (a) indicates that the number of terms is zero.

α	equations
1×10^{-6}	$\dot{\tau}_{xx} = -1.265\tau_{xx} - 1.235\tau_{yy} + 0.583\mathrm{tr}\boldsymbol{\tau}\tau_{xx} + 1.208\mathrm{tr}\boldsymbol{\tau}\tau_{yy} + 2.279\tau_{xy}\kappa_{xy} - 0.714(\mathrm{tr}\boldsymbol{\tau})^2 + 0.822\tau_{yy}\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = -0.110\tau_{xx} - 1.411\tau_{yy} - 0.990\tau_{zz} + 0.165\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{zz} = 0.000$ $\dot{\tau}_{xy} = -1.208\tau_{xy} + 1.140\kappa_{xy} - 0.102\mathrm{tr}\boldsymbol{\tau}\tau_{xy} + 2.290\tau_{yy}\kappa_{xy}$
3×10^{-4}	$\dot{\tau}_{xx} = -1.477\tau_{xx} + 2.232\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{zz} = \dot{\tau}_{yy}$ $\dot{\tau}_{xy} = -1.320\tau_{xy} + 1.158\kappa_{xy}$

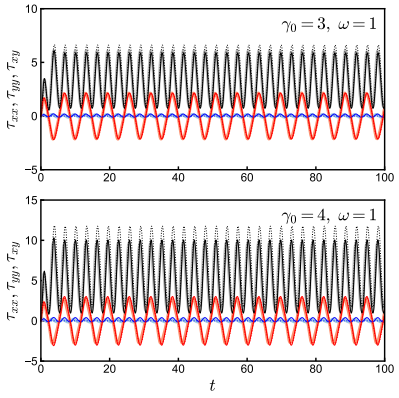


Figure 11. The predicted constitutive equations for the FENE dumbbell model (left) and the test simulation results (right). Here, the a-Lasso was utilized to obtain the approximate constitutive equations. For test simulations, we solved the constitutive equations under the oscillatory shear flows with $\gamma_0 = 3$ and $\omega = 1$ (right upper panel) and $\gamma_0 = 4$ and $\omega = 1$ (right lower panel). The bold lines show the exact solutions, and the thin solid and short-dashed lines show the results with the smaller α value ($\alpha = 1 \times 10^{-6}$) and the larger α value ($\alpha = 3 \times 10^{-4}$).

Thus, we here develop an *approximate* constitutive equation using *Rheo*-SINDy.

To obtain dynamical equations by *Rheo*-SINDy, one first needs to design an appropriate library $\boldsymbol{\Theta}$. To prepare $\boldsymbol{\Theta}$ for the FENE dumbbell model, we utilize the physical insights obtained from the analytical expression of the FENE-P dumbbell model. We here assume the constitutive equation of the FENE-P dumbbell model is *similar* to that of the FENE dumbbell model. Since the FENE-P dumbbell model is

a simplified version of the FENE dumbbell model, we believe that this is a reasonable assumption. Here, we note that the stress expression shown in Eq. (23) is no longer applicable to the FENE dumbbell model since the values of $h(t)$ differ for each individual dumbbell. Thus, it is invalid to obtain stress through the conformation tensor $\boldsymbol{C}$. Based on the above considerations, we decided to use the custom library presented in Eq. (33), which was also used in the constitutive equation of the FENE-P dumbbell model.

Figure 10 compares (a) the total number of terms and (b) the error rate predicted by the STRidge and a-Lasso. Similar to the previous discussions, we can obtain sparse solutions over a wide range of α values with the a-Lasso, whereas the STRidge gives sparse solutions only within a limited range of α. The left table in Fig. 11 shows the equations obtained by the a-Lasso with two α values chosen from the viewpoints of the sparsity and error rate in the same way as Fig. 5. We note that the predictions obtained by the STRidge are inferior to those obtained by the a-Lasso shown in Fig. 11, which is discussed in Sec. S6 in the supporting information. From the left table in Fig. 11, if α is appropriately chosen, the a-Lasso can give sparse equations with coefficients of reasonable (not excessively large) magnitudes. Comparing the equations for the FENE-P model obtained by the a-Lasso with $\alpha = 3 \times 10^{-8}$ (Fig. 8) and those for the FENE model obtained by the a-Lasso with $\alpha = 1 \times 10^{-6}$ (Fig. 11), the appearing terms are almost identical, which demonstrates the similarity between these models. The difference in the coefficients thus represents the difference between these models. The right panels in Fig. 11 show the test simulation results obtained by the equations shown in the left table. We found that the equations obtained with $\alpha = 1 \times 10^{-6}$ can reproduce well the BD simulation results outside the range of the training data within the investigated parameters, including the oscillatory behavior of τ_{yy}. (With the large α ($\alpha = 3 \times 10^{-4}$), the identified equation for τ_{yy} becomes $\dot{\tau}_{yy} = 0$, which fails to reproduce the oscillatory behavior of τ_{yy}.) This success suggests that *Rheo*-SINDy with the a-Lasso is effective for discovering *unknown* constitutive equations. Nevertheless, we note that the equations presented in Fig. 11 may fail to predict test data significantly outside the range of the training data. Reproducing such highly nonlinear data would require the nonlinear terms dropped in Fig. 11. In this sense, the constitutive equations for the FENE dumbbell model obtained here are appropriately referred to as the *approximate* constitutive equations.

Thanks to the equations obtained using *Rheo*-SINDy, it is possible to provide a physical interpretation with the assistance of rheological knowledge. For example, from the comparison of the equations obtained for the FENE-P dumbbell model (cf. Fig. 8) and those for the FENE dumbbell model (cf. Fig. 11), the equations for larger α value ($\alpha = 1 \times 10^{-4}$ for the FENE-P dumbbell model and $\alpha = 3 \times 10^{-4}$ for the FENE dumbbell model) are similar except for the coefficient values. Furthermore, the terms in these equations are the same as those for the UCM model (and thus the Hookean dumbbell model). This indicates that all of these models share the same origin based on the dumbbell model. The linear term of stress in the constitutive equation represents the relaxation of stress (see Eq. (S5) in the supporting information). Since the relaxation

time at equilibrium ($\lambda = \zeta/4h_{\rm eq}$) is taken as the unit time in this study, the coefficient of this term should be -1 at equilibrium (and thus for the UCM model, see Eqs. (S6)–(S8)). From Figs. 8 and 11, the coefficient of the linear term of stress is smaller than -1, which indicates $\lambda_{\rm sf} < \lambda_{\rm eq}$ with the subscript "sf" and "eq" standing for "shear flow" and "equiliblium", respectively. This indicates that under shear flow, the values of spring strength for the FENE-P and FENE dumbbell models become larger than $h_{\rm eq}$, which implies the appearance of the FENE effects under flow. From this discussion, it is evident that *Rheo*-SINDy can provide physically interpretable constitutive equations.

5. Concluding Remarks

We tested that the sparse identification for nonlinear dynamics (SINDy) modified for nonlinear rheological data, which we call *Rheo*-SINDy, is effective in finding constitutive equations of complex fluids. We found that *Rheo*-SINDy can successfully identify correct equations from training data generated from *known* constitutive equations, as well as provide approximate constitutive equations (or reduced order models) from training data generated by mesoscopic models when constitutive equations are analytically *unknown*.

Rheo-SINDy for two phenomenological constitutive equations (i.e., the upper convected Maxwell model and Giesekus model) revealed the following two things. First, compared to constant shear tests, oscillatory shear tests are appropriate for generating training data. Second, the sequentially thresholded Ridge regression (STRidge) and adaptive Lasso (a-Lasso) are effective in finding appropriate constitutive equations. We then examined the commonly used mesoscopic model, namely the dumbbell model with three different representations of spring strength: the Hookean, FENE-P, and FENE springs. Although the Hookean and FENE-P dumbbell models have analytical constitutive equations, for the FENE dumbbell model, there is no analytical expression of the constitutive equation. We confirmed through the Hookean dumbbell model that even in the presence of noise, the a-Lasso provides the correct solution over a wide range of the hyperparameter α. *Rheo*-SINDy was also effective in discovering the complex constitutive equations of the FENE-P dumbbell model. This case study revealed that the identification of complex equations requires the preparation of an appropriate custom library based on prior physical knowledge. Using physical insights obtained from the Hookean and FENE-P dumbbell models, we attempted to find *approximate* constitutive equations for the FENE dumbbell model. We found that the a-Lasso can successfully give the approximate constitutive equations, which can be used in predictions beyond the range of the training data.

From our investigation, *Rheo*-SINDy with the STRidge or a-Lasso is effective for discovering constitutive equations from nonlinear rheological data. We found that the STRidge is generally superior in terms of retaining correct terms, while the a-Lasso is more robust to the selection of α than the STRidge. To obtain correct constitutive equations, in addition to selecting the appropriate optimization method, we are required to design an appropriate library by using physical insights, namely *domain knowledge*.

Designing such a proper library necessitates not only including necessary terms but also excluding unnecessary terms.

This research is expected to have an impact on fields such as rheology and fluid mechanics. From a rheological perspective, for several systems such as entangled polymers [42, 43] and wormlike micellar solutions [44, 45], sophisticated mesoscopic models suitable for numerical simulations under flow have been proposed. These mesoscopic models can generate reasonable training data not only under shear flow but also under extensional flow. Finding new approximate models from the data obtained by these mesoscopic simulations would be an interesting research subject. Furthermore, it would be desirable to conduct *Rheo*-SINDy for experimental data obtained by Large Amplitude Oscillatory Shear (LAOS) experiments [46]. Since the LAOS measurements do not provide all the major stress components under shear flow, exploring methods for discovering the constitutive equations from experimental data would be a future challenge. When approximate constitutive models are identified, those models can be employed for predictions of complex flows, which would deepen our understanding of complex fluids. We will continue our research in these directions.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We thank Prof. Yoshinobu Kawahara for his valuable comments on data-driven methods. We also thank Dr. John J. Molina for carefully reading the manuscript and providing insightful comments. TS was partially supported by JST PRESTO Grant Number JPMJPR22O3. SM was financially supported by the Kyoto University Science and Technology Innovation Creation Fellowship (FSMAT), Grant Number JPMJFS2123.

References

[1] Brunton S L and Kutz J M 2022 *Data-Driven Science and Engineering* 2nd ed (Cambridge University Press)
[2] Brunton S L, Proctor J L and Kutz J N 2016 *Proc. Natl. Acad. Sci. USA* **113** 3932–3937
[3] de Silva B M, Higdon D M, Brunton S L and Kutz J N 2020 *Front. Artif. Intell.* **3** 25
[4] Kaheman K, Brunton S L and Nathan Kutz J 2022 *Mach. Learn.: Sci. Technol.* **3** 015031
[5] Fasel U, Kutz J N, Brunton B W and Brunton S L 2022 *Proc. Math. Phys. Eng. Sci.* **478** 20210904
[6] Schmidt M and Lipson H 2009 *Science* **324** 81–85
[7] Bongard J and Lipson H 2007 *Proc. Natl. Acad. Sci. USA* **104** 9943–9948
[8] Reinbold P A K, Kageorge L M, Schatz M F and Grigoriev R O 2021 *Nat. Commun.* **12** 3219
[9] Udrescu S M and Tegmark M 2020 *Sci. Adv.* **6** eaay2631
[10] Cranmer M, Sanchez Gonzalez A, Battaglia P, Xu R, Cranmer K, Spergel D and Ho S 2020 Discovering symbolic models from deep learning with inductive biases *NeurIPS* vol 33 ed Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H pp 17429–17442

[11] Lemos P, Jeffrey N, Cranmer M, Ho S and Battaglia P 2023 *Mach. Learn.: Sci. Technol.* **4** 045002
[12] Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S and Yang L 2021 *Nat. Rev. Phys.* **3** 422–440
[13] Raissi M, Perdikaris P and Karniadakis G E 2019 *Journal of computational physics* **378** 686–707 ISSN 0021-9991
[14] Jia X, Willard J, Karpatne A, Read J S, Zwart J A, Steinbach M and Kumar V 2021 *ACM/IMS Trans. Data Sci.* **2** 1–26
[15] Rosofsky S G, Al Majed H and Huerta E A 2023 *Machine Learning: Science and Technology* **4** 025022
[16] Larson R G 1988 *Constitutive Equations for Polymer Melts and Solutions* (Butterworths Series in Chemical Engineering)
[17] Rouse P E 1953 *J. Chem. Phys.* **21** 1272–1280
[18] Doi M and Edwards S F 1986 *The Theory of Polymer Dynamics* (Oxford University Press)
[19] Masubuchi Y, Takimoto J I, Koyama K, Ianniruberto G, Marrucci G and Greco F 2001 *J. Chem. Phys.* **115** 4387–4394
[20] Doi M and Takimoto J 2003 *Phil. Trans. R. Soc. A.* **361** 641–652
[21] Likhtman A E 2005 *Macromolecules* **38** 6128–6139
[22] Jamali S 2023 *Rheology Bulletin* **92**(1) 20–24
[23] Miyamoto S 2024 *Nihon Reoroji Gakkaishi (J. Soc. Rheol. Jpn.)* **52** 15–19
[24] Fang L, Ge P, Zhang L, E W and Lei H 2022 *J. Mach. Learn.* **1** 114–140
[25] Mahmoudabadbozchelou M, Kamani K M, Rogers S A and Jamali S 2022 *Proc. Natl. Acad. Sci. USA* **119** e2202234119
[26] Jin H, Yoon S, Park F C and Ahn K H 2023 *Rheol. Acta* **62** 569–586
[27] Mahmoudabadbozchelou M and Jamali S 2021 *Sci. Rep.* **11** 12015
[28] Mahmoudabadbozchelou M, Karniadakis G E and Jamali S 2022 *Soft Matter* **18**(1) 172–185
[29] Saadat M, Mahmoudabadbozchelou M and Jamali S 2022 *Rheol. Acta* **61** 721–732
[30] Mahmoudabadbozchelou M, Caggioni M, Shahsavari S, Hartt W H, Em Karniadakis G and Jamali S 2021 *J. Rheol.* **65** 179–198
[31] Lennon K R, McKinley G H and Swan J W 2023 *Proc. Natl. Acad. Sci. USA* **120** e2304669120
[32] Zhao L, Li Z, Caswell B, Ouyang J and Karniadakis G E 2018 *J. Comput. Phys.* **363** 116–127
[33] Zhao L, Li Z, Wang Z, Caswell B, Ouyang J and Karniadakis G E 2021 *J. Comput. Phys.* **427** 110069
[34] Seryo N, Sato T, Molina J J and Taniguchi T 2020 *Phys. Rev. Res.* **2** 033107
[35] Seryo N, Molina J J and Taniguchi T 2021 *Nihon Reoroji Gakkaishi (J. Soc. Rheol. Jpn.)* **49** 97–113
[36] Miyamoto S, Molina J J and Taniguchi T 2023 *Phys. Fluids* **35** 063113
[37] Fukami K, Murata T, Zhang K and Fukagata K 2021 *J. Fluid Mech.* **926** A10
[38] Mahmoudabadbozchelou M, Kamani K M, Rogers S A and Jamali S 2024 *Proc. Natl. Acad. Sci. USA* **121** e2313658121
[39] Bird R B, Armstrong R C and Hassager O 1987 *Dynamics of Polymeric Liquids* 2nd ed vol 2 (Oxford University Press)
[40] Graham M D 2014 *Phys. Fluids* **26** 101301
[41] Mochimaru Y 1983 *J. Non-Newtonian Fluid Mech.* **12** 135–152
[42] Sato T and Taniguchi T 2019 *Macromolecules* **52** 3951–3964
[43] Miyamoto S, Sato T and Taniguchi T 2023 *Rheol. Acta* **62** 57–70
[44] Sato T, Moghadam S, Tan G and Larson R G 2020 *J. Rheol.* **64** 1045–1061
[45] Sato T and Larson R G 2022 *J. Rheol.* **66** 639–656
[46] Hyun K, Wilhelm M, Klein C O, Cho K S, Nam J G, Ahn K H, Lee S J, Ewoldt R H and McKinley G H 2011 *Prog. Polym. Sci.* **36** 1697–1753

Supporting Information for: Rheo-SINDy: Finding a constitutive model from rheological data for complex fluids using sparse identification for nonlinear dynamics

Takeshi Sato[1,*,†], Souta Miyamoto[2,†], and Shota Kato[3]

[1]Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan
[2]Department of Chemical Engineering, Graduate School of Engineering, Kyoto University, Kyoto 615-8510, Japan
[3]Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

E-mail: takeshis@scl.kyoto-u.ac.jp

*Author to whom any correspondence should be addressed.
†These authors contributed equally to this work.
Keywords: sparse identification, rheology, constitutive model

Table S1. The regularization term $R(\boldsymbol{\xi}_{\mu\nu})$ for the sparse regression methods.

Method	Regularization term $R(\boldsymbol{\xi}_{\mu\nu})$
STLSQ	$\lambda_0 \|\|\boldsymbol{\xi}_{\mu\nu}\|\|_0$
STRidge	$\lambda_0 \|\|\boldsymbol{\xi}_{\mu\nu}\|\|_0 + \lambda_2 \|\|\boldsymbol{\xi}_{\mu\nu}\|\|_2^2$
Lasso	$\lambda_1 \|\|\boldsymbol{\xi}_{\mu\nu}\|\|_1$
E-Net	$\lambda_1 \|\|\boldsymbol{\xi}_{\mu\nu}\|\|_1 + \lambda_2 \|\|\boldsymbol{\xi}_{\mu\nu}\|\|_2^2$
a-Lasso	$\lambda_1 \|\|\boldsymbol{\xi}'_{\mu\nu}\|\|_1$

S1. Sparse Regression Methods

To solve the optimization problem in Eq. (9) in the main text, we used five sparse regression methods: (i) the sequentially thresholded least squares (STLSQ), (ii) sequentially thresholded Ridge regression (STRidge), (iii) least absolute shrinkage and selection operator (Lasso), (iv) Elastic-Net (E-Net), and (v) adaptive Lasso (a-Lasso).

The differences among these methods lie in the regularization term $R(\boldsymbol{\xi}_{\mu\nu})$ as shown in Table S1. The hyperparameters of ℓ_i norm ($i = 0, 1, 2$) are denoted as λ_i (> 0). The ℓ_0 and ℓ_1 norms are defined as

$$||\boldsymbol{\xi}_{\mu\nu}||_0 = \sum_j \delta(\xi_{\mu\nu,j}) \tag{S1}$$

and

$$||\boldsymbol{\xi}_{\mu\nu}||_1 = \sum_j |\xi_{\mu\nu,j}|, \tag{S2}$$

where $\delta(\xi_{\mu\nu,j})$ is the Kronecker delta function, which is equal to 1 if $\xi_{\mu\nu,j} \neq 0$ and 0 otherwise. The vector $\boldsymbol{\xi}'_{\mu\nu}$ in the a-Lasso is defined as $\boldsymbol{\xi}'_{\mu\nu} = \boldsymbol{w}_{\mu\nu} \otimes \boldsymbol{\xi}_{\mu\nu}$, where $\otimes$ is the element-wise product and $\boldsymbol{w}_{\mu\nu}$ is the adaptive weight vector and its j-th element is defined as $w_{\mu\nu,j} = |\xi_{\mu\nu,j}|^{-\delta}$ with δ being the positive constant.

The STLSQ and STRidge were implemented by iteratively conducting the least square regression and the Ridge regression, respectively, while setting the coefficients with smaller absolute values than a certain threshold α (> 0) to zero based on the original papers [S1, S2]. In the STRidge, the hyperparameter λ_2 was set to 0.05. The Lasso, E-Net, and a-Lasso were implemented using the scikit-learn library [S3]. In this library, the loss functions for the Lasso and E-Net are respectively defined as

$$\hat{\boldsymbol{\xi}}_{\mu\nu} = \underset{\boldsymbol{\xi}_{\mu\nu}}{\operatorname{argmin}} \frac{1}{2n} ||\dot{\boldsymbol{t}}_{\mu\nu} - \boldsymbol{\Theta}\boldsymbol{\xi}_{\mu\nu}||_2^2 + \alpha ||\boldsymbol{\xi}_{\mu\nu}||_1, \tag{S3}$$

and

$$\hat{\boldsymbol{\xi}}_{\mu\nu} = \underset{\boldsymbol{\xi}_{\mu\nu}}{\operatorname{argmin}} \frac{1}{2n} ||\dot{\boldsymbol{t}}_{\mu\nu} - \boldsymbol{\Theta}\boldsymbol{\xi}_{\mu\nu}||_2^2 + \alpha\beta ||\boldsymbol{\xi}_{\mu\nu}||_1 + \frac{\alpha(1-\beta)}{2} ||\boldsymbol{\xi}_{\mu\nu}||_2^2, \tag{S4}$$

where n is the number of data points, β is the ℓ_1 ratio, and α and β are the hyperparameters. The loss function for the Lasso is obtained by setting $\beta = 1$ in Eq. (S4). In this study, β was set to 0.5 for the E-Net. According to the original paper

of a-Lasso [S4], the a-Lasso can be implemented as the Lasso problem as the following steps:

(i) Define $\xi'_{\mu\nu,j} = \xi_{\mu\nu,j}/w_{\mu\nu,j}$, $j = 1, \ldots, J$.

(ii) Solve the Lasso problem for $\boldsymbol{\xi}'_{\mu\nu}$ using Eq. (S3).

(iii) Output $\hat{\xi}_{\mu\nu,j} = \hat{\xi}'_{\mu\nu,j}/w_{\mu\nu,j}$, $j = 1, \ldots, J$.

The adaptive weight $w_{\mu\nu,j}$ depends on the coefficients, and thereby the output coefficients can be varied in each iteration. To obtain the converged solution, we initialized the weights as unit vectors $\boldsymbol{w} = \mathbf{1}$ and repeated the above steps until the coefficients $\hat{\xi}_{\mu\nu,j}$ no longer change [S5]. Here, the hyperparameter δ was set to 3 (see Sec. S4 for the effect of δ).

S2. Case Studies for Phenomenological Constitutive Equations

S2.1. Upper Convected Maxwell (UCM) Model

The simplest constitutive equation for viscoelastic fluids is the upper convected Maxwell (UCM) model [S6] shown as

$$\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}t} - \boldsymbol{\tau}\cdot\boldsymbol{\kappa}^{+} - \boldsymbol{\kappa}\cdot\boldsymbol{\tau} = -\frac{1}{\lambda}\boldsymbol{\tau} + 2G\boldsymbol{D}. \tag{S5}$$

Here, the left-hand side of Eq. (S5) is the upper-convected time derivative of $\boldsymbol{\tau}$, λ is the relaxation time, G is the modulus, and $\boldsymbol{D}$ is the deformation rate tensor defined as $\boldsymbol{D} = (\boldsymbol{\kappa} + \boldsymbol{\kappa}^{+})/2$. Using λ as the unit time and G as the unit stress (i.e., $\lambda = G = 1$), we can obtain dimensionless expressions for time $\tilde{t} = t/\lambda$, velocity gradient tensor $\tilde{\boldsymbol{\kappa}} = \lambda\boldsymbol{\kappa}$, and stress $\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau}/G$. In what follows, we omit the tilde in dimensionless variables for simplicity. The dimensionless form of the UCM model under shear flow is thus written as

$$\dot{\tau}_{xx} = -\tau_{xx} + 2\tau_{xy}\kappa_{xy}, \tag{S6}$$

$$\dot{\tau}_{yy/zz} = \tau_{yy/zz} = 0, \tag{S7}$$

$$\dot{\tau}_{xy} = -\tau_{xy} + \kappa_{xy} + \tau_{yy}\kappa_{xy} = -\tau_{xy} + \kappa_{xy}. \tag{S8}$$

Here, since the initial conditions for $\boldsymbol{\tau}$ are set to the values of $\boldsymbol{\tau}$ at equilibrium, namely $\boldsymbol{\tau} = \mathbf{0}$, $\tau_{yy/zz}$ of the UCM model is zero under shear flow.

For the UCM model, we generate training data by numerically solving Eqs. (S6)–(S8) under two shear flow scenarios: simple shear and oscillatory shear tests. For the simple shear test, the shear rate is kept constant ($\kappa_{xy} = \dot{\gamma}$) across various values ($\dot{\gamma} \in \{1, 1.7, 2.8, 4.6, 7.7, 13, 22, 36, 60, 100\}$) with simulations running from $t = 0$ to $t = 10$ using a time step of $\Delta t = 1.0 \times 10^{-4}$. The oscillatory shear test introduces a time-dependent oscillatory shear strain, $\gamma(t) = \gamma_0 \sin(\omega t)$, with $\gamma_0 = 2$ and $\omega = 1$, over a period from $t = 0$ to $t = 100$, employing the same time step. In both tests, data are collected at intervals of $\Delta t_{\mathrm{train}} = 1 \times 10^{-2}$, resulting in a total of 10^4 data points for the training data.

S2.2. Giesekus Model

The Giesekus model, which is one of the most popular phenomenological constitutive equations [S7], shows typical shear rheological properties and is used to fit various complex fluids, including polymer solutions and wormlike micellar solutions. The tensorial form of the Giesekus constitutive equation can be written as

$$\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}t} - \boldsymbol{\tau}\cdot\boldsymbol{\kappa}^{+} - \boldsymbol{\kappa}\cdot\boldsymbol{\tau} = -\frac{1}{\lambda}\boldsymbol{\tau} - \frac{\alpha_{\mathrm{G}}}{G\lambda}\boldsymbol{\tau}\cdot\boldsymbol{\tau} + 2G\boldsymbol{D}, \tag{S9}$$

where α_{G} is the parameter governing the nonlinear response of the Giesekus model. The Giesekus equation under shear flow is thus given by

$$\dot{\tau}_{xx} = -\tau_{xx} - \alpha_{\mathrm{G}}(\tau_{xx}^2 + \tau_{xy}^2) + 2\tau_{xy}\kappa_{xy}, \tag{S10}$$

$$\dot{\tau}_{yy} = -\tau_{yy} - \alpha_{\mathrm{G}}(\tau_{yy}^2 + \tau_{xy}^2), \tag{S11}$$

$$\dot{\tau}_{zz} = 0, \tag{S12}$$

$$\dot{\tau}_{xy} = -\tau_{xy} - \alpha_{\mathrm{G}}(\tau_{xx} + \tau_{yy})\tau_{xy} + \tau_{yy}\kappa_{xy} + \kappa_{xy}. \tag{S13}$$

Here, all quantities are non-dimensionalized by using λ as the unit time and G as the unit stress. From Eqs. (S10)–(S13), the total number of collect terms in the Giesekus model is 12.

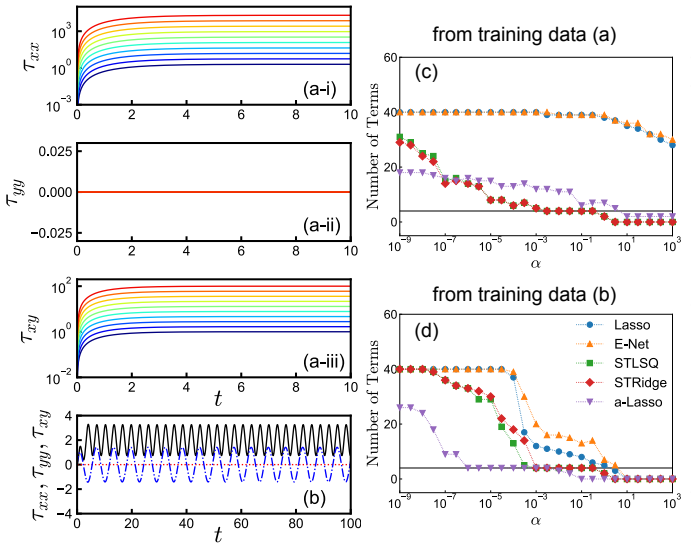
We generate the training data by solving Eqs. (S10)–(S13) numerically with $\alpha_{\mathrm{G}} = 0.5$ and $\Delta t = 1 \times 10^{-4}$. We note that the Giesekus model with $\alpha_{\mathrm{G}} = 0.5$ gives sufficient nonlinear features under shear flow. We applied the oscillatory shear flow with $\gamma_0 = 2$ and various ω values ($\omega \in \{0.1, 0.2, \ldots, 1\}$) for $0 \leq t \leq 100$. From the computed stress data, we collected data at the interval of $\Delta t_{\mathrm{train}} = 1 \times 10^{-2}$.

S3. Rheo-SINDy Results for Phenomenological Constitutive Equations

S3.1. Upper Convected Maxwell Model

Through this case study, we first check the appropriate methods to take the shear rheological data for *Rheo*-SINDy. Figure S1 shows the training data and results for the UCM model. Figure S1(a) and (b) are the stress data under the simple shear flows with the various shear rates and those under the oscillatory shear flow.

We conducted the *Rheo*-SINDy regressions by using the polynomial library that includes up to third order terms of τ_{xx}, τ_{yy}, τ_{xy}, and κ_{xy}. Thus, there were 35 candidate terms for each component of the constitutive equation. The terms related to τ_{zz} were excluded because they do not contribute to the UCM dynamics. The correct number of terms is four, as shown in Eqs. (S6)–(S8). Figures S1(c) and (d) present the number of total terms varying with the hyperparameter α obtained by *Rheo*-SINDy using the training data (a) and (b), respectively (for the detail of the hyperparameter α, see Sec. S1). Figure S1(c) indicates that the sparse solutions can be obtained by the STLSQ, STRidge, and a-Lasso, but not by the Lasso and E-Net. Moreover, regarding the number of terms, the STLSQ and STRdge exhibit similar behavior. Specifically, we confirm that the correct *number* of terms (cf. Eqs. (S6)–(S8)) are obtained by the STLSQ and



(e) SINDy results

data	method	α	equations
(a)	STRidge	3×10^{-3}	$\dot{\tau}_{xx} = -1.000\tau_{xx} + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy}$
(a)	a-Lasso	1×10^{-1}	$\dot{\tau}_{xx} = -1.239\tau_{xx} + 0.983\tau_{xy}^2$ $+ 1.385\tau_{xy}\kappa_{xy} + 0.100\kappa_{xy}^2$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -0.997\tau_{xy} + 0.997\kappa_{xy}$
(b)	STRidge	3×10^{-3}	$\dot{\tau}_{xx} = -1.000\tau_{xx} + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy}$
(b)	a-Lasso	1×10^{-6}	$\dot{\tau}_{xx} = -1.000\tau_{xx} + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0.000$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy}$

Figure S1. Training data obtained by the UCM model (a) under simple shear flow ($\kappa_{xy} = \dot{\gamma}$) and (b) under oscillatory shear flow ($\kappa_{xy} = \gamma_0 \omega \cos(\omega t)$). The number of total terms obtained by (c) the training data (a) (i.e., simple shear flow) and by (d) the training data (b) (i.e., oscillatory shear flow). (e) The constitutive equations obtained by *Rheo*-SINDy. Here, the exact equations for the UCM model under shear flow are shown in Eqs. (S6)–(S8). The parameters for the applied shear flows to obtain the training data are summarized in Sec. S2.1. In (b), xx-, yy-, and xy-components of the stress tensor are plotted with the black solid, red dotted, and blue dash-dotted lines, respectively. In (c) and (d), the number of total terms for five different optimization methods is plotted against the hyperparameter α. The black horizontal lines in (c) and (d) indicate the correct number of the terms in the UCM model.

STRidge with $3\times10^{-3} \leq \alpha \leq 3\times10^{-1}$. Figure S1(d) indicates that the STLSQ, STRidge, and a-Lasso yielded the correct number of terms, though all five methods gave sparse solutions. In most of the cases where the number of terms obtained was correct, the obtained coefficients were also correct for the UCM model. These results suggest that the oscillatory shear test is more appropriate than the simple shear test to obtain the correct constitutive equations for the UCM model. Figure S1(e) lists the constitutive equations obtained by the STRidge and a-Lasso. We can see that the STRidge and a-Lasso can give the correct constitutive equations, except for the a-Lasso in the simple shear test. Furthermore, we confirmed that the correct equations were obtained even for α values not shown in Fig. S1(e) in the case of the UCM model. These findings show the basic validity of finding the constitutive equations from the rheological data by *Rheo*-SINDy. Figure S1 indicates that the STLSQ, STRidge, and a-Lasso demonstrate better performance in discovering the correct constitutive equations compared to the Lasso and E-Net; thus, we use the former three methods in the following discussion.

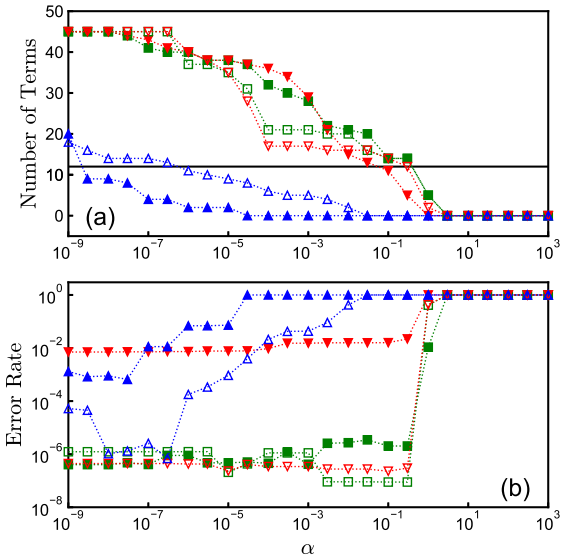


Figure S2. (a) The number of total terms and (b) the error rate obtained for the Giesekus model. The optimization methods include the STLSQ (green squares), STRidge (red reverse triangles), and a-Lasso (blue triangles). The filled and open symbols indicate the results with the single data trajectory of $\kappa_{xy} = \gamma_0\omega\cos(\omega t)$ with $\gamma_0 = 2$ and $\omega = 0.1$ for $0 \le t \le 100$ and those with the multiple (10) data trajectories of $\kappa_{xy} = \gamma_0\omega\cos(\omega t)$ with $\gamma_0 = 2$ and $\omega \in \{0.1, 0.2, \ldots, 1\}$ for $0 \le t \le 100$, respectively.

S3.2. Giesekus Model

We here explain the results of *Rheo*-SINDy for the Giesekus model. This case used the polynomial library consisting of up to second order terms of τ_{xx}, τ_{yy}, τ_{xy}, and κ_{xy}, which is the sufficient candidate terms to obtain the exact equations. Figure S2 shows (a) the total number of terms and (b) the error rate obtained by *Rheo*-SINDy for the training data of the Giesekus model. The error rate is defined as the sum of the mean squared errors (MSEs) of $\dot{\boldsymbol{t}}_{\mu\nu} - \boldsymbol{\Theta}\hat{\boldsymbol{\xi}}_{\mu\nu}$. The MSEs were scaled so that the maximum value of each method was 1. We show results using a single data trajectory with $\omega = 0.1$ and multiple data trajectories with $\omega \in \{0.1, 0.2, \ldots, 1.0\}$ as the training data. Figure S2(a) indicates that the a-Lasso evidently provides a sparser solution compared to the other two methods. Furthermore, Fig. S2(b) demonstrates that the regressions using the multiple data trajectories give solutions with smaller errors than those using the single data trajectory. We note that, similar to the number of terms obtained by *Rheo*-SINDy, coefficient values generally depend on α.

Figures S3(a) and (b) show the constitutive equations found by *Rheo*-SINDy and the test simulation results, respectively. Here, we used the training data of the multiple data trajectories. The α value for each method was chosen considering the sparsity indicated in Fig. S2(a) and the small loss indicated in Fig. S2(b). For test simulations shown in Fig. S3(b), we employed the oscillatory shear flow with $\gamma_0 = 4$ and $\omega = 0.5$, which is outside of the parameters in the training data described in Sec. S2.2. Figure S3(a) reveals that the STRidge with $\alpha = 3\times10^{-1}$ can give almost exact constitutive equations,

(a) SINDy results

method	α	equations
STLSQ	3×10^{-1}	$\dot{\tau}_{xx} = +1.404\tau_{xx} + 2.404\tau_{yy} - 0.500\tau_{xx}^2 + 2.405\tau_{xx}\tau_{yy} - 2.904\tau_{xy}^2 + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = -1.000\tau_{yy} - 0.500\tau_{yy}^2 - 0.500\tau_{xy}^2$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy} - 0.500\tau_{xx}\tau_{xy} - 0.499\tau_{yy}\tau_{xy} + 1.000\tau_{yy}\kappa_{xy}$
STRidge	3×10^{-1}	$\dot{\tau}_{xx} = -1.001\tau_{xx} - 0.500\tau_{xx}^2 - 0.499\tau_{xy}^2 + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = -0.999\tau_{yy} - 0.496\tau_{yy}^2 - 0.500\tau_{xy}^2$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy} - 0.500\tau_{xx}\tau_{xy} - 0.497\tau_{yy}\tau_{xy} + 0.998\tau_{yy}\kappa_{xy}$
a-Lasso	3×10^{-7}	$\dot{\tau}_{xx} = -1.498\tau_{xx} - 0.502\tau_{yy} - 0.495\tau_{xx}^2 - 0.475\tau_{xx}\tau_{yy} + 1.998\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = -0.993\tau_{yy} - 0.472\tau_{yy}^2 - 0.500\tau_{xy}^2$ $\dot{\tau}_{xy} = -0.998\tau_{xy} + 0.999\kappa_{xy} - 0.499\tau_{xx}\tau_{xy} - 0.484\tau_{yy}\tau_{xy} + 0.995\tau_{yy}\kappa_{xy}$

(b) test simulations

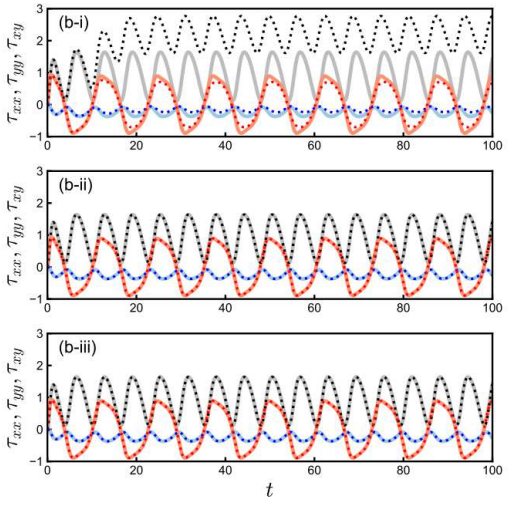


Figure S3. (a) The obtained constitutive equations for three optimization methods and (b) test simulation results under the oscillatory shear flow with $\gamma_0 = 4$ and $\omega = 0.5$ for (i) the STLSQ, (ii) STRidge, (iii) and a-Lasso. The training data are the same as those in Fig. S2. The exact equations for Giesekus model under shear flow are shown in Eqs. (S10)–(S13). In (a), the constitutive equations obtained by the multiple data trajectories are shown. In (b), the xx-, yy-, and xy-components of the stress tensor are shown with black, blue, and red lines, respectively. The dotted and solid lines in (b) denote the predictions by the equations shown in (a) and those by the exact Giesekus model, respectively.

including the value of α_{G} (cf. Eqs. (S10)–(S13)). As inferred from this, the predictions based on the constitutive equations obtained by the STRidge demonstrate a good agreement with the test data as shown in Fig. S3(b-ii). In contrast to the success of the STRidge, the STLSQ and a-Lasso failed to identify the correct solution, as indicated in Fig. S3(a). The constitutive equation obtained by the STLSQ with $\alpha = 3 \times 10^{-1}$ has a low error rate as shown in Fig. S2(b), but its predicted τ_{xx} significantly deviates from the test data as seen in Fig. S3(b-i). In contrast, although the a-Lasso did not provide the correct solution for τ_{xx}, the test simulations with the obtained constitutive equations exhibit a good agreement with the test data. These test simulations demonstrate that the STRidge and a-Lasso are promising approaches for *Rheo*-SINDy.

S4. Hyperparameter of the Adaptive Lasso

Here, we shortly note the effect of changing the hyperparameter δ of the a-Lasso, which determines the adaptive weight. Figure S4 compares the total number of terms for the Giesekus model obtained by the a-Lasso with three different δ values. The training data include the multiple trajectories, which are the same as those in Fig. S3. Here, the results for the STLSQ are also shown for comparison. As shown in Fig. S4, the solutions obtained by the a-Lasso with $\delta = 1$, 3, and 5 are sparser than those obtained by the STLSQ. Due to the increased effects of weights, the solutions for $\delta = 3$ and 5

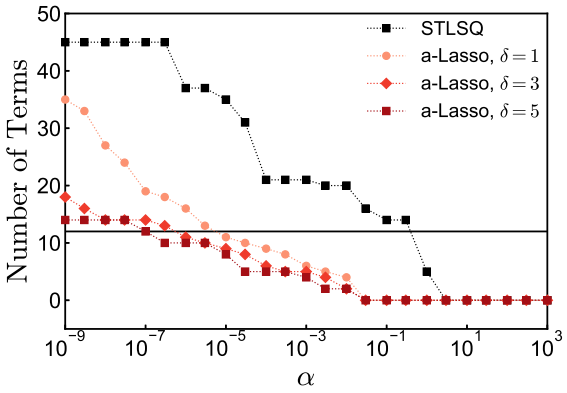


Figure S4. The total number of terms obtained by the STLSQ (black symbols) and a-Lasso (red symbols) for the Giesekus model. Here, the circles, diamonds, and squares in the red series indicate the results with $\delta = 1$, 3, and 5 for the adaptive weight $w_{\mu\nu,j}$, respectively.

are sparser compared to the solutions for $\delta = 1$. Moreover, the results with $\delta = 3$ are almost the same as those with $\delta = 5$, although the a-Lasso with $\delta = 5$ provides slightly sparser solutions. Thus, the hyperparameter $\delta = 3$ can be considered sufficiently large to obtain sparse solutions. We note, in general, that a sparser solution is superior from the perspective of overfitting and helps prevent unexpected divergence during test simulations. From these discussions, in this study, we used $\delta = 3$ as the adaptive weight in the a-Lasso.

S5. Stress Expressions for the FENE-P Dumbbell Model

As noted in Sec. 3.3 in the main text, the constitutive equation for the FENE-P dumbbell model can be expressed in terms of the stress (cf. Eqs. (29)–(31)). We here show the derivation of the constitutive equation for the FENE-P model [S8].

To improve clarity, let us rewrite the stress $\boldsymbol{\tau}$ in Eq. (15) in the main text as follows:

$$\boldsymbol{\tau}(t) = \rho h_{\mathrm{eq}} Z_{\mathrm{eq}}^{-1} Z \langle \boldsymbol{R}(t)\boldsymbol{R}(t)\rangle - G\boldsymbol{I}, \tag{S14}$$

where Z has already been defined in Eq. (28) in the main text and Z_{eq} indicates Z at equilibrium. In what follows, we express all variables in dimensionless forms by using the unit time λ and the unit stress $\rho k_{\mathrm{B}}T$. Additionally, for simplicity, we omit the tilde representing dimensionless quantities. Taking the trace of both sides of Eq. (S14) and using the relation $\langle \boldsymbol{R}^2(t)\rangle = R_{\max}^2(1 - Z^{-1})$, we can rewrite Z as a function of $\boldsymbol{\tau}$:

$$Z = 1 + \frac{1}{3n_{\mathrm{K}} Z_{\mathrm{eq}}^{-1}}(\mathrm{tr}\boldsymbol{\tau} + 3). \tag{S15}$$

Taking the convected derivative of $\boldsymbol{\tau}/Z$, the time evolution of stress can be expressed as

$$\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}t} - \boldsymbol{\tau}\cdot\boldsymbol{\kappa}^{+} - \boldsymbol{\kappa}\cdot\boldsymbol{\tau} = -Z_{\mathrm{eq}}^{-1} Z\boldsymbol{\tau} + 2\boldsymbol{D} + \frac{\mathrm{D}\ln Z}{\mathrm{D}t}(\boldsymbol{\tau} + \boldsymbol{I}), \tag{S16}$$

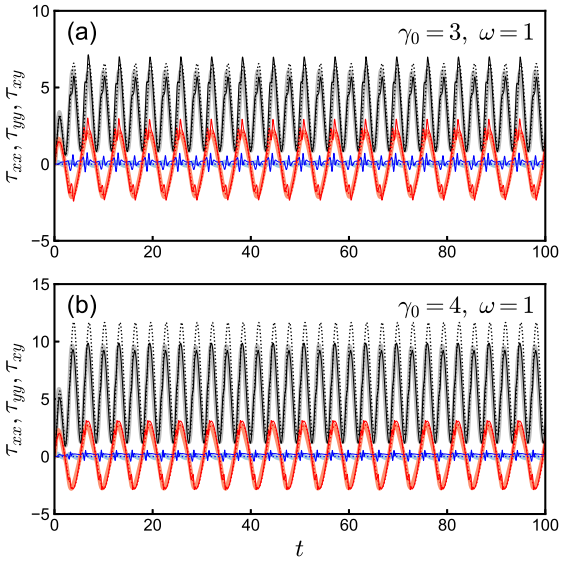


Figure S5. Test simulation results obtained by *Rheo*-SINDy with the STRidge for the library shown in Eq. (33) in the main text. The test simulations are conducted under the oscillatory shear flows with (a) $\gamma_0 = 3$ and $\omega = 1$ and (b) $\gamma_0 = 4$ and $\omega = 1$. The bold lines show the exact solutions, and the thin solid and short-dashed lines show the results with the smaller α value ($\alpha = 1 \times 10^{-1}$) and the larger α value ($\alpha = 1$).

which is the same as in Eq. (27) in the main text. Since we do not address the spatial gradient in rheological calculations, $\mathrm{D}(\cdots)/\mathrm{D}t$ simply reduces $\mathrm{d}(\cdots)/\mathrm{d}t$. To obtain Eq. (S16), we used the following relation that can be obtained by Eqs. (22) and (23) in the main text:

$$\frac{\mathrm{d}\boldsymbol{C}}{\mathrm{d}t} - \boldsymbol{C}\cdot\boldsymbol{\kappa}^{+} - \boldsymbol{\kappa}\cdot\boldsymbol{C} = -\frac{n_{\mathrm{K}}}{3}\boldsymbol{\tau}. \tag{S17}$$

From Eq. (S15), the time evolution of $\ln Z$ can be expressed in terms of $\mathrm{tr}\,\boldsymbol{\tau}$ as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}\boldsymbol{\tau} = \left\{3n_{\mathrm{K}}Z_{\mathrm{eq}}^{-1} + (\mathrm{tr}\boldsymbol{\tau} + 3)\right\}\frac{\mathrm{d}\ln Z}{\mathrm{d}t}. \tag{S18}$$

Furthermore, taking trace of Eq. (S16) and using Eq. (S18), we can have

$$\frac{\mathrm{d}\ln Z}{\mathrm{d}t} = \frac{1}{3n_{\mathrm{K}}Z_{\mathrm{eq}}^{-1}}\left\{-Z_{\mathrm{eq}}^{-1}Z\mathrm{tr}\boldsymbol{\tau} + 2\mathrm{tr}\boldsymbol{D} + \mathrm{tr}\left(\boldsymbol{\tau}\cdot\boldsymbol{\kappa}^{+} + \boldsymbol{\kappa}\cdot\boldsymbol{\tau}\right)\right\}. \tag{S19}$$

Combining Eqs. (S16) and (S19), we can express the time evolution of $\boldsymbol{\tau}$ (i.e., $\dot{\boldsymbol{\tau}}$) as a function of $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$. Specifically, Eqs. (S16) and (S19) reduce to Eqs. (29)–(31) in the main text under shear flow.

S6. STRidge Regressions for the FENE Dumbbell Model

Figures S5 and S6 show the test simulation results for the FENE dumbbell model using the approximate constitutive equations obtained by *Rheo*-SINDy with the STRidge. Here, in Fig. S5, we employed the custom library shown in Eq. (33) in the main text ($N_{\boldsymbol{\Theta}} = 29$), while in Fig. S6, we utilized a polynomial library including polynomial terms

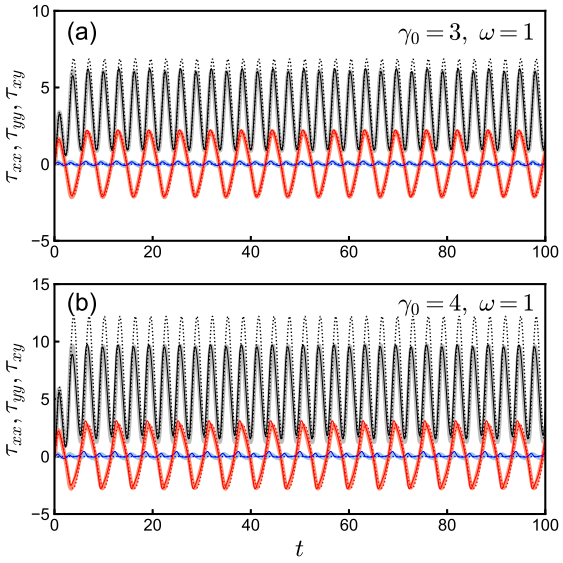


Figure S6. Test simulation results obtained by the STRidge for the library including polynomial terms up to the third order of $\{\tau_{xx}, \tau_{yy}, \tau_{zz}, \tau_{xy}, \kappa_{xy}\}$. The flow parameters for the test simulations are the same as those in Fig. S5. The bold lines show the exact solutions, and the thin solid and short-dashed lines show the results with the smaller α value ($\alpha = 3 \times 10^{-2}$) and the larger α value ($\alpha = 1$).

up to the third order of $\{\tau_{xx}, \tau_{yy}, \tau_{zz}, \tau_{xy}, \kappa_{xy}\}$ ($N_{\boldsymbol{\Theta}} = 56$). From the thin solid lines in Fig. S5, which show the results with the smaller $\alpha = 1 \times 10^{-1}$, while the magnitudes of the predicted stress components almost match the results of the BD simulation, spike-like predictions are occasionally observed. When using the third order polynomial library, the solutions for the small α, indicated by thin solid lines in Fig. S6, closely resemble the results of the BD simulations. This is likely attributed to the fact that the larger number of candidate terms included in the library improves the predictive ability of the model. Nevertheless, we note that increasing the number of terms in the library without careful consideration does not necessarily lead to an improvement in the model performance. By increasing the number of terms in the library, overfitting issues may arise. For example, when *Rheo*-SINDy chooses terms that are likely to be significantly large under shear flow, such as $\tau_{xx}\kappa_{xy}^2$, there is an increased possibility that the differential equations may fail to be solved when conducting test simulations for the parameters outside of the training data.

References

[S1] S. L. Brunton and J. L. Proctor, and J. N. Kutz, *Proc. Nat. Acad. Sci.*, **113**, 3932 (2016).
[S2] S. H. Rudy and S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Sci. Adv.*, **3**, e1602614 (2017).
[S3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *J. Mach. Learn. Res.*, **12**, 2825 (2011).
[S4] H. Zou, *J. Am. Stat. Assoc.*, **101**, 1418 (2006).

[S5] K. Fukami, T. Murata, K. Zhang, and K. Fukagata, *J. Fluid Mech.*, **926**, A10 (2021).
[S6] R. G. Larson, *Constitutive Equations for Polymer Melts and Solutions*, Butterworths Series in Chemical Engineering (1988).
[S7] H. Giesekus, *J. Non-Newtonian Fluid Mech.*, **11**, 69 (1982).
[S8] Y. Mochimaru, *J. Non-Newtonian Fluid Mech.*, **12**, 135 (1983).