Computer Methods in Applied Mechanics and Engineering 432 (2024) 117379

Contents lists available at ScienceDirect

Comput. Methods Appl. Mech. Engrg.

journal homepage: www.elsevier.com/locate/cma

Learning macroscopic equations of motion from dissipative particle dynamics simulations of fluids

Matevž Jug [a,b], Daniel Svenšek [a,b], Tilen Potisk [a,b,*], Matej Praprotnik [a,b]

[a] Theory Department, National Institute of Chemistry, Hajdrihova 19, Ljubljana, SI-1001, Slovenia
[b] Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, Ljubljana, SI-1000, Slovenia

ARTICLE INFO

Keywords:
Sparsity
Model selection
Particle simulations
Macroscopic dynamics
Regression

ABSTRACT

Macroscopic descriptions of both natural and engineered materials usually include a number of phenomenological parameters that have to be estimated from experiments or large-scale microscopic simulations. When dealing with advanced complex materials, these descriptions are sometimes not *a priori* available or not even known. Using sparsity-promoting techniques one can extract macroscopic dynamic models directly from particle-based simulations. In this work, we showcase such an approach on a simple fluid and test its robustness. We introduce a novel measure for automatic macroscopic model selection that combines stability and accuracy of a model. Using this measure and employing only a few physics-based assumptions, we are able to infer both the mass continuity equation and an equation for the conservation of linear momentum. Moreover, the extracted phenomenological and non-phenomenological parameters agree well with their numerically measured values and the well-known semi-empirical estimates. The presented model selection framework can be applied to simulations or experimental data of more complex systems, described in general by a rich set of coupled nonlinear macroscopic equations.

1. Introduction

Macroscopic descriptions of materials provide a systematic framework for predicting their dynamic behavior on large (industrial) spatio-temporal scales. Methods for deriving macroscopic dynamic equations are usually based on a symmetry-based approach [1,2] and provide dynamic laws for collective variables connected with conservation laws or symmetry breaking [3]. Although these methods are thermodynamically consistent [4], they lack a bridge that connects the macroscopic phenomenological parameters with microscopic properties.

In contrast, microscopic methods, such as molecular dynamics (MD) [5,6], provide a detailed picture of the material, but are computationally intensive. The number of degrees of freedom in such a simulation can be reduced by the Mori–Zwanzig formalism [7,8], resulting in coarse-grained mesoscopic methods [9–13] with variable degree of accuracy, which are governed by dynamics in the form of a generalized Langevin equation. However, the extraction of phenomenological equilibrium or transport coefficients requires prior knowledge of a complete macroscopic description, which is not always known in advance, especially when dealing with a novel material. Moreover, the determination of transport coefficients is prone to noise when using equilibrium methods [14], or requires a careful simulation setup to measure a specific phenomenological parameter when using non-equilibrium methods [15,16].

[*] Corresponding author at: Theory Department, National Institute of Chemistry, Hajdrihova 19, Ljubljana, SI-1001, Slovenia.
E-mail addresses: matevz.jug@ki.si (M. Jug), daniel.svensek@fmf.uni-lj.si (D. Svenšek), tilen.potisk@ki.si (T. Potisk), praprot@cmm.ki.si (M. Praprotnik).

https://doi.org/10.1016/j.cma.2024.117379
Received 14 June 2024; Received in revised form 8 August 2024; Accepted 7 September 2024
Available online 12 September 2024

Data-driven techniques are becoming increasingly popular for discovering dynamic systems from data. Examples include equation-free modeling [17], learning effective dynamics [18,19], modeling emergent dynamics [20], detecting causal relationships between time-dependent variables [21], symbolic regression [22,23], equation learning based on Gaussian processes [24] and neural network based learning [25–28]. Sparse Identification of Nonlinear Dynamics (SINDy) [29] is a framework based on regularized least squares regression, formally with an additional sparsity-promoting term, which has been shown to be effective in discovering explicit and parsimonious dynamic laws. It has already been applied to a variety of problems, such as chemical reaction dynamics [30], plasma physics [31], nonlinear optics [32,33], mesoscale ocean eddies [34], oscillations in the tropical atmosphere [35], and COVID-19 transmission dynamics [36]. In the case of partial differential equation discovery [37], the weak formulation of SINDy [38,39] has proved to be particularly resistant to noise and could even handle experimental data [40,41].

However, the application of sparsity-promoting techniques to particle simulation data, *i.e.* particle trajectories, is rare, with the exception of Refs. [42–44]. In Ref. [42], SINDy was applied to a particle-based description of active matter in the form of interactive self-propelled chiral particles, as well as to experimental video data on a driven colloidal system and collective motion of sunbleak fish. By projecting the spatio-temporal data onto a spectral basis composed of Chebyshev polynomials in time and Fourier modes in space, they were able to extract macroscopic dynamic models. In Ref. [43], the weak formulation of SINDy was applied to a system of interacting particles, described by stochastic differential equations. In contrast to Ref. [42], where a Gaussian kernel is used to obtain the particle number density and the polarization density field, no smoothing of the input fields is required when using the weak SINDy approach. In Ref. [44], sparse regression was used to extract the Vlasov equation as well as the single-fluid and multi-fluid magnetohydrodynamic models from particle-in-cell simulations of plasma. Therein, the weak formulation was crucial to capture slow and large-scale phenomena in the presence of noisy simulation data.

In this work, we develop a novel macroscopic model selection measure and test it, together with SINDy, on simulation datasets describing the transient dynamics of a simple fluid in two different simulation configurations. We describe the dynamics of the fluid using dissipative particle dynamics (DPD) [45,46], which is a state-of-the-art mesoscopic particle-based technique. DPD is based on the representation of multiple atoms or even molecules as single beads and is therefore suitable for the study of complex materials [47], such as polymers [9], colloidal dynamics [48], magnetic fluids [49], biological membranes [50] or red blood cells [51–54]. It has also recently been used to model ultrasound propagation in simple fluids [55]. We choose to simulate a one-component isothermal DPD fluid, *i.e.* a simple fluid, since the macroscopic transport coefficients of such a system have a solid theoretical and numerical foundation. We use the weak formulation of SINDy, whose superior handling of noisy data makes it ideal for studying large-scale MD or DPD simulation datasets of complex materials. Our new non-parametric measure automatically selects the most suitable model according to its accuracy and stability and is, in combination with the weak SINDy, robust both to high-levels of noise and large model libraries. For an overview of our approach see Fig. 1.

2. Learning framework

2.1. Sparsity-promoting regression method

In general, a macroscopic dynamic model of a material is given by a set of M fields $\{u_1, u_2, \dots, u_M\}$ equipped with coupled nonlinear partial differential equations. We assume that these equations can be written as follows:

$$\partial_t u_i = \sum_j c_j^{u_i} D_j f_j(u_1, u_2, \dots, u_M) \,, \tag{1}$$

where $c_j^{u_i}$ denotes a macroscopic coefficient corresponding to the jth term of the dynamics of the field u_i, and D_j is either identity or a general spatial derivative acting on some function of the fields f_j.

The description of complex materials in terms of fields, as in Eq. (1), is used in several well-established field-theoretical methods, such as classical density functional theory (DFT) [56], its dynamic extensions dynamic density functional theory (DDFT) [57] and power functional theory (PFT) [58], Mori–Zwanzig projection formalism [59] and macroscopic dynamics [1,2]. These methods have been applied to a wide range of complex systems, modeling both phase behavior [60] and dynamics of polymers [7,61–65], protein absorption on nanoparticles [66], colloidal fluids [67,68], active matter [69,70], nematic liquid crystals [3], magnetic fluids [71] and gels [72].

The problem of finding, for each field u_i, an equation of the form (1) that best describes a given dataset of known values of the fields u_1, u_2, …, u_M at some sample points $\{\mathbf{x}_k\}$ in space and time can be tackled by linear regression. Constructing a suitable library of candidate terms $D_j f_j$ and evaluating Eq. (1) at every sample point yields a linear system for each field: $\mathbf{u}_i^t = \Phi \mathbf{c}^{u_i}$, where $\mathbf{c}^{u_i}$ contains the coefficients $[\mathbf{c}^{u_i}]_j = c_j^{u_i}$, the library matrix Φ contains values of candidate terms at sample points $[\Phi]_{jk} = D_j f_j(u_1(\mathbf{x}_k), u_2(\mathbf{x}_k), \dots, u_M(\mathbf{x}_k))$, and $\mathbf{u}_i^t$ contains time derivatives of field u_i at the same points $[\mathbf{u}_i^t]_k = \partial_t u_i(\mathbf{x}_k)$. The system can then be used to find optimal coefficients that fit the dataset as closely as possible.

When discovering dynamic laws of a material about which we have only limited knowledge, it is reasonable to include a large number of terms into the library that could contribute to the unknown dynamics. We assume that only a few of these candidate terms are actually necessary for a good description of its macroscopic dynamics. To achieve a sparse solution to the regression problem, we use the Sequentially Thresholded Least-Squares (STLSQ) method, which was first presented in [29]. This is an iterative algorithm, in which each iteration, enumerated with the index l, consists of two steps. First, a least-squares fit is calculated: $\mathbf{c}_l^{u_i} = \mathrm{argmin}_{\mathbf{c}^{u_i}} \|\mathbf{u}_i^t - \Phi_l \mathbf{c}^{u_i}\|_2^2$, using the current library matrix Φ_l with normalized columns. Then, all terms whose coefficients in $\mathbf{c}_l^{u_i}$ are below some predetermined threshold λ, are removed from the library. After no new terms are removed in this step, a final

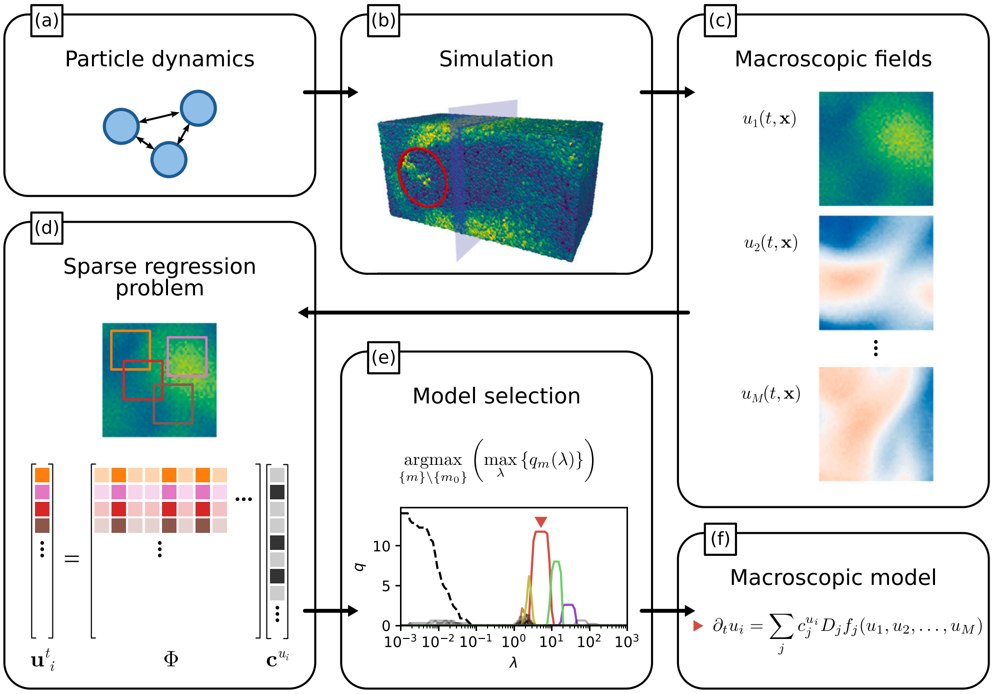

Fig. 1. Overview of data acquisition and the learning framework: (a) The fluid is described as a system of particles that evolve according to some force-field, in our case given by dissipative particle dynamics (Section 3.1). (b) The transient dynamics of the fluid are stored in particle trajectory data (Section 3.3). (c) Macroscopic fields $u_1(t,\mathbf{x})$, $u_2(t,\mathbf{x})$, …, $u_M(t,\mathbf{x})$ are calculated from particle trajectories; in our case the mass density $\rho(t,\mathbf{x})$ and the two components of the velocity field $v_x(t,\mathbf{x})$ and $v_y(t,\mathbf{x})$. (d) Random spatio-temporal regions are sampled from the coarse-grained fields and used to construct a sparse regression problem with a large amount of candidate terms (Section 2.1). (e) The regression is performed across a wide range of values of the sparsity control parameter λ and the best macroscopic model (f) is selected among the resulting models $\{m\}$ according to our novel measure q that balances stability and accuracy (Section 2.2).

least-squares fit is performed with the resulting sparse library without normalization. The threshold λ controls the sparsity of the result and is the only control parameter of this method. Its effect on the form of discovered models is discussed in Section 2.2.

The discrete approximations of spatio-temporal derivatives of noisy data are usually of low quality [37]. To mitigate the effects of noise in the estimation of these derivatives, several techniques have been proposed: spectral denoising [73], polynomial interpolation [37,74], and weak formulation of the problem [38,39]. Under the assumption that the given datapoints are arranged on a grid, we use the latter. We obtain a weak formulation of SINDy by multiplying Eq. (1) with a sample function $\psi_k(t,\mathbf{x})$ and subsequent spatio-temporal integration over the dataset domain:

$$\iint \partial_t u_i \, \psi_k \mathrm{d}t\mathrm{d}\mathbf{x} = \iint \sum_j c_j D_j f_j \, \psi_k \mathrm{d}t\mathrm{d}\mathbf{x} \, . \tag{2}$$

Eq. (2) is in principle valid for any choice of $\psi_k(t,\mathbf{x})$. In our work, we use local polynomial bumps that are nonzero only in cuboidal regions of $h_x \times h_y$ spatial and h_t temporal gridpoints and are centered around a given point in space–time (t_k, x_k, y_k):

$$\psi_k(t,x,y) = \tilde{\psi}\left(\frac{2}{h_t}\left\{t - t_k\right\}\right) \cdot \tilde{\psi}\left(\frac{2}{h_x}\left\{x - x_k\right\}\right) \cdot \tilde{\psi}\left(\frac{2}{h_y}\left\{y - y_k\right\}\right), \tag{3}$$

where $\tilde{\psi}$ is defined as [75]

$$\tilde{\psi}(x) = \begin{cases} (1-x^2)^{\mu+1} & ; -1 < x < 1 \\ 0 & ; \text{otherwise} \end{cases} . \tag{4}$$

Here, μ is the maximum order of a derivative in the library.

Integrating Eq. (2) by parts and taking into account that sample functions $\psi_k(t,\mathbf{x})$ are zero at the dataset domain boundary, leads to

$$-\iint u_i \, \partial_t \psi_k \mathrm{d}t\mathrm{d}\mathbf{x} = \sum_j c_j^{u_i} (-1)^{|\alpha_j|} \iint f_j \, D_j \psi_k \mathrm{d}t\mathrm{d}\mathbf{x} \, , \tag{5}$$

where $|\alpha_j|$ denotes the order of the corresponding derivative. As can be seen from Eq. (5), the weak formulation converts spatio-temporal derivatives of the noisy macroscopic fields into spatio-temporal derivatives of our sample functions $\psi_k(t,\mathbf{x})$, which are analytically known. The assumed form of the sought Eqs. (1) ensures that all derivatives can be converted in this way.

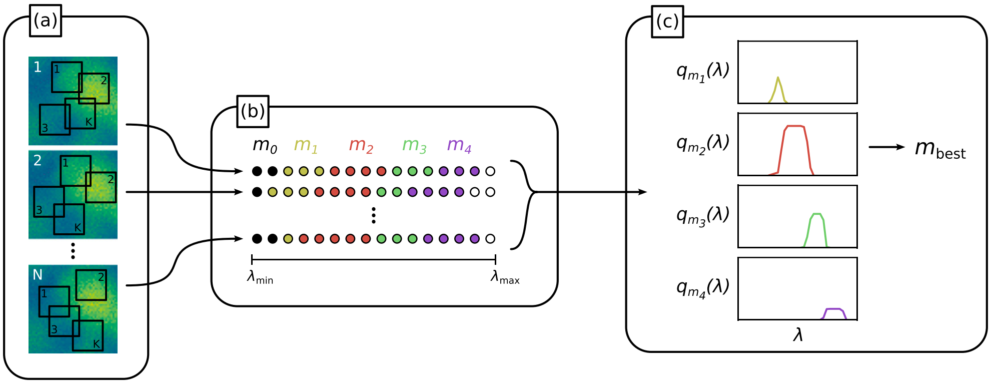

Fig. 2. The process of model selection: (a) N random sets of K spatio-temporal samples are taken from the dataset. (b) For each sample set, sparse regression is performed across a range of values of λ, resulting in various models m_i, flanked on one side by the model that contains all terms m_0 (black dots) and on the other by the empty model $m = \{\}$ (white dots). Only four other models, m_1, m_2, m_3 and m_4, are depicted to signify the fact that typically only a handful of them are discovered consistently, *i.e.* for most of the N random sets of samples. (c) The measure q, see Eq. (6), is calculated for each model as a function of λ and the model with the highest peak, in this case m_2, is selected as the optimal one, m_{best}, according to Eq. (7). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Just as Eq. (1) can be represented as a linear system for the unknown coefficients $c_j^{u_i}$ on a set of chosen sample points $\{\mathbf{x}_k\}$, Eq. (5) can be translated into a linear system on a set of chosen sample functions $\{\psi_k\}$ centered around $\mathbf{x}_k$. To this end, we redefine the quantities $[\mathbf{u}_i^t]_k$ and $[\Phi]_{jk}$ as integrals over space–time, $[\mathbf{u}_i^t]_k = -\iint u_i\, \partial_t \psi_k \mathrm{d}t\mathrm{d}\mathbf{x}$ and $[\Phi]_{jk} = (-1)^{|\alpha_j|} \iint f_j\, D_j \psi_k \mathrm{d}t\mathrm{d}\mathbf{x}$. In practice, these integrals are evaluated as sums over the $h_t \times h_x \times h_y$ gridpoints where the sample function is nonzero. A sparse solution for the unknown coefficients $c_j^{u_i}$ can then be extracted using the STLSQ method.

2.2. Model selection

Changing the sparsity control parameter λ produces models with varying degrees of complexity. With limited knowledge of the system, the choice of the optimal model is usually not clear, necessitating the use of some kind of selection criterion. A good model of macroscopic dynamics should first and foremost fit the data well and secondly, its form should be stable against resampling of the data. This ensures that the model describes the underlying physics and not just specific samples.

To identify models that are both accurate and stable, we introduce a selection measure q as the ratio of the probability p_m that a model with a certain set of terms $m = \{j : c_j \neq 0\}$ appears as the result and the average mean squared error (MSE) of models with this set of terms:

$$q_m = \frac{p_m}{\langle \text{MSE}_m \rangle} = \frac{n_m^2 K}{N \sum_{v=1}^{n_m} \|\mathbf{u}_{iv}^t - \Phi_v \mathbf{c}^{u_i}\|_2^2} \,. \tag{6}$$

The probability p_m is calculated as the number of times n_m that a set of terms m is extracted from some set of samples, divided by the number of these sample sets N, $p_m = \frac{n_m}{N}$. A particular set of samples consists of K sample functions $\psi_k(t, \mathbf{x})$, whose nonzero regions are randomly distributed in the space–time of the dataset. In principle, any measure $q = f(p, \text{MSE})$, where f is a strictly increasing function of p and a strictly decreasing function of MSE, would work as a selection measure — the ratio of the two quantities is merely the simplest possible choice.

To find the best model, we perform the regression on each of these sample sets for a range of values of the sparsity control parameter, spanning from $\lambda_{\min}$, where the STLSQ algorithm consistently removes no terms, to $\lambda_{\max}$, where all terms are removed. The model that achieves the highest value of this MSE-weighted probability q, which means that it is both stable (high probability p) and accurate (low MSE), is then selected as the best candidate for the true macroscopic model:

$$m_{\text{best}} = \underset{\{m\}\setminus\{m_0\}}{\operatorname{argmax}} \left(\max_{\lambda} \left\{ q_m(\lambda) \right\} \right) \,. \tag{7}$$

The model with no terms removed, m_0, must be excluded from the selection process, as it will always be stable for sufficiently low λ. The approach is summarized schematically in Fig. 2. The notion of stability against resampling of the data originates in stability selection [76], which has already been successfully applied to partial differential equation discovery [42,77]. In contrast to our work, stability selection considers probabilities of individual terms, regardless of which identified model they belong to, instead of entire models, which does not allow for an obvious way of incorporating a measure of accuracy into the criterion. Moreover, stability selection still requires manual selection of some threshold probability to determine the terms from which the final model is to be constructed. For a comparison of our approach with some of the other common methods of model selection, see Section 4.3.

3. Simulations

3.1. Dissipative particle dynamics

We model the simple fluid using DPD, where individual particles represent several atoms or molecules. To have more control over the viscosity of the fluid, we use both the standard [46] and the transverse [78] DPD thermostats. The force between particles i and j consists of a conservative force $\mathbf{F}_{ij}^C$, a dissipative force $\mathbf{F}_{ij}^D$ and a random force $\mathbf{F}_{ij}^R$, which gives a total force $\mathbf{F}_i$ on particle i:

$$\mathbf{F}_i = \sum_{j\neq i} \mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R, \tag{8}$$

where

$$\mathbf{F}_{ij}^C = a_{ij}\omega_C(r_{ij})\hat{\mathbf{r}}_{ij}, \tag{9}$$

$$\mathbf{F}_{ij}^D = -\gamma^{\parallel}\omega_D^{\parallel}(r_{ij})\left(\hat{\mathbf{r}}_{ij}\otimes\hat{\mathbf{r}}_{ij}\right)\cdot\mathbf{v}_{ij} - \gamma^{\perp}\omega_D^{\perp}(r_{ij})\left(\mathrm{I}-\hat{\mathbf{r}}_{ij}\otimes\hat{\mathbf{r}}_{ij}\right)\cdot\mathbf{v}_{ij}, \tag{10}$$

$$\mathbf{F}_{ij}^R = \sigma^{\parallel}\omega_R^{\parallel}(r_{ij})\left(\hat{\mathbf{r}}_{ij}\otimes\hat{\mathbf{r}}_{ij}\right)\cdot\boldsymbol{\xi}_{ij} + \sigma^{\perp}\omega_R^{\perp}(r_{ij})\left(\mathrm{I}-\hat{\mathbf{r}}_{ij}\otimes\hat{\mathbf{r}}_{ij}\right)\cdot\boldsymbol{\xi}_{ij}, \tag{11}$$

where a_{ij} is the interaction parameter between beads i and j, $\gamma^{\parallel}$ and $\gamma^{\perp}$ are the friction parameters, $\sigma^{\parallel}$ and $\sigma^{\perp}$ are the noise strengths, ω_C, $\omega_D^{\parallel}$, $\omega_D^{\perp}$, $\omega_R^{\parallel}$ and $\omega_R^{\perp}$ are the weight functions, $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ is the relative velocity of the two interacting particles, $\hat{\mathbf{r}}_{ij} = \frac{\mathbf{r}_i-\mathbf{r}_j}{|\mathbf{r}_i-\mathbf{r}_j|}$ is the normalized vector along the inter-particle axis, and $\boldsymbol{\xi}_{ij}$ denotes a random unit vector. $\boldsymbol{\xi}_{ij}$ is uncorrelated with respect to different pairs of particles i and j and antisymmetric with respect to a swap of i and j ($\boldsymbol{\xi}_{ij} = -\boldsymbol{\xi}_{ji}$):

$$\langle \boldsymbol{\xi}_{ij}(t)\otimes\boldsymbol{\xi}_{kl}(t')\rangle = \mathrm{I}(\delta_{ik}\delta_{jl} - \delta_{jk}\delta_{il})\delta(t-t'), \tag{12}$$

which ensures that linear momentum is conserved.

The conservative part of the force is a linearly decreasing function of r_{ij}:

$$\omega_C(r_{ij}) = \begin{cases} 1-\frac{r_{ij}}{r_c}, & r_{ij} < r_c \\ 0, & r_{ij} \geq r_c \end{cases}, \tag{13}$$

where r_c is the cutoff and is chosen as the length scale in our simulations. To keep the system at a constant temperature T_0, the following relations must hold [46,78]:

$$\omega_D^{\parallel} = {\omega_R^{\parallel}}^2, \qquad 2\gamma^{\parallel}k_BT_0 = {\sigma^{\parallel}}^2, \tag{14}$$

$$\omega_D^{\perp} = {\omega_R^{\perp}}^2, \qquad 2\gamma^{\perp}k_BT_0 = {\sigma^{\perp}}^2, \tag{15}$$

where k_B is the Boltzmann constant. Typically, $\omega_R^{\parallel}(r_{ij})$ and $\omega_R^{\perp}(r_{ij})$ are set equal to $\omega_C(r_{ij})$.

DPD closely reproduces hydrodynamic behavior of a viscous fluid on the mesoscale. As shown in Refs. [79,80] for the $\gamma^{\perp} = 0$ case, local conservation laws for the mass density and the density of linear momentum can be recovered using a Fokker–Planck approach. The same cannot be done for the energy density, which means that standard DPD cannot support temperature gradients. This can be remedied by adding an internal energy variable to each particle [81,82]. In this work, however, we restrict ourselves to standard isothermal DPD simulations and focus on the mass and linear momentum densities.

Using DPD, we model the fluid in a cuboidal cell with periodic boundary conditions, see Fig. 3. We simulate a one-component fluid where all particles have the same mass m and all conservative interactions have the same strength $a_{ij} = a$. We set the DPD parameters to commonly used values [83]: interaction amplitudes to $a = 25\,k_BT_0/r_c$ and $\gamma^{\parallel} = \gamma^{\perp} = 4.5\,m/\tau$ and mass density to $\rho_0 = 3\,m/r_c^3$. With the chosen DPD parameter values, a timestep of $\Delta t = 0.01\,\tau$ leads to a thermally stable simulation. We take the interaction cutoff radius r_c, the particle mass m and the equilibrium thermal energy k_BT_0 as units of length, mass and energy, respectively. The characteristic unit of time is therefore $\tau = r_c\sqrt{m/k_BT_0}$.

3.2. Measuring shear viscosity

To compare the value of the extracted shear viscosity η with its actual value, we measure it using the periodic Poiseuille flow method [84]. We simulate the periodic Poiseuille flow in a simulation cell with dimensions $L_x\times L_y\times L_z = 40\times20\times20\,r_c^3$. We apply a constant external force $\mathbf{F}^{\mathrm{ext}} = \mathrm{sgn}(L_x/2-x)F_y\hat{\mathbf{e}}_y$ with amplitude $F_y = 0.002\,\frac{k_BT_0}{r_c}$ along the y axis to all particles. The sign reversal in the middle of the simulation cell and the periodic boundary conditions mimic infinite parallel plates in the yz plane. After running the simulation for $500\,\tau$, we calculate the viscosity from the y component of the resulting velocity field via [84]

$$\eta = \frac{\rho_0F_yL_x^2}{12\;m\langle v_y\rangle}, \tag{16}$$

where $\langle v_y\rangle = \frac{1}{L_xL_yL_z}\int \mathrm{sgn}(L_x/2-x)v_y\mathrm{d}x\mathrm{d}y\mathrm{d}z$. We determine the accuracy of the result by repeating the simulation 10 times with different initial positions of the particles. The average measured shear viscosity is then equal to $\eta = (1.332 \pm 0.005)\,\frac{m}{r_c\tau}$.

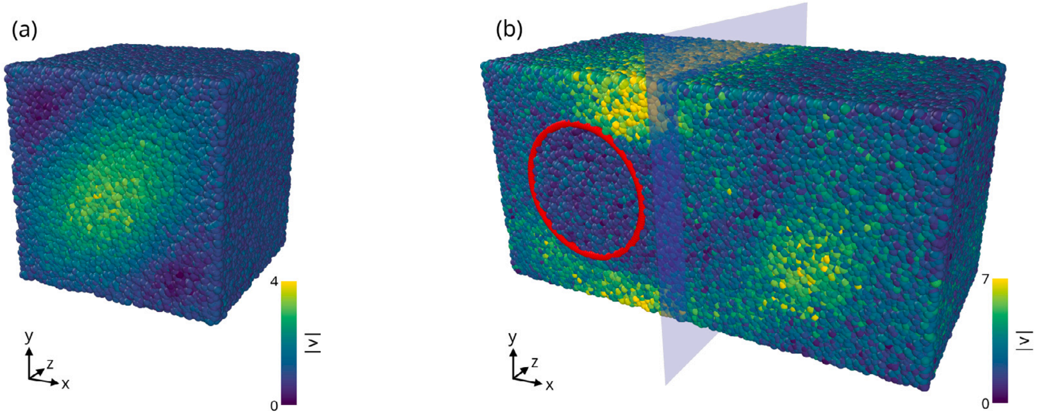

Fig. 3. State of the two simulations at the start of (a) or during (b) data acquisition. In Simulation A (a), the initial condition is a velocity profile given by Eq. (17), while in Simulation B (b), the flow is induced by an external force acting on all particles in a narrow region on the left side of the simulation cell, $x < 1\,r_c$. The particles are colored according to the magnitude of their velocity, except for the frozen particles, which are marked in red. In Simulation B, the data to the left of the blue plane is discarded. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Datasets for model discovery

Since the learning framework relies on both temporal and spatial derivatives of the coarse-grained fields, it is crucial to simulate and capture the transient flow dynamics of the fluid. The two simulation setups that we use to generate suitable datasets are therefore: relaxation of an initial velocity profile (Simulation A) and flow past a cylinder (Simulation B), see Fig. 3. In both simulations, initial and boundary conditions are translationally invariant along the z axis and an averaging is performed along this axis when computing the datasets. This allows us to control the amount of noise in the data by simply varying the size of the simulation cell in this direction.

For the relaxing velocity profile, the simulation cell is a cube with a side length of $L_x = L_y = L_z = 50\,r_c$. The initial velocity profile of the particles is set to

$$\mathbf{v}(x, y) = v_0 \begin{bmatrix} \sin(2\pi x/L_x) + \sin(2\pi y/L_y) - \sin(2\pi\{x/L_x + y/L_y\}) \\ \cos(2\pi y/L_y) + \cos(2\pi x/L_x) - \cos(2\pi\{x/L_x + y/L_y\}) \end{bmatrix}, \tag{17}$$

with $v_0 = 1\,\frac{r_c}{\tau}$, and allowed to relax for $50\,\tau$.

For the flow past a cylinder, the simulation cell has dimensions $L_x \times L_y \times L_z = 100 \times 50 \times 50\,r_c^3$. The cylinder has a diameter of $d = 30\,r_c$, is centered at $x = 25\,r_c$, $y = 25\,r_c$ and oriented along the z axis. It is represented by a layer of frozen DPD particles surrounding the structure [85,86], uniformly distributed with a surface density of $3\,m/r_c^2$. The flow is induced by a constant external force $\mathbf{F}^{\text{ext}} = F_x\hat{\mathbf{e}}_x$ with amplitude $F_x = 5\,\frac{k_B T_0}{r_c}$, which acts along the x axis on all particles between $x = 0$ and $x = 1\,r_c$. Before the data acquisition begins, we simulate the system for $200\,\tau$. During this time, the flow reaches an average speed of about $\langle v_x \rangle \approx 1.6\,\frac{r_c}{\tau}$ and vortices start to shed from the cylinder. With the Reynolds number of about $\text{Re} = \frac{\rho_0 d \langle v_x \rangle}{\eta} \approx 120$, this is expected and is in line with literature [87]. To make the resulting dataset comparable with that of Simulation A, the simulation is then continued for $50\,\tau$ and only the right half ($x > 50\,r_c$) of the simulation cell is considered.

The datasets of both simulations, illustrated in Fig. 4, are then obtained by binning the particles in a 50×50 square grid in x and y coordinates, and recorded at every 10th timestep, resulting in time intervals of $10\Delta t = 0.1\,\tau$. Three fields are calculated: the mass density ρ, the horizontal velocity v_x and the vertical velocity v_y. As a general rule the dataset should capture both the slowest (*e.g.* viscous) and the fastest (*e.g.* sonic) timescales of the systems. To do that one has to perform sufficiently long simulations and record the trajectories frequently enough. For our case the duration of the simulations of $50\,\tau$ and time intervals of $0.1\,\tau$ were sufficient to discover the correct dynamic equations.

4. Results and discussion

The first step in discovering the macroscopic dynamics is to select the set of relevant variables to appear on the left-hand side of Eq. (1). Here, we choose the mass density ρ and the density of linear momentum $\mathbf{g} = (g_x, g_y) = (\rho v_x, \rho v_y)$, which are the only variables in standard DPD whose dynamics can be written as a conservation law [79,80].

The next step is to construct the library of candidate terms that appear on the right-hand side of Eq. (1). We choose to include all terms that consist of derivatives up to the second order acting on products of the three fields (ρ, v_x and v_y) up to third order, so that the library contains convective terms, such as $\partial_y \rho v_x v_y$, as well as dissipative terms, such as $\partial_x^2 v_x$. We also include product terms without derivatives. On the other hand, products of derivatives, such as $\partial_y v_x \partial_x v_y$, cannot be included in the library, as they

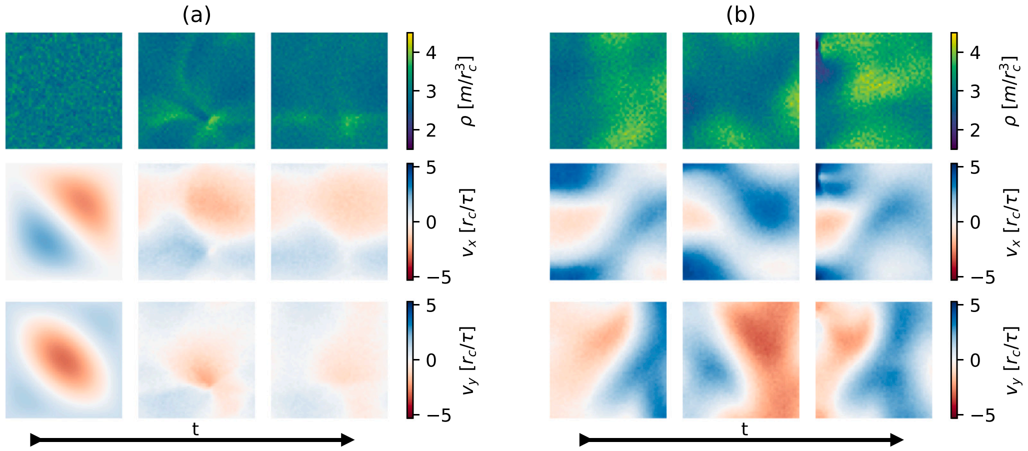

Fig. 4. Three timeslices of density (ρ) and velocity (v_x and v_y) profiles for Simulation A, velocity relaxation, (a) and Simulation B, flow behind a cylinder, (b), recorded at $t = 0$, $t = 25\,\tau$ and $t = 50\,\tau$ after the start of data acquisition. The passage of time is indicated by the arrows.

cannot be converted as described in Section 2.1. The total number of candidate terms, 114, could be reduced by taking the spatial symmetries of the system into account [39,42]. However, we choose to use the full, unconstrained library for all three equations — as a test of the algorithm's ability to identify the symmetries on its own.

Since the mass density remains close to its mean value ρ_0, we use the zero mean density $\tilde{\rho} = \rho - \rho_0$ when constructing the library. This reduces the correlation between the terms, at the cost of potentially increasing the number of terms required to describe the dynamics.

To run the sparse regression algorithm on our datasets, we use the open-source Python package pySINDy [75].

4.1. Model discovery

To extract the dynamic models for the mass density ρ and the density of linear momentum $\mathbf{g}$, we perform the sparse regression on $N = 50$ sets of $K = 2000$ sample functions with $h_x = h_y = 18$ and $h_t = 8$, for 100 values of the sparsity control parameter, logarithmically spaced from $\lambda_{\min} = 10^{-3}$ to $\lambda_{\max} = 10^3$. These values remain constant throughout this work. We have observed that the peak discovery probability of many models falls sharply for h_t less than 8 and h_x, h_y less than 18. A more systematic study of the effects of these quantities is envisioned in future research. We assume that the system is sufficiently ergodic, so that the samples, whose temporal dimension h_t is much smaller than the total duration of the numerical simulations, are not meaningfully statistically dependent.

Calculating first just the probabilities of identification of specific models (Fig. 5a, c, e, g, i and k), we observe the same general behavior when increasing λ for all three equations and both datasets. Initially, the probability of identifying the model with no terms removed (dashed line) decreases steadily and reaches zero at $\lambda \approx 0.03$ in the case of the density equation, or $\lambda \approx 0.1$ in the case of the momentum equation. This is followed by an intermediate region, in which no model is consistently identified. Finally, at $\lambda \approx 0.2$ in the case of the density equation, or $\lambda \approx 1$ in the case of the momentum equation, stable sparse models begin to appear, many of which reach $p = 1$.

In the case of the density equation, for both datasets, two nonzero models achieve $p = 1$ (red and green in Fig. 5a and b). Among them, the MSE-weighted probability q (Fig. 5c and d) clearly selects the more complex one. Its four terms correspond exactly to the mass continuity equation

$$\partial_t \rho = -\nabla \cdot (\rho \mathbf{v}) = -\rho_0 \partial_x v_x - \rho_0 \partial_y v_y - \partial_x \tilde{\rho} v_x - \partial_y \tilde{\rho} v_y \, . \tag{18}$$

As shown in Table 1, the averages of its four coefficients also agree, up to at least two decimal places. The other stable model corresponds to a simplification of the mass continuity equation $\partial_t \rho = -\rho_0 \nabla \cdot \mathbf{v}$ where spatial fluctuations of density have been neglected.

For both components of the momentum equation, the procedure again selects the same model (red in Fig. 5f, h, j and l) for each of the two datasets. Both of these selected models have eight terms, which can be arranged into the corresponding component of a form of the compressible Navier–Stokes equation:

$$\partial_t \mathbf{g} = -\nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) + \eta \nabla^2 \mathbf{v} - \nabla P(\rho) \tag{19}$$

$$\partial_t g_i = -\rho_0 \left(\partial_x v_x v_i + \partial_y v_y v_i \right) - \left(\partial_x \tilde{\rho} v_x v_i + \partial_y \tilde{\rho} v_y v_i \right) + \eta \left(\partial_x^2 v_i + \partial_y^2 v_i \right) - \frac{1}{m} \left(k_B T + 2\alpha a \frac{\rho_0}{m} \right) \partial_i \tilde{\rho} - \frac{\alpha a}{m^2} \partial_i \tilde{\rho}^2 \, , \tag{20}$$

Table 1
Discovered model terms for the time derivative of mass density, together with expected and discovered values of the corresponding coefficients. The distributions of the discovered values are represented by their mean value and standard deviation. For clarity, units have been left out.

c_j^ρ	$D_j f_j$	Expected value	Simulation A	Simulation B
$-\rho_0$	$\partial_x v_x$	-3	-3.0001 ± 0.0023	-2.9980 ± 0.0047
$-\rho_0$	$\partial_y v_y$	-3	-3.0002 ± 0.0019	-2.9984 ± 0.0041
-1	$\partial_x \tilde{\rho} v_x$	-1	-1.0005 ± 0.0027	-0.9995 ± 0.0021
-1	$\partial_y \tilde{\rho} v_y$	-1	-1.0017 ± 0.0036	-0.9993 ± 0.0026

Table 2
Discovered model terms for the time derivative of each component of linear momentum density, together with expected and discovered values of the corresponding coefficients. The distributions of the discovered values are represented by their mean value and standard deviation. For clarity, units have been left out.

$c_j^{g_x}$	$D_j f_j$	Expected value	Simulation A	Simulation B
$-\rho_0$	$\partial_x v_x^2$	-3	-2.993 ± 0.006	-2.992 ± 0.002
$-\rho_0$	$\partial_y v_x v_y$	-3	-2.978 ± 0.006	-3.009 ± 0.002
-1	$\partial_x \tilde{\rho} v_x^2$	-1	-1.020 ± 0.010	-0.995 ± 0.003
-1	$\partial_y \tilde{\rho} v_x v_y$	-1	-1.031 ± 0.012	-1.013 ± 0.004
η	$\partial_x^2 v_x$	1.332	2.279 ± 0.047	2.244 ± 0.064
η	$\partial_y^2 v_x$	1.332	1.344 ± 0.031	1.419 ± 0.024
$-(1+2\alpha a \rho_0)$	$\partial_x \tilde{\rho}$	-16.150	-16.043 ± 0.017	-16.132 ± 0.036
$-\alpha a$	$\partial_x \tilde{\rho}^2$	-2.525	-2.678 ± 0.017	-2.578 ± 0.035
$c_j^{g_y}$	$D_j f_j$	Expected value	Simulation A	Simulation B
$-\rho_0$	$\partial_x v_x v_y$	-3	-3.004 ± 0.007	-3.007 ± 0.002
$-\rho_0$	$\partial_y v_y^2$	-3	-2.997 ± 0.007	-3.020 ± 0.004
-1	$\partial_x \tilde{\rho} v_x v_y$	-1	-1.009 ± 0.016	-1.000 ± 0.003
-1	$\partial_y \tilde{\rho} v_y^2$	-1	-1.021 ± 0.014	-1.004 ± 0.006
η	$\partial_y^2 v_y$	1.332	2.248 ± 0.042	1.677 ± 0.050
η	$\partial_x^2 v_y$	1.332	1.498 ± 0.059	1.834 ± 0.036
$-(1+2\alpha a \rho_0)$	$\partial_y \tilde{\rho}$	-16.150	-16.084 ± 0.012	-16.143 ± 0.024
$-\alpha a$	$\partial_y \tilde{\rho}^2$	-2.525	-2.624 ± 0.015	-2.668 ± 0.026

where i is either x or y. The last two terms in Eq. (20), which describe the pressure gradient, correspond to the well-known semi-empirical equation of state for one-component DPD systems [83]:

$$P = \rho_n k_B T + \alpha a \rho_n^2, \tag{21}$$

where $\alpha = (0.101 \pm 0.001)\, r_c^4$ is an empirical parameter and $\rho_n = \frac{\rho}{m}$ is the number density of particles.

Comparing the probability plots for the two datasets (Fig. 5e, g, i and k), it seems that the identification of the Navier–Stokes equation is less probable using data from Simulation B, as the interval of λ, for which this model is always discovered is much narrower. The next model that is common to both components and both datasets is the six-term model (marked in purple) corresponding to Eq. (20) with $\eta = 0$. The fact that the viscosity terms are always the first to be thresholded out suggests that viscosity effects contribute the least to the overall dynamics of these two simulations and are therefore the most difficult to discover.

The coefficients of the two discovered eight-term dynamic equations are collected in Table 2. The averages of the four non-phenomenological parameters ($-\rho_0$ or -1) agree with the theory in at least two significant digits for both components and both datasets. The extracted values for the viscosity η are slightly larger than the value calculated from the Poiseuille flow in Section 3.1. Interestingly, in both datasets, the coefficient in front of $\partial_x^2 v_x$ ($\partial_y^2 v_y$) is noticeably larger than the coefficient in front of $\partial_y^2 v_x$ ($\partial_x^2 v_y$), perhaps indicating a contribution from the bulk viscosity of the fluid. Indeed, a model with nonzero bulk viscosity, containing an additional, mixed derivative term $\partial_x \partial_y v_x$ is identified in the data from Simulation A (yellow in Fig. 5f and j), but not consistently enough to be selected by our measure. A different simulation setup, in which the fluid is compressed more, might make the identification of this model more probable.

We find an excellent agreement between the extracted values for the pressure gradient terms and the DPD equation of state, Eq. (21). The physical significance of the extracted terms related to the pressure gradient is the following: the coefficient of the linear term $\nabla_x \tilde{\rho}$ is equal to the squared speed of sound $c^2 = \frac{1}{m}(k_B T + 2\alpha a \frac{\rho_0}{m})$, while the coefficient of the nonlinear term $\nabla_x \tilde{\rho}^2$ corresponds to one half of the so-called Beyer's nonlinear acoustic parameter [88]: $\frac{B}{\rho_0^2} = \left(\frac{\partial^2 P}{\partial \rho^2}\right)_{\rho=\rho_0} = \frac{2\alpha a}{m^2}$. This parameter plays an important role in theranostic biomedical applications of ultrasound, where the assumption of linearity is often not valid [89].

We also note that most of the coefficients of the two components of the same vector term (*e.g.* $\partial_x \tilde{\rho}$ and $\partial_y \tilde{\rho}$) agree even better with each other than with their expected value, even though we have not provided any explicit symmetry constraints. This implies that the algorithm has independently discovered the vectorial nature of momentum.

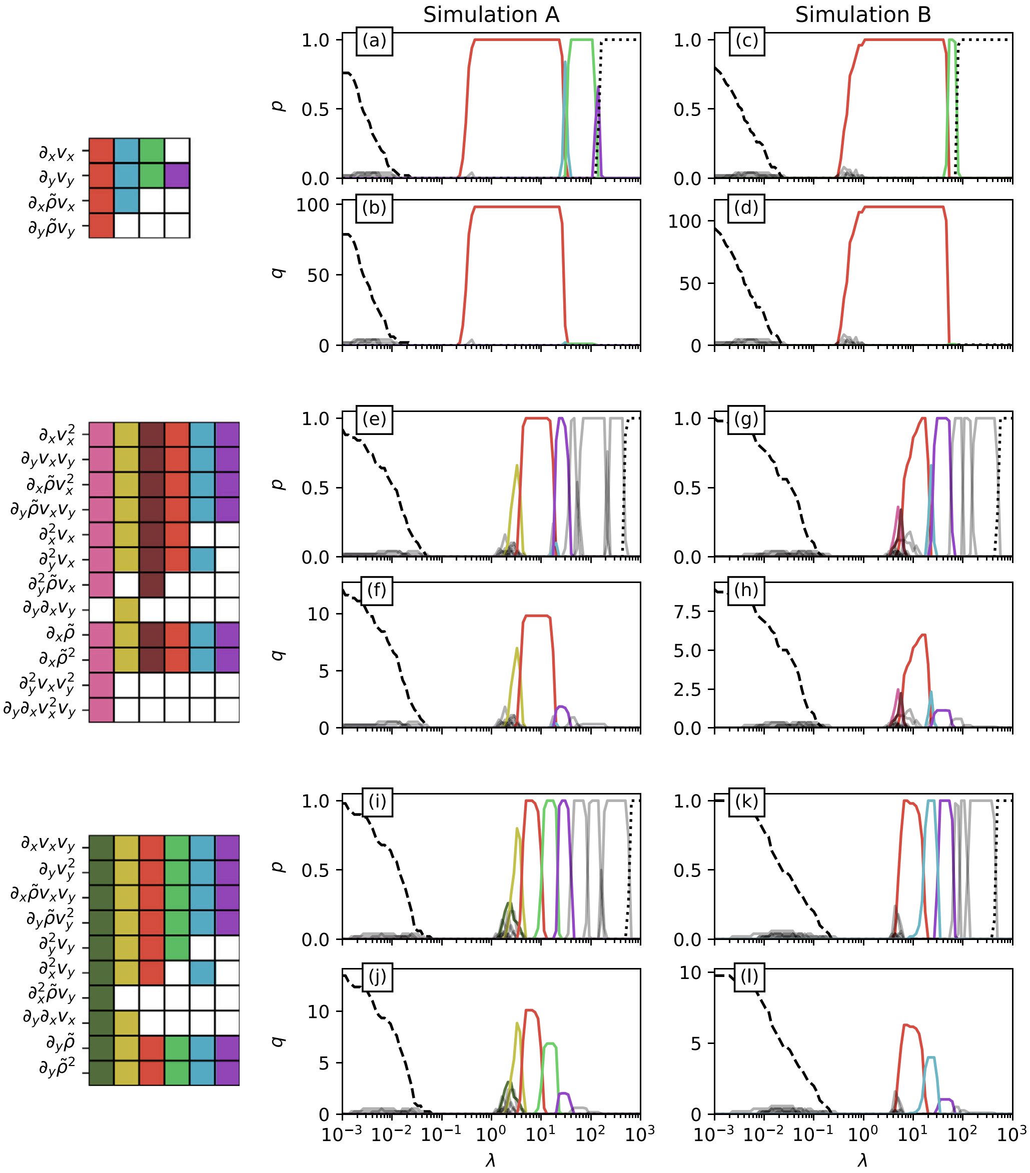

Fig. 5. Identification probabilities of the discovered models and their MSE-weighted counterparts for the density equation (a, b, c, d), the x-component of the momentum equation (e, f, g, h) and the y-component of the momentum equation (i, j, k, l) for different values of the sparsity control parameter λ. Only lines of models that appear at least twice at any one λ are shown. The dashed line marks the model with all possible terms, while the dotted line marks the model with all terms thresholded out. Apart from these two models, up to six models with the largest number of terms that reach at least $p = 0.25$ are marked in color and have their terms listed in the legend. All others are marked by transparent black lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Robustness

To test the effect of noise on the results of the presented method, we run Simulation A in shallower (lower L_z) simulation cells, thus reducing the number of particles used in the calculation of input fields.

As can be seen from Fig. 6, the algorithm successfully identifies and selects the continuity equation at all tested cell thicknesses, with its maximum q being twice as high as that of the next best model even at $L_z = 1\,r_c$. The x-component of the Navier–Stokes equation is identified up to $L_z = 2\,r_c$, although the q-based selection chooses the inviscid model for $L_z < 4\,r_c$. The y-component of the Navier–Stokes equation is similarly identified for all cell thicknesses except $L_z = 1\,r_c$, but is only selected as the best candidate up to $L_z = 14\,r_c$. Then, it is replaced by the model without the $\partial_x^2 v_y$ term, which in turn is later superseded, again, by the model without any viscous terms.

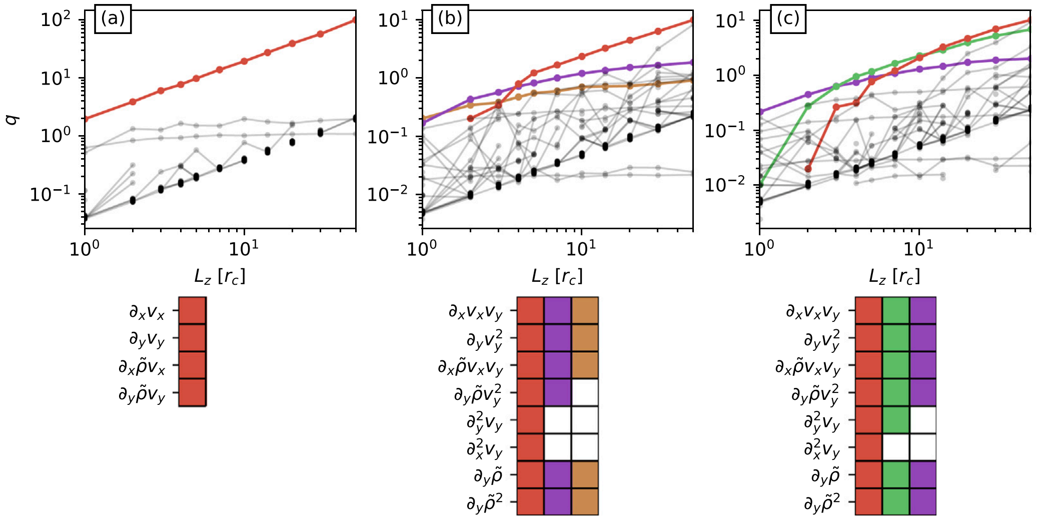

Fig. 6. Maximum values of q of particular models for the density equation (a) x-component of the momentum equation (b) and y-component of the momentum equation (c) for different thicknesses L_z of the simulation cell when running Simulation A. Only models that have the highest maximum q at any L_z are marked in color and have their terms listed in the legend. The model with all possible terms and the empty model are not shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

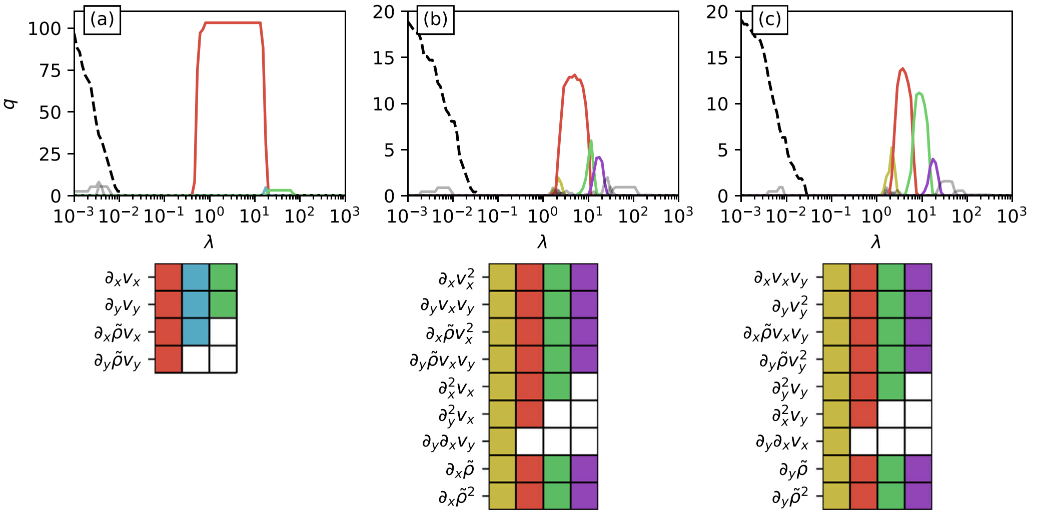

Fig. 7. MSE-weighted probabilities of discovered models for the density equation (a), the x-component of the momentum equation (b) and the y-component of the momentum equation (c) for different values of the sparsity control parameter λ, using data from Simulation A and a 510-term library. Only lines of models that appear at least twice at any one λ are shown. The dashed line marks the model with all possible terms. For each plot, the three or four sparse models with highest maximum values of q are marked in color and have their terms listed in the legend. All others are marked by transparent black lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The presented method also works well with libraries that are much larger than the one we have been using so far. To demonstrate this, we perform a sweep of the sparsity control parameter on the data from Simulation A with a library containing products of input fields up to the fourth order and derivatives up to the fourth order, amounting to 510 different terms. As it can be seen in Fig. 7, even with this larger library the continuity and the Navier–Stokes equations remain the selected macroscopic models. For an even larger library, the seven-term model for the y-component of the momentum equation (Fig. 7c) achieves a higher maximum q than the eight-term model. Interestingly, and perhaps more importantly, although a larger library enables the discovery of more complex dynamics, no such models appear.

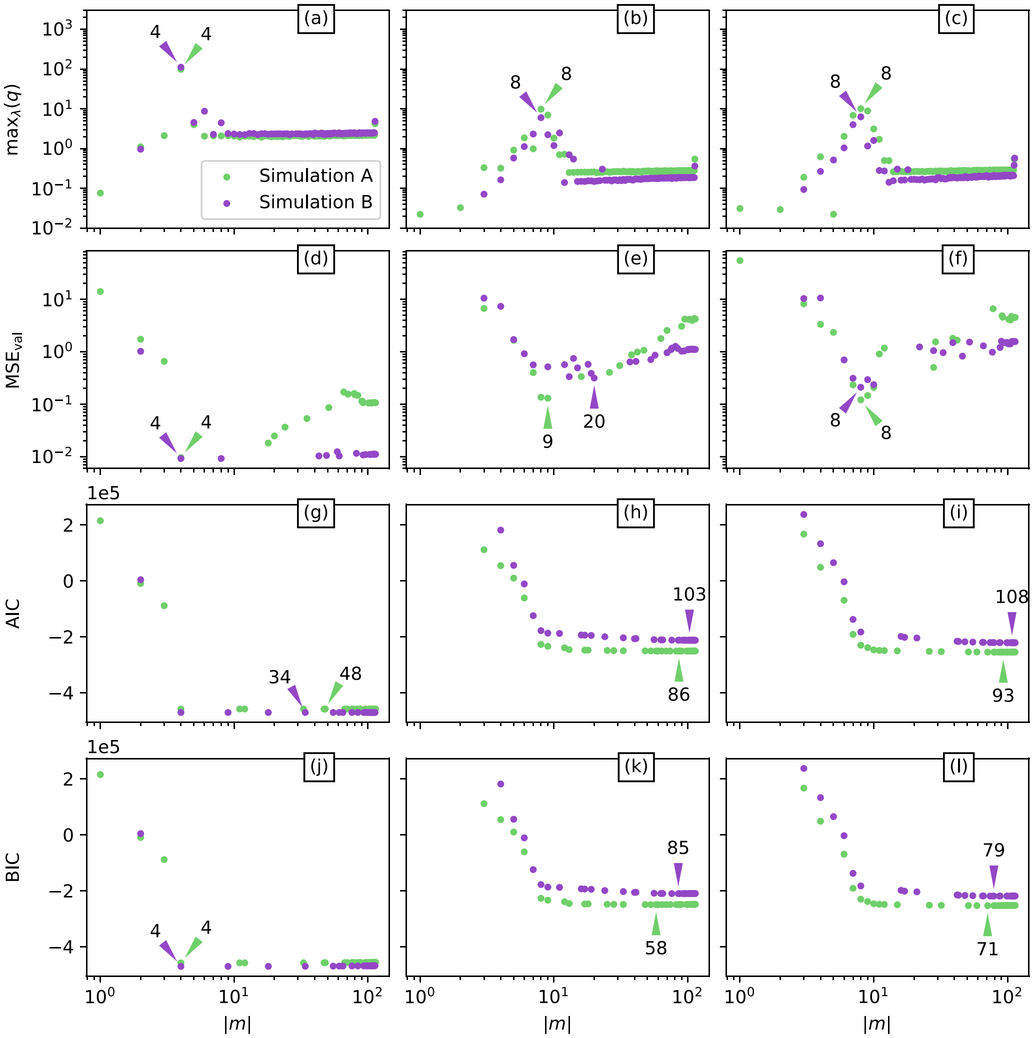

Fig. 8. Dependence of maximum q, the validation MSE, the AIC and the BIC on the number of terms in the models found for the density equation (first column, a, d, g, j), the x-component of the momentum equation (second column, b, e, h, k) and the y-component of the momentum equation (third column, c, f, i, l) for both datasets. If several models with the same number of terms were found, the better value (higher in the case of $\max(q)$ and lower in the other three cases) is displayed. The number of terms for which a particular criterion reaches its extremum is marked with an arrow. For the linear momentum equations, the minima of AIC and BIC (h, i, k, l) select a much larger set of terms than our measure (b, c).

4.3. Other model selection methods

We compare our model selection method (Eq. (7)) with some of the other commonly used automatic approaches: validation on a separate dataset, Akaike information criterion (AIC) [90] and Bayesian information criterion (BIC) [91]. To facilitate the comparison, the results from Fig. 5 are presented compactly in Fig. 8a, b and c.

To generate two independent sets of samples, one for regression and one for validation, we split our dataset spatially. We take $NK = 10^5$ samples only from the top half of the data ($y > 25r_c$), run the regression, and compute, for each discovered model, the MSE on NK samples taken from the bottom half of the data. For the density equation (Fig. 8d) and the y-component of the momentum equation (Fig. 8f), the models with the minimal validation error are the correct, four- and eight-term models for both datasets. However, for the x-component of the momentum equation (Fig. 8e), the two selected models both have too many terms.

Information criteria are model selection measures that explicitly penalize the number of terms in a model and, like our method, do not require a separate validation dataset. AIC-based selection has already proved successful on some equation-finding problems [92], while a custom but similar criterion was used in Refs. [38,43]. We again perform model discovery with NK samples, this time taken from the entire dataset, and compute the two criteria as follows [93]:

$$\mathrm{AIC} = NK \ln \frac{\|\mathbf{u}_i^t - \Phi \mathbf{c}^{u_i}\|_2^2}{NK} + 2|m| \,, \tag{22}$$

$$\mathrm{BIC} = NK \ln \frac{\|\mathbf{u}_i^t - \Phi \mathbf{c}^{u_i}\|_2^2}{NK} + 2|m| \ln(NK)\,, \tag{23}$$

where $|m|$ denotes the number of terms in m. As can be seen in Fig. 8g, h, i, j, k and l, both criteria struggle to select the correct model, opting for a much larger model most of the time. Incidentally, a clear kink can be seen at the correct number of terms for both criteria, indicating that a manual Pareto analysis, similar to Ref. [44], could also be used for model selection.

5. Conclusions and outlook

In this work, we extracted macroscopic dynamic equations from particle trajectory data acquired from DPD simulations. We considered a simple fluid where the relationship between the model parameters and the macroscopic phenomenological coefficients is well understood. Prior knowledge of the sought macroscopic law, *i.e.* the compressible Navier–Stokes equations, allowed us to introduce and test a new model selection measure. Weighting the stability of a model with its average mean squared error produced a criterion that can select suitable parsimonious models without the need to manually set either the sparsity control parameter or any other threshold. For both mass density and linear momentum density, this measure correctly selected the full continuity equation and momentum conservation, respectively. In contrast, model selection based on a validation dataset or information criteria was unsuccessful.

The learning framework correctly identified the pressure equation of state. The obtained parameters of the equation of state, *i.e.* the speed of sound and the Beyer's nonlinear acoustic parameter, are in excellent quantitative agreement with those derived from the semi-empirical relation for one-component DPD systems. This method could therefore serve as an alternative, potentially more efficient method for measuring the pressure equation of state. The claim for efficiency can be argued from the fact that only a single simulation is required to discover the equation of state. In contrast, current methods for measuring the equation of state involve performing large number of simulations at different densities and measuring the equilibrium pressure or using open boundary molecular dynamics [94] and measuring the equilibrium density at different values of the external pressure.

The presented method is extremely robust to noise. It is able to identify momentum conservation at low simulation cell thicknesses, while mass conservation is reliably discovered even at the almost absurdly low thickness of $L_z = 1\,r_c$, where each dataset bin contains on average only about 3 particles. It also works quantitatively well even when using an extensive library with very few physics-informed assumptions to constrain its size. In particular, no symmetry assumptions were made at all.

One of the drawbacks of the weak formulation approach as used in this work is that it can only discover terms expressed as derivatives of some functions of the input macroscopic variables. This has no influence on the discovery of linearized dynamic equations. On the other hand, nonlinear effects, which cannot be expressed in this way, *e.g.* the dependence of viscosity on density, can only be discovered by including the required derivatives in the set of input fields. Such derivatives must then be evaluated in advance using some other method, and cannot benefit directly from the weak formulation.

The fact that the presented scheme requires only a few physics-based assumptions, is able to automatically select a suitable macroscopic model and is highly robust to noise makes it ideal for application to particle simulations or experimental data of more complex systems — as long as the data is either in the form of fields or such fields can be reconstructed from it. Examples of systems, where macroscopic dynamic laws could be extracted in such a way, include the motility of cells, bacteria and other objects captured by biomedical imaging tools, the collective dynamics of proteins, and systems with variables arising from spontaneous symmetry breaking, *e.g.* liquid crystal phases or ferromagnetic fluids. The learning framework could also be applied to heterogeneous systems with a spatially varying density field, such as metastable fluids or gels, where a 2-fluid description could come into play [95,96]. In this work we have focused on discovering the bulk dynamics of the fluid, disregarding the dynamics at the boundary of the system (*e.g.* on the surface of the cylindrical obstacle). To discover the form of the boundary conditions one could use the approach described in Ref. [97], where a physics-informed neural network (PINN) was used to infer the macroscopic fields on the boundary given some measurement points in the bulk and at or near the boundary. In a more recent work [98] PINNs were used in a subdomain of a given system and coupled to classical numerical methods in a unified framework. Studies along these lines will be pursued in our future work.

CRediT authorship contribution statement

Matevž Jug: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Daniel Svenšek:** Writing – review & editing, Methodology, Conceptualization. **Tilen Potisk:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization, Software, Funding acquisition. **Matej Praprotnik:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge the financial support under Grant Nos. P1-0002 and J1-50035 from the Slovenian Research and Innovation Agency. The financial support through ERC Advanced Grant (Grant No. 885155) from the European Research Council is also gratefully acknowledged.

References

[1] P.C. Martin, O. Parodi, P.S. Pershan, Unified hydrodynamic theory for crystals, liquid crystals, and normal fluids, Phys. Rev. A 6 (6) (1972) 2401–2420, http://dx.doi.org/10.1103/PhysRevA.6.2401.

[2] H. Pleiner, H.R. Brand, Hydrodynamics and Electrohydrodynamics of Liquid Crystals, in: A. Buka, L. Kramer (Eds.), Pattern Formation in Liquid Crystals, Springer New York, New York, NY, 1996, pp. 15–67, http://dx.doi.org/10.1007/978-1-4612-3994-9_2.

[3] D. Forster, Hydrodynamic fluctuations, broken symmetry, and correlation functions, in: W.A. Benjamin (Ed.), Advanced Book Program, 1975.

[4] S. Groot, P. Mazur, Non-Equilibrium Thermodynamics, North-Holland Publishing Company, 1962.

[5] M. Allen, D. Tildesley, Computer Simulation of Liquids, Oxford University Press, 2017.

[6] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, Nat. Struct. Biol. 9 (9) (2002) 646–652, http://dx.doi.org/10.1038/nsb0902-646.

[7] C. Hijón, P. Español, E. Vanden-Eijnden, R. Delgado-Buscalioni, Mori–Zwanzig formalism as a practical computational tool, Faraday Discuss. 144 (2010) 301–322, http://dx.doi.org/10.1039/B902479B.

[8] J. Jin, A.J. Pak, A.E.P. Durumeric, T.D. Loose, G.A. Voth, Bottom-up Coarse-Graining: Principles and Perspectives, J. Chem. Theory Comput. 18 (10) (2022) 5759–5791, http://dx.doi.org/10.1021/acs.jctc.2c00643.

[9] R.D. Groot, P.B. Warren, Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation, J. Chem. Phys. 107 (11) (1997) 4423–4435, http://dx.doi.org/10.1063/1.474784.

[10] I. Pagonabarraga, D. Frenkel, Dissipative particle dynamics for interacting systems, J. Chem. Phys. 115 (11) (2001) 5015–5026, http://dx.doi.org/10.1063/1.1396848.

[11] P. Español, Fluid particle model, Phys. Rev. E 57 (1998) 2930–2948, http://dx.doi.org/10.1103/PhysRevE.57.2930.

[12] A. Malevanets, R. Kapral, Mesoscopic model for solvent dynamics, J. Chem. Phys. 110 (17) (1999) 8605–8613, http://dx.doi.org/10.1063/1.478857.

[13] W.G. Noid, Perspective: Coarse-grained models for biomolecular systems, J. Chem. Phys. 139 (9) (2013) 090901, http://dx.doi.org/10.1063/1.4818908.

[14] Y. Zhang, A. Otani, E.J. Maginn, Reliable Viscosity Calculation from Equilibrium Molecular Dynamics Simulations: A Time Decomposition Method, J. Chem. Theory Comput. 11 (8) (2015) 3537–3546, http://dx.doi.org/10.1021/acs.jctc.5b00351.

[15] A. Boromand, S. Jamali, J.M. Maia, Viscosity measurement techniques in Dissipative Particle Dynamics, Comput. Phys. Comm. 196 (2015) 149–160, http://dx.doi.org/10.1016/j.cpc.2015.05.027.

[16] G. Jung, F. Schmid, Computing bulk and shear viscosities from simulations of fluids with dissipative and stochastic interactions, J. Chem. Phys. 144 (20) (2016) 204104, http://dx.doi.org/10.1063/1.4950760.

[17] C.W. Gear, J.M. Hyman, P.G. Kevrekidid, I.G. Kevrekidis, O. Runborg, C. Theodoropoulos, Equation-Free, Coarse-Grained Multiscale Computation: Enabling Microscopic Simulators to Perform System-Level Analysis, Commun. Math. Sci. 1 (4) (2003) 715–762.

[18] P.R. Vlachas, G. Arampatzis, C. Uhler, P. Koumoutsakos, Multiscale simulations of complex systems by learning their effective dynamics, Nat. Mach. Intell. 4 (4) (2022) 359–366, http://dx.doi.org/10.1038/s42256-022-00464-w.

[19] P.R. Vlachas, J. Zavadlav, M. Praprotnik, P. Koumoutsakos, Accelerated simulations of molecular systems through Learning of Effective Dynamics, J. Chem. Theory Comput. 18 (2022) 538–549.

[20] A.J. Roberts, Model Emergent Dynamics in Complex Systems, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014, http://dx.doi.org/10.1137/1.9781611973563.

[21] V. Del Tatto, G. Fortunato, D. Bueti, A. Laio, Robust inference of causality in high-dimensional dynamical processes from the information imbalance of distance ranks, Proc. Natl. Acad. Sci. USA 121 (19) (2024) e2317256121.

[22] J. Bongard, H. Lipson, Automated reverse engineering of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA 104 (24) (2007) 9943–9948, http://dx.doi.org/10.1073/pnas.0609476104.

[23] M. Schmidt, H. Lipson, Distilling Free-Form Natural Laws from Experimental Data, Science 324 (5923) (2009) 81–85, http://dx.doi.org/10.1126/science.1165893.

[24] M. Raissi, G.E. Karniadakis, Hidden physics models: Machine learning of nonlinear partial differential equations, J. Comput. Phys. 357 (2018) 125–141, http://dx.doi.org/10.1016/j.jcp.2017.11.039.

[25] R. González-García, R. Rico-Martínez, I. Kevrekidis, Identification of distributed parameter systems: A neural net based approach, Comput. Chem. Eng. 22 (1998) S965–S968, http://dx.doi.org/10.1016/S0098-1354(98)00191-4.

[26] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707, http://dx.doi.org/10.1016/j.jcp.2018.10.045.

[27] L. Lu, P. Jin, G. Pang, Z. Zhang, G.E. Karniadakis, Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators, Nat. Mach. Intell. 3 (3) (2021) 218–229, http://dx.doi.org/10.1038/s42256-021-00302-5.

[28] E. Kiyani, K. Shukla, G.E. Karniadakis, M. Karttunen, A framework based on symbolic regression coupled with eXtended Physics-Informed Neural Networks for gray-box learning of equations of motion from data, Comput. Methods Appl. Mech. Engrg. 415 (2023) 116258, http://dx.doi.org/10.1016/j.cma.2023.116258.

[29] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA 113 (15) (2016) 3932–3937, http://dx.doi.org/10.1073/pnas.1517384113.

[30] M. Hoffmann, C. Fröhner, F. Noé, Reactive SINDy: Discovering governing reactions from concentration data, J. Chem. Phys. 150 (2) (2019) 025101, http://dx.doi.org/10.1063/1.5066099.

[31] M. Dam, M. Brøns, J. Juul Rasmussen, V. Naulin, J.S. Hesthaven, Sparse identification of a predator–prey system from simulation data of a convection model, Phys. Plasmas 24 (2) (2017) 022310, http://dx.doi.org/10.1063/1.4977057.

[32] M. Sorokina, S. Sygletos, S. Turitsyn, Sparse identification for nonlinear optical communication systems: SINO method, Opt. Express 24 (26) (2016) 30433–30443, http://dx.doi.org/10.1364/OE.24.030433.

[33] A.V. Ermolaev, A. Sheveleva, G. Genty, C. Finot, J.M. Dudley, Data-driven model discovery of ideal four-wave mixing in nonlinear fibre optics, Sci. Rep. 12 (1) (2022) 12711, http://dx.doi.org/10.1038/s41598-022-16586-5.

[34] L. Zanna, T. Bolton, Data-Driven Equation Discovery of Ocean Mesoscale Closures, Geophys. Res. Lett. 47 (17) (2020) e2020GL088376, http://dx.doi.org/10.1029/2020GL088376.

[35] N. Díaz, M. Barreiro, N. Rubido, Data driven models of the Madden-Julian Oscillation: understanding its evolution and ENSO modulation, Npj Clim. Atmos. Sci. 6 (1) (2023) 203, http://dx.doi.org/10.1038/s41612-023-00527-8.

[36] Y.-X. Jiang, X. Xiong, S. Zhang, J.-X. Wang, J.-C. Li, L. Du, Modeling and prediction of the transmission dynamics of COVID-19 based on the SINDy-LM method, Nonlinear Dynam. 105 (3) (2021) 2775–2794, http://dx.doi.org/10.1007/s11071-021-06707-6.

[37] S.H. Rudy, S.L. Brunton, J.L. Proctor, J.N. Kutz, Data-driven discovery of partial differential equations, Sci. Adv. 3 (4) (2017) e1602614, http://dx.doi.org/10.1126/sciadv.1602614.

[38] D.A. Messenger, D.M. Bortz, Weak SINDy for partial differential equations, J. Comput. Phys. 443 (2021) 110525, http://dx.doi.org/10.1016/j.jcp.2021.110525.

[39] D.R. Gurevich, P.A.K. Reinbold, R.O. Grigoriev, Robust and optimal sparse regression for nonlinear PDE models, Chaos 29 (10) (2019) 103113, http://dx.doi.org/10.1063/1.5120861.

[40] C. Joshi, S. Ray, L.M. Lemma, M. Varghese, G. Sharp, Z. Dogic, A. Baskaran, M.F. Hagan, Data-Driven Discovery of Active Nematic Hydrodynamics, Phys. Rev. Lett. 129 (25) (2022) 258001, http://dx.doi.org/10.1103/PhysRevLett.129.258001.

[41] M. Golden, R.O. Grigoriev, J. Nambisan, A. Fernandez-Nieves, Physically informed data-driven modeling of active nematics, Sci. Adv. 9 (27) (2023) eabq6120, http://dx.doi.org/10.1126/sciadv.abq6120.

[42] R. Supekar, B. Song, A. Hastewell, G.P.T. Choi, A. Mietke, J. Dunkel, Learning hydrodynamic equations for active matter from particle simulations and experiments, Proc. Natl. Acad. Sci. USA 120 (7) (2023) e2206994120, http://dx.doi.org/10.1073/pnas.2206994120.

[43] D.A. Messenger, D.M. Bortz, Learning mean-field equations from particle data using WSINDy, Phys. D 439 (2022) 133406, http://dx.doi.org/10.1016/j.physd.2022.133406.

[44] E.P. Alves, F. Fiuza, Data-driven discovery of reduced plasma physics models from fully kinetic simulations, Phys. Rev. Res. 4 (3) (2022) 033192, http://dx.doi.org/10.1103/PhysRevResearch.4.033192.

[45] P.J. Hoogerbrugge, J.M.V.A. Koelman, Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics, Europhys. Lett. 19 (3) (1992) 155, http://dx.doi.org/10.1209/0295-5075/19/3/001.

[46] P. Español, P. Warren, Statistical Mechanics of Dissipative Particle Dynamics, Europhys. Lett. 30 (4) (1995) 191, http://dx.doi.org/10.1209/0295-5075/30/4/001.

[47] K.P. Santo, A.V. Neimark, Dissipative particle dynamics simulations in colloid and Interface science: a review, Adv. Colloid Interface 298 (2021) 102545, http://dx.doi.org/10.1016/j.cis.2021.102545.

[48] W. Pan, B. Caswell, G.E. Karniadakis, Rheology, Microstructure and Migration in Brownian Colloidal Suspensions, Langmuir 26 (1) (2010) 133–142, http://dx.doi.org/10.1021/la902205x.

[49] J.O. Wuming Li, X. Zhuang, Dissipative particle dynamics simulation for the microstructures of ferromagnetic fluids, Soft Mater. 14 (2) (2016) 87–95, http://dx.doi.org/10.1080/1539445X.2016.1150293.

[50] R. Groot, K. Rabone, Mesoscopic Simulation of Cell Membrane Damage, Morphology Change and Rupture by Nonionic Surfactants, Biophys. J. 81 (2) (2001) 725–736, http://dx.doi.org/10.1016/S0006-3495(01)75737-2.

[51] I.V. Pivkin, G.E. Karniadakis, Accurate Coarse-Grained Modeling of Red Blood Cells, Phys. Rev. Lett. 101 (2008) 118105, http://dx.doi.org/10.1103/PhysRevLett.101.118105.

[52] D.A. Fedosov, H. Noguchi, G. Gompper, Multiscale modeling of blood flow: from single cells to blood rheology, Biomech. Model. Mech. 13 (2) (2013) 239–258, http://dx.doi.org/10.1007/s10237-013-0497-9.

[53] J. Mauer, M. Peltomäki, S. Poblete, G. Gompper, D.A. Fedosov, Static and dynamic light scattering by red blood cells: A numerical study, PLoS One 12 (5) (2017) e0176799, http://dx.doi.org/10.1371/journal.pone.0176799.

[54] A. Economides, G. Arampatzis, D. Alexeev, S. Litvinov, L. Amoudruz, L. Kulakova, C. Papadimitriou, P. Koumoutsakos, Hierarchical Bayesian Uncertainty Quantification for a Model of the Red Blood Cell, Phys. Rev. Appl. 15 (2021) 034062, http://dx.doi.org/10.1103/PhysRevApplied.15.034062.

[55] P. Papež, M. Praprotnik, Dissipative particle dynamics simulation of ultrasound propagation through liquid water, J. Chem. Theory Comput. 18 (2022) 1227–1240.

[56] C. Ebner, W.F. Saam, D. Stroud, Density-functional theory of simple classical fluids. I. Surfaces, Phys. Rev. A 14 (1976) 2264–2273, http://dx.doi.org/10.1103/PhysRevA.14.2264.

[57] U.M.B. Marconi, P. Tarazona, Dynamic density functional theory of fluids, J. Chem. Phys. 110 (16) (1999) 8032–8044, http://dx.doi.org/10.1063/1.478705.

[58] M. Schmidt, Power functional theory for many-body dynamics, Rev. Modern Phys. 94 (2022) 015007, http://dx.doi.org/10.1103/RevModPhys.94.015007.

[59] H. Grabert, Projection Operator Techniques in Nonequilibrium Statistical Mechanics, Springer Berlin Heidelberg, 1982, http://dx.doi.org/10.1007/bfb0044591.

[60] K.C. Daoulas, A. Cavallo, R. Shenhar, M. Müller, Phase behaviour of quasi-block copolymers: A DFT-based Monte-Carlo study, Soft Matter 5 (2009) 4499–4509, http://dx.doi.org/10.1039/B911364A.

[61] B. Li, K. Daoulas, F. Schmid, Dynamic coarse-graining of polymer systems using mobility functions, J. Phys.: Condens. Matter. 33 (19) (2021) 194004, http://dx.doi.org/10.1088/1361-648X/abed1b.

[62] M. Müller, Memory in the relaxation of a polymer density modulation, J. Chem. Phys. 156 (12) (2022) 124902, http://dx.doi.org/10.1063/5.0084602.

[63] S. Mantha, S. Qi, F. Schmid, Bottom-up construction of dynamic density functional theories for inhomogeneous polymer systems from microscopic simulations, Macromolecules 53 (9) (2020) 3409–3423, http://dx.doi.org/10.1021/acs.macromol.0c00130.

[64] G.H. Fredrickson, H. Orland, Dynamics of polymers: A mean-field theory, J. Chem. Phys. 140 (8) (2014) 084902, http://dx.doi.org/10.1063/1.4865911.

[65] Z. Li, X. Bian, B. Caswell, G.E. Karniadakis, Construction of dissipative particle dynamics models for complex fluids via the Mori–Zwanzig formulation, Soft Matter 10 (2014) 8659–8672, http://dx.doi.org/10.1039/C4SM01387E.

[66] S. Angioletti-Uberti, M. Ballauff, J. Dzubiella, Dynamic density functional theory of protein adsorption on polymer-coated nanoparticles, Soft Matter 10 (2014) 7932–7945, http://dx.doi.org/10.1039/C4SM01170H.

[67] M. Rex, H.H. Wensink, H. Löwen, Dynamical density functional theory for anisotropic colloidal particles, Phys. Rev. E 76 (2007) 021403, http://dx.doi.org/10.1103/PhysRevE.76.021403.

[68] A.J. Archer, Dynamical density functional theory for molecular and colloidal fluids: A microscopic approach to fluid mechanics, J. Chem. Phys. 130 (1) (2009) 014509, http://dx.doi.org/10.1063/1.3054633.

[69] A.M. Menzel, A. Saha, C. Hoell, H. Löwen, Dynamical density functional theory for microswimmers, J. Chem. Phys. 144 (2) (2016) 024115, http://dx.doi.org/10.1063/1.4939630.

[70] F. Jülicher, S.W. Grill, G. Salbreux, Hydrodynamic theory of active matter, Rep. Progr. Phys. 81 (7) (2018) 076601, http://dx.doi.org/10.1088/1361-6633/aab6bb.

[71] T. Potisk, D. Svenšek, H. Pleiner, H.R. Brand, Continuum model of magnetic field induced viscoelasticity in magnetorheological fluids, J. Chem. Phys. 150 (17) (2019) 174901, http://dx.doi.org/10.1063/1.5090337.

[72] S. Bohlius, H.R. Brand, H. Pleiner, Macroscopic dynamics of uniaxial magnetic gels, Phys. Rev. E 70 (2004) 061411, http://dx.doi.org/10.1103/PhysRevE.70.061411.

[73] H. Schaeffer, Learning partial differential equations via data discovery and sparse optimization, Proc. R. Soc. A: Math. Phys. Eng. Sci. 473 (2197) (2017) 20160446, http://dx.doi.org/10.1098/rspa.2016.0446.

[74] P.A.K. Reinbold, R.O. Grigoriev, Data-driven discovery of partial differential equation models with latent variables, Phys. Rev. E 100 (2) (2019) 022219, http://dx.doi.org/10.1103/PhysRevE.100.022219.

[75] A. Kaptanoglu, B. de Silva, U. Fasel, K. Kaheman, A. Goldschmidt, J. Callaham, C. Delahunt, Z. Nicolaou, K. Champion, J.-C. Loiseau, J. Kutz, S. Brunton, PySINDy: A comprehensive Python package for robust sparse system identification, J. Open Sour. Softw. 7 (69) (2022) 3994, http://dx.doi.org/10.21105/joss.03994.

[76] N. Meinshausen, P. Bühlmann, Stability Selection, J. R. Stat. Soc. B: Stat. Methodol. 72 (4) (2010) 417–473, http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x.

[77] S. Maddu, B.L. Cheeseman, I.F. Sbalzarini, C.L. Müller, Stability selection enables robust learning of differential equations from limited noisy data, Proc. R. Soc. A: Math. Phys. Eng. Sci. 478 (2262) (2022) 20210916, http://dx.doi.org/10.1098/rspa.2021.0916.

[78] C. Junghans, M. Praprotnik, K. Kremer, Transport properties controlled by a thermostat: An extended dissipative particle dynamics thermostat, Soft Matter 4 (1) (2008) 156–161, http://dx.doi.org/10.1039/B713568H.

[79] P. Español, Hydrodynamics from dissipative particle dynamics, Phys. Rev. E 52 (2) (1995) 1734–1742, http://dx.doi.org/10.1103/PhysRevE.52.1734.

[80] C.A. Marsh, G. Backx, M.H. Ernst, Static and dynamic properties of dissipative particle dynamics, Phys. Rev. E 56 (2) (1997) 1676–1691, http://dx.doi.org/10.1103/physreve.56.1676.

[81] P. Espanol, Dissipative particle dynamics with energy conservation, Europhys. Lett. 40 (6) (1997) 631.

[82] J.B. Avalos, A. Mackie, Dissipative particle dynamics with energy conservation, Europhys. Lett. 40 (2) (1997) 141.

[83] R.D. Groot, P.B. Warren, Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation, J. Chem. Phys. 107 (11) (1997) 4423–4435, http://dx.doi.org/10.1063/1.474784.

[84] J.A. Backer, C.P. Lowe, H.C.J. Hoefsloot, P.D. Iedema, Poiseuille flow to measure the viscosity of particle model fluids, J. Chem. Phys. 122 (15) (2005) 154503, http://dx.doi.org/10.1063/1.1883163.

[85] I.V. Pivkin, G.E. Karniadakis, A new method to impose no-slip boundary conditions in dissipative particle dynamics, J. Comput. Phys. 207 (1) (2005) 114–128, http://dx.doi.org/10.1016/j.jcp.2005.01.006.

[86] I.V. Pivkin, G.E. Karniadakis, Coarse-graining limits in open and wall-bounded dissipative particle dynamics systems, J. Chem. Phys. 124 (18) (2006) 184101, http://dx.doi.org/10.1063/1.2191050.

[87] D.J. Tritton, Flow Past a Circular Cylinder, in: D.J. Tritton (Ed.), Physical Fluid Dynamics, Springer Netherlands, Dordrecht, 1977, pp. 18–29, http://dx.doi.org/10.1007/978-94-009-9992-3_3.

[88] R.T. Beyer, Parameter of Nonlinearity in Fluids, J. Acoust. Soc. Am. 32 (6) (1960) 719–721, http://dx.doi.org/10.1121/1.1908195.

[89] R. Cobbold, Foundations of Biomedical Ultrasound, Oxford University Press, 2007.

[90] H. Akaike, A New Look at the Statistical Model Identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723, http://dx.doi.org/10.1109/TAC.1974.1100705.

[91] G. Schwarz, Estimating the Dimension of a Model, Ann. Statist. 6 (2) (1978) 461–464, http://dx.doi.org/10.1214/aos/1176344136.

[92] N.M. Mangan, J.N. Kutz, S.L. Brunton, J.L. Proctor, Model selection for dynamical systems via sparse regression and information criteria, Proc. R. Soc. A: Math. Phys. Eng. Sci. 473 (2204) (2017) 20170009, http://dx.doi.org/10.1098/rspa.2017.0009.

[93] K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed., Springer, New York, 2002.

[94] R. Delgado-Buscalioni, J. Sablić, M. Praprotnik, Open boundary molecular dynamics, Eur. Phys. J.: Spec. Top. 224 (12) (2015) 2331–2349.

[95] H. Pleiner, J.L. Harden, General Nonlinear 2-Fluid Hydrodynamics of Complex Fluids and Soft Matter, AIP Conf. Proc. 708 (1) (2004) 46–51.

[96] D. Drew, S. Passman, Theory of Multicomponent Fluids, Springer, 2014.

[97] S. Cai, Z. Wang, S. Wang, P. Perdikaris, G.E. Karniadakis, Physics-Informed Neural Networks for Heat Transfer Problems, J. Heat Transfer 143 (6) (2021) 060801.

[98] K. Shukla, Z. Zou, C.H. Chan, A. Pandey, Z. Wang, G.E. Karniadakis, Neurosem: A hybrid framework for simulating multiphysics problems by coupling pinns and spectral elements, 2024, arXiv:2407.21217.