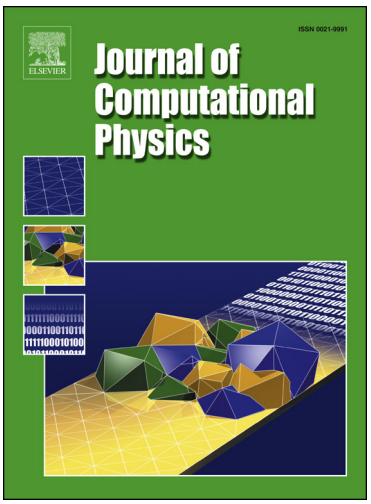
Journal Pre-proof



Incorporating physical constraints in a deep probabilistic machine learning framework for coarse-graining dynamical systems

Sebastian Kaltenbach, Phaedon-Stelios Koutsourelakis

PII: S0021-9991(20)30447-2
DOI: https://doi.org/10.1016/j.jcp.2020.109673
Reference: YJCPH 109673

To appear in: *Journal of Computational Physics*

Received date: 6 January 2020
Revised date: 9 May 2020
Accepted date: 16 June 2020

Please cite this article as: S. Kaltenbach, P.-S. Koutsourelakis, Incorporating physical constraints in a deep probabilistic machine learning framework for coarse-graining dynamical systems, *J. Comput. Phys.* (2020), 109673, doi: https://doi.org/10.1016/j.jcp.2020.109673.

Highlights

- A generative model for the automated discovery of CG dynamics.
- The target density is augmented by virtual observables which reflect physical constraints.
- The incorporation of physical constraints leads to a reduction of the training data.
- A probabilistic formulation that is capable of quantifying predictive uncertainty.
- Full reconstruction of futures of the entire FG state vector as well as any FG observable.

Incorporating physical constraints in a deep probabilistic machine learning framework for coarse-graining dynamical systems

Sebastian Kaltenbach[a], Phaedon-Stelios Koutsourelakis[a,*]

[a]*Professorship of Continuum Mechanics, Technical University of Munich*

Abstract

Data-based discovery of effective, coarse-grained (CG) models of high-dimensional dynamical systems presents a unique challenge in computational physics and particularly in the context of multiscale problems. The present paper offers a data-based, probabilistic perspective that enables the quantification of predictive uncertainties. One of the outstanding problems has been the introduction of physical constraints in the probabilistic machine learning objectives. The primary utility of such constraints stems from the undisputed physical laws such as conservation of mass, energy etc. that they represent. Furthermore and apart from leading to physically realistic predictions, they can significantly reduce the requisite amount of training data which for high-dimensional, multiscale systems are expensive to obtain (Small Data regime). We formulate the coarse-graining process by employing a probabilistic state-space model and account for the aforementioned equality constraints as virtual observables in the associated densities. We demonstrate how deep neural nets in combination with probabilistic inference tools can be employed to identify the coarse-grained variables and their evolution model without ever needing to define a fine-to-coarse (restriction) projection and without needing time-derivatives of state variables.

We advocate a sparse Bayesian learning perspective which avoids overfitting and reveals the most salient features in the CG evolution law. The formulation adopted enables the quantification of a crucial, and often ne-

*Corresponding author

Email addresses: sebastian.kaltenbach@tum.de (Sebastian Kaltenbach), p.s.koutsourelakis@tum.de (Phaedon-Stelios Koutsourelakis)

glected, component in the CG process, i.e. the predictive uncertainty due to information loss. Furthermore, it is capable of reconstructing the evolution of the full, fine-scale system and therefore the observables of interest need not be selected a priori. We demonstrate the efficacy of the proposed framework by applying it to systems of interacting particles and a series of images of a nonlinear pendulum. In both cases we identify the underlying coarse dynamics and can generate extrapolative predictions including the forming and propagation of a shock for the particle systems and a stable trajectory in the phase space for the pendulum.

Keywords: Bayesian machine learning, virtual observables, multiscale modeling, reduced order modeling, coarse graining

1. Introduction

High-dimensional, nonlinear dynamical systems are ubiquitous in applied physics and engineering. The computational resources needed for their solution can grow exponentially with the dimension of the state-space as well as with the smallest time-scale that needs to be resolved and which determines the discretization time-step. Hence the ability to construct reduced, *coarse-grained* descriptions and models that are nevertheless predictive of various observables and at time-scales much larger than the inherent ones, is an important task (Givon et al., 2004).

One strategy for learning such coarse-grained (CG) models is based on data generated by simulations of the fine-grained (FG) system. This can yield an automated solution especially in cases where domain knowledge is limited or absent. The derivation of CG models from data is also particularly relevant in domains where FG models are not available, such as in social sciences or biophysics, but data abound (Bialek, 2012; Alber et al., 2019). Data-based methodologies have also been fueled by recent advances in statistical- (Ghahramani, 2015) or machine-learning (LeCun et al., 2015) which, in large part, have been enabled by large datasets (and the computational means to leverage them). We note nevertheless that coarse-graining tasks based on FG simulation data exhibit some fundamental differences (Koutsourelakis et al., 2016). Firstly, the acquisition of FG simulation data is by definition expensive and the reduction of the required FG simulations is one of the objectives of CG model development. Secondly, in physical applications, significant information about the underlying physical/mathematical

structure of the problem, and of the CG model in particular, is available. This information might come in the form of constraints that reflect e.g undisputed physical principles such as conservation laws (e.g. mass, momentum, energy). Injecting this prior information into the CG models in combination with FG data in an automated fashion represents a significant challenge (Marcus and Davis, 2019), especially in the context of *probabilistic* models (Stinis et al., 2019). Such a capability would be instrumental not only in reducing the required amount of FG data, but more importantly, in enabling predictions under *extrapolative* settings as those arising e.g. when the initial conditions of the FG system are different from the ones in the training data.

In this paper, we propose a generative, probabilistic (Bayesian) machine learning framework (Koutsourelakis and Bilionis, 2011) which employs FG simulation data augmented by *virtual observables* to account for constraints. The latter concept which we elucidate in the sequel, enables the incorporation of domain knowledge in probabilistic models and represents, in our opinion the most novel contribution of this paper. Furthermore and within the Bayesian framework advocated, it allows us to introduce appropriate priors that promote the discovery of *slow-varying* CG state-variables which is a highly-desirable feature for multiscale systems (Kevrekidis et al., 2003). In contrast to most existing techniques which consider the problems of CG state variable discovery and CG model construction in two or more steps (Schmid, 2010; Williams et al., 2015; Wu and Noé, 2017; Froyland et al., 2014), we address both simultaneously (Felsberger and Koutsourelakis, 2019). The framework proposed consists of two building blocks: a probabilistic coarse-to-fine map (Schöberl et al., 2017) and an evolution law for the CG dynamics. The former can be endowed with great flexibility in discovering appropriate CG variables when combined with deep neural nets (Raissi et al., 2017, 2019; Yang and Perdikaris, 2019), which is especially challenging if the number of training data is small[1]. We demonstrate nevertheless the efficacy of such an approach when physical information is incorporated a-priori into the model. The CG variables identified are not restricted to indicator functions of subdomains of the state-space as in other generative models (Mardt et al., 2018; Wu and Noé, 2017; Wu et al., 2018) and which are difficult to learn when the simulation data is limited and has not sufficiently populated all important

[1]In the dynamical systems investigated the size of the dataset depends on the length of the FG time-sequences as well as the number of such sequences employed for training.

regions of the state-space.

The second component of the proposed framework pertains to the discovery of the CG evolution law which is learned by employing a large vocabulary of feature functions and sparsity-inducing priors. This leads to interpretable solutions (Duncker et al., 2019), even in the *Small Data* regime that avoid overfitting and reveal salient characteristics of the CG system (Grigo and Koutsourelakis, 2019). The premise of sparsity (Pantazis and Tsamardinos, 2019) has been employed in the past for the discovery of the CG dynamics as e.g. in the SINDy method (Brunton et al., 2016; Kaiser et al., 2018; Champion et al., 2019). This however requires the availability of time-derivatives of the CG variables and does not directly lead to a posterior on the model parameters that can reflect inferential uncertainties. Nonparametric models for the CG dynamics have also been proposed (Ohkubo, 2011) but have been restricted to low dimensions. The learned CG dynamics are in general nonlinear in contrast to efforts based on transfer operators (Klus et al., 2018) and particularly the Koopman operator (Koopman, 1931; Mezić, 2005; Brunton et al., 2016). While the associated theory guarantees the existence of a linear operator, this is possible in the infinite dimensional space of observables, it does not specify how many should be used to obtain a good approximation, and more importantly, how one can predict future FG states given predictions on the evolution of those observables i.e. the reconstruction step.

The latter constitutes the main difference of the proposed model with non-generative ones based e.g. on information-theoretic concepts (Katsoulakis and Plecháč, 2013; Harmandaris et al., 2016; Katsoulakis and Vilanova, 2019) or on the Mori-Zwanzig (MZ) formalism (Mori, 1965; Zwanzig, 1973; Chorin and Stinis, 2007). Apart from the difficulties in approximating the right-hand-side of the MZ-prescribed CG dynamics, and particularly the memory term (Lei et al., 2016; Zhu et al., 2018), this can only guarantee correct predictions of the CG variables' evolution. If observables not depending on CG variables are of interest, then a reconstruction operator would need to be added. In contrast, in the proposed model this reconstruction operator is represented by the probabilistic coarse-to-fine map which is simultaneously learned from the data and can quantify predictive uncertainties associated with the information loss that unavoidably takes place in any CG process as well as due to the fact that finite (and preferably, small) data has been used for training.

The enabling computational technology for training the proposed model

is based on probabilistic inference. In order to resolve the intractable posterior on latent variables and model parameters in our Bayesian framework, we make use of Stochastic Variational Inference (Hoffman et al., 2013) as MCMC is cumbersome in high dimensions. We operate on the discretized time domain (Archambeau and Opper, 2011) and demonstrate how amortized (Krishnan et al., 2017; Fortuin et al., 2019) and non-amortized approximations can be employed.

The remainder of the paper is structured as follows: In Section 2 we present the general methodological framework with special attention on the two building blocks of the state-space model proposed i.e. the transition law for the CG dynamics and the incorporation of virtual observables (section 2.2), as well as the the emission law which provides the link between CG and FG description through a probabilistic *coarse-to-fine* map (section 2.3). Computational aspects related to inference and prediction are discussed in sections 2.4 and 2.5 respectively. Section 3 contains illustrative applications involving coarse-graining of high-dimensional systems of interacting particles (section 3.1) as well as learning the dynamics of a nonlinear pendulum (section 3.2) from a sequence of images. We conclude in section 4 which also contains a discussion on possible extensions.

2. Methodology

In general, we use the subscript f or lower-case letters to denote variables associated with the (high-dimensional) fine-grained(FG)/full-order model and the subscript c or upper-case letters for quantities of the (lower-dimensional) coarse-grained(CG)/reduced-order description. We also use a circumflex $\hat{}$ to denote observed/known variables. We begin with the presentation of the FG and the CG model and subsequently explain the essential ingredients of the proposed formulation.

2.1. The FG and CG models

We consider a, generally high-dimensional, FG system with state variables $\boldsymbol{x}$ of dimension d_f ($d_f >> 1$) such that $\boldsymbol{x} \in \mathcal{X}_f \subset \mathbf{R}^{d_f}$. The dynamics of the FG system are dictated by system of deterministic or stochastic ODEs i.e.,

$$\dot{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t, t), \quad t > 0 \tag{1}$$

The initial condition $\boldsymbol{x}_0$ might be deterministic or drawn from a specified distribution. In the following we do not make explicit use of the FG dynamics but rely purely on FG data i.e. time sequences simulated from Equation (1) with a time-step, say δt. That is, our observables consists of n data sequences over $T+1$ FG time-steps δt i.e.,

$$\mathcal{D}_{T,n} = \{\hat{\boldsymbol{x}}^{(1:n)}_{0:T\delta t}\} \tag{2}$$

We denote the (unknown) CG state variables by $\boldsymbol{X}$ and assume $\boldsymbol{X} \in \mathcal{X}_c \subset \mathbf{R}^{d_c}$, where d_c is the dimension of the CG system. We presuppose Markovian dynamics[2] for the CG system of the form:

$$\dot{\boldsymbol{X}}_t = \boldsymbol{F}(\boldsymbol{X}_t, t) \tag{3}$$

which we discretize using a linear multistep method and a CG time step Δt:

$$\boldsymbol{R}_l(\boldsymbol{X}) = \sum_{k=0}^{K} \left(\alpha_k \boldsymbol{X}_{(l-k)\Delta t} + \Delta t \beta_k \boldsymbol{F}(\boldsymbol{X}_{(l-k)\Delta t})\right) = 0, \qquad l = K, K+1, \ldots \tag{4}$$

where α_k, β_k are the parameters of the discretization scheme and $\boldsymbol{R}_l$ the corresponding residual at time step l (Butcher, 2016). We note that depending on the values of the parameters K, α_k, β_k, several of the well-known, explicit/implicit, numerical time-integration schemes can be recovered. In this work, our goal is two-fold:

a) to identify the CG state-variables $\boldsymbol{X}$ and their relation with the FG description $\boldsymbol{x}$,

b) to identify the right-hand side of Equation (3),

in view of enabling predictions of the FG system over longer time horizons. Traditionally, the aforementioned tasks are *not* considered simultaneously. Usually the CG state variables are specified a priori using domain-knowledge (physical insight) or based on the observables of interest (Harmandaris et al., 2016). In other efforts, linear or non-linear dimensionality reduction procedures are first employed in order to identify such a lower-dimensional set of collective variables $\boldsymbol{X}$ (e.g. (Coifman et al., 2008)). In both of these cases,

[2] As discussed in section 3, this assumption can be relaxed.

$\boldsymbol{X}$ are defined using a *fine-to-coarse*, projection map e.g. $\boldsymbol{X} = \Pi(\boldsymbol{x})$ where $\Pi : \mathcal{X}_f \subset \mathbb{R}^{d_f} \rightarrow \mathcal{X}_c \subset \mathbb{R}^{d_c}$. Irrespective of whether this map is prescribed from the physics or learned from data, it is generally a many-to-one function that does not have an inverse i.e. if the CG states $\boldsymbol{X}$ are known one cannot readily reconstruct $\boldsymbol{x}$ (Trashorras and Tsagkarogiannis, 2010).

We note that that this has nothing to do with the quality of the CG evolution law (problem b) above). Even if the Mori-Zwanzig (MZ) formalism were employed, which in principle provides an exact, closed system of evolution equations for any observable of the FG states and therefore for $\boldsymbol{X} = \Pi(\boldsymbol{x})$, even if all the terms in the right-hand side were available, one would simply be able to predict the future evolution of $\boldsymbol{X}$ but not $\boldsymbol{x}$. This might be sufficient for a lot of problems of practical interest where the CG variables (or observables thereof) are of sole interest. Our goal however is a bit more ambitious, i.e. we seek to find a $\boldsymbol{X}$ that would allow us to reconstruct as accurately as possible the whole FG vector $\boldsymbol{x}$ into the future. As with any coarse-graining process, we recognize that this would unavoidably imply some information loss which in turn will give rise to predictive uncertainty (Katsoulakis and Trashorras, 2006). In this work, we advocate a probabilistic framework that quantifies this uncertainty.

With regards to problem b) above, we note that its solution hinges upon the CG variables $\boldsymbol{X}$ employed (problem a)). Irrespective of the breadth of the model forms considered (i.e. functions $\boldsymbol{F}$ in Equation (3)), the evolution of some $\boldsymbol{X}$ might fall outside this realm. For example, it is known from MZ theory that memory terms can become significant for certain observables. It is well-known that such memory terms can be substituted or approximated by additional variables (Kondrashov et al., 2015) which would in turn imply an augmented CG description $\boldsymbol{X}$ in Equation (3) that contains these auxiliary internal state variables (Coleman and Gurtin, 1967).

We address problems a) and b) in the coarse-graining process *simultaneously* by employing a probabilistic state-space model. This consists of two densities i.e.

- the transition law which dictates the evolution of the CG variables $\boldsymbol{X}$ (section 2.2). Special attention is paid to the definition of *virtual observables* with which the CG states and their dynamics can be injected with physical information.

- the emission law which provides the link between CG and FG description through a probabilistic *coarse-to-fine* map (section 2.3, (Felsberger

and Koutsourelakis, 2019)).

We emphasize that in our formulation, the CG state-variables $\boldsymbol{X}$ are implicitly defined as latent generators of the FG description $\boldsymbol{x}$. As discussed in detail in the sequel, this enables a straightforward, *probabilistic* reconstruction of $\boldsymbol{x}$ when $\boldsymbol{X}$ is known. The inverse map (analogous to Π above) arises naturally through probabilistic inference as explained in section 2.4. An overview of the essential elements of the proposed model can be seen in the probabilistic graphical model of Figure 1.

2.2. Transition Law: CG dynamics and virtual observables

Typical state-space models (Cappe et al., 2005; Ghahramani, 2004; Durstewitz, 2017; Krishnan et al., 2017) postulate Markovian, stochastic dynamics for the hidden variables $\boldsymbol{X}$, in the form of a diffusion process, which are subsequently discretized *explicitly* using e.g. a Euler-Maruyama scheme with time step Δt. This gives rise to a, generally Gaussian, conditional density $p(\boldsymbol{X}_{(l+1)\Delta t}|\boldsymbol{X}_{l\Delta t})$ which can be stacked over multiple time-instants in order to formulate a generalized prior on the CG-space.

When the CG state-variables $\boldsymbol{X}$ are given (in part or in whole) physical meaning (e.g. as thermodynamic state variables), then some of the equations for their evolution are prescribed by associated physical principles e.g. conservation of mass, momentum, energy. These can be reflected in the residuals $\boldsymbol{R}_l$ of the governing equations as in Equation (4) or alternatively as equality constraints of the form:

$$\boldsymbol{c}_l(\boldsymbol{X}_{l\Delta t}) = \boldsymbol{0}, \quad l = 0, 1, \dots \tag{5}$$

which must hold at each time-step. The function $\boldsymbol{c}_l : \mathcal{X}_c \subset \mathbb{R}^{d_c} \to \mathbb{R}^{M_c}$ enforces these known constraints at each time-step (see specific examples in section 3) and the only requirement we will impose is that of differentiability of $\boldsymbol{c}_l$ (see section 2.4). In order to account for the aforementioned constraints in the transition law of the CG state variables, we employ the novel (to the best of our knowledge) concept of *virtual observables*. In particular for each of the residuals $\boldsymbol{R}_l$ in Equation (4), we define a new variable/vector $\hat{\boldsymbol{R}}_l$ which relates to $\boldsymbol{R}_l$ as follows:

$$\hat{\boldsymbol{R}}_l = \boldsymbol{R}_l(\boldsymbol{X}) + \sigma_R \boldsymbol{\epsilon}_R, \qquad \boldsymbol{\epsilon}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{6}$$

We further assume that $\hat{\boldsymbol{R}}_l$ have been *virtually observed* and $\hat{\boldsymbol{R}}_l = 0$ leading to an augmented version of the data in Equation (2), by a set of virtual

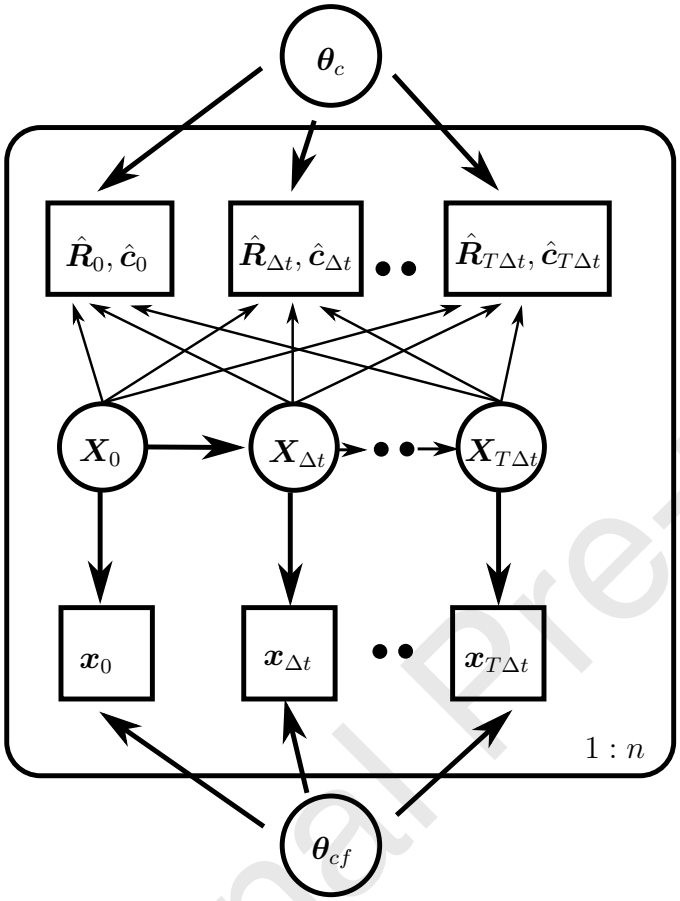


Figure 1: Proposed probabilistic graphical model. The CG variables $\boldsymbol{X}$ are latent and are inferred together with the parameters $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_{cf}$. Apart from the the FG states $\boldsymbol{x}$, the observables are augmented by *virtual observables* $\hat{\boldsymbol{R}}, \hat{\boldsymbol{c}}$ (see section 2.2). These virtual observables can depend on all CG variables but more often this dependence is restricted to only a few of them.

observations and therefore virtual likelihoods of the type:

$$p(\hat{\boldsymbol{R}}_l = \mathbf{0} \mid \boldsymbol{X}, \sigma_R) = \mathcal{N}(\mathbf{0} \mid \boldsymbol{R}_l(\boldsymbol{X}), \sigma_R^2 \boldsymbol{I}) \tag{7}$$

The "noise" parameter σ_R determines the intensity of the enforcement of the virtual observations and is analogous to the tolerance parameter with which residuals are enforced in a deterministic solution of the dynamics. Similarly, for constraints of the form of Equation (5), additional variables and virtual observables of the type:

$$\mathbf{0} = \hat{\boldsymbol{c}}_l = \boldsymbol{c}_l(\boldsymbol{X}_{l\Delta t}) + \sigma_c \boldsymbol{\epsilon}_c, \quad \boldsymbol{\epsilon}_c \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{8}$$

can be defined which would lead to an augmented (virtual) likelihood with terms of the type:

$$p\left(\hat{\boldsymbol{c}}_l = \mathbf{0} \mid \boldsymbol{X}_{l\Delta t}, \sigma_c\right) = \mathcal{N}\left(\mathbf{0} \mid \boldsymbol{c}_l(\boldsymbol{X}_{l\Delta t}), \sigma_c^2 \boldsymbol{I}\right) \tag{9}$$

where the role of σ_c^2 is analogous to σ_R^2 above.

Since the goal is to identify the right-hand side of the evolution laws in Equation (3), we denote by $\boldsymbol{\theta}_c$ the parameters appearing in $\boldsymbol{F}$ i.e. $\boldsymbol{F}(\boldsymbol{X}_t, t;\, \boldsymbol{\theta}_c)$. Accordingly, the virtual observations in Equation (6) or Equation (8) would depend on $\boldsymbol{\theta}_c$. We defer until section 3 a detailed discussion on the form, the parametrization as well as the prior specifications in the Bayesian setting adopted. The latter plays an important role as with sparsity-inducing priors we can avoid overfitting and obtain a parsimonious and physically-interpretable solution for $\boldsymbol{F}$. We finally remark that physical information taking the form of equalities can also be available for the FG states $\boldsymbol{x}$. While this can be incorporated using appropriate virtual observables as above, the inference framework would exhibit significant differences (in brief, FG states would need to be inferred as well) and in order to avoid confusion we do not discuss such cases here.

2.3. Emission law: Coarse-to-Fine map

We make use of a probabilistic *generative* model in the definition of the CG state-variables through a *coarse-to-fine* map (Felsberger and Koutsourelakis, 2019) as opposed to traditional, many-to-one maps from the FG description to the CG one. We denote the associated (conditional) density by:

$$p_{cf}(\boldsymbol{x}_t \mid \boldsymbol{X}_t;\, \boldsymbol{\theta}_{cf}) \tag{10}$$

Observables $\mathcal{D}$	$\hat{\boldsymbol{x}}^{(1:n)}_{0:T\Delta t}$	FG simulation Data
	$\hat{\boldsymbol{R}}^{(1:n)}_{0:T}$	Virtual Observables corresponding to CG model residuals
	$\hat{\boldsymbol{c}}^{(1:n)}_{0:T}$	Virtual Observables corresponding to CG constraints
Latent variables	$\boldsymbol{X}^{(1:n)}_{0:T\Delta t}$	CG state variable
Model parameters $\boldsymbol{\theta}$	$\boldsymbol{\theta}_{cf}$	parameters in the coarse-to-fine mapping
	$\boldsymbol{\theta}_{c}$	parameters in the CG evolution law

Table 1: Data, latent variables and model parameters

where $\boldsymbol{\theta}_{cf}$ denote the (unknown) parameters that will be learned from the data. The form of p_{cf} can be adapted to the particulars of the problem and can be endowed with various levels of domain knowledge. In section 3, we provide various examples, from particle-systems where p_{cf} is fully determined by the physics, to a more abstract case where deep neural networks are employed in order to learn the full p_{cf}. We note finally that a (probabilistic) fine-to-coarse map can still be learned in the current setting, and would correspond to the *posterior* of $\boldsymbol{X}_t$ given $\boldsymbol{x}_t$. We discuss this as well as all aspects pertaining to inference and learning in the next section.

2.4. Inference and Learning

We start this section by summarizing the main elements of the model presented (i.e. data, latent variables and parameters - see also Table 1) and subsequently describe a fully Bayesian inference scheme based on Stochastic Variational Inference (SVI, (Hoffman et al., 2013)) tools.

We adopt an enlarged definition of *data* which we cumulatively denote by $\mathcal{D}$ and which encompasses:

- FG simulation data as in Equation (2) consisting of n sequences of the FG state-variables. As the likelihood model implied by the p_{cf} in Equation (10) involves only the observables at each *coarse* time-step we denote those by $\{\hat{\boldsymbol{x}}^{(1:n)}_{0:T\Delta t}\}$. We assume that the number of observations in each sequence is the same although this is not necessary. In fact, the length of each time-sequence and the number of time-sequences needed could be the subject of an active learning scheme. This would be particularly important in cases where very expensive, high-dimensional FG simulators are employed. The generative, proposed formulation can account for any type of (in)direct or (in)complete/partial, experimental or computational observations relating to FG states which we omit

here for simplicity of the presentation. We nevertheless illustrate this capability of the model in the example of section 3.2.

- Virtual observables relating to the CG states $\boldsymbol{X}$ at each time-step l consisting of residuals $\hat{\boldsymbol{R}}_l^{(1:n)}$ as in Equation (6) and/or constraints $\hat{\boldsymbol{c}}_l^{(1:n)}$ as in Equation (8) (the superscript pertains to the time sequence $i = 1, \dots, n$). Assuming they pertain to all time-steps, we denote them by $\left\{\hat{\boldsymbol{R}}_{0:T}^{(1:n)}, \hat{\boldsymbol{c}}_{0:T}^{(1:n)}\right\}$.

The latent (unobserved) variables of the model are represented by the CG state-variables $\left\{\boldsymbol{X}_{0:T\Delta t}^{(1:n)}\right\}$ which relate to the FG data through the p_{cf} (in Equation (10)) and to the virtual observables through Equation (7) or Equation (9).

Finally, the (unknown) parameters of the model which we denote cumulatively by $\boldsymbol{\theta}$ consist of[3]:

- $\boldsymbol{\theta}_c$ which parametrize the right-hand-side of the CG evolution law (see end of section 2.2),
- $\boldsymbol{\theta}_{cf}$ which parametrize the probabilistic coarse-to-fine map (Equation (10)),
- σ_R, σ_c involved in the enforcement of virtual observables in Equation (6) and Equation (8) respectively, and,
- *hyperparameters* associated with the priors employed on the latent variables or the previous parameters.

Following a fully-Bayesian formulation, we can express the posterior of the unknowns (i.e. latent variables and parameters) as follows:

$$p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta})\ p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta})}{p(\mathcal{D})} \tag{11}$$

where $p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta})$ denotes the prior on the latent variables and parameters.

[3]If any of the parameters in this list are prescribed, then they are omitted from $\boldsymbol{\theta}$.

We discuss first the likelihood term $p(\mathcal{D}|\boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta})$ which can be decomposed into the product of three (conditionally) independent terms, one for each data-type, i.e.:

$$p(\mathcal{D} \mid \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta}) = p(\hat{\boldsymbol{x}}_{0:T\Delta t}^{(1:n)} \mid \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta})\, p(\hat{\boldsymbol{R}}_{0:T}^{(1:n)} \mid \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta})\, p(\hat{\boldsymbol{c}}_{0:T}^{(1:n)} \mid \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta}) \tag{12}$$

We further note that (from Equation (10)):

$$p(\hat{\boldsymbol{x}}_{0:T\Delta t}^{(1:n)} \mid \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{l=0}^{T} p_{cf}(\boldsymbol{x}_{l\ \Delta t}^{(i)} \mid \boldsymbol{X}_{l\ \Delta t}^{(i)}, \boldsymbol{\theta}_{cf}) \tag{13}$$

and (from Equation (7)):

$$\begin{aligned} p(\hat{\boldsymbol{R}}_{0:T}^{(1:n)} | \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta}) &= \textstyle\prod_{i=1}^{n} \prod_{l=0}^{T} \mathcal{N}\left(\mathbf{0} | \boldsymbol{R}_l(\boldsymbol{X}^{(i)}), \sigma_R^2 \boldsymbol{I}\right) \\ &\propto \textstyle\prod_{i=1}^{n} \prod_{l=0}^{T} \frac{1}{\sigma_R^{dim(\boldsymbol{R})}} \exp\left\{-\frac{1}{2\sigma_R^2} \left|\boldsymbol{R}_l(\boldsymbol{X}^{(i)})\right|^2\right\} \end{aligned} \tag{14}$$

and (from Equation (9)):

$$\begin{aligned} p(\hat{\boldsymbol{c}}_{0:T}^{(1:n)} | \boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta}) &= \textstyle\prod_{i=1}^{n} \prod_{l=0}^{T} \mathcal{N}(\mathbf{0} | \boldsymbol{c}_l(\boldsymbol{X}_{l\ \Delta t}^{(i)}), \sigma_c^2 \boldsymbol{I}) \\ &\propto \textstyle\prod_{i=1}^{n} \prod_{l=0}^{T} \frac{1}{\sigma_c^{dim(\boldsymbol{c})}} \exp\left\{-\frac{1}{2\sigma_c^2} \left|\boldsymbol{c}_l(\boldsymbol{X}_{l\ \Delta t}^{(i)})\right|^2\right\} \end{aligned} \tag{15}$$

While the complexity of the expressions involved imply a non-analytic solution for the posterior, we emphasize that the terms above encode actual and virtual observables (constraints) and they are differentiable, a property that is crucial for carrying out Variational Inference.

Before presenting the inference procedure, we mention an interesting possibility for encoding prior information for the latent CG states $\boldsymbol{X}_{0:T\Delta t}^{(1:n)}$ through the prior term $p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)})$. A desirable property of the CG state-variables is that of *slowness* i.e. that they should capture features of the system that evolve over (much) larger time-scales (Kevrekidis et al., 2003). The discovery of such features has been the goal of several statistical analysis procedures (e.g. Slow Feature Analysis (Wiskott and Sejnowski, 2002)) as well as in physics/chemistry literature (see a recent review in (Klus et al., 2018)). In this work we promote the discovery of such slow features by appropriate prior selection, and in particular by penalizing the jumps between two successive

time-instants, i.e.:

$$\begin{aligned} p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)}) &= \prod_{i=1}^{n} p_{c,0}(\boldsymbol{X}_0^{(i)}) \prod_{l=0}^{T-1} p(\boldsymbol{X}_{(l+1)\ \Delta t}^{(i)} | \boldsymbol{X}_{l\ \Delta t}^{(i)}, \sigma_X^2 \boldsymbol{I}) \\ &= \prod_{i=1}^{n} p_{c,0}(\boldsymbol{X}_0^{(i)}) \prod_{l=0}^{T-1} \mathcal{N}(\boldsymbol{X}_{(l+1)\ \Delta t}^{(i)} | \boldsymbol{X}_{l\ \Delta t}^{(i)}, \sigma_X^2 \boldsymbol{I}) \\ &\propto \prod_{i=1}^{n} p_{c,0}(\boldsymbol{X}_0^{(i)}) \prod_{l=0}^{T-1} \frac{1}{\sigma_X^{dc}} \exp\left\{ -\frac{1}{\sigma_X^2} \left| \boldsymbol{X}_{(l+1)\ \Delta t}^{(i)} - \boldsymbol{X}_{l\ \Delta t}^{(i)} \right|^2 \right\} \end{aligned} \tag{16}$$

where $p_{c,0}$ is a prior density for the initial CG state. We observe that the strength of the penalty is inversely proportional to the hyperparameter σ_X^2 and in the limit $\sigma_X^2 \to 0$ it implies a constant time history of $\boldsymbol{X}_t$. As the appropriate value for σ_X^2 depends on the problem, we include this in the parameter vector $\boldsymbol{\theta}$ that is inferred/learned from the data.

Given the intractability of the actual posterior, we advocate in this work Variational Inference. This operates on a parameterized family of densities, say $q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})$ and attempts to find the one (i.e. the value of $\boldsymbol{\phi}$) that most closely approximates the posterior by minimizing their Kullback-Leibler divergence. It can be readily shown (Bishop, 2006), that the optimal $q_{\boldsymbol{\phi}}$, maximizes the Evidence Lower Bound (ELBO) $\mathcal{F}(q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta}))$ below:

$$\begin{aligned} \log p(\mathcal{D}) &= \log \int p(\mathcal{D},\ \boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})\ d\boldsymbol{X}_{0:T\Delta t}^{(1:n)}\ d\boldsymbol{\theta} \\ &= \log \int \frac{p(\mathcal{D}|\ \boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta}) p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})}{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})} q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})\ d\boldsymbol{X}_{0:T\Delta t}^{(1:n)}\ d\boldsymbol{\theta} \\ &\geq \int \log \frac{p(\mathcal{D}|\ \boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta}) p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})}{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})} q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})\ d\boldsymbol{X}_{0:T\Delta t}^{(1:n)}\ d\boldsymbol{\theta} \\ &= \mathcal{F}(q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})) \end{aligned} \tag{17}$$

In the examples analyzed we decompose the approximate posterior as:

$$\begin{aligned} q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta}) &= q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)})\ q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \\ &= \left[\prod_{i=0}^{n} q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(i)}) \right]\ \ q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \end{aligned} \tag{18}$$

where the first line is the so-called mean-field approximation and the second is a direct consequence of the (conditional) independence of the time sequences in the likelihood. We note that evaluations of the ELBO $\mathcal{F}$ involve

expectations with respect to $q_{\boldsymbol{\phi}}$ i.e.:

$$\mathcal{F}\left(q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})\right) = \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p(\mathcal{D}|\ \boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log \frac{p(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})}{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)},\ \boldsymbol{\theta})}\right] \tag{19}$$

and in order to maximize it (with respect to $\boldsymbol{\phi}$), gradients of those are needed. Given the intractability of these expectations and their derivatives, we make use of Monte Carlo estimates in combination with stochastic gradient ascent for the $\boldsymbol{\phi}$-updates. In order to reduce the Monte Carlo error in these estimates, we make use of the reparametrization trick (Kingma and Welling, 2014), for which the differentiability of the residuals/constraints is necessary. We specify the particulars of the algorithm more precisely in the numerical illustration section (see e.g. Algorithm 3 or 4).

We note that maximum likelihood or maximum-a-posteriori (MAP) point estimates for any of the parameters involved can be obtained as a special case of the aforementioned scheme by employing a $q_{\boldsymbol{\phi}}$ that is equal to a Dirac-delta function. Furthermore, amortized versions of the approximate posterior $q_{\boldsymbol{\phi}}$ i.e. forms that explicitly account on the dependence on the data values, can be employed in part or in whole. These have the capability of being able to transfer information across data points and are necessary in the realm of Big Data. We note though that we operate in the *Small Data* regime, i.e. the number of time sequences n (and time-steps T) is not particularly large. Hybrid versions between amortized and non-amortized posteriors could also be employed (Kim et al., 2018).

We note finally that while the ELBO $\mathcal{F}$ is used purely as the objective function for the determination of the approximate posterior, its role can be quite significant in model validation and refinement. In particular since $\mathcal{F}$ approximates the model evidence (denominator of Equation (11)), once evaluated, it can be used to comparatively assess different models. These could have different CG states $\boldsymbol{X}$ (in type and/or number) or different parametrizations $\boldsymbol{\theta}$. In this regard, the ELBO $\mathcal{F}$ could serve as the primary driver for the adaptive refinement of the CG model (Grigo and Koutsourelakis, 2019) in order to better explain the observables and lead to superior predictions.

2.5. Prediction

An essential feature of the proposed modeling framework is the ability to produce *probabilistic* predictive estimates. These encompass the information-loss due to the coarse-graining process as well as the epistemic

uncertainty arising from finite (and small) datasets. We distinguish between two settings:

a) the *"interpolative"* i.e. predictions into the future of a sequence i observed up to time-step T i.e. $\hat{\boldsymbol{x}}^{(i)}_{0:T\Delta t}$ which was used in the training phase - see section 3, or

b) the *"extrapolative"* i.e. predictions for a completely new initial condition $\hat{\boldsymbol{x}}_0$ - see section 3.

We note that any predictions should account for the domain knowledge incorporated in the training through the residuals $\boldsymbol{R}_l$ or constraints $\boldsymbol{c}_l$. Formally that is, one should enlarge the posterior density defined in Equation (11), in order to account for the residuals and/or constraints at future time-steps. This would in turn imply, that future (FG or CG) states should be inferred from such an augmented posterior i.e. prediction would imply an enlarged inference process. In the examples presented we adopt a simpler procedure that retains the essential features (i.e. probabilistic nature) but is more computationally expedient. In particular, for case a) above and if $q_{\boldsymbol{\phi}}(\boldsymbol{X}^{(i)}_{T\Delta t})$ is the (marginal) posterior of the last, hidden CG state and $q(\boldsymbol{\theta})$ the posterior of the model parameters, then we (see also Agorithm 1):

- sample from $q(\boldsymbol{X}^{(i)}_{T\Delta t}), q(\boldsymbol{\theta})$
- for each sample, we propagate the CG dynamics dynamics of Equation (3) (e.g. by solving the corresponding residual Equations (4)) in order to obtain $\boldsymbol{X}^{(i)}_{(T+1)\Delta t}, \boldsymbol{X}^{(i)}_{(T+2)\Delta t}, \ldots$, and,
- we sample $\boldsymbol{x}^{(i)}_{(T+1)\Delta t}$ from $p_{cf}(\boldsymbol{x}^{(i)}_{(T+1)\Delta t}|\boldsymbol{X}^{(i)}_{(T+1)\Delta t}, \boldsymbol{\theta}_{cf})$, $\boldsymbol{x}^{(i)}_{(T+2)\Delta t}$ from $p_{cf}(\boldsymbol{x}^{(i)}_{(T+2)\Delta t}|\boldsymbol{X}^{(i)}_{(T+2)\Delta t}, \boldsymbol{\theta}_{cf})$ etc.

We note that this procedure does not necessarily ensure enforcement of the constraints by future CG states. Nevertheless it gives rise to samples of the full FG state evolution from which any observable of interest as well as the predictive uncertainty can be computed.

For the *extrapolative* setting above, i.e. for a new FG initial condition $\hat{\boldsymbol{x}}_0$, the evolution equations of the CG states as well as the emission density p_{cf} can be employed as long as the initial state $\boldsymbol{X}_0$ is specified or better

Algorithm 1: Prediction - Algorithm for *interoplative* setting

Result: Sample of $\boldsymbol{x}^{(i)}_{(T+P)\Delta t}$
Data: $q_{\boldsymbol{\phi}}(\boldsymbol{X}_{T\Delta t}), q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$
Sample from $q_{\boldsymbol{\phi}}(\boldsymbol{X}^{(i)}_{T\Delta t})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$;
while *Time-step* $(T+P)\Delta t$ *of interest not reached* **do**
  Apply the CG evolution law as described in Equation (4);
end
Sample from $p_{cf}(\boldsymbol{x}_{(T+P)\Delta t} \mid \boldsymbol{X}_{(T+P)\Delta t}, \boldsymbol{\theta})$

Algorithm 2: Prediction - Algorithm for *extrapolative* setting

Result: Sample of $\boldsymbol{x}_{P\Delta t}$
Data: $p_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_0), q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$
Apply Bayesian Inference as described in Equation (20) to infer $p(\boldsymbol{X}_0|\hat{\boldsymbol{x}}_0)$;
Sample from $p(\boldsymbol{X}_0|\hat{\boldsymbol{x}}_0)$ and $q(\boldsymbol{\theta})$;
while *Time-step* $P\Delta t$ *of interest not reached* **do**
  Apply the CG evolution law as described in Equation (4);
end
Sample from $p_{cf}(\boldsymbol{x}_{P\Delta t}|\boldsymbol{X}_{P\Delta t}, \boldsymbol{\theta})$

yet inferred. For that purpose, the posterior $p(\boldsymbol{X}_0|\hat{\boldsymbol{x}}_0)$ of $\boldsymbol{X}_0$ needs to be determined which according to Bayes rule will be proportional to:

$$p(\boldsymbol{X}_0 \mid \hat{\boldsymbol{x}}_0) \propto p_{cf}(\hat{\boldsymbol{x}}_0 \mid \boldsymbol{X}_0, \boldsymbol{\theta}_{cf})\ p_{c,0}(\boldsymbol{X}_0) \tag{20}$$

where $p_{c,0}(\boldsymbol{X}_0)$ is the initial state's prior (see also Equation (16)). For each sample of $\boldsymbol{\theta}_{cf}$ from the (approximate) posterior $q_{\boldsymbol{\phi}}(\boldsymbol{\theta}_{cf})$, samples of $\boldsymbol{X}_0$ must be drawn from $p(\boldsymbol{X}_0|\hat{\boldsymbol{x}}_0)$ and subsequently propagated as in the 3 steps above in order to obtain predictive samples of the full FG state vector (see Algorithm 2).

2.6. Computational considerations

We note that in multiscale dynamical systems of physical interest, the computational cost stems primarily from the simulation of the FG system due to its generally very high-dimensional state-vector $\boldsymbol{x}$ and very small time-step δt. Hence, one of the main objectives of this work is to enable the

learning of the CG dynamics with the fewest possible and shortest possible FG time-sequences.

We note that once such FG simulation (or experimental) data have been obtained, neither the training phase of the CG model (section 2.4) nor the prediction phase (section 2.5) require any additional FG simulations. The cost of training depends on the dimension of the CG states $\boldsymbol{X}$ as well as the number of parameters $\boldsymbol{\theta}_c$ (for the CG dynamics), $\boldsymbol{\theta}_{cf}$ (for the coarse-to-fine map) and $\boldsymbol{\phi}$ (for the approximate posterior).

We emphasize that this is a one-time, offline cost i.e. once the CG model has been trained, it can be used to produce probabilistic predictive estimates of the whole FG state-vector into the future without any further recourse to the FG model. One needs only to simulate in such case the CG dynamics which due to the lower-dimensional state-vector $\boldsymbol{X}$ and the much larger CG time-step Δt are much less cumbersome than the FG system.

Finally, if more FG data (e.g. longer or new sequences) become available at a later stage, the SVI algorithm can be re-initialized from the previous values and incorporate the new likelihood terms. If a modest amount of data is introduced, one would expect small (or even no changes for faraway states) changes and therefore rapid convergence. Naturally the introduction of observables at new time instants would introduce additional latent variables for the corresponding CG states.

3. Numerical Illustrations

We demonstrate the capabilities of the proposed framework in discovering predictive, coarse-grained evolution laws as well as effective coarse-grained descriptions, on three examples. Two of those involve very high-dimensional systems of stochastically interacting particles (section 3.1, (Felsberger and Koutsourelakis, 2019)) and the third, a nonlinear pendulum, the dynamics of which we attempt to identify simply from sequences of images (section 3.2, (Champion et al., 2019)). In the sequel, we specify the elements of the proposed model that were presented generically in the previous sections and concretize parametrizations and their meaning. The goals of the numerical illustrations are:

- to assess the predictive performance of the model under "interpolative" and "extrapolative" conditions (see section 2.5). By "interpolative" we mean the ability to predict the evolution of an FG states-sequence when data from this sequence has been used for training. By "extrapolative",

we mean the ability to predict the full FG state evolution from new initial conditions that were *not* used in training.

- to examine the effect of the number n and length T of the data sequences and assess the model's ability to learn the correct structure with small n, T and partial observations.
- to examine the enforcement of the residuals/constraints (e.g. conservation of mass) in the inferred and predicted states.
- to examine the ability of the model to identify *sparse*, interpretable solutions for the CG dynamics.
- to assess the magnitude and time evolution of the predictive uncertainty estimates.
- to assess the ability of the model to learn effective CG state variables and accurate coarse-to-fine maps.

Some of the simulation results as well as the corresponding code will be made available at the following github repository[4] upon publication.

3.1. Particle systems

3.1.1. FG model

The FG model consists of d_f identical particles which can move in the bounded one-dimensional domain $[-1,\ 1]$ (under periodic boundary conditions). The FG variables $\boldsymbol{x}_t$ consist therefore of the coordinates of the particles at each time instant t and the dimension of the system d_f is equal to the number of particles. We consider two types of stochastic dynamics that correspond to an advection-diffusion-type (section 3.1.5) and an inviscid-Burgers-type behavior (section 3.1.6). The particulars of the microscopic dynamics are described in the corresponding sections. In the following, we discuss common aspects of both problems that pertain to the CG description, the CG evolution law and the inference procedures.

[4] `https://github.com/SebastianKaltenbach/PhysicalConstraints_ProbabilisticCG.git`

3.1.2. CG variables and coarse-to-fine mapping

For the CG representation, we employ the normalized particle density $\rho(s,t),\ s \in [-1,\ 1]$ (Li et al., 2007) which we discretize in d_c bins. The state vector $\boldsymbol{X}_t = \{X_{t,j}\}_{j=1}^{d_c}$ contains the particle density values in each of the bins j, i.e. $\sum_{j=1}^{d_c} X_{t,j} = 1$ and $X_{t,j} \geq 0\ \forall t, j$. We emphasize that CG and FG variables are of a different nature (i.e. proportion of particles in each bin vs. coordinates of particles) and, more importantly for practical purposes, of very different dimension.

The nature of the CG variables $\boldsymbol{X}_t$ suggests a *multinomial* for the coarse-to-fine density p_{cf} (section 2.3) i.e.:

$$p_{cf}(\boldsymbol{x}_t|\boldsymbol{X}_t) = \frac{d_f!}{m_1(\boldsymbol{x}_t)!\ m_2(\boldsymbol{x}_t)! \ldots m_{d_c}(\boldsymbol{x}_t)!} \prod_{j=1}^{d_c} X_{t,j}^{m_j(\boldsymbol{x}_t)}, \tag{21}$$

where $m_j(\boldsymbol{x}_t)$ is the number of particles in bin j. The underlying assumption is that, given the CG state $\boldsymbol{X}_t$, the coordinates of the particles $\boldsymbol{x}_t$ are *conditionally* independent. This does *not* imply that they move independently nor that they cannot exhibit coherent behavior (Felsberger and Koutsourelakis, 2019). The practical consequence of Equation (21) is that no parameters need to be learned for p_{cf} (in contrast to section 3.2).

3.1.3. The CG evolution law and the virtual observables

With regards to the evolution law of the CG states (Equation (3)), we postulate a right-hand side $\boldsymbol{F}(\boldsymbol{X}_t;\boldsymbol{\theta}_c) = \{F_j(\boldsymbol{X}_t;\boldsymbol{\theta}_c)\}_{j=1}^{d_c}$ of the form:

$$\begin{aligned} F_j(\mathbf{X}_t,\boldsymbol{\theta}_c) &= \textstyle\sum_{m=1}^{M} \theta_{c,m}\ \psi_m^{(j)}(\mathbf{X}_t) \\ &= \underbrace{\sum_{h=-H}^{H} \theta_{c,h}^{(1)} X_{t,j+h}}_{1^{st} order} + \underbrace{\sum_{h_1=-H}^{H} \sum_{h_2 \geq h_1}^{H} \theta_{c,\ (h_1,h_2)}^{(2)} X_{t,j+h_1} X_{t,j+h_2}}_{2^{nd} order} \end{aligned} \tag{22}$$

which consists of first- and second-order interactions over a window of size H with $\boldsymbol{\theta}_c^{(1)}$ and $\boldsymbol{\theta}_c^{(2)}$ denoting the vectors of the corresponding unknown coefficients. In this case, the total number of unknown coefficients $\boldsymbol{\theta}_c$, is $M = dim(\boldsymbol{\theta}_c) = (2H+1) + (H+1)(2H+1)$ and grows quadratically with the neighborhood-size H. Since each of the CG variables $X_{t,j}$ refers to the particle density at bin j (and at time t), the neighborhood size H corresponds to the number of bins to the left or to the right of bin j that affect its

evolution in time The feature functions that we generically denote with $\psi_m^{(j)}$ in Equation (22) can also involve higher-order interactions or be of non-polynomial type. Non-Markovian models could be accommodated as well by accounting for memory terms. It is obviously impossible to know a priori which feature functions are relevant in the evolution of the CG states or what types of interactions are essential (e.g. first, second-order etc). At the same time, and especially in the Small Data regime considered, employing a large vocabulary of feature functions can lead to *overfitting*, lack of interpretability and poor predictions, particularly under "extrapolative" conditions. This highly-important *model selection* issue has been of concern in several coarse-graining studies (Noid, 2013). We propose of automatically addressing this within the Bayesian framework advocated by employing appropriate sparsity-inducing priors for $\boldsymbol{\theta}_c$ (Felsberger and Koutsourelakis, 2019). In particular, we make use of the Automatic Relevance Determination (ARD, (Mackay, 1995)) model according to which

$$p(\theta_{c,m} \mid \tau_m) = \mathcal{N}(\theta_{c,m} \mid 0, \tau_m^{-1}), \qquad m = 1, 2, \ldots, M = dim(\boldsymbol{\theta}_c). \tag{23}$$

The following hyperprior for the precision hyperparameters $\boldsymbol{\tau} = \{\tau_m\}_{m=1}^M$ was used:

$$p(\tau_k \mid \gamma_0, \delta_0) = Gamma(\tau_k \mid \gamma_0, \delta_0) \tag{24}$$

The hyperparameters γ_0 and δ_0 are set to very small values 10^{-9} in all ensuing studies (Bishop and Tipping, 2000). As we demonstrate in the sequel, the hypeprior proposed can give rise to parsimonious solutions for the CG dynamics even in the Small Data setting considered.

A discretized version of the CG evolution law (Equation (3) and Equation (22)) with time step Δt is considered by employing a forward Euler scheme[5] which implies the following residual vector $\boldsymbol{R}_l$ at each time-step l (Equation (4)):

$$\boldsymbol{R}_l(\boldsymbol{X}) = \boldsymbol{X}_{(l+1)\Delta t,j} - \boldsymbol{X}_{l\Delta t,j} - \Delta t\ \boldsymbol{F}(\boldsymbol{X}_{l\Delta t,j}, \boldsymbol{\theta}_c), \quad \forall\ l \tag{25}$$

and the corresponding virtual observables $\hat{\boldsymbol{R}}_l$ (Equation (6)).

More importantly, the nature of the CG variables suggests a *conservation of mass* constraint that has to be fulfilled at each time step l. In view of

[5]This corresponds to a multistep method in Equation (4) with $K = 1$, $a_0 = 1, a_1 = -1, \beta_0 = 0$ and $\beta_1 = -1$.

the discussion of section 2.2, this suggests the scalar constraint function as in Equation (5):

$$c_l(\boldsymbol{X}_{l\Delta t}) = \sum_{j=1}^{d_c} X_{l\Delta t,j} \; - 1 = 0, \quad \forall \; l \tag{26}$$

and the corresponding virtual observables $\hat{c}_l$ (Equation (8)).

3.1.4. Inference and Learning

Given the multinomial p_{cf} in Equation (21), we employed the following procedure for generating training data which consists of n numerical experiments in which the FG model is randomly initialized and propagated for one coarse time-step Δt i.e. for $T = \frac{\Delta t}{\delta t}$ microscopic time-steps. In particular:

- For $i = 1, \ldots, n$, we:
  - sample CG initial state $\hat{\boldsymbol{X}}_0^{(i)}$ from a density $p_{c,0}(\hat{\boldsymbol{X}}_0^{(i)})$.
  - sample FG initial state $\hat{\boldsymbol{x}}_0^{(i)}$ from $p_{cf}(\hat{\boldsymbol{x}}_0^{(i)}|\boldsymbol{X}_0^{(i)})$.
  - solve the (discretized) FG model for $\frac{\Delta t}{\delta t}$ microscopic time-steps and record final state $\hat{\boldsymbol{x}}_{\Delta t}^{(i)}$

The generated FG data $\{\hat{\boldsymbol{x}}_{\Delta t}^{(i)}\}_{i=1}^n$ over a *single* CG time-step are used subsequently to draw inferences on the CG model states and parameters (section 2.4). We note that longer time sequences could readily be generated (albeit at an increased cost). The number of samples n is also something that can be selected adaptively since inferences and predictions can be updated as soon as more data become available. The density $p_{c,0}(\boldsymbol{X}_0^{(i)})$ from which initial CG states are drawn, can be selected quite flexibly and some indicative samples are shown in Figure 2 for the advection-diffusion case, and in Figure 12 for the inviscid-Burgers' case. In summary, the data $\mathcal{D}$ employed, apart from $\{\hat{\boldsymbol{x}}_{\Delta t}^{(i)}\}_{i=1}^n$ above consists of the virtual observables $\{\hat{\boldsymbol{R}}_0^{(1:n)}, \hat{\boldsymbol{c}}_1^{(1:n)}\}$.

As a result of the data employed and the parametrization adopted, we have $\boldsymbol{X}_{\Delta t}^{(1:n)}$ as the sole latent vector and $\boldsymbol{\theta}_c, \boldsymbol{\tau}$ as the unknown (hyper)parameters. Since only a single CG time-step was considered, we omitted the slowness prior (see Equation (16)). Hence we sought an approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}, \boldsymbol{\theta}_c, \boldsymbol{\tau})$ (Equation (17)) which we factorized as in Equation (18) as follows:

$$q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(1:n)}, \boldsymbol{\theta}_c, \boldsymbol{\tau}) = \left[\prod_{i=1}^{n} q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(i)})\right] q(\boldsymbol{\theta}_c) q(\boldsymbol{\tau}) \tag{27}$$

Upon substitution in Equation (19), this yields the following ELBO:

$$\begin{aligned}\mathcal{F}(q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(1:n)},\ \boldsymbol{\theta}_c, \boldsymbol{\tau})) \ &= \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p(\mathcal{D}|\ \boldsymbol{X}_{\Delta t}^{(1:n)},\ \boldsymbol{\theta}_c)\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p(\ \boldsymbol{\theta}_c \mid \boldsymbol{\tau})\right] \\ &+\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p(\boldsymbol{\tau})\right] - \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log q_{\boldsymbol{\phi}}\right]\end{aligned} \tag{28}$$

where:

$$p(\mathcal{D}|\boldsymbol{X}_{\Delta t}^{(1:n)}, \boldsymbol{\theta}_c) = p(\hat{\boldsymbol{x}}_{\Delta t}^{(1:n)}|\boldsymbol{X}_{\Delta t}^{(1:n)})\ p(\hat{\boldsymbol{R}}_0^{(1:n)}|\boldsymbol{X}_{\Delta t}^{(1:n)}, \boldsymbol{\theta}_c)\ p(\hat{\boldsymbol{c}}_1^{(1:n)}|\boldsymbol{X}_{\Delta t}^{(1:n)}) \tag{29}$$

Based on Equation (28) the optimal variational posterior densities can be obtained as:

$$\log q^{opt}(\boldsymbol{\theta}_c) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(1:n)})}\left[\log\ p(\hat{\boldsymbol{R}}_0^{(1:n)}|\boldsymbol{X}_{0:1\Delta t}^{(1:n)}, \boldsymbol{\theta}_c)\right] + \mathbb{E}_{q(\boldsymbol{\tau})}\left[\log\ p(\boldsymbol{\theta}_c \mid \boldsymbol{\tau})\right] \tag{30}$$

$$\log q^{opt}(\boldsymbol{\tau}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta}_c)}\left[\log\ p(\boldsymbol{\theta}_c \mid \boldsymbol{\tau})\right] + \log\ p(\boldsymbol{\tau}) \tag{31}$$

$$\begin{aligned}\log q_{\boldsymbol{\phi}}^{opt}(X_{\Delta t}^{(i)}) \ &= \log\ p_{cf}(\boldsymbol{x}_{\Delta t}^{i} \mid \boldsymbol{X}_{\Delta t}^{i}) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta}_c)}\left[\log\ p(\hat{\boldsymbol{R}}_0^{(i)}|\boldsymbol{X}_{0:1\Delta t}^{(i)}, \boldsymbol{\theta}_c)\right] \\ &+\log\ p(\hat{\boldsymbol{c}}_1^{(i)}|\boldsymbol{X}_{\Delta t}^{(i)})\end{aligned} \tag{32}$$

The equations above are coupled and a closed-form solution can be obtained only for the first two. In particular, the optimal posterior approximation for $\boldsymbol{\theta}_c$ is a multivariate normal with mean $\mu_{\boldsymbol{\theta}_c}$ and covariance $\boldsymbol{S}_{\boldsymbol{\theta}_c}$.

$$\mathbf{S}_{\boldsymbol{\theta}_c}^{-1} = \sigma_R^{-2}\sum_{i=1}^{n}\sum_{j=1}^{d_c}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(i)})}\left[\boldsymbol{\psi}^{(j)}(\boldsymbol{X}_{\Delta t}^{(i)})\left(\boldsymbol{\psi}^{(j)}(\boldsymbol{X}_{\Delta t}^{(i)})\right)^T\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\tau})}[diag(\boldsymbol{\tau})] \tag{33}$$

$$\mathbf{S}_{\boldsymbol{\theta}_c}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}_c} = \sigma_R^{-2}\sum_{i=1}^{n}\sum_{j=1}^{d_c}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(i)})}\left[\boldsymbol{\psi}^{(j)}(\boldsymbol{X}_{\Delta t}^{(i)})\right] \tag{34}$$

where the vector $\boldsymbol{\psi}^{(j)}$ consists of the M feature functions $\psi_m^{(j)}$ in Equation (22). The optimal posterior approximation for the vector $\boldsymbol{\tau}$ of the hyperparameters $\{\tau_m\}_{m=1}^{M}$ reduces to a product of independent Gamma-densities (Bishop and Tipping, 2000) with parameters γ_m and δ_m which are given by:

$$\gamma_m = \gamma_0 + 0.5, \qquad \delta_m = \delta_0 + \frac{1}{2}\left(\mu_{\boldsymbol{\theta}_c,m} + S_{\boldsymbol{\theta}_c,(m,m)}\right), \quad m = 0, 1, \ldots, M = dim(\boldsymbol{\theta}_c) \tag{35}$$

Algorithm 3: Inference algorithm for particle systems

Result: $\{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(i)})\}_{i=1}^{n},\ q(\boldsymbol{\theta}_c),\ q(\boldsymbol{\tau})$
Data: $\{\boldsymbol{X}_0^{(i)}, \hat{\boldsymbol{x}}_{\Delta t}^{(i)}\}_{i=1}^{n}$
Initialize the parameters for the variational distributions;
Set iteration counter w to zero;
Set convergence limit ϵ;
while $||parameters_w - parameters_{w-1}||^2 > \epsilon$ **do**
  for $i \leftarrow 1$ **to** n **do**
    Update $q_{\boldsymbol{\phi}}(\boldsymbol{X}_{\Delta t}^{(i)})$ by maximizing the ELBO (see Equation (28))
  end
  update $q(\boldsymbol{\theta}_c)$ according to Equation (33) and Equation (34) ;
  update $q(\boldsymbol{\tau})$ according to Equation (35) ;
  update the iteration counter by one ;
end

Finally and since closed-form updates for the optimal posterior $q_{\boldsymbol{\phi}}^{opt}(X_{\Delta t}^{(i)})$ are impossible, we employed Stochastic Variational Inference (SVI) as detailed in section 2.4 by assuming a multivariate lognormal (in order to ensure positivity of $X_{\Delta t,j}$) with parameters $\boldsymbol{\phi} = \{\boldsymbol{\mu}_i, \boldsymbol{S}_i\}_{i=1}^{n}$[6]. Noisy gradients with respect to the parameters $\boldsymbol{\phi}$ were estimated with Monte Carlo and the reparametrization trick (Kingma and Welling, 2014) and $\boldsymbol{\phi}$ were updated using stochastic gradient ascent (the ADAM algorithm of (Kingma and Ba, 2014) in particular). The inference steps are summarized in Algorithm 3.

	$d_f = dim(\boldsymbol{x})$	$d_c = dim(\boldsymbol{X})$	FG time-step δt	CG time-step Δt
Advection-Diffusion	250×10^3	≤ 64	2.5×10^{-3}	2
inviscid Burgers	250×10^3	≤ 128	2.5×10^{-3}	4

Table 2: FG/CG state-space dimensions and FG/CG time-steps for particle systems investigated.

[6]Diagonal covariances $\boldsymbol{S}_i$ were employed.

3.1.5. Advection-Diffusion system

For the simulations presented in this section $d_f = 250 \times 10^3$ particles were used, which, at each microscopic time step $\delta t = 2.5 \times 10^{-3}$ performed random, non-interacting, jumps of size $\delta s = \frac{1}{640}$, either to the left with probability $p_{left} = 0.1875$ or to the right with probability $p_{right} = 0.2125$. The positions were restricted in $[-1, 1]$ with periodic boundary conditions. It is well-known (Cottet and Koumoutsakos, 2000) that in the limit (i.e. $d_f \to \infty$) the particle density $\rho(s, t)$ can be modeled with an advection-diffusion PDE with diffusion constant $D = (p_{left} + p_{right})\frac{\delta s^2}{2\delta t}$ and velocity $v = (p_{right} - p_{left})\frac{\delta s}{\delta t}$:

$$\frac{\partial \rho}{\partial t} + v\frac{\partial \rho}{\partial s} = D\frac{\partial^2 \rho}{\partial s^2}, \qquad s \in (-1, 1).. \tag{36}$$

For the CG description, 64 bins were employed i.e. $d_c = 64$ and a time step $\Delta t = 2$ (see Table 2). Furthermore we employed first- and second-order feature function as in Equation (22) with a neighborhood size $H = 5$ which implies a total of $M = 77$ unknown parameters $\boldsymbol{\theta}_c$. We incorporate virtual observables pertaining to the residuals $\hat{\boldsymbol{R}}_0$ with $\sigma_R^2 = 10^{-6}$ (Equation (7))[7] and the virtual observables $\hat{\boldsymbol{c}}_1$ pertaining to the conservation-of-mass constraint with $\sigma_c^2 = 10^{-10}$ (Equation (9)).

We employed $n = 32$ and $n = 64$ time sequences for training that were generated as detailed in section 3.1.4 with initial conditions $\{\boldsymbol{X}_0^{(i)}\}_{i=1}^n$ such as the ones seen in Figure 2. The initial conditions were generated by sampling the amplitude of a *sine* function, which was shifted up to ensure all values are positive and then normalized.

Figure 3 provides a histogram of the function values of the conservation-of-mass constraint $\left\{c_1(\boldsymbol{X}_{\Delta t}^{(i)})\right\}_{i=1}^n$ upon convergence. The small values suggest that this has been softly incorporated in the CG states. A similar histogram for the norm of the residuals $\left\{\boldsymbol{R}_0(\boldsymbol{X}^{(i)})\right\}_{i=1}^n$ is depicted in Figure 4 which also suggests enforcement of the CG evolution with the parameters $\boldsymbol{\theta}_c$ learned from the data. The evolution of the posterior mean $\boldsymbol{\mu}_{\boldsymbol{\theta}_c}$ (Equation (34)) of (a subset of) these parameters over the iterations of the SVI is depicted in Figure 5. Therein, and more clearly in Figure 6, one can observe the

[7] A very interesting possibility which is not explored here would be to learn σ_R^2 i.e. the strength of the enforcement of the CG evolution law from the data. This would increase the flexibility of the model in cases where the vocabulary of the feature functions selected in the right-hand side of the CG dynamics is not rich enough.

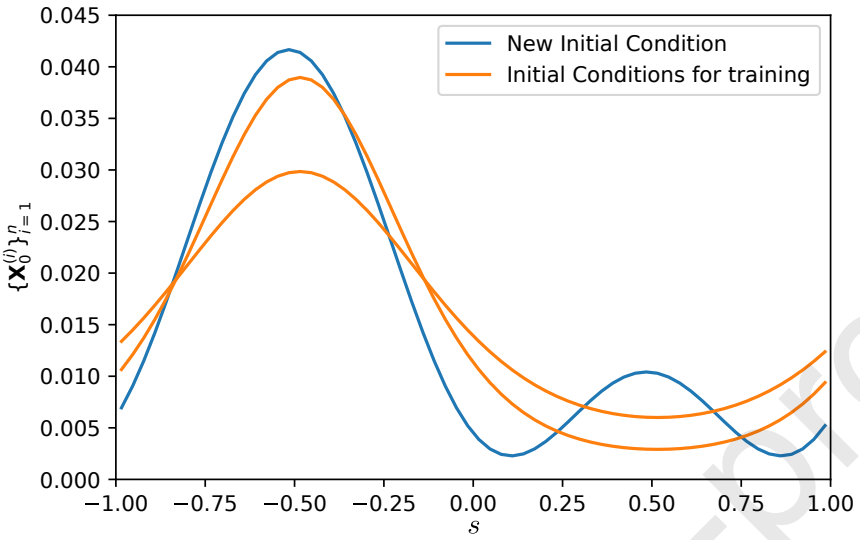


Figure 2: Sample initial conditions $\left\{\boldsymbol{X}_0^{(i)}\right\}_{i=1}^n$ for the Advection-Diffusion problem (orange) and an initial condition (blue) used for "extrapolative" predictions.

ability of the ARD prior to deactivate the vast majority of the right-hand-side feature functions and reveal a small subset of non-zero, salient terms. Both with $n = 32$ and $n = 64$ training data sequences, only parameters $\boldsymbol{\theta}_c$ associated with first-order-interactions (Equation (22)) are activated. In particular, these are $\theta_{c,-3}^{(1)}$ and $\theta_{c,1}^{(1)}$ which are associated with the feature functions $X_{t,j-3}$ and $X_{t,j+1}$ respectively in Equation (22). This shares similarities with a finite-difference discretization scheme for the advection-diffusion and could be considered as an upwind scheme. The two identified coefficients do not form a centered difference operator but the center of the operator is shifted to the left and therefore takes into account the direction of the particle movement. As the value of the coefficients is not exactly the same the diffusive part is also captured.

Figure 7 depicts one of the inferred CG states $\boldsymbol{X}_{\Delta t}^{(i)}$ as well as the associated posterior uncertainty. Once the CG evolution law is learned, this state can be propagated into the future as detailed in section 2.5 in order to generate predictions. Indicative predictions (under "interpolative" conditions)

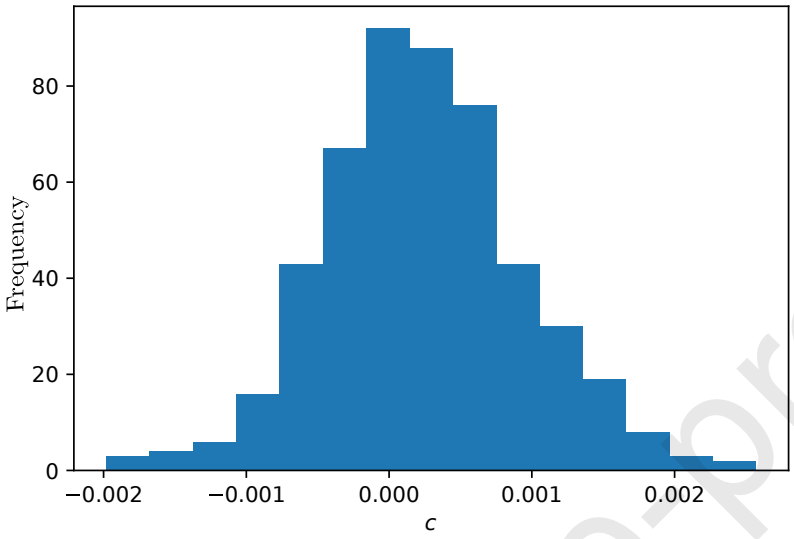


Figure 3: Histogram of the mass constraint c_1

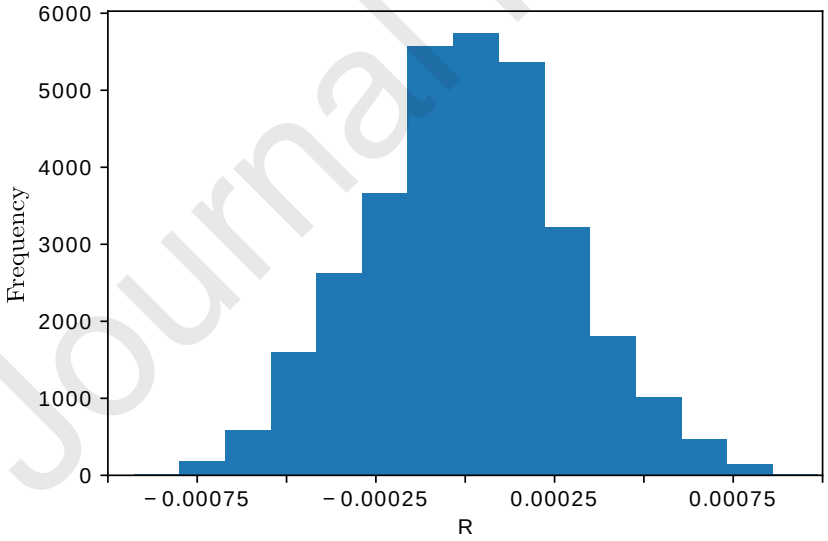


Figure 4: Histogram of the norm of the residual $\boldsymbol{R}_0$

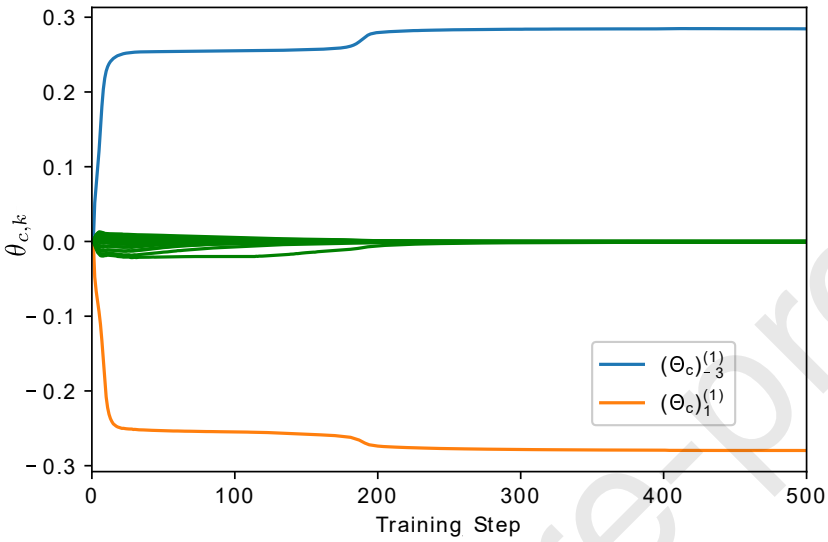


Figure 5: Evolution of a subset of $\boldsymbol{\theta}_c$ parameters with respect to the iterations of the SVI for $n = 64$.

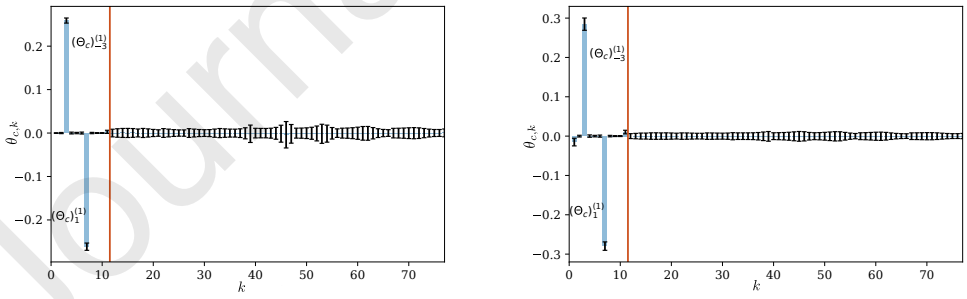


Figure 6: Comparison of the inferred parameters $\boldsymbol{\theta}_c$ for $n = 32$ (left) and $n = 64$ (right) training data sequences. The black bars indicate +/- 1 standard deviation. The red vertical line separates first- from second-order coefficients.

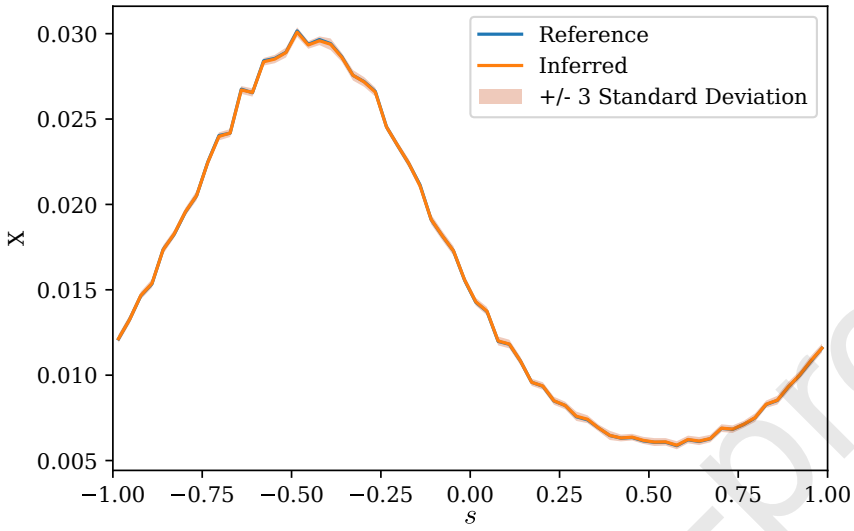


Figure 7: Inferred CG state $\boldsymbol{X}_{\Delta t}^{(i)}$ for a data sequence i. Reference is obtained by sorting the particles into bins according to their position.

can be seen in Figure 8 where the particle density $\rho_x(t, s)$ up to $25\Delta t$ into the future is drawn. The latter as well as the associated uncertainty bounds are estimated directly from the reconstructed FG states. As one would expect, the predictive uncertainty grows, the further into the future one tries to predict. Figure 9 compares the predictive performance as a function of the training data used i.e. $n = 32$ or $n = 64$. In both cases, the ground truth is envelopped and as one would expect, more training data lead to smaller uncertainty bounds.

We also tested the trained model (on $n = 64$) under "extrapolative" conditions i.e. for a different initial condition than the ones included in the training data (Figure 2). The predictive estimates in Figure 10 show very good agreement with the reference solution. It is important to point out that the model can correctly advect and diffuse the particle-bump initially introduced around $s = 0.5$ which suggests that the CG dynamics learned reflect the most important features of the problem.

Finally, in Figure 11, the evolution of the mass constraint into the future is depicted and good agreement with the target value is observed.

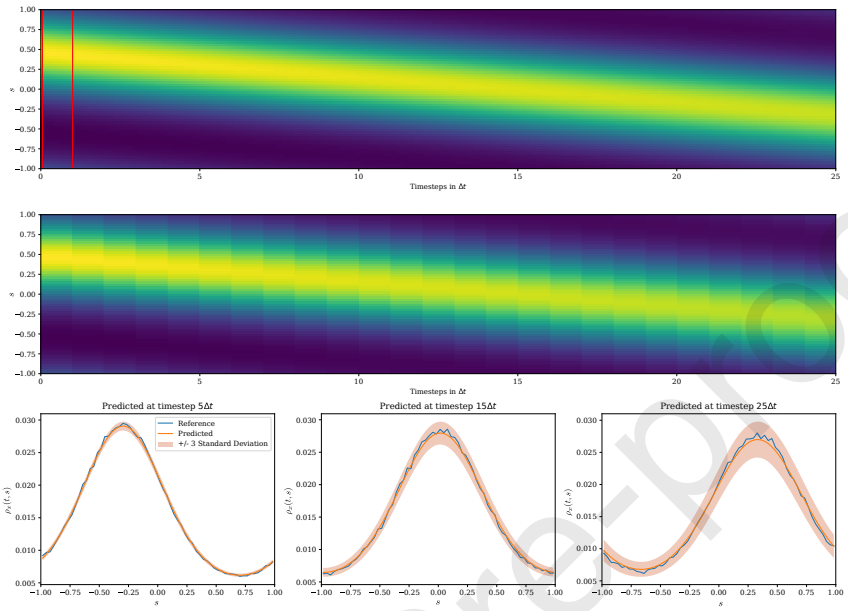


Figure 8: Prediction based on an initial condition contained in the training data. Top: Reference data (the vertical lines indicate the time instances with given data), Middle: Predictive posterior mean, Bottom: snapshots at three different time instances.

3.1.6. Burgers' system

The second test-case involved an FG system of $d_f = 500 \times 10^3$ particles which perform *interactive* random walks i.e. the jump performed at each fine-scale time-step $\delta t = 2.5 \times 10^{-3}$ depends on the positions of the other walkers. In particular we adopted interactions as described in Roberts (1989); Chertock and Levy (2001); Li et al. (2007) so as, in the limit (i.e. when $d_f \to \infty, \delta t \to 0, \delta s \to 0$), the particle density $\rho(s,t)$ follows the inviscid Burgers' equation:

$$\frac{\partial \rho}{\partial t} + \frac{1}{2}\frac{\partial \rho^2}{\partial s} = 0, \qquad s \in (-1,1). \tag{37}$$

For the CG description, 128 bins were employed i.e. $d_c = 128$ and a time step $\Delta t = 4$ (see Table 2). As compared with the previous case, we

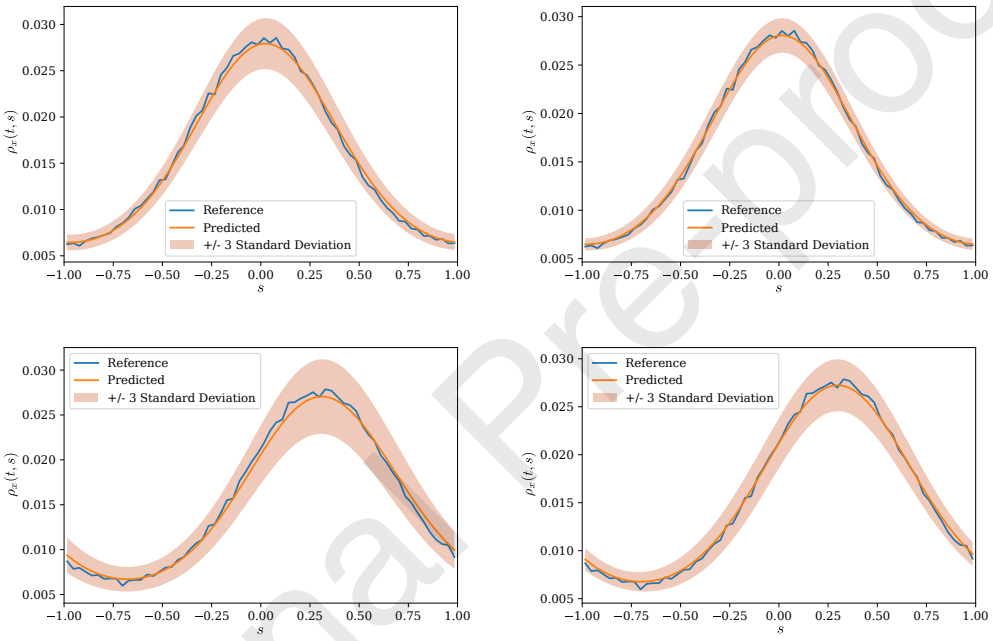


Figure 9: Comparison of the predictions for $n = 32$ (left) and $n = 64$ (right) at $15\Delta t$ (top) and $25\Delta t$ (bottom).

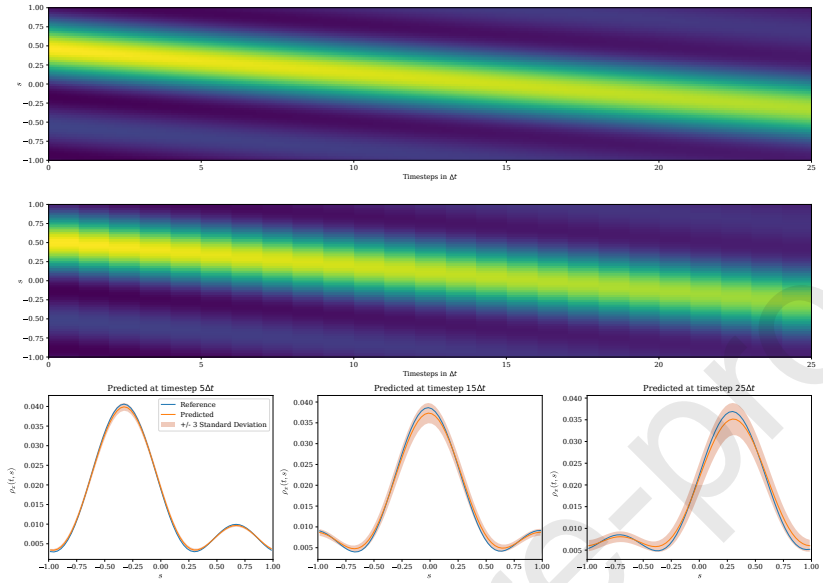


Figure 10: Prediction based on an initial condition NOT contained in the training data. Top: Reference data, Middle: Predictive posterior mean, Bottom: snapshots at three different time instances

enlarged the neighborhood size H in the first- and second-order interactions to $H = 8$, which yielded $M = 170$ right-hand-side terms in Equation (22). We incorporate virtual observables pertaining to the residuals $\hat{\boldsymbol{R}}_0$ with $\sigma_R^2 = 10^{-7}$ (Equation (7)) and the virtual observables $\hat{\boldsymbol{c}}_1$ pertaining to conservation-of-mass constraint with $\sigma_c^2 = 10^{-10}$ (Equation (9)).

We employed $n = 32$, $n = 64$ and $n = 128$ time sequences for training that were generated as detailed in section 3.1.4 with initial conditions $\{\boldsymbol{X}_0^{(i)}\}_{i=1}^n$ such as the ones seen in Figure 12. They were generated by randomizing the width and height of a triangular profile.

Figure 13 provides a histogram of the function values of the conservation-of-mass constraint $\left\{c_1(\boldsymbol{X}_{\Delta t}^{(i)})\right\}_{i=1}^n$ upon convergence. The small values suggest that this has been softly incorporated in the CG states. A similar histogram for the norm of the residuals $\left\{\boldsymbol{R}_0(\boldsymbol{X}^{(i)})\right\}_{i=1}^n$ is depicted in Figure 14 which

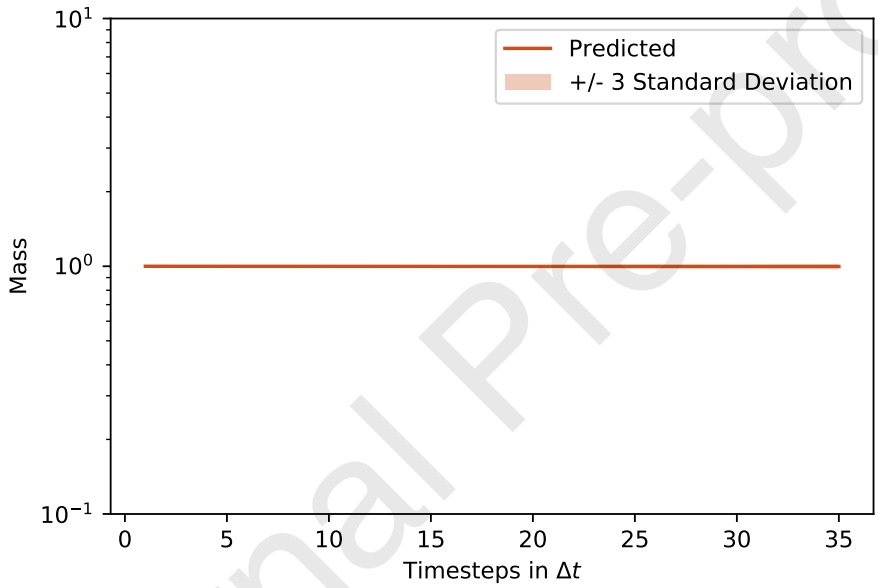


Figure 11: Evolution of the mass constraint (target value is 1) in time including future time-instants. "Predicted" corresponds to the posterior mean.

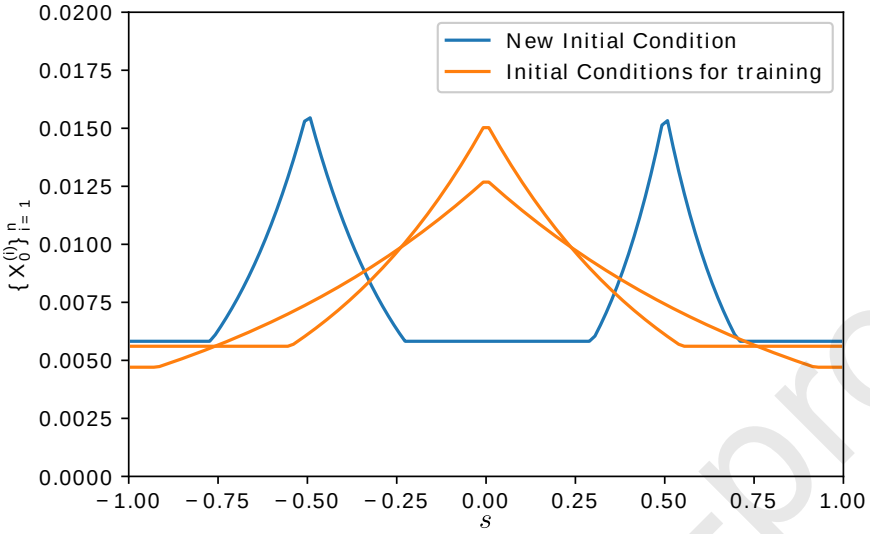


Figure 12: Sample initial conditions $\left\{\boldsymbol{X}_0^{(i)}\right\}_{i=1}^{n}$ for the Burgers' problem (orange) and an initial condition (blue) used for "extrapolative" predictions.

also suggests enforcement of the CG evolution with the parameters $\boldsymbol{\theta}_c$ learned from the data. The evolution of the posterior mean $\boldsymbol{\mu}_{\boldsymbol{\theta}_c}$ (Equation (34)) of (a subset of) these parameters over the iterations of the SVI is depicted in Figure 15. As in the previous example, in Figure 16 one can observe the ability of the ARD prior model to yield sparse solutions for the right-hand side of the CG evolution law. For all three training datasets with $n = 32, 64, 128$ time-sequences, only parameters $\boldsymbol{\theta}_c$ associated with second-order-interactions (Equation (22)) are activated. In particular, these are the negative coefficient $\boldsymbol{\theta}^{(2)}_{c,(0,0)}$ (in all three cases) as well as different second-order coefficients. In the cases of $n = 32$ and $n = 64$ two coefficients are found with positive mean and high posterior uncertainty, but they also have negative posterior correlation (correlation coefficient of -0.88). As all activated coefficients pertain to feature-functions involving the actual bin or bins to the left, the learned evolution law could be interpreted as an upwind scheme, which takes the direction of the Burgers' flow into account. Such schemes are considered advantageous for numerical simulations of fluid flows.

Figure 17 depicts one of the inferred CG states $\boldsymbol{X}^{(i)}_{\Delta t}$ as well as the asso-

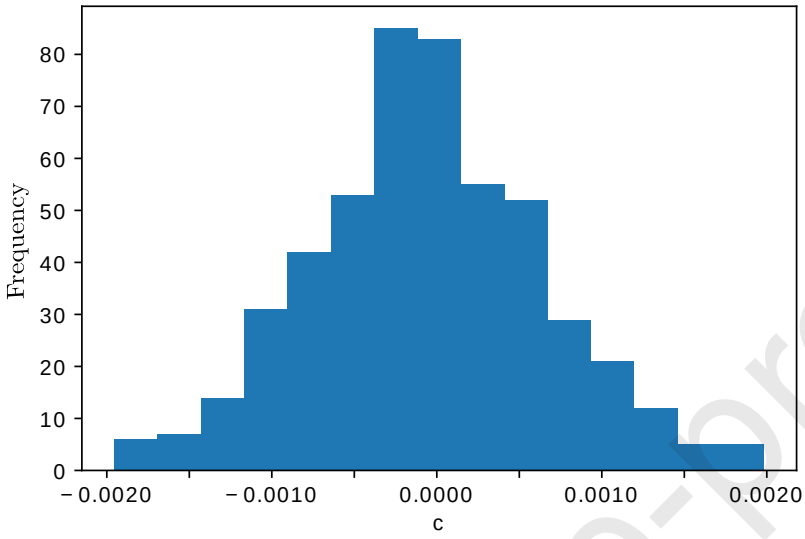


Figure 13: Histogram of the mass constraint c_1

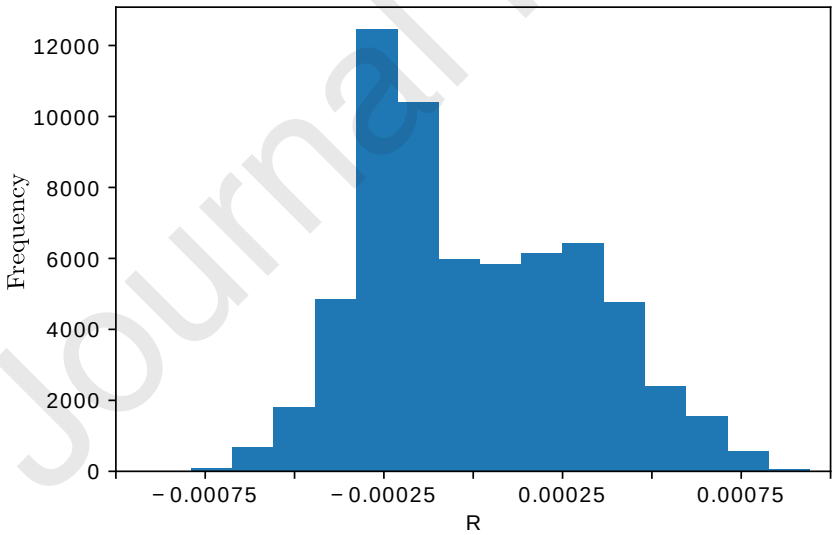


Figure 14: Histogram of the norm of the residual constraint $\boldsymbol{R}_0$

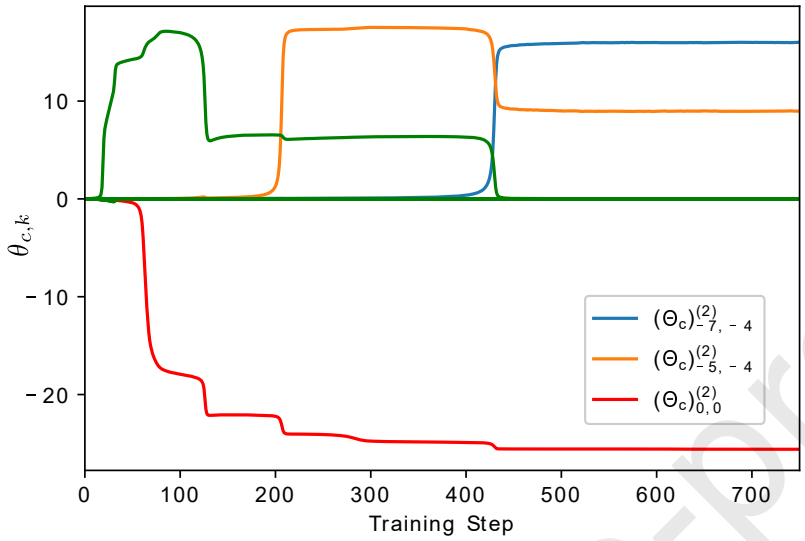


Figure 15: Evolution of a subset of $\boldsymbol{\theta}_c$ parameters with respect to the iterations of the SVI for $n = 64$.

ciated posterior uncertainty. Given the learned CG dynamics, this state can be propagated into the future as detailed in section 2.5 in order to generate predictions. Indicative predictions (under "interpolative" conditions) can be seen in Figure 18 where the particle density up to $25\Delta t$ into the future is drawn. The latter as well as the associated uncertainty bounds are estimated directly from the reconstructed FG states. As in the previous example, the predictive uncertainty grows, the further into the future one tries to predict. Figure 19 compares the predictive performance as a function of the training data used i.e. $n = 32$ or $n = 64$. The increase in data leads for this example to a better fit of the posterior mean to the reference, which captures the location of the shock more precisely. The predictive uncertainty bounds are particularly large at the location of the shock which is the most challenging component in such systems.

We also test the trained model (on $n = 64$) under "extrapolative" conditions i.e. for a "bimodal" initial condition which was quite different from the ones included in the training data (Figure 12). The predictive estimates in Figure 20 show very good agreement with the reference solution. We want

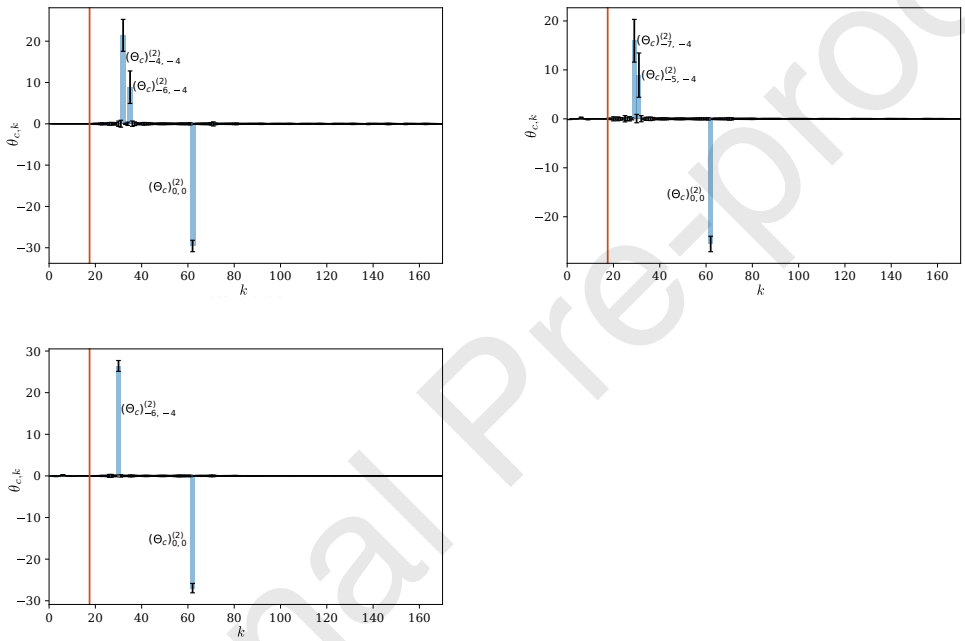


Figure 16: Comparison of the inferred parameters $\boldsymbol{\theta}_c$ for $n = 32$ (top-left), $n = 64$ (top-right) and $n = 128$ (bottom-left) training data. The black bars indicate +/- 1 standard deviation. The red vertical line separates first- from second-order coefficients.

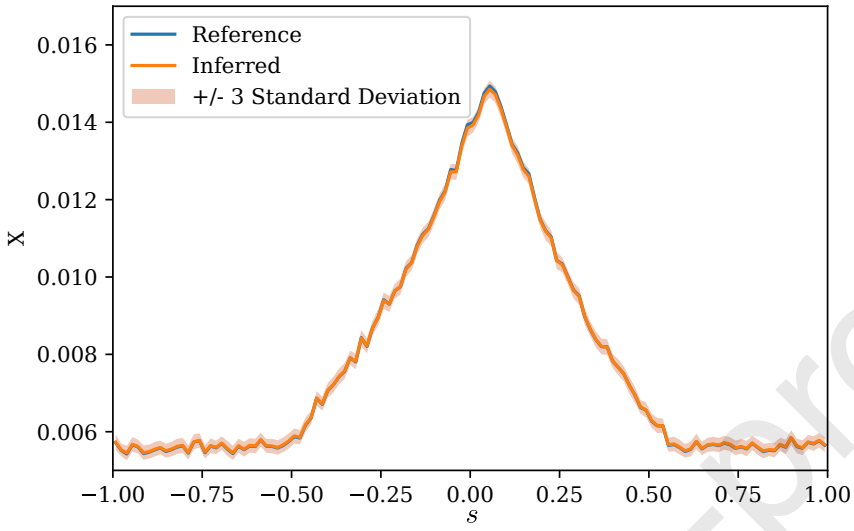


Figure 17: Example of inferred CG state $\boldsymbol{X}_{\Delta t}^{(i)}$ for data sequence i.

to point out that the trained model is capable of capturing the development, the position as well as the propagation of a shock front. Finally, in Figure 21, the evolution of the mass constraint into the future is depicted and good agreement with the target value is observed.

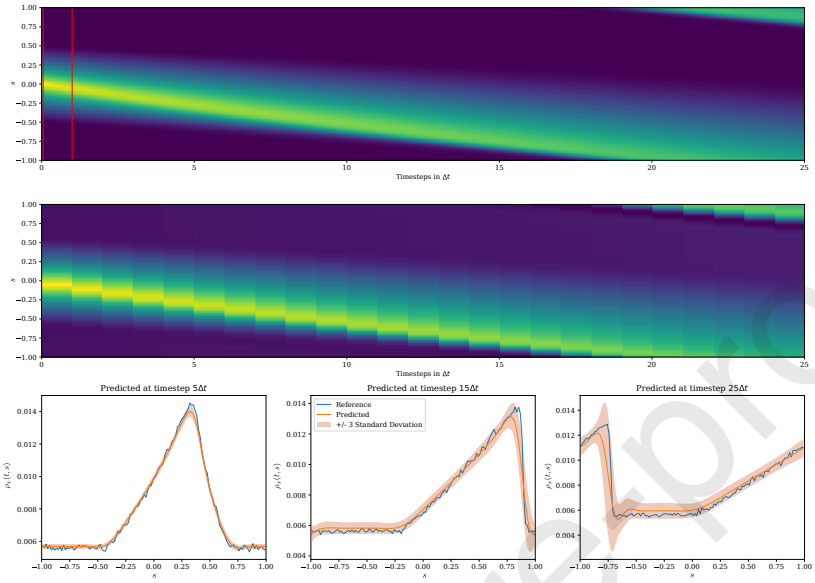


Figure 18: Prediction based on an initial condition contained in the training data. Top: Reference data (the vertical lines indicate the time instances with given data), Middle: Predictive posterior mean, Bottom: snapshots at three different time instances

3.2. Nonlinear Pendulum

In this final example we consider time sequences of images of a nonlinear pendulum in two dimensions as in (Champion et al., 2019).

3.2.1. FG model

For the FG data we generate a series of black-and-white images of a moving disc tied on a string and forming a pendulum (see Figure 31). Each image consists of 29×29 pixels each and each pixel's value was either 1 (occupied) or -1 (unoccupied). Hence $\boldsymbol{x}_t$ was a $d_f = 29^2 = 581$-dimensional vector of binary variables. The dynamics of the pendulum can be fully described by the rotation angle y_t which follows a nonlinear, second-order ODE of the form:

$$\ddot{y}_t + \sin(y_t) = 0 \tag{38}$$

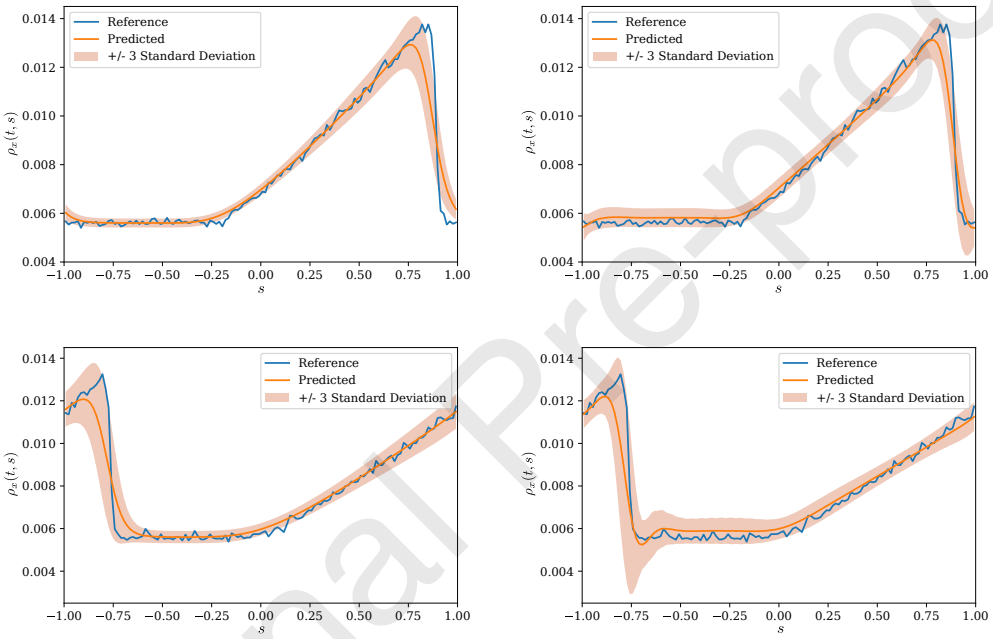


Figure 19: Comparison of the predictions for $n = 32$ (left) and $n = 64$ (right) training data at $15\Delta t$ (top) and $25\Delta t$ (bottom).

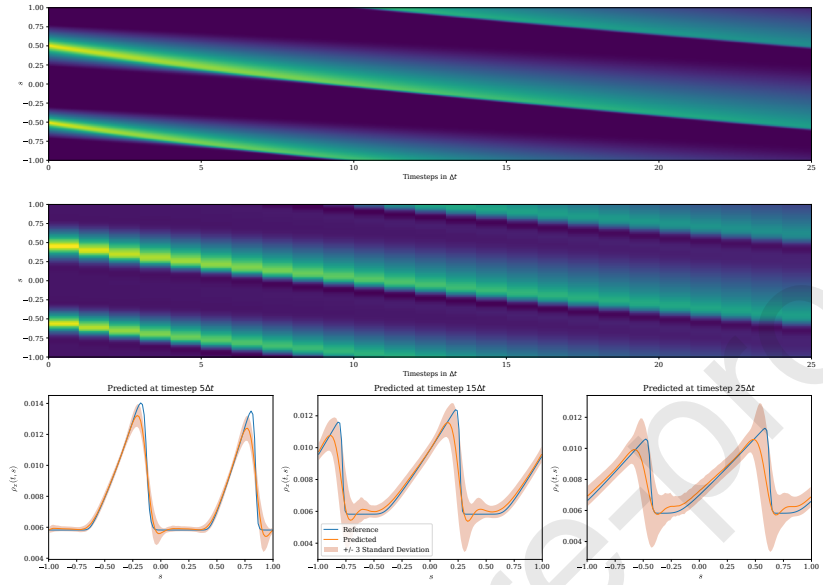


Figure 20: Prediction based on an initial condition NOT contained in the training data. Top: Reference data, Middle: Predictive posterior mean, Bottom: snapshots at three different time instances

The primary goal is to identify the right CG variables as well as CG dynamics solely from image data i.e. binary vectors $\{\hat{\boldsymbol{x}}^{(i)}_{0:T\Delta t}\}_{i=1}^{n}$ collected over T time-steps as the pendulum is initialized from n states/positions. The length of time sequences in the following numerical results was $T = 74$ and the CG time-step $\Delta t = 0.05$[8]. We also considered the effect of missing data i.e. only observing a subset of the $T+1$ values in each sequence and present respective results in Section 3.2.5.

3.2.2. CG variables and coarse-to-fine mapping

The only knowledge introduced a priori with regards to the CG variables $\boldsymbol{X}_t$ is that $dim(\boldsymbol{X}) = d_c = 2$. We intend to investigate procedures that can

[8]For the generation of images a *microscopic* time-step $\delta t = 0.01$ for the integration of Equation (38) was used.

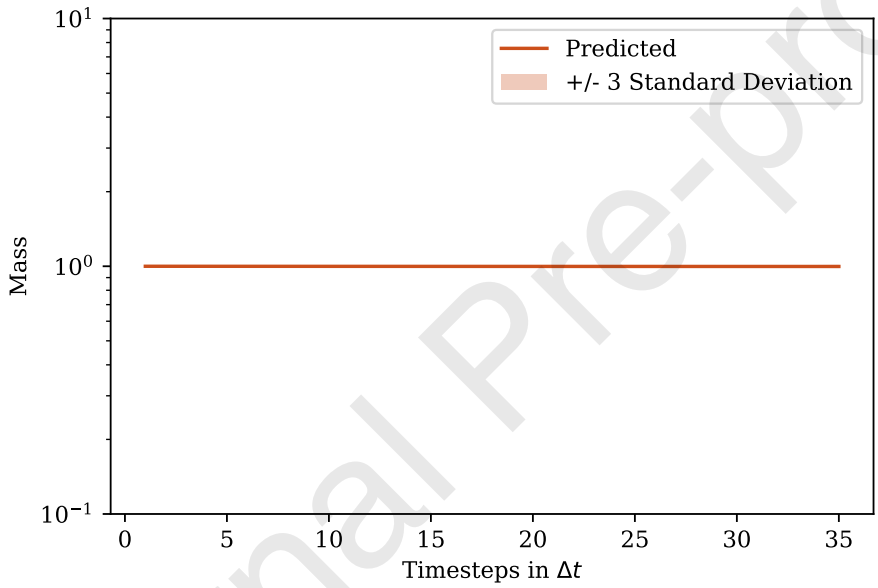


Figure 21: Evolution of the mass constraint (target value is 1) in time including future time-instants. "Predicted" corresponds to the posterior mean.

automatically identify d_c i.e. the number of CG variables. We note at this stage that such efforts could be guided by the ELBO $\mathcal{F}$ (e.g. Equation (19)) which approximates the model evidence and therefore provides a natural Bayesian score for comparing models with different numbers of CG variables.

The other pertinent model component is the coarse-to-fine map which is enabled by the $p_{cf}(\boldsymbol{x}_t|\boldsymbol{X}_t)$ (section 2.3). To that end, we employed the following logistic model[9]:

$$p_{cf}(\boldsymbol{x}|\boldsymbol{X}) = \prod_{s=1}^{d_f} p_{cf}(x_s|\boldsymbol{X}) \tag{39}$$

with

$$p_{cf}(x_s|\boldsymbol{X}) = \begin{cases} \dfrac{1}{1+\exp(-G_s(\boldsymbol{X};\boldsymbol{\theta}_{cf}))} & \text{for } x_s = 1 \\ \dfrac{1}{1+\exp(+G_s(\boldsymbol{X};\boldsymbol{\theta}_{cf}))} & \text{for } x_s = 0 \end{cases} \tag{40}$$

where x_s is the value $(1, 0)$ of each of the pixels $s = 1, \ldots, d_f$. For the link functions $\{G_s\}_{s=1}^{d_f}$, we employed a deep neural net with weights $\boldsymbol{\theta}_{cf}$, the details of which are shown in Figure 22. One fully connected layer followed by two transposed convolutional layers were found to be flexible enough to accurately represent the functions G_s. The CNNs were specifically chosen because of their ability to extract/map features from/to images.

3.2.3. The CG evolution law and the virtual observables

With regards to the evolution law of the CG states $\boldsymbol{X}_t = \{X_{t,1}, X_{t,2}\}$, we postulate the following form:

$$\begin{aligned} \dot{X}_{t,1} &= F_1(\boldsymbol{X}_t, \boldsymbol{\theta}_c) = X_{t,2} \\ \dot{X}_{t,2} &= F_2(\mathbf{X}_t, \boldsymbol{\theta}_c) = \boldsymbol{\theta_c}^T \boldsymbol{\psi}(X_{t,1}) = \textstyle\sum_{m=0}^{M} \theta_{c,m}\ \psi_m(X_{t,1}) \end{aligned} \tag{41}$$

where $\boldsymbol{\theta}_c$ denote the associated parameters. In total we employed $M = 101$ feature functions of the following type:

$$\psi_m(X) = \begin{cases} 1, & m = 0 \\ \sin(mX), & m = 1, \ldots, M/2 = 50 \\ \cos((m-50)X), & m = 51, \ldots, M = 100 \end{cases} \tag{42}$$

[9] We omit the time-index t for clarity.

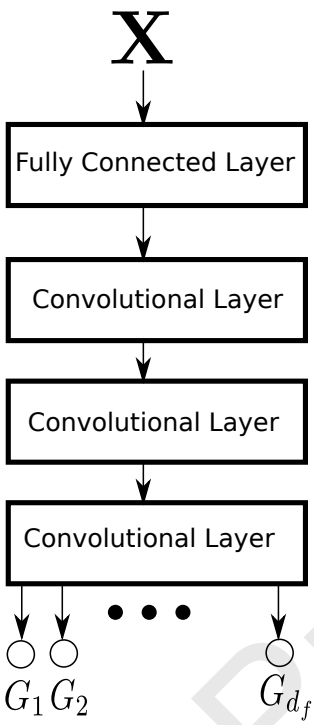


Figure 22: Deep neural net employed for the link functions G_s (Equation (39)). After one dense layer which $32 \cdot 7 \cdot 7$ nodes and rectified linear unit activation function (ReLU), two two-dimensional transposed convolutional layers with 32 filters and a kernel size of 3 as well as a ReLU activation function are applied followed by one-last two-dimensional transposed convolutional layers with one filter, kernel size 3 and without activation to generate the functions G_s .

The form of Equation (41) implies a second-order ODE where the second CG variable plays the role of the velocity. With regards to the parameters $\boldsymbol{\theta}_c$, the sparsity-inducing ARD prior detailed in section 3.1.2 was employed.

To enforce the associated dynamics, we made use of the sympletic Euler time-discretization scheme, which is a first-order integrator, that is explicit in the first variable $(X_{t,1})$ and *implicit* in the other $(X_{t,2})$[10]. The associated

[10]This corresponds to a multistep method in Equation (4) with $K = 1$, $a_0 = 1, a_1 = -1, \beta_0 = 0$ and $\beta_1 = -1$ for the explicit part and $K = 1$, $a_0 = 1, a_1 = -1, \beta_0 = -1$ and $\beta_1 = 0$ for the implicit part.

virtual observables (see Equation (6)) were enforced with $\sigma_R^2 = 10^{-5}$.

3.2.4. Inference and Learning

As in the previous examples (Equation (27)), the approximate posterior was factorized as:

$$q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(1:n)}, \boldsymbol{\theta}_c, \boldsymbol{\tau}) = \left[\prod_{i=1}^{n} q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(i)})\right] q(\boldsymbol{\theta}_c) q(\boldsymbol{\tau}) \tag{43}$$

and closed-form updates were used for $q(\boldsymbol{\theta}_c)$ (see Equations (33) and (34)) and $q(\boldsymbol{\tau})$ (see Equation (35)).

SVI was applied for the posterior densities $q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(i)})$ on the vector of the latent CG states $\boldsymbol{X}_{0:T\Delta t}^{(i)}$ which we approximated with multivariate Gaussians. Since the posterior reveals the fine-to-coarse map which apart from insight can be used for predictive purposes as well, we employed an *amortized* version of SVI ((Kingma and Welling, 2014)) i.e. explicitly accounted for the dependence of each $q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(i)})$ on the corresponding FG observables $\hat{\boldsymbol{x}}_{0:T\Delta t}^{(i)}$ i.e.:

$$q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\ \Delta t}^{(i)}) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_{0:T\Delta t}^{(i)})\ ,\ \boldsymbol{S}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_{0:T\Delta t}^{(i)})\right) \tag{44}$$

The parameters $\boldsymbol{\phi}$ were the weights of a deep convolutional neural net, the architecture of which is shown in Figure 23. This was chosen because it mirrors the DNN architecture employed for the coarse-to-fine map in Figure 22.

Finally it should be mentioned that the "slowness" prior was employed on the hidden states $\boldsymbol{X}_{0:T\Delta t}^{(1:n)}$ as described in Equation (16)[11]. Maximum-likelihood estimates for the hyperparameter $\sigma_{\boldsymbol{X}}^2$ were employed which readily arise by differentiating the ELBO $\mathcal{F}$ and which yield the following update equation:

$$\sigma_{\boldsymbol{X}}^2 = \frac{1}{n\ T\ d_c} \sum_{i=1}^{n} \sum_{l=0}^{T-1} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}_{0:T\Delta t}^{(i)})} \left[\left| \boldsymbol{X}_{(l+1)\ \Delta t}^{(i)} - \boldsymbol{X}_{l\ \Delta t}^{(i)} \right|^2\right] \tag{45}$$

Maximum likelihood estimates were also obtained for the parameters $\boldsymbol{\theta}_{cf}$ (Equation (39)) by numerically differentiating the ELBO $\mathcal{F}$ and performing Stochastic Gradient Ascent (SGA).

[11] For the prior distribution $p_{c,0}(\boldsymbol{X}_0^{(i)})$ a Gaussian mixture distribution with means +1.5 and -1.5 and standard deviation 1.5 was used.

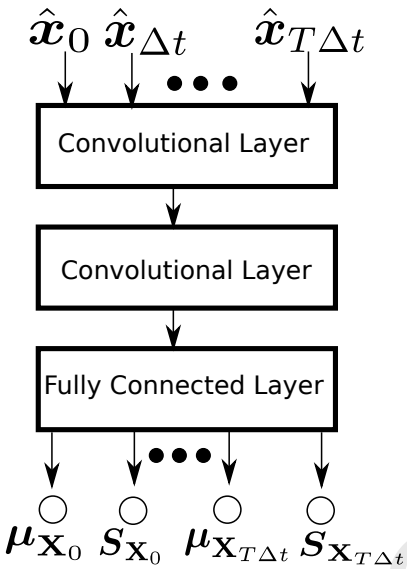


Figure 23: DNN architecture for approximate posterior $q_{\boldsymbol{\phi}}$. The input consists of a time series of pictures of the pendulum and can therefore considered to be three-dimensional, where the first and second dimension are the number of pixels and the third dimension is the number of time steps available for training. This input is given to a three-dimensional convolutional layer with kernel size $(3, 3, 2)$, 32 filters and a ReLU activation followod by another three-dimensional convolutional layer with kernel size 2 in each dimension, 64 filters and a ReLU activation. The last layer is a fully connected layer with $2d_c \cdot T$ nodes and without activation to generate the mean and variance values for each time step of the inferred $\boldsymbol{X}$ coordinates.

A general summary of the steps involved for the inference procedure is-
784 can be found in Algorithm 4. For the implementation we made use of the Tensorflow framework (Abadi et al., 2016).

Algorithm 4: Algorithm for the Pendulum system

Result: $\boldsymbol{\phi}$,$q(\boldsymbol{\theta}_c)$,$q(\boldsymbol{\tau})$,$\boldsymbol{\theta}_{cf}$,$\sigma_{\boldsymbol{X}}$

Data: $\hat{\boldsymbol{x}}_{0:T\Delta t}^{(1:n)}$

Initialize all required parameters;

Set iteration counter w to zero;

while $||ELBO_w - ELBO_{w-1}||^2 > \epsilon$ **do**

  Update the parameters $\boldsymbol{\theta}_{cf}$ and $\boldsymbol{\phi}$ by SGA of the ELBO (Equation (19)) ;

  update $q(\boldsymbol{\theta}_c)$ according to Equation (33) and Equation (34) ;

  update $q(\boldsymbol{\tau})$ according to Equation (35) ;

  update the parameter $\sigma_{\boldsymbol{X}}$ according to Equation (45);

  update the iteration counter by one;

end

3.2.5. Results

Each data sequence $\hat{\boldsymbol{x}}_{0:T\Delta t}^{(i)}$ used consisted of 75 images, i.e. $T = 74$, generated with a time-step $\Delta t = 0.05$ (Figure 24). We investigated two cases for the number of data sequences i.e. $n = 16$ and $n = 64$. The data generation involved sampling uniformly the initial angle $y_0 \in [-\pi, \pi]$ and assuming zero initial velocity i.e $\dot{y}_0 = 0$. We emphasize that none of the data sequences contained a complete oscillation of the pendulum i.e. always partial trajectories were observed.

Figure 25 indicates the posterior means of the inferred $\boldsymbol{\theta}_c$ that parametrize the CG evolution law (Equation (41)) for $n = 16$ and $n = 64$. Of the 101 possible terms, only 2 are activated due the ARD prior.

Figure 26 illustrates trajectories in the two-dimensional CG state-space obtained with various initial conditions for the CG model identified with $n = 16$ and $n = 64$ data sequences. The blue curves correspond to "interporlative" settings i.e. to the CG states of an observed sequence of images, whereas the orange curves to "extrapolative settings" i.e. to the CG states inferred by initializing the pendulum from an arbitrary position *not* contained in the training data. In Figure 27 the predicted evolution in time of both coarse-grained variables is shown. The periodic nature of the CG dynamics is obvious, even though the CG state variables implicitly identified do not correspond to the natural ones i.e. y_t and $\dot{y}_t$.

This can be seen in Figure 28 where for data-sequences $\boldsymbol{x}_{0:T\Delta t}^{(i)}$ (corresponding to the pendulum at various positions i.e. angles $y_{0:T\Delta t}$), we compute

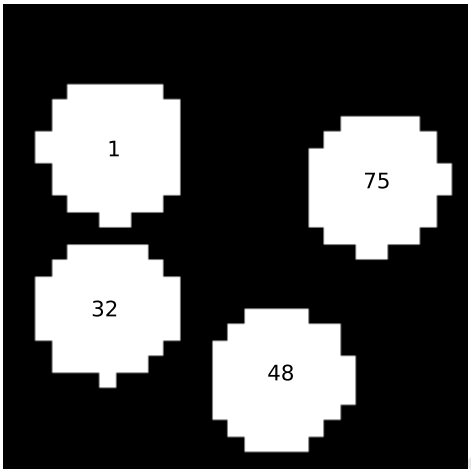


Figure 24: Indicative positions of the pendulum in a data sequence $\hat{\boldsymbol{x}}^{(i)}_{0:T\Delta t}$. The number indicates the corresponding time-step.

from the approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{X}^{(i)}_{0:T\Delta t}|\boldsymbol{x}^{(i)}_{0:T\Delta t})$ (Equation (44)) the mean of the corresponding CG states $\boldsymbol{X}^{(i)}_{0:T\ \Delta t}$ as well as the (in this case negligible) standard deviation. For each time instant $l = 0, 1, \ldots, T$, we plot the pairs of $y_{l\Delta t}$ and (the mean of) $X_{l\Delta t,1}$ (i.e. the first of the CG variables identified) to show the relation between the two variables. While it is obvious from the scales that the first CG variable identified is *not* the angle, it appears to be isomorphic to y. The latter property persists for $n = 64$ even though the sign of the relation has been reversed. The difference between the first CG variable identified and the natural angle y explains the difference between the CG evolution law identified (Figure 25) and the reference one Equation (38).

Figure 29 provides predictive estimates of the position of the center of mass in time. These were obtained by propagating the CG variables in time and for each time instant, sampling p_{cf} for corresponding images $\boldsymbol{x}$. From the latter, the center of mass was computed from the activated pixels i.e. the pixels with value 1. Naturally, predictive uncertainty arises due the stochasticity in the initial conditions of $\boldsymbol{X}$ as well as in p_{cf}. The latter is quantified by the standard deviation and plotted in Figure 29. As in the

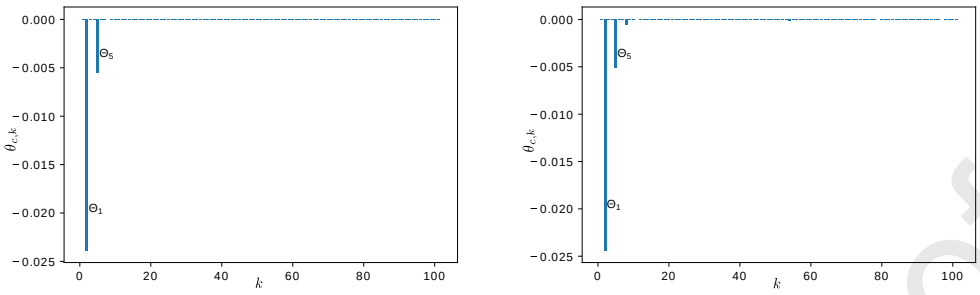


Figure 25: Posterior means of the inferred $\boldsymbol{\theta}_c$ that parametrize the CG evolution law (Equation (41)) for $n = 16$ (left) and $n = 64$ (right) training data.

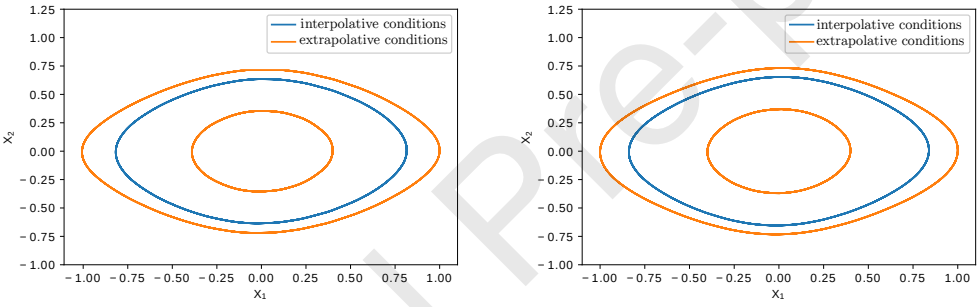


Figure 26: Comparison of trajectories in state space $\boldsymbol{X}$ of the CG dynamics learned for $n = 16$ (left) and $n = 64$ (right) training data.

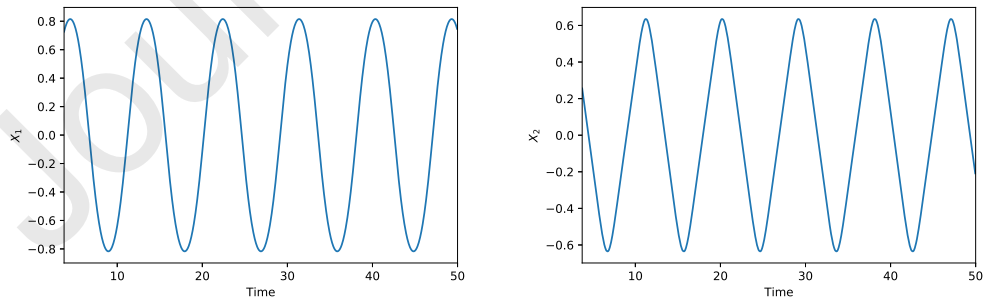


Figure 27: Predicted posterior mean of CG state variables $\boldsymbol{X}_t$

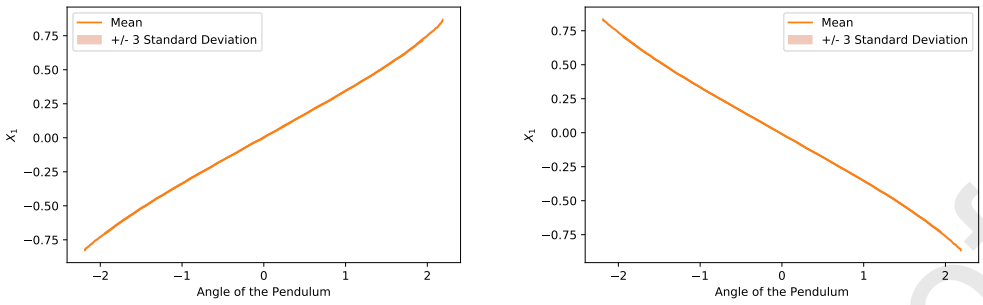


Figure 28: Mapping between the angle of the pendulum and the coarse-grained coordinates for 32 training data and 64 (right) training data.

previous examples, the predictive uncertainty grows, albeit modestly, with time.

Figure 30 depicts predictions in time for two pixels in the image. One can clearly distinguish the change-points i.e. when the pendulum crosses the pixel and its value is changed from 0 to 1 as well as the predictive uncertainty which is concentrated at those change-points. This demonstrates one of the strengths of our approach as due to the coarse-to-fine mapping the whole FG state is reconstructed and every observable can be computed together with the associated predictive uncertainty.

Finally, Figure 31 compares actual images obtained by the reference dynamics of the pendulum with the predictive posterior mean obtained by the CG model and p_{cf} trained on the data. Even though these extend up to 875 time-steps i.e. more than 11 times longer than the time-window over which observations were available, they match the reference quite accurately, a strong indication that the right CG variables and CG dynamics have been identified. An animation containing all frames can be found by following this link.

3.2.6. Missing data

The generative nature of the proposed model makes it highly suitable for handling missing FG data either in the form of partial observations of the FG state vector $\boldsymbol{x}_t$ or observations over a portion/subset of the time-sequence considered. We investigate the latter case in this section but note that in both situations the only modification required is removing the likelihood terms corresponding to the missing data from Equation (13).

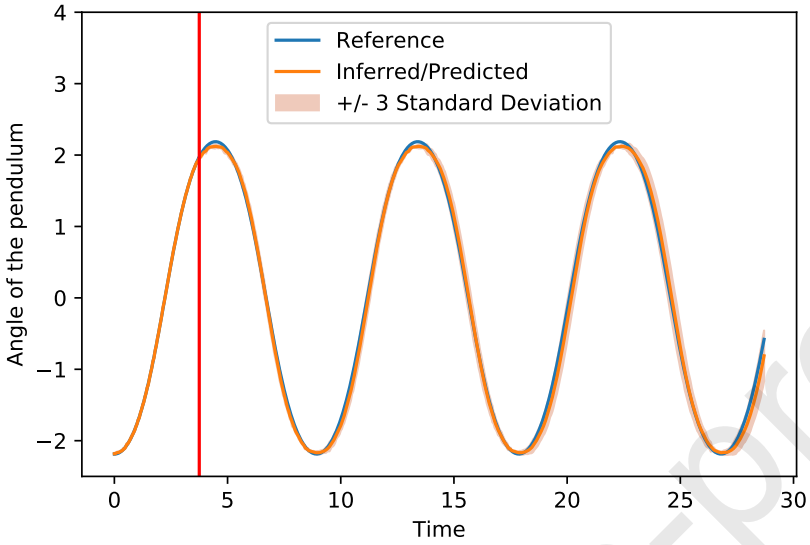


Figure 29: Inferred/Predicted evolution of the center of mass of the pendulum. The vertical line separates the inferred states from the predictions

In particular, we investigated the performance of the model when every second FG state $\boldsymbol{x}_t$ in the training sequences was *not* observed i.e. the FG observables consisted of $\{\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_{2\Delta t}^{(i)}, \boldsymbol{x}_{4\Delta t}^{(i)}, \ldots, \boldsymbol{x}_{T\Delta t}^{(i)}\}$ for each data sequence i (where $T = 74$ as before). As one would expect, fewer observations lead to higher inferential uncertainties as seen when comparing Figure 28 (fully observed case) with Figure 32 (partially observed case). More importantly, fewer observations lead to higher predictive uncertainty as seen when comparing the predictions for the center of pendulum in Figure 29 (fully observed case) with Figure 33 (partially observed case).

4. Conclusions

We proposed a probabilistic generative model for the automated discovery of coarse-grained variables and dynamics based on fine-grained simulation data. The FG simulation data are augmented in a fully Bayesian fashion by virtual observables that enable the incorporation of physical constraints at the CG level that appear in the form of equalities. These could be residuals of the CG evolution law or more importantly conservation laws that are available when CG variables have physical meaning. This is particu-

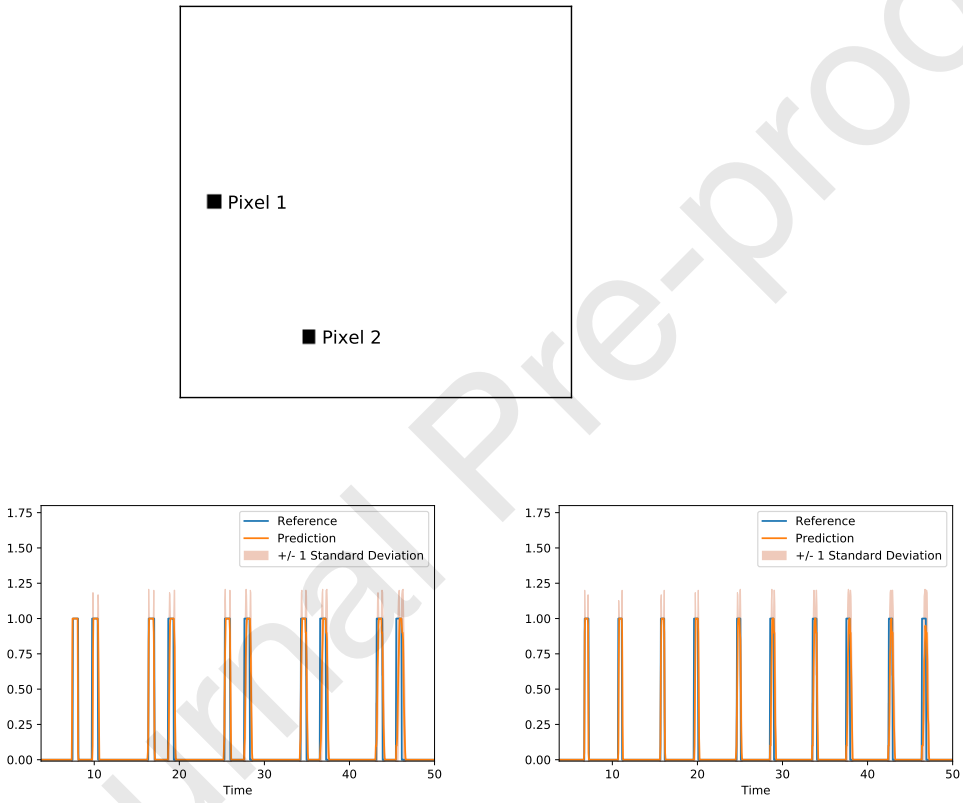
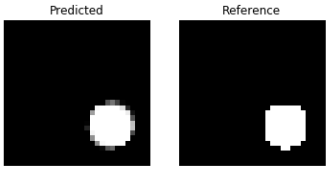
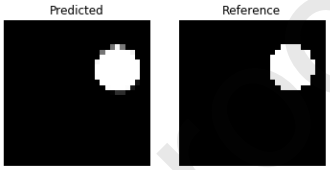


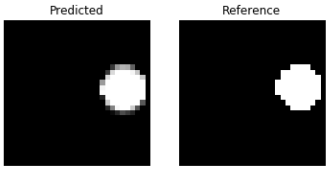
Figure 30: Predicted time history of a single pixel: Pixel 1 (left) and Pixel 2 (right)



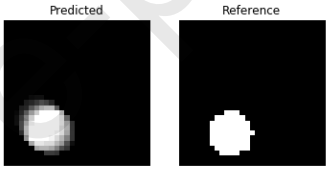
(a) Time step 125



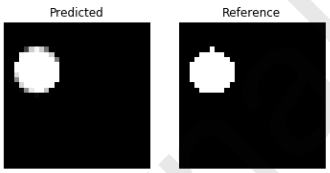
(b) Time step 275



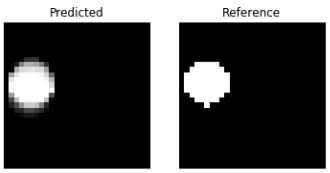
(c) Time step 425



(d) Time step 575



(e) Time step 725



(f) Time step 875

Figure 31: Predictive posterior means of images of the pendulum compared to the reference data

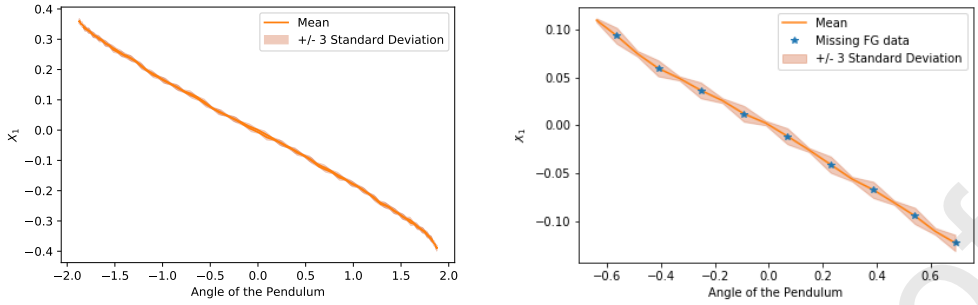


Figure 32: Effect of missing data on the CG variables.The figure on the right is zoomed-in to show the higher uncertainty associated with CG states with missing data

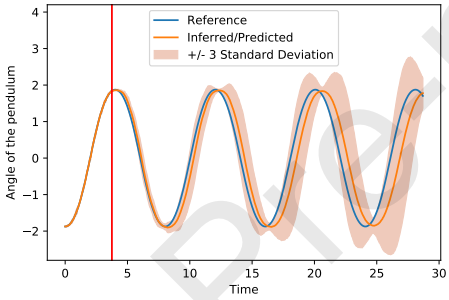


Figure 33: Inferred/Predicted evolution of the center of mass of the pendulum for the missing data case. The vertical line separates the inferred states from the predictions

larly important in the context of physical modeling as in many cases such domain knowledge is a priori available and its inclusion can, not only reduce the amount of training data, but endow the CG model learned with the necessary features that would allow it to provide accurate predictions in out-of-distribution settings. Our approach learns simultaneously a coarse-to-fine mapping and an evolution law for the coarse-grained dynamics by employing probabilistic inference tools for the latent variables and model parameters. The use of deep neural nets for the former component can endow great expressiveness and flexibility. The concept of sparsity, which is invoked in learning CG dynamics from a large vocabulary of right-hand-side terms, is readily incorporated using sparsity-inducing Bayesian priors without any hyperparameter tuning. Furthermore, appropriate priors can promote the discovery of slow-varying CG variables which better capture the macroscopic features

of the system. As a result of the aforementioned characteristics, the framework can learn from *Small Data* (i.e. shorter and fewer FG time-sequences) which is a crucial advantage in multiscale models where the simulation of the FG dynamics is expensive and slow in exploring the state-space. The model proposed was successfully tested on coarse-graining tasks from different areas. In all three examples, the method performed well under interpolative, and more importantly under extrapolative settings i.e. in cases where initial conditions different from the ones seen during training, are prescribed. Partial or incomplete FG observations can readily be handled due to its generative nature. Moreover, as it is able to reconstruct the entire FG state vector at any future time instant, it is capable of producing predictions of any FG observable of interest as well as quantify the associated predictive uncertainty.

There exists various possibilities to extend the proposed framework, both methodologically as well as in terms of applications. In the latter case and apart from using it for predictive purposes, the CG model learned could also be employed in optimization and control applications. On the methodological front an obvious extension would be to account for the virtual observables at future time-instants as well. This would ensure their enforcement by future CG states but would unavoidably complicate their simulation as a probabilistic inference scheme would need to be employed in order to draw samples.

Another important question pertains to the stability of the CG dynamics identified (Pan and Duraisamy, 2020). This is not currently guaranteed in the discretized nor in the continuous version. This could potentially be achieved by an a-priori parametrization of the CG dynamics in a way that guarantees stability which could in turn reduce the expressivity of the model. Finally, we note that, in our opinion, the most difficult question in coarse-graining multiscale systems, is finding the number of CG state variables that are needed. In physics problems, very often one has an idea of which variables would be suitable either based on the analysis-objectives and/or physical insight. Almost never though does one have a guarantee that these variables are sufficient. Assuming they are, the problem then reduces to finding the appropriate closures (i.e. right-hand sides in the CG dynamics) which is the problem we try to address in this paper. The discovery of additional, potentially non-physical CG state variables, would require additional advances for which we believe the ELBO, i.e. the (approximate) model evidence, could serve as the guiding objective.

References

D. Givon, R. Kupferman, A. Stuart, Extracting Macroscopic Dynamics: Model Problems and Algorithms, Nonlinearity (2004).

W. Bialek, Biophysics: Searching for Principles, Princeton University Press, 2012.

M. Alber, A. Buganza Tepole, W. R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W. W. Lytton, P. Perdikaris, L. Petzold, E. Kuhl, Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences, NPJ digital medicine 2 (2019) 115. doi:10.1038/s41746-019-0193-y.

Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452–459. URL: http://www.nature.com/nature/journal/v521/n7553/full/nature14541.html. doi:10.1038/nature14541.

Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

P.-S. Koutsourelakis, N. Zabaras, M. Girolami, Big data and predictive computational modeling, Journal of Computational Physics 321 (2016) 1252–1254.

G. Marcus, E. Davis, Rebooting AI: Building Artificial Intelligence We Can Trust, Pantheon, 2019.

P. Stinis, T. Hagge, A. M. Tartakovsky, E. Yeung, Enforcing constraints for interpolation and extrapolation in generative adversarial networks, Journal of Computational Physics 397 (2019) 108844.

P.-S. Koutsourelakis, E. Bilionis, Scalable Bayesian Reduced-Order Models for Simulating High-Dimensional Multiscale Dynamical Systems, Multiscale Modeling & Simulation 9 (2011) 449–485. doi:10.1137/100783790.

I. Kevrekidis, C. Gear, J. Hyman, P. Kevrekidis, O. Runborg, K. Theodoropoulos, Equation-free multiscale computation: enabling microscopic simulators to perform system-level tasks, Communications in Mathematical Sciences 1 (2003) 715–762.

P. J. Schmid, Dynamic mode decomposition of numerical and experimental data, Journal of Fluid Mechanics 656 (2010) 5–28. URL: https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/dynamic-mode-decomposition-of-numerical-and-experimental-data/AA4C763B525515AD4521A6CC5E10DBD4. doi:10.1017/S0022112010001217.

M. O. Williams, I. G. Kevrekidis, C. W. Rowley, A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition, Journal of Nonlinear Science 25 (2015) 1307–1346. doi:10.1007/s00332-015-9258-5.

H. Wu, F. Noé, Variational approach for learning Markov processes from time series data, arXiv:1707.04659 [math, stat] (2017). URL: http://arxiv.org/abs/1707.04659.

G. Froyland, G. A. Gottwald, A. Hammerlindl, A Computational Method to Extract Macroscopic Variables and Their Dynamics in Multiscale Systems, SIAM Journal on Applied Dynamical Systems 13 (2014) 1816–1846. URL: https://epubs.siam.org/doi/abs/10.1137/130943637. doi:10.1137/130943637.

L. Felsberger, P. Koutsourelakis, Physics-constrained, data-driven discovery of coarse-grained dynamics, Communications in Computational Physics 25 (2019) 1259–1301. doi:10.4208/cicp.OA-2018-0174.

M. Schöberl, N. Zabaras, P.-S. Koutsourelakis, Predictive coarse-graining, Journal of Computational Physics 333 (2017) 49–77.

M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations, arXiv preprint arXiv:1711.10561 (2017).

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, Journal of Computational Physics 378 (2019) 686–707.

Y. Yang, P. Perdikaris, Conditional deep surrogate models for stochastic, high-dimensional, and multi-fidelity systems, Computational Mechanics (2019) 1–18.

A. Mardt, L. Pasquali, H. Wu, F. Noé, VAMPnets for deep learning of molecular kinetics, Nature Communications 9 (2018). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5750224/. doi:10.1038/s41467-017-02388-1.

H. Wu, A. Mardt, L. Pasquali, F. Noe, Deep generative markov state models, in: Advances in Neural Information Processing Systems, 2018, pp. 3975–3984.

L. Duncker, G. Bohner, J. Boussard, M. Sahani, Learning interpretable continuous-time models of latent stochastic dynamical systems, arXiv preprint arXiv:1902.04420 (2019).

C. Grigo, P.-S. Koutsourelakis, A physics-aware, probabilistic machine learning framework for coarse-graining high-dimensional systems in the small data regime, arXiv preprint arXiv:1902.03968 (2019).

Y. Pantazis, I. Tsamardinos, A unified approach for sparse dynamical system inference from temporal measurements, Bioinformatics 35 (2019) 3387–3396. URL: https://academic.oup.com/bioinformatics/article/35/18/3387/5305020. doi:10.1093/bioinformatics/btz065.

S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences 113 (2016) 3932–3937.

E. Kaiser, J. N. Kutz, S. L. Brunton, Sparse identification of nonlinear dynamics for model predictive control in the low-data limit, Proceedings of the Royal Society A 474 (2018) 20180335.

K. Champion, B. Lusch, J. N. Kutz, S. L. Brunton, Data-driven discovery of coordinates and governing equations, arXiv preprint arXiv:1904.02107 (2019).

J. Ohkubo, Nonparametric model reconstruction for stochastic differential equations from discretely observed time-series data, Physical Review E 84 (2011) 066702. URL: https://link.aps.org/doi/10.1103/PhysRevE.84.066702. doi:10.1103/PhysRevE.84.066702.

S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, F. Noé, Data-Driven Model Reduction and Transfer Operator Approximation, Journal of Nonlinear Science 28 (2018) 985–1010. URL: https://doi.org/10.1007/s00332-017-9437-7. doi:10.1007/s00332-017-9437-7.

B. O. Koopman, Hamiltonian Systems and Transformations in Hilbert Space, Proceedings of the National Academy of Sciences of the United States of America 17 (1931) 315–318. URL: https://www.jstor.org/stable/86114.

I. Mezić, Spectral properties of dynamical systems, model reduction and decompositions, Nonlinear Dynamics 41 (2005) 309–325.

S. L. Brunton, B. W. Brunton, J. L. Proctor, J. N. Kutz, Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control, PloS one 11 (2016) e0150171.

M. A. Katsoulakis, P. Plecháč, Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems, The Journal of chemical physics 139 (2013) 074115.

V. Harmandaris, E. Kalligiannaki, M. Katsoulakis, P. Plecháč, Path-space variational inference for non-equilibrium coarse-grained systems, Journal of Computational Physics 314 (2016) 355–383. URL: http://www.sciencedirect.com/science/article/pii/S002199911600173X. doi:10.1016/j.jcp.2016.03.021.

M. A. Katsoulakis, P. Vilanova, Data-driven, variational model reduction of high-dimensional reaction networks, Journal of Computational Physics (2019) 108997.

H. Mori, Transport, collective motion, and brownian motion, Progress of theoretical physics 33 (1965) 423–455.

R. Zwanzig, Nonlinear generalized langevin equations, Journal of Statistical Physics 9 (1973) 215–220.

A. Chorin, P. Stinis, Problem reduction, renormalization, and memory, Communications in Applied Mathematics and Computational Science 1 (2007) 1–27.

H. Lei, N. A. Baker, X. Li, Data-driven parameterization of the generalized langevin equation, Proceedings of the National Academy of Sciences 113 (2016) 14183–14188.

Y. Zhu, J. M. Dominy, D. Venturi, On the estimation of the Mori-Zwanzig memory integral, Journal of Mathematical Physics 59 (2018) 103501. URL: https://aip.scitation.org/doi/10.1063/1.5003467. doi:10.1063/1.5003467.

M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference, The Journal of Machine Learning Research 14 (2013) 1303–1347.

C. Archambeau, M. Opper, Approximate inference for continuous-time Markov processes, Bayesian Time Series Models (2011) 125–140.

R. G. Krishnan, U. Shalit, D. Sontag, Structured inference networks for nonlinear state space models, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

V. Fortuin, G. Rätsch, S. Mandt, Multivariate time series imputation with variational autoencoders, arXiv preprint arXiv:1907.04155 (2019).

J. C. Butcher, Numerical methods for ordinary differential equations, John Wiley & Sons, 2016.

R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni, B. Nadler, Diffusion maps, reduction coordinates and low dimensional representation of stochastic systems, Multiscale Modeling & Simulation 7 (2008) 842 – 864.

J. Trashorras, D. Tsagkarogiannis, From Mesoscale Back to Microscale: Reconstruction Schemes for Coarse-Grained Stochastic Lattice Systems, SIAM Journal on Numerical Analysis 48 (2010) 1647–1677. URL: http://epubs.siam.org/doi/abs/10.1137/080722382. doi:10.1137/080722382.

M. A. Katsoulakis, J. Trashorras, Information loss in coarse-graining of stochastic particle dynamics, Journal of statistical physics 122 (2006) 115–135. URL: http://link.springer.com/article/10.1007/s10955-005-8063-1.

D. Kondrashov, M. D. Chekroun, M. Ghil, Data-driven non-Markovian closure models, Physica D: Nonlinear Phenomena 297 (2015) 33–55. URL: http://www.sciencedirect.com/science/article/pii/S0167278914002413. doi:10.1016/j.physd.2014.12.005.

B. D. Coleman, M. E. Gurtin, Thermodynamics with Internal State Variables, The Journal of Chemical Physics 47 (1967) 597–613. URL: http://scitation.aip.org/content/aip/journal/jcp/47/2/10.1063/1.1711937. doi:10.1063/1.1711937.

O. Cappe, E. Moulines, T. Ryden, Inference in Hidden Markov Models, Springer-Verlag, 2005.

Z. Ghahramani, Unsupervised Learning, in: O. Bousquet, G. Raetsch, U. von Luxburg (Eds.), Advanced Lectures on Machine Learning LNAI 3176, Springer-Verlag, 2004.

D. Durstewitz, A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements, PLOS Computational Biology 13 (2017) e1005542. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005542. doi:10.1371/journal.pcbi.1005542.

L. Wiskott, T. J. Sejnowski, Slow feature analysis: Unsupervised learning of invariances, Neural Computation 14 (2002) 715–770. doi:10.1162/089976602317318938.

C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, in: The International Conference on Learning Representations (ICLR), volume abs/1312.6114, Banff, Alberta, Canada, 2014. URL: http://arxiv.org/abs/1312.6114.

Y. Kim, S. Wiseman, A. C. Miller, D. Sontag, A. M. Rush, Semi-amortized variational autoencoders, arXiv preprint arXiv:1802.02550 (2018).

C. Grigo, P.-S. Koutsourelakis, Bayesian model and dimension reduction for uncertainty propagation: applications in random media, SIAM/ASA Journal on Uncertainty Quantification 7 (2019) 292–323.

J. Li, P. G. Kevrekidis, C. W. Gear, I. G. Kevrekidis, Deciding the Nature of the Coarse Equation Through Microscopic Simulations: The Baby-Bathwater Scheme, SIAM Rev. 49 (2007) 469–487. URL: http://dx.doi.org/10.1137/070692303. doi:10.1137/070692303.

W. G. Noid, Perspective: Coarse-grained models for biomolecular systems, The Journal of Chemical Physics 139 (2013). URL: http://scitation.aip.org/content/aip/journal/jcp/139/9/10.1063/1.4818908. doi:http://dx.doi.org/10.1063/1.4818908.

D. Mackay, Probable Networks and Plausible Predictions - a Review of Practical Bayesian Methods for Supervised Neural Networks, Network-Computation in Neural Systems 6 (1995) 469–505. doi:10.1088/0954-898X/6/3/011.

C. M. Bishop, M. E. Tipping, Variational relevance vector machines, in: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 2000, pp. 46–53.

D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

G.-H. Cottet, P. D. Koumoutsakos, Vortex Methods: Theory and Practice, 2 edition ed., Cambridge University Press, Cambridge ; New York, 2000.

S. Roberts, Convergence of a Random Walk Method for the Burgers Equation, Mathematics of Computation 52 (1989) 647–673. URL: http://www.jstor.org/stable/2008486. doi:10.2307/2008486.

A. Chertock, D. Levy, Particle Methods for Dispersive Equations, Journal of Computational Physics 171 (2001) 708–730. URL: http://www.sciencedirect.com/science/article/pii/S0021999101968032. doi:10.1006/jcph.2001.6803.

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467 (2016).

S. Pan, K. Duraisamy, Physics-informed probabilistic learning of linear embeddings of nonlinear dynamics with guaranteed stability, SIAM Journal on Applied Dynamical Systems 19 (2020) 480–509.

Declaration of interests

X 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: