Integrating Expert ODEs into Neural ODEs: Pharmacology and Disease Progression

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modeling a system's temporal behaviour in reaction to external stimuli is a fundamental problem in many areas. Pure Machine Learning (ML) approaches often fail in the small sample regime and cannot provide actionable insights beyond predictions. A promising modification has been to incorporate expert domain knowledge into ML models. The application we consider is predicting the progression of disease under medications, where a plethora of domain knowledge is available from pharmacology. Pharmacological models describe the dynamics of carefully-chosen medically meaningful variables in terms of systems of Ordinary Differential Equations (ODEs). However, these models only describe a limited collection of variables, and these variables are often not observable in clinical environments. To close this gap, we propose the latent hybridisation model (LHM) that integrates a system of expert-designed ODEs with machine-learned Neural ODEs to fully describe the dynamics of the system and to link the expert and latent variables to observable quantities. We evaluated LHM on synthetic data as well as real-world intensive care data of COVID-19 patients. LHM consistently outperforms previous works, especially when few training samples are available such as at the beginning of the pandemic.

1 Introduction

Understanding the evolution of a system in response to stimuli is *the* central problem in many areas. The Machine Learning (ML) approach to this problem has been to learn a collection of latent variables and construct a dynamical model of the system directly from observational data. While ML has achieved strong predictive performance in some applications, it has two central weaknesses. The first is that it requires large datasets. The second is that the latent variables that the ML approach identifies often have no physical interpretation and do not correspond to any previously-identified quantities.

One approach to dealing with these weaknesses has been to incorporate expert domain knowledge into ML models. Most of the work using this approach has focused on incorporating *high-level* knowledge about the underlying physical system, such as conservation of energy [7, 27, 76], independence of mechanism [48], monotonicity [45], or linearity [28]. In addition, there have been attempts to integrate domain-specific "expert models" into ML models to create "hybrid" models. Most of this work has employed expert models that directly issue predictions [42, 64, 68, 70] or extract useful features from the raw measurements [33].

The approach taken in this paper begins with an expert model in the form of a system of Ordinary Differential Equations (ODEs) and integrates that expert model into a system of Neural ODEs [8]. The specific problem we address is that of predicting disease progression under medications; the specific expert model(s) come from Pharmacology [34] – but we think our approach may be much more widely applicable. For a number of diseases, available pharmacological models, built on the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

basis of specialized knowledge and laboratory experiments, provide a description of the dynamics of carefully-chosen medically meaningful variables in terms of ODEs that govern the evolution of these states [14, 2, 24]. However, these models are typically not directly applicable in clinical environments, because they involve too few variables to fully describe a patient's health state [60, 31], because the expert variables which the models employ may be observable in the laboratory setting but not in clinical environments [23, 5], and because the relationships between the expert variables and clinically observable quantities is not known [19]. We will give a example later in Section 3.5.

This paper proposes a novel hybrid modeling framework, the Latent Hybridisation Model (LHM), that imbeds a given pharmacological model (a collection of expert variables and the ODEs that describe the evolution of these variables) into a larger latent variable ML model (a system of Neural ODEs). In the larger model, we use observational data to learn *both* the evolution of the unobservable latent variables and the relationship between measurements and *all* the latent variables – the expert variables from the pharmacological model *and* the latent variables in the larger model. The machine learning component provides links between the expert variables and the clinical measurements, the underlying pharmacological model improves sample efficiency, and the expert variables provide additional insights to the clinicians. A variety of experiments (using synthetic and real data) demonstrate the effectiveness of our hybrid approach.

2 Problem setting

We consider a set of hospitalized patients $[N] = \{1, \ldots, N\}$ over a time horizon $[0, T]$; $t = 0$ represents the time of admission and $t = T$ represents the maximal length of stay. The health status of each patient i is characterized by a collection of *observable* physiological variables $\mathbf{x}_i(t) \in \mathbb{R}^D$, $D \in \mathbb{N}^+$; because the physiological variables may include vital signs, bloodwork values, biomarkers, etc., $\mathbf{x}_i(t)$ is typically a high-dimensional vector. Although the physiological variables are observable, they are typically *measured* only at discrete times, and with error. To avoid confusion, we distinguish the measurements of these variables $\mathbf{y}(t)$ from the true values; i.e.

$$\mathbf{y}_i(t) = \mathbf{x}_i(t) + \epsilon_{it} \tag{1}$$

where the independent noise term ϵ_{it} accommodates the measurement error (modeling ϵ_t as an autocorrelated stochastic process is left as a future work). For illustrative purposes, we also assume that ϵ_t follows a Normal distribution $N(0, \sigma_i^2)$, but any parametric distribution could be easily accommodated. We denote the measurement times for each patient as $\mathcal{T}_i = \{t_{i1}, t_{i2}, \ldots\}$. We write $\mathbf{a}_i(t) \in \mathbb{R}^A$, $A \in \mathbb{N}^+$ for the *treatments* the patient receives. Some treatments (e.g. intravenous medications) are continuous; others (e.g. surgical interventions) are discrete, so some components of $\mathbf{a}_i(t)$ may be continuous functions but others are (discontinuous) step functions.

It is convenient to write $\mathcal{A}_i[t_1 : t_2] = \{\mathbf{a}_i(t) | t_1 \leq t \leq t_2\}$ and $\mathcal{Y}_i[t_1 : t_2] = \{\mathbf{y}_i(t) | t_1 \leq t \leq t_2, t \in \mathcal{T}_i\}$ for the treatments and measurements (respectively) during the the time window $[t_1, t_2]$. Note that $\mathcal{Y}_i[0 : t]$ and $\mathcal{A}_i[0 : t]$ represent *histories* at time t while $\mathcal{A}_i[t : T]$ and $\mathcal{Y}_i[t : T]$ represent *treatment plans* and *predictions*, respectively. Our objective is to *predict* the future measurements under a given treatment plan $\mathcal{A}_i[t_0 : T]$ given the history:

$$\mathbb{P}(\mathcal{Y}_i[t_0 : T] \mid \underbrace{\mathcal{Y}_i[0 : t_0], \mathcal{A}_i[0 : t_0]}_{\text{Historical observations}}, \underbrace{\mathcal{A}_i[t_0 : T]}_{\text{Treatment plan}}). \tag{2}$$

Understanding this distribution will allow us to compute both point estimates and credible intervals (reflecting uncertainty). Note that it is important for the clinician to understand uncertainty in order to balance risk and reward. When the context is clear, we will omit the subscript i and the time index t.

3 Method

3.1 The pharmacological model

We begin with a pharmacological model which describes the dynamics of a collection of "expert" variables $\mathbf{z}^e(t) \in \mathbb{R}^E$. Each expert variable captures a distinct and medically-meaningful aspect of the human body, e.g. the activation of immune system. The pharmacological model describes the dynamics as a system of Ordinary Differential Equations (ODEs):

$$\dot{\mathbf{z}}^e(t) = f^e(\mathbf{z}^e(t), \mathbf{a}(t); \theta^e), \tag{3}$$

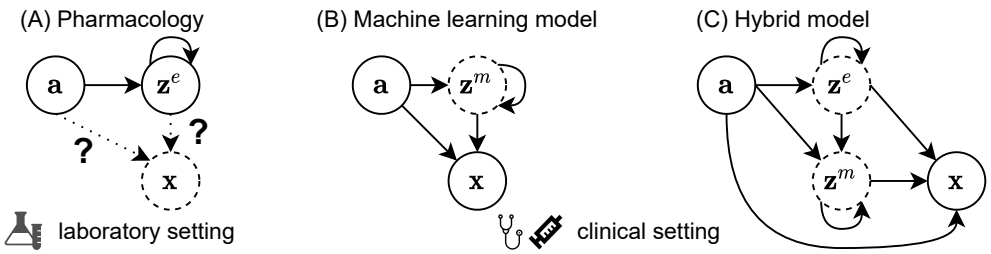

Figure 1: Dependency structure of the three models designed for the laboratory or clinical settings. Dashed nodes represent unobservable variables. The expert variables $\mathbf{z}^e$ are observable in the laboratory setting but not in the clinical setting. The pharmacological model does not contain the links to the clinical variables $\mathbf{x}$.

where we have written $\dot{\mathbf{z}}^e(t)$ for the time derivative of $\mathbf{z}^e$. The functional form of $f^e : \mathbb{R}^E \times \mathbb{R}^A \to \mathbb{R}^E$ is specified but the unknown parameters θ^e (e.g., coefficients) need to be estimated from data.[1,2]

It is important to note that the system of ODEs (3) describes dynamics that are *self-contained*, in the sense that the time derivatives $\dot{\mathbf{z}}^e(t)$ depend only on the *current* values of the expert variables $\mathbf{z}^e(t)$ and the *current* treatments $\mathbf{a}(t)$, and not on histories or on other variables. To ensure that this obtains, it may be necessary to limit the scope of the model and limit attention to a single system of the body (or perhaps to several closely related systems) [13]. As a consequence of these limitations, the expert variables will usually not give a full picture of the health status of the patient and will usually not account for the full array of *observable* physiological variables $\mathbf{x}(t)$ [23].

3.2 The latent hybridisation model: linking expert variables with measurements

As we have already noted, pharmacological models are typically developed and calibrated in the laboratory, where the expert variables can be directly measured – in patients, in laboratory animals, or even in vitro (Figure 1 A). In clinical environments, the expert variables are frequently not observed (Figure 1 B and C). To use the pharmacological models in clinical environments, we must establish links between the expert variables $\mathbf{z}^e(t)$ and the clinical measurements $\mathbf{y}(t)$. To do this we introduce additional latent variables $\mathbf{z}^m(t) \in \mathbb{R}^M$ and posit the following relationship between the latent variables $\mathbf{z}^e, \mathbf{z}^m$ and the observable physiological variables $\mathbf{x}(t)$:

$$\mathbf{x}(t) = g(\mathbf{z}^e(t), \mathbf{z}^m(t), \mathbf{a}(t);\ \gamma) \tag{4}$$

The function $g : \mathbb{R}^{E \times M \times A} \to \mathbb{R}^D$ is a neural network with (unknown) weights γ, and maps the latent space to the "physiological space". We also posit that the dynamics of the latent variables $\mathbf{z}^m(t)$ follow a system of ODEs governed by its current values $\mathbf{z}^m(t)$, the treatments $\mathbf{a}(t)$ *and* the current values of the expert variables $\mathbf{z}^e(t)$.

$$\dot{\mathbf{z}}^m(t) = f^m(\mathbf{z}^m(t), \mathbf{z}^e(t), \mathbf{a}(t);\ \theta^m), \tag{5}$$

The function $f^m : \mathbb{R}^{M \times E \times A} \to \mathbb{R}^M$ is a neural network with (unknown) weights θ^m. Equations (3)-(5) specify the dynamics of LHM. It is convenient to write $\mathbf{z}(t) = [\mathbf{z}^e(t)\ \mathbf{z}^m(t)]$ for the vector of all latent variables and $\Theta = (\theta^e, \theta^m, \gamma, \sigma)$ for the set of all (unknown) coefficients.

The coefficients Θ will be learned from data. However, even after these coefficients are learned, the initial state of the patient $\mathbf{z}_i(0)$ is still *unknown*. In fact, the variation in initial states reflects the heterogeneity of the patient population. If the coefficients and the initial state were known, the entire trajectory of $\mathbf{z}_i$ given the treatments could be computed (numerically). Because we have assumed that the noise/errors ϵ_t are independent, we have

$$\mathcal{Y}_i[t_0 : T] \perp\!\!\!\perp \mathcal{Y}_i[0 : t_0] \mid \mathbf{z}_i(0), \mathcal{A}_i[0 : T], \Theta, \quad \forall t_0 < T \tag{6}$$

However, because the initial state is unknown, it must be *learned* from the measurements $\mathbf{y}_i(t)$.

[1] The system in Equation (3) is quite general; appropriate choices of expert variables allow it to capture both high-order ODEs and time-dependent ODEs [50].

[2] Some care must be taken because systems such as (3) do not always admit unique global solutions. In practice, the pharmacological models are sufficiently well-behaved that global solutions exist and are unique. Although closed-form solutions may not be available, there are various efficient numerical methods for solution.

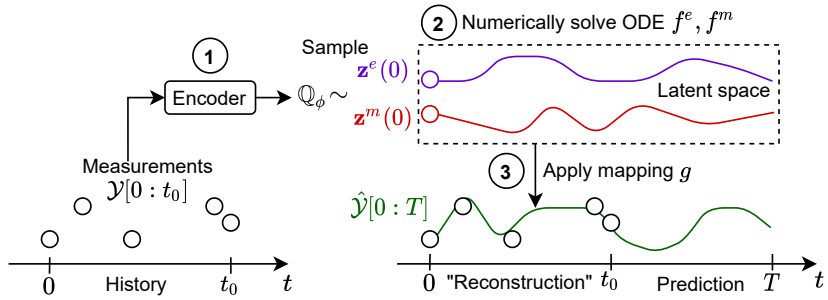

Figure 2: Illustration of the training and prediction procedure.

LHM would reduce to a pure latent neural ODE model [8, 55] if we omitted the expert variables (Figure 1 B). However, that would amount to *discarding prior (expert) information* and so is evidently undesirable. Indeed, as we have noted in the introduction, our approach is driven by the idea of incorporating this prior (expert) information into our hybrid model.

In the current work, we assume that the pharmacological model in Equation 3 is correct. In practice, the model might be wrong in two ways. The obvious way is that the functional form of f^e might be misspecified (e.g. a linear model might be specified when the truth is actually nonlinear). Many existing techniques can address such misspecification and could be integrated into LHM [29, 49, 75]; see the discussion in Appendix A.7. Alternatively, it might be that the system of expert variables is *not* self-contained, and that their evolution actually depends on additional *latent* variables, we leave this more challenging problem for future work.

Practical extensions to LHM such as including static covariates and modeling informative sampling are discussed in Appendix A.7.

3.3 Independent and informative priors

It may be challenging to pinpoint the exact value of the latent variables $\mathbf{z}^e$ based on observations (e.g. due to measurement noise or sampling irregularity). For this reason, we quantify the uncertainty around $\mathbf{z}^e$ using Bayesian inference. In what follows, we assume the initial states $\mathbf{z}_i(0)$ of patients are independently sampled from a prior distribution $\mathbf{z}_i(0) \sim \mathbb{P}_0$. Two points are worth noting.

Independent Priors. We use *independent* prior distributions on the expert variables $\mathbf{z}^e$ and the latent variables $\mathbf{z}^m$, i.e. $\mathbb{P}(\mathbf{z}(0)) = \mathbb{P}(\mathbf{z}^e(0)) \times \mathbb{P}(\mathbf{z}^m(0))$. This guarantees that information in $\mathbf{z}^m(0)$ does not duplicate (any of the) information in $\mathbf{z}^e(0)$, which captures our belief that the latent variables are incremental to the expert variables. In addition, independent priors are also commonly used in Bayesian latent variable models such as variational autoencoders (VAEs) [38, 30].

Informative Priors The prior distribution on the expert variables $\mathbb{P}(\mathbf{z}^e(0))$ should reflect domain knowledge. Such knowledge is usually available from previous studies in Pharmacology [31]. Using an informative prior tends to improve the estimation of latent variables, especially in small-sample settings [41]. Moreover, the expert variables usually take values in specific ranges (e.g. $[0, 10]$ [34]) and going beyond the valid range may lead to divergence. The informative prior can encode such prior knowledge to stabilize training.

3.4 Model training and prediction via amortized variational inference

Given the training dataset $\mathcal{D} = \{(\mathcal{Y}_i[0:T], \mathcal{A}_i[0:T])\}_{i\in[N]}$, we use amortized variational inference (AVI) to estimate the global parameters Θ and the unknown initial condition $\mathbf{z}_i(0)$ [74]. Figure 2 presents a diagram of the training procedure. We start by learning a *variational distribution* to approximate the posterior $\mathbb{P}(\mathbf{z}_i(0)|\mathcal{Y}_i[0:T], \mathcal{A}_i[0:T])$. As is standard in AVI [38, 74], we use a Normal distribution with diagonal covariance matrix as approximation:

$$\mathbb{Q}(\mathbf{z}_i(0)|\mathcal{Y}_i[0:T], \mathcal{A}_i[0:T]) = N(\mu_i, \Sigma_i); \quad \mu_i, \Sigma_i = e(\mathcal{Y}_i[0:T], \mathcal{A}_i[0:T];\ \phi) \tag{7}$$

Here the parameters μ_i, Σ_i are produced by an inference network (also known as an encoder) $e(\cdot)$ with trainable weights ϕ. When the context is clear, we will denote the variational distribution defined

by Equations (7) as $\mathbb{Q}_\phi$. The evidence lower bound (ELBO) for the global parameter Θ and the inference network parameters ϕ is defined as

$$\text{ELBO}(\Theta, \phi) = \mathbb{E}_{\mathbf{z}(0)\sim\mathbb{Q}_\phi}\left[\log\mathbb{P}(\mathcal{Y}_i[0:T]|\mathcal{A}_i[0:T], \mathbf{z}(0), \Theta)\right] - \text{KL}[\mathbb{Q}_\phi|\mathbb{P}_0] \tag{8}$$

To compute the ELBO for a given Θ and ϕ, we sample $\mathbf{z}(0) \sim \mathbb{Q}_\phi$ and numerically solve the ODEs to obtain $\mathbf{z}(t), \forall t \in [0, T]$ (Figure 2; Steps 1, 2). Then, we compute the inner log-likelihood function using the mapping g and the noise distribution (Equations (1) and (4)). Finally, we use Monte Carlo sampling to evaluate the KL divergence term: $\mathbb{E}_{\mathbf{z}(0)\sim\mathbb{Q}_\phi}[\log\mathbb{Q}_\phi(\mathbf{z}(0)) - \log\mathbb{P}_0(\mathbf{z}(0))]$. This is because the informative prior $\mathbb{P}_0$ may not have an analytical KL divergence (unlike the standard Normal prior used in previous works [38, 55]). We optimize ELBO by stochastic gradient ascent and update all parameters jointly in an end-to-end manner (detailed in Appendix A.3).

The prediction procedure follows the same steps as illustrated in Figure 2. For a new patient with history $\mathcal{Y}_i[0:t_0], \mathcal{A}_i[0:t_0]$, we first estimate the variational posterior $\mathbb{Q}_\phi$ using the trained encoder. From Equation (6), we can estimate the target distribution in Equation (2) as:

$$\mathbb{E}_{\mathbf{z}(0)\sim\mathbb{Q}_\phi}\left[\mathbb{P}\big(\mathbf{y}(t)|\mathbf{z}(0), \mathcal{A}_i[0:t_0], \mathcal{A}_i[t_0:T]\big)\right],\ \forall t > t_0. \tag{9}$$

where $\mathcal{A}_i[t_0:T]$ is a future treatment plan. The outer expectation can be approximated by Monte Carlo sampling from $\mathbb{Q}_\phi$ and the inner probability is given by the likelihood function.

Choice of variational distribution and encoder. The training procedure above is agnostic to the exact choice of variational distribution and encoder architecture. We choose the Normal distribution to make fair comparisons with the previous works [8, 55] . For the same reason, we use the reversed time-aware LSTM encoder proposed in [8]. In the Appendix A.4, we show additional experiments with more complex variational distributions, i.e. Normalizing Flows [53].

3.5 Using LHM to provide clinical decision support

In order for clinicians to properly treat patients, they need to predict the progression of disease given the treatments. Although machine learning models may demonstrate feature importance [10, 4], they do not uncover the relationships between those features and the underlying pathophysiology. LHM can provide the missing link between clinical observations and disease mechanisms. In combination with clinical reasoning, this can provide treating clinicians with decision support in several complementary ways.

First, LHM can inform the clinicians about the values of the expert variables $\mathbf{z}^e(t)$ that cannot be observed in the clinical environment but are important for prognosis, choice of treatment, and anticipation of complications. For example, understanding and predicting immune response is pivotal when deciding on immunosuppresive therapy in the treatment of COVID-19: an extreme immune response may lead to a potentially fatal cytokine storm [18], but a suppressed immune response may be equally dangerous in case of (secondary) infection [63, 39]. However, because immune response is not directly observable in the clinical environment, clinicians must rely on proxies such as C-reactive protein (CRP) for inflammation [44]; by their very nature, such proxies are noisy and highly imperfect measures of the desired values.

Secondly, LHM can provide the clinician with predictions of the disease progression given the treatments, enabling the clinicians to design the best treatment plan for the patient at hand.

Finally, LHM can bridge the gap between the laboratory and clinical environments, helping to align model output with clinical reasoning, and thus to bring models to the patient bedside and also to foster translational research [59, 22].

4 Related works

Hybrid models. Hybrid models combine a given expert model with ML [66]. Depending on the type and functionality of the expert model, various approaches have been proposed. *Residual Models* and *Ensembles* use expert models that can issue predictions directly [42, 64, 67, 68, 70]. A Residual Model fits a ML model to the residuals of the expert model while an Ensemble averages the ML and expert predictions. *Feature Extraction* makes use of an expert model that extracts useful features from the measurements [33]; an ML model then uses these features to make predictions. These methods

are not suitable for our setting because our expert model is an ODE that governs the latent variables (5); it does not issue predictions of measurements nor does it extract features that the ML model can use. Appendix A.5 Table 2 summarizes these approaches.

ML inspired by physics uses physical laws to guide the design of architectures [58, 73, 62], loss functions [71, 20], and weight initialization [51]. Examples include Hamiltonian neural networks [27, 7, 76], which reflect the conservation of energy. These models utilize general physical laws rather than a specific expert model, and are rather different than the hybrid models discussed above.

Neural ODEs. Neural ODEs approximate unknown ODEs by a neural network [8], frequently using standard feed-forward networks. ODE^{2}VAE uses an architecture with identity blocks to approximate second-order ODEs [72] and GRU-ODE uses an architecture inspired by the Gated Recurrent Unit (GRU) [15, 9]. Neural ODEs and extensions have achieved state-of-the-art performance in a variety of problems involving irregularly-sampled time series data [55, 15, 37]. We discuss other approaches to learning unknown ODEs from data in Appendix A.5.

Mechanistic models. Mechanistic models are widely applied in sciences such as Physics [61], Epidemiology [69, 26], and Pharmacology [24, 34]. These models use ODEs to describe a system's continuous-time evolution under interventions. The intervention affects the system *deterministically* through the governing ODEs; e.g., Equation (3). This deterministic notion of intervention effect is different from the probabilistic one adopted by the statistical causal inference literature [56] (for a detailed discussion, see [57] and Appendix A.5). LHM uses two ODEs to describe the disease progression under treatment: the known expert ODE f^e and the data-driven Neural ODE f^m.

Latent variable models. Latent variable models are widely used in disease progression modeling [65, 4]. These models attempt to infer a set of latent variables to predict complex high-dimensional measurements. The latent variables sometimes have high-level interpretations (e.g. cluster membership), but do not usually correspond to any well-defined and clinically meaningful physiological variable. Moreover, without informative priors, the latent variables can usually be identified only up to certain transformations (e.g. permutation of cluster labels [47]). By contrast, LHM involves medically meaningful expert variables driven by known governing equations and following informative priors.

5 Experiment and evaluation

Here we present the results of two experimental studies, one with simulated data and one with real data. In both experiments, we study the effect of dexamethasone treatment for COVID-19 patients. Both studies are modeled on the real-life treatment of COVID-19 patients in the ICU.

5.1 Simulation study

In this simulation, we use LHM to predict the results of a single dexamethasone treatment. Each patient i will receive a one-time treatment; with dosage $d_i \sim \text{uniform}[0, 10]$ mg and time $s_i \sim \text{uniform}[0, 14]$. Our objective is to predict future measurements.

Datasets. We generated a variety of datasets to evaluate the model performance under different scenarios. To evaluate how the number of clinical measurements affects performance, we generated datasets with $D = 20, 40$ or 80 observable physiological variables $\mathbf{x}$. For each dataset, we set the number of un-modeled states $\mathbf{z}^m$ according to the number of variables in $\mathbf{x}$ to be $M = D/10 = 2, 4$ or 8 (respectively). (We made this choice to reflect the fact that a larger number of physiological variables often necessitates a larger number of un-modeled states.) We consider a time horizon of $T = 14$ days; this is the median length of stay in hospital for Covid-19 patients [52]. After setting M and D, we generate the data points within a dataset independently.

We use a pharmacological model adapted from [12] that describes five expert variables ($E = 5$) under dexamethasone treatment for COVID-19 patients. We use the same model in the real-data experiment. We specify the model and the expert variables in Appendix A.1. The un-modeled states $\mathbf{z}^m$ are governed by the nonlinear ODE f^m shown in equation (5); the true physiological variables $\mathbf{x}$ are generated by the function g in Equation (4). The specifications of f^m and g are provided in Appendix A.4. For each patient i, each of the components of its initial condition $\mathbf{z}_i(0)$ were independently drawn from an exponential distribution with rate $\lambda = 100$. Measurement noises are drawn independently from $\epsilon_{it} \sim N(0, \sigma^2)$, for $\sigma = 0.2, 0.4$ or 0.8; Equation (1). We first simulate all daily measurements at $t = 1, 2, \ldots, T$, and then randomly remove measurements with probability 0.5 to proxy the fact that measurements are made irregularly.

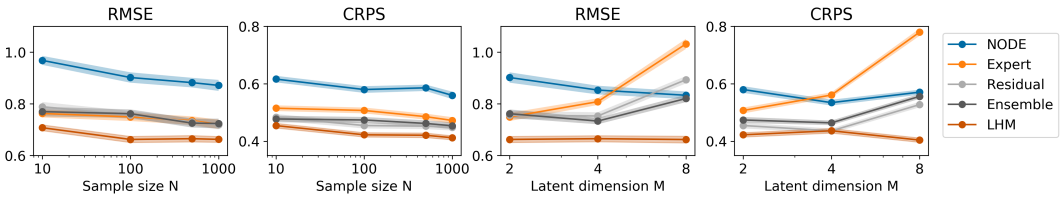

Figure 3: **Simulation Results**. Prediction accuracy on future measurements $\mathcal{Y}[2:T]$ given the observed history $\mathcal{Y}[0:2]$ as measured by RMSE and CRPS. The left 2 panels show the results under different training sample sizes N_0. The right 2 panels show the results for different numbers of un-modeled variables M. The shaded areas represent 95% confidence intervals.

Prediction task For a given patient i, we use the measurements $\mathcal{Y}[0:t_0]$ up to some time t_0 and the treatment plan for that particular patient to *predict* the future measurements $\mathcal{Y}[t_0:T]$. (Note that treatment may have occurred prior to time t_0 or may be planned following time t_0.) To evaluate the performance under different lengths of observed history, we use $t_0 = 2, 5$ or 10 days. We evaluate the prediction accuracy according to Root Mean Squared Error (RMSE) and the uncertainty calibration using the Continuous Ranked Probability Score (CRPS).

Training and Evaluation. We partition each dataset into a training set, a validation set, and a testing set. We consider training sets consisting of $N_0 = 10, 100, 500$ or 1000 data points; each validation set has 100 data points and each testing set has 1000 data points.

Benchmarks. We compare the performance of our method (LHM) with the performance of four other methods: latent Neural ODE (NODE), the original Pharmacology model (Expert), the residual model (Residual), and the ensemble model (Ensemble) of Expert and NODE, described below. The details of the optimization and hyper-parameter settings are reported in Appendix A.4.

NODE involves Z latent variables $\mathbf{z}(t) \in \mathbb{R}^Z$ whose evolutions are described by $\dot{\mathbf{z}}(t) = f^m(\mathbf{z}(t), \mathbf{a}(t);\ \theta^m)$, where f^m is a neural network with trainable weights θ^m. The number of latent variables Z is a *hyper-parameter* that is set to be greater than $M+E$, which is the number of *true* latent variables.[3] NODE predict the physiological variables as $\hat{\mathbf{y}}_N(t) = g_N(\mathbf{z}(t), \mathbf{a}(t);\ \gamma_N)$, where g_N is a a neural network with trainable weights γ_N. **Expert** is given the true governing equation (3), which describes the expert variables $\mathbf{z}^e(t)$, but no un-modeled latent variables. We use a neural network g_E with trainable weights γ_E to predict the physiological variables: $\hat{\mathbf{y}}_E(t) = g_E(\mathbf{z}^e(t), \mathbf{a}(t);\ \gamma_E)$. **Residual** Given a trained Expert model, we calculate its residuals $\mathbf{r}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_E(t)$. Then a NODE is trained to predict the residuals. The final prediction is $\hat{\mathbf{y}}_E(t) + \hat{\mathbf{r}}(t)$. **Ensemble** makes prediction as $w_{1t}\hat{\mathbf{y}}_N(t) + w_{2t}\hat{\mathbf{y}}_E(t)$, where $\hat{\mathbf{y}}_N(t)$ and $\hat{\mathbf{y}}_E(t)$ are the predictions issued by NODE and Expert respectively. The ensemble weights w_{1t} and w_{2t} are learned on the validation set to minimize the prediction error.

Results. Figure 3 shows the predictive performance with $t_0 = 2$ and $\sigma = 0.2$ (results for other settings are reported in Appendix A.4). LHM achieves the best overall performance. Expert does not perform well because it leaves out $\mathbf{z}^m(t)$. As expected, its performance significantly degrades as the number M of latent variables increases. NODE is more robust to increases in M because it treats the number of latent variables as a hyper-parameter. Nevertheless, NODE is less sample efficient and it achieves worse performance for sample sizes up to 1000 (left half). Both Residual and Ensemble achieve performance gains over NODE and Expert alone, but they under-perform LHM because they perform averaging directly in the output space rather than trying to infer $\mathbf{z}^m$ in the latent space.

5.2 Real-data experiments

In this experiment, we use real data to evaluate the predictive performance of LHM and to illustrate its utility for decision support in a realistic clinical setting that closely tracks the actual treatment of COVID-19 patients in ICU.

Dataset. We used data from the Dutch Data Warehouse (DDW), a multicenter and full-admission anonymized electronic health records database of critically ill COVID-19 patients [21]. Up until

[3] We found that the performance of NODE is not sensitive to the exact choice of Z so long as it is sufficiently larger than $M + E$. (This is consistent with findings reported in the literature [16].)

Table 1: Prediction accuracy (RMSE) on COVID-19 intensive care data under different training sample sizes N. Prediction horizon $H = 24$ hours. The standard deviations are shown in the brackets.

Method $\backslash N_0$	100	250	500	1000
Expert	0.718 (0.71)	0.704 (0.02)	0.702 (0.02)	0.713 (0.01)
Residual	0.958 (0.63)	1.003 (0.03)	0.717 (0.05)	0.635 (0.04)
Ensemble	0.707 (0.60)	0.657 (0.05)	0.628 (0.05)	0.599 (0.05)
NODE	0.662 (0.65)	0.659 (0.02)	0.644 (0.05)	0.650 (0.04)
ODE2VAE	0.674 (0.62)	0.666 (0.02)	0.643 (0.02)	0.619 (0.02)
GRU-ODE	0.722 (0.60)	0.673 (0.05)	0.623 (0.05)	0.601 (0.05)
Time LSTM	0.706 (0.63)	0.649 (0.03)	0.600 (0.03)	0.631 (0.02)
LHM	**0.633 (0.51)**	**0.605 (0.02)**	**0.529 (0.02)**	**0.511 (0.02)**

March 2021, DDW has collected the health trajectories for 3464 patients admitted to intensive care units (ICU) across the Netherlands. However, even if we use the entire DDW for training, the sample size is still relatively small compared to what is typically used by ML (tens or hundreds of thousands of samples [32]). Furthermore, patients are even scarcer at the early stage of pandemic, arguably when a decision support tool is most needed: only 607 patients were admitted at the first peak (by April 2020). After applying the eligibility criterion detailed in Appendix A.6, we obtained a dataset of **2097** patients whose disease progression is characterized by an irregularly-sampled time series of **27** physiological variables (Appendix A.6). These variables capture the vital signals, respiratory mechanics, and biomarkers that are crucial for clinical decisions. In addition, we also included **11** static variables that are known to affect the progression of COVID-19, e.g. BMI (Appendix A.6).

Prediction task. Denote t_0 as the time when the patient received the first dose of dexamethasone (we set $t_0 = 24$ for untreated patients). We use the history up to 24 hours before t_0, $\mathcal{Y}[t_0 - 24 : t_0]$, to predict the future $\mathcal{Y}[t_0 : t_0 + 24H]$ over a time horizon $H = 1, 3$ or 7 days. We use $N_0 = 100, 250, 500$ or 1000 patients for training, 97 for validation, and 1000 for testing. The pharmacological model and the prior distribution of expert variables are detailed in Appendix A.1.

Benchmarks. In addition to all the benchmarks introduced in Section 5.1, we compared the results with two extensions of NODE, GRU-ODE and ODE^{2}VAE, which achieved strong performance in medical time series prediction [72, 15]. We also used the Time LSTM as a strong baseline [6].

Results. The main results are shown in Table 1 (additional results are shown in Appendix A.6). LHM consistently outperformed the benchmarks. Its performance with $N_0 = 100$ samples is close to the pure ML approaches' performance with $N_0 = 500$ samples. As the sample size increases from 100 to 1000, the predictive accuracy of LHM improves by 19% while Time LSTM improves by 11% and NODE by less than 5%. A larger improvement rate suggests LHM adapts to the newly available data faster, which is important when the samples are scarce. As expected, the standalone expert model achieved poor performance because it is unable to capture the full array of clinical measurements.

LHM in action. Here we show how LHM can support clinical decisions beyond predicting future clinical measurements. Managing the level of immune response is pivotal when deciding on immunosuppresive therapy for COVID-19 patients [18, 63, 39]. This is a challenging task because the "right" level of immune response varies across patients [35]: for most patients, we would like to reduce the immune response to avoid cytokine storm and consequent organ failure [18], but for patients with other infections (e.g., a secondary bacterial or fungal infection), we would like to keep their immune systems activated [39, 11]. Because immune response is not directly observable in the clinical settings, clinicians resort to unspecific inflammatory markers such as C-Reactive Protein (CRP) [44]. However, better markers of the immune response such as the cytokine Type I IFNs can be measured in the laboratory setting and have been included in the pharmacological model as an expert variable z_1^e. Moreover, the immune response is affected by dexamethasone concentrations in the lung tissue (the expert variable z_2^e). These concentrations are not easily or routinely measured in a clinical setting, and are therefore not available to clinicians.

Figure 4 shows the *measurements* of CRP $\mathbf{x}_1$, and the *inferred* immune response z_1^e and dexamethasone concentration z_2^e. The two expert variables are inferred by a trained LHM using the first five days of observations. We selected three representative patients based on the treatment regimen they received. Patient A is representative of the 59.8% of the population who did not receive dexamethasone; Patient C is representative of the 12.2% who received dexamethasone according to the guidelines

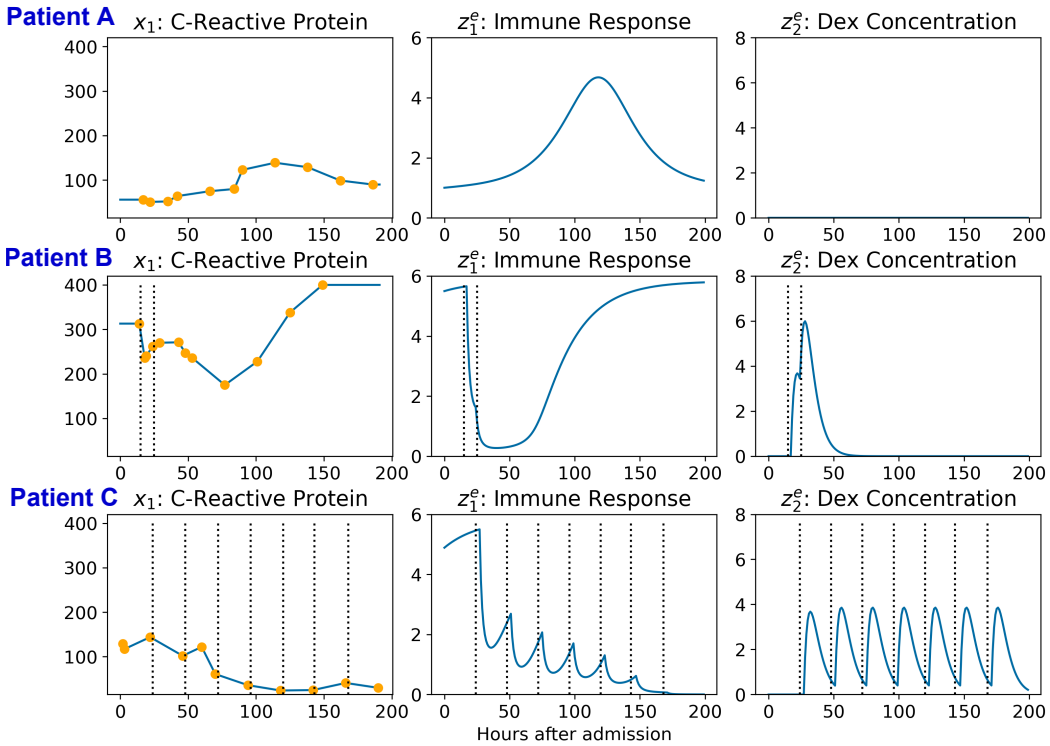

Figure 4: **The observed measurements and the inferred expert variables for three illustrative patients**. Left: The observed physiological variable $x_1(t)$: C-Reactive Protein. Middle: The inferred expert variable $z_1^e(t)$: the immune response to viral infection. Right: The inferred expert variable $z_2^e(t)$: dexamethasone concentration. Vertical dotted lines mark the times of dexamethasone injections.

[46], and Patient B is representative of the 28.0% of the population who received dexamethasone but whose treatment was not according to the guidelines.

For patient A, the initial level of CRP was moderately high, but then it rose and peaked at about 100 hours after admission. In the absence of contraindications, a clinician might begin dexamethasone treatment at this point, but LHM predicts that immune activity z_1^e would decrease afterwards even without treatment. (The right panel is blank because dexamethasone was never administered.)

Patient B was admitted to the ICU with a very high level of the inflammatory marker CRP. Two doses of dexamethasone were given in rapid succession, preceding a decline in both CRP x_1 and immune activity z_1^e. However, after the dexamethasone depleted in the patient's body, the expert model predicts that the immune response will pick up again. This is reflected in the re-occurrence of the high CRP level.

Patient C was admitted to the ICU with a moderately high level of CRP. LHM also inferred a high level of immune activity z_1^e at the initial stage. Inflammation was greatly reduced after dexamethasone treatments, which has an immunosuppressive effect; so x_1 and z_1^e display the same downward trend. Because dexamethasone concentrations falls rapidly within 24 hours after treatment, it was repeated at 24 hour intervals, as seen in the right panel. This pattern is clinically expected under the treatment regimen.

6 Discussion and future work

This paper has focused on a single disease (COVID-19), a single treatment (dexamethasone), and a single expert model. The ultimate goal is to build a model that encompasses a variety of diseases, a variety of treatments and multiple expert models. This is a challenge for future work.

References

[1] Abul K Abbas, Andrew H Lichtman, and Shiv Pillai. *Cellular and molecular immunology E-book*. Elsevier Health Sciences, 2014.

[2] Balaji M Agoram, Steven W Martin, and Piet H van der Graaf. The role of mechanism-based pharmacokinetic–pharmacodynamic (pk–pd) modelling in translational research of biologics. *Drug discovery today*, 12(23-24):1018–1024, 2007.

[3] Ahmed M Alaa, Scott Hu, and Mihaela Schaar. Learning from clinical judgments: Semi-markov-modulated marked hawkes processes for risk prognosis. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2017.

[4] Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[5] Jeffrey K Aronson and Robin E Ferner. Biomarkers—a general review. *Current protocols in pharmacology*, 76(1):9–23, 2017.

[6] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74, 2017.

[7] Tom Bertalan, Felix Dietrich, Igor Mezić, and Ioannis G Kevrekidis. On learning hamiltonian systems from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):121107, 2019.

[8] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *CoRR, abs/1406.1078*, 2014.

[10] Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, pages 3512–3520, 2016.

[11] Michael J Cox, Nicholas Loman, Debby Bogaert, and Justin O'Grady. Co-infections: potentially lethal and unexplored in covid-19. *The Lancet Microbe*, 1(1):e11, 2020.

[12] Wei Dai, Rohit Rao, Anna Sher, Nessy Tania, Cynthia J Musante, and Richard Allen. A prototype qsp model of the immune response to sars-cov-2 for community development. *CPT: pharmacometrics & systems pharmacology*, 10(1):18–29, 2021.

[13] Meindert Danhof. Systems pharmacology–towards the modeling of network interactions. *European Journal of Pharmaceutical Sciences*, 94:4–14, 2016.

[14] Meindert Danhof, Joost de Jongh, Elizabeth CM De Lange, Oscar Della Pasqua, Bart A Ploeger, and Rob A Voskuyl. Mechanism-based pharmacokinetic-pharmacodynamic modeling: biophase distribution, receptor theory, and dynamical systems analysis. *Annu. Rev. Pharmacol. Toxicol.*, 47:357–400, 2007.

[15] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[16] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[17] Justin C Earp, Nancy A Pyszczynski, Diana S Molano, and William J Jusko. Pharmacokinetics of dexamethasone in a rat model of rheumatoid arthritis. *Biopharmaceutics & drug disposition*, 29(6):366–372, 2008.

[18] David C Fajgenbaum and Carl H June. Cytokine storm. *New England Journal of Medicine*, 383(23):2255–2273, 2020.

[19] James D Falvey, Teagan Hoskin, Berrie Meijer, Anna Ashcroft, Russell Walmsley, Andrew S Day, and Richard B Gearry. Disease activity assessment in ibd: clinical indices and biomarkers fail to predict endoscopic remission. *Inflammatory bowel diseases*, 21(4):824–831, 2015.

[20] Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. Predicting ac optimal power flows: Combining deep learning and lagrangian dual methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 01, pages 630–637, 2020.

[21] Lucas M Fleuren, Daan P de Bruin, Michele Tonutti, Robbert CA Lalisang, and Paul WG Elbers. Large-scale icu data sharing for global collaboration: the first 1633 critically ill covid-19 patients in the dutch data warehouse. *Intensive care medicine*, 47(4):478–481, 2021.

[22] Lucas M Fleuren, Patrick Thoral, Duncan Shillan, Ari Ercole, Paul WG Elbers, and Right Data Right Now Collaborators Mark Hoogendoorn Ben Gibbison Thomas LT Klausch Tingjie Guo Luca F. Roggeveen Eleonora L. Swart Armand RJ Girbes. Machine learning in intensive care medicine: ready for take-off? *Intensive care medicine*, 46:1486–1488, 2020.

[23] Richard Frank and Richard Hargreaves. Clinical biomarkers in drug discovery and development. *Nature reviews Drug discovery*, 2(7):566–580, 2003.

[24] Rudolf Gesztelyi, Judit Zsuga, Adam Kemeny-Beke, Balazs Varga, Bela Juhasz, and Arpad Tosaki. The hill equation and the origin of quantitative pharmacology. *Archive for history of exact sciences*, 66(4):427–438, 2012.

[25] Amir Gholami, Kurt Keutzer, and George Biros. Anode: Unconditionally accurate memory-efficient gradients for neural odes. *arXiv preprint arXiv:1902.10298*, 2019.

[26] Michael A Gilchrist and Akira Sasaki. Modeling host–parasite coevolution: a nested approach based on mechanistic models. *Journal of Theoretical Biology*, 218(3):289–308, 2002.

[27] Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[28] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[29] Franz Hamilton, Alun L Lloyd, and Kevin B Flores. Hybrid modeling and prediction of dynamical systems. *PLoS computational biology*, 13(7):e1005655, 2017.

[30] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[31] Nick Holford, Young-A Heo, and Brian Anderson. A pharmacokinetic standard for babies and adults. *Journal of pharmaceutical sciences*, 102(9):2941–2952, 2013.

[32] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[33] Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*, 2017.

[34] Bertram G Katzung. *Basic and clinical pharmacology*. Mc Graw Hill, 2012.

[35] Marla J Keller, Elizabeth A Kitsis, Shitij Arora, Jen-Ting Chen, Shivani Agarwal, Michael J Ross, Yaron Tomer, and William Southern. Effect of systemic glucocorticoids on mortality or mechanical ventilation in patients with covid-19. *Journal of hospital medicine*, 15(8):489–493, 2020.

[36] Shabaana A Khader, Maziar Divangahi, Willem Hanekom, Philip C Hill, Markus Maeurer, Karen W Makar, Katrin D Mayer-Barber, Musa M Mhlanga, Elisa Nemes, Larry S Schlesinger, et al. Targeting innate immunity for tuberculosis vaccination. *The Journal of clinical investigation*, 129(9):3482–3491, 2020.

[37] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[38] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[39] Philipp Koehler, Matteo Bassetti, Arunaloke Chakrabarti, Sharon CA Chen, Arnaldo Lopes Colombo, Martin Hoenigl, Nikolay Klimko, Cornelia Lass-Flörl, Rita O Oladele, Donald C Vinh, et al. Defining and managing covid-19-associated pulmonary aspergillosis: the 2020 ecmm/isham consensus criteria for research and clinical guidance. *The Lancet Infectious Diseases*, 2020.

[40] William G Kramer, Richard P Lewis, Tyson C Cobb, Wilbur F Forester, James A Visconti, Lee A Wanke, Harold G Boxenbaum, and Richard H Reuning. Pharmacokinetics of digoxin: comparison of a two-and a three-compartment model in man. *Journal of pharmacokinetics and biopharmaceutics*, 2(4):299–312, 1974.

[41] Michael D Lee and Wolf Vanpaemel. Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, 25(1):114–127, 2018.

[42] Dehao Liu and Yan Wang. Multi-fidelity physics-constrained neural network and its application in materials modeling. *Journal of Mechanical Design*, 141(12), 2019.

[43] Carl M Metzler. Usefulness of the two-compartment open model in pharmacokinetics. *Journal of the American Statistical Association*, 66(333):49–53, 1971.

[44] Richard F Mortensen. C-reactive protein, inflammation, and innate immunity. *Immunologic research*, 24(2):163–176, 2001.

[45] Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, and Naren Ramakrishnan. Incorporating prior domain knowledge into deep neural networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE, 2018.

[46] NIH. Therapeutic management of adults with covid-19. https://www.covid19treatmentguidelines.nih.gov/therapeutic-management/. Accessed: 2021-05-26.

[47] Ari Pakman, Yueqi Wang, Catalin Mitelut, JinHyung Lee, and Liam Paninski. Neural clustering processes. In *International Conference on Machine Learning*, pages 7455–7465. PMLR, 2020.

[48] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[49] Eric J Parish and Karthik Duraisamy. A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics*, 305:758–774, 2016.

[50] Lawrence Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

[51] Jordan S Read, Xiaowei Jia, Jared Willard, Alison P Appling, Jacob A Zwart, Samantha K Oliver, Anuj Karpatne, Gretchen JA Hansen, Paul C Hanson, William Watkins, et al. Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, 55(11):9173–9190, 2019.

[52] Eleanor M Rees, Emily S Nightingale, Yalda Jafari, Naomi R Waterlow, Samuel Clifford, Carl AB Pearson, Thibaut Jombart, Simon R Procter, Gwenan M Knight, CMMID Working Group, et al. Covid-19 length of hospital stay: a systematic review and data synthesis. *BMC medicine*, 18(1):1–22, 2020.

[53] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[54] Maria Isabel Ribeiro. Kalman and extended kalman filters: Concept, derivation and properties. *Institute for Systems and Robotics*, 43:46, 2004.

[55] Yulia Rubanova, Ricky TQ Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *33th Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[56] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[57] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[58] Kristof T Schütt, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706.08566*, 2017.

[59] Duncan Shillan, Jonathan AC Sterne, Alan Champneys, and Ben Gibbison. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical Care*, 23(1):1–11, 2019.

[60] Simone MC Spoorenberg, Vera HM Deneer, Jan C Grutters, Astrid E Pulles, GP Voorn, Ger T Rijkers, Willem Jan W Bos, and Ewoudt MW van de Garde. Pharmacokinetics of oral vs. intravenous dexamethasone in patients hospitalized with community-acquired pneumonia. *British journal of clinical pharmacology*, 78(1):78–83, 2014.

[61] William James Stronge. *Impact mechanics*. Cambridge university press, 2018.

[62] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

[63] Judith van Paassen, Jeroen S Vos, Eva M Hoekstra, Katinka MI Neumann, Pauline C Boot, and Sesmu M Arbous. Corticosteroid use in covid-19 patients: a systematic review and meta-analysis on clinical outcomes. *Critical Care*, 24(1):1–22, 2020.

[64] Jian-Xun Wang, Jin-Long Wu, and Heng Xiao. Physics-informed machine learning approach for reconstructing reynolds stress modeling discrepancies based on dns data. *Physical Review Fluids*, 2(3):034603, 2017.

[65] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94, 2014.

[66] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 2020.

[67] Jin-Long Wu, Heng Xiao, and Eric Paterson. Physics-informed machine learning approach for augmenting turbulence models: A comprehensive framework. *Physical Review Fluids*, 3(7):074602, 2018.

[68] Tianfang Xu and Albert J Valocchi. Data-driven methods to improve baseflow prediction of a regional groundwater model. *Computers & Geosciences*, 85:124–136, 2015.

[69] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of thoracic disease*, 12(3):165, 2020.

[70] Kun Yao, John E Herr, David W Toth, Ryker Mckintyre, and John Parkhill. The tensormol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical science*, 9(8):2261–2269, 2018.

[71] Alireza Yazdani, Lu Lu, Maziar Raissi, and George Em Karniadakis. Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLoS computational biology*, 16(11):e1007575, 2020.

[72] Cagatay Yildiz, Markus Heinonen, and Harri Lähdesmäki. Ode2vae: Deep generative second order odes with bayesian neural networks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[73] Leonardo Zepeda-Núñez, Yixiao Chen, Jiefu Zhang, Weile Jia, Linfeng Zhang, and Lin Lin. Deep density: circumventing the kohn-sham equations via symmetry preserving neural networks. *arXiv preprint arXiv:1912.00775*, 2019.

[74] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[75] Liang Zhang, Gang Wang, and Georgios B Giannakis. Real-time power system state estimation via deep unrolled neural networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 907–911. IEEE, 2018.

[76] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. In *International Conference on Learning Representations*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Appendix A.2
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Access to the DDW is regulated. Researcher have to sign an end user license before access to the data is granted. The experiment code will be released after acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 5.1, 5.2, Appendix 5.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Section 5.1, 5.2, Appendix 5.1.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Appendix 5.1
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Section 5.2
   (b) Did you mention the license of the assets? [Yes] Appendix A.6

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Appendix A.6

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Appendix A.6

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]