A Appendix

Summary of Appendices.

Each section can be read independently.

- A.1: The pharmacological model for dexamethasone.
- A.2: Medical importance and impact.
- A.3: Optimization and gradient calculation.
- A.4: Details about the simulation study.
- A.5: Extended related works.
- A.6: Details about the real data experiment.
- A.7: Practical extensions of LHM.

A.1 The pharmacological model for dexamethasone

The expert ODE we used is adapted from [17]. As illustrated in Figure 5, it involves five expert variables z_1 to z_5 (the superscript e is omitted for brevity). The z_1 represents the innate immune response to viral infection (measured in the laboratory using Type I IFNs [43, 1]). The z_2 and z_3 represent the concentration of dexamethasone in lung tissue and plasma respectively. The z_4 represents the viral load and z_5 represents the adaptive immune response (measured in the laboratory using Cytotoxic T cells [1]).

The expert model that describes these variables are developed based on specialized knowledge and laboratory experiments. Firstly, the immune responses and viral replication are modeled as:

$$\dot{z_1} = k_{IR} \cdot z_4 + k_{PF} \cdot z_4 \cdot z_1 - k_O \cdot z_1 + \frac{E_{max} \cdot z_1^{h_P}}{EC_{50}^{h_P} + z_1^{h_P}} - k_{Dex} \cdot z_1 \cdot z_2 \tag{10}$$

$$\dot{z_4} = k_{DP} \cdot z_4 - k_{IIR} \cdot z_4 \cdot z_1 - k_{DC} \cdot z_4 \cdot z_5^{h_C} \tag{11}$$

$$\dot{z_5} = k_1 \cdot z_1 \tag{12}$$

The five terms in the first Equation for z_1 captures the initial immune reaction to the virus, the physiological positive feedback, the immune cell mortality, the pathological positive feedback, and the effect of dexamethasone [17]. The second equation of z_4 captures the viral replication, and the effect of innate and adaptive immune systems on the virus. The last equation captures the adaptive immune response triggered by the innate immune response [1]. The unknown coefficients k_{IR}, k_{PF}, k_O, E_{max}, h_P, k_{Dex}, k_{DP}, k_{IIR}, k_{DC}, h_C are positive real numbers.

The concentration of dexamethasone (z_2, z_3) is described by a standard two-compartmental pharmacokinetics model [52, 48]:

$$\dot{z_2} = -k_2 \cdot z_2 + k_3 \cdot z_3 \tag{13}$$

$$\dot{z_3} = -k_3 \cdot z_3 \tag{14}$$

The coefficients k_2, k_3 are positive real numbers. In the literature, it is often assumed for simplicity that the treatment is given at time $t = 0$, and the initial condition of the plasma concentration $z_3(0)$ corresponds to the dosage [22]. Since the plasma concentration z_3 decays exponentially over time, we can equivalently express it as a sum of exponentials: $z_3(t) = \sum_i d_i \cdot I(t > t_i) \cdot \exp(k_3(t_i - t))$ when dosages d_i are given at time t_i, $i \geq 1$. The function $I(\cdot)$ is an indicator function.

Prior distribution for real-data experiment. The initial condition $\mathbf{z}(0)$ corresponds to the patient state at the time of ICU admission. Since dexamethasone is generally administered *during* the ICU stay [57], its concentration at admission should be very close to zero. Hence we use an exponential distribution with rate $\lambda = 100$ as the prior of $z_2(0)$ and $z_3(0)$. On the other hand, the immune response and viral load may vary across patients greatly. To allows for more heterogeneity, we use an exponential distribution with rate $\lambda = 0.1$ as the prior of $z_1(0)$, $z_4(0)$, and $z_5(0)$. The exponential distribution also reflects the positivity of the expert variables because it has a positive support.

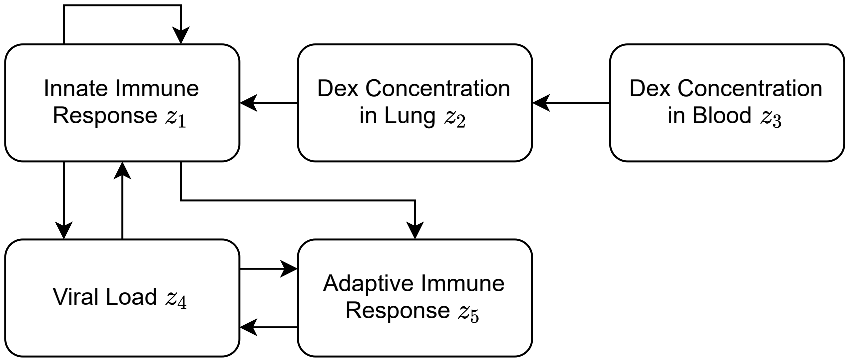
Innate Immune Response z_1
Dex Concentration in Lung z_2
Dex Concentration in Blood z_3
Viral Load z_4
Adaptive Immune Response z_5

Figure 5: The expert variables and their temporal interactions as described by the expert ODE.

A.2 Medical importance and impact

A.2.1 Medical importance of LHM

Integration of machine learning (ML) and pathophysiology is a major challenge in adopting ML models in a clinical setting. While clinicians seek to understand the mechanisms that drive disease progression for prognosis and treatment allocation, machine learning models do not currently provide such disease dynamics. With these dynamics, however, model results would translate to clinically interpretable concepts, would resonate with clinicians, and could then support clinical decision making.

In addition, disease dynamics are indispensable for the clinical interpretation of predictive modeling results. Predictive modeling has taken flight in the medical field, but many models are left stranded because clinicians can solely rely on feature importance and no mechanistic interpretation of the results. Such an interpretation, however, will increase clinicians' trust in these models and expedite their use in clinical practice.

Lastly, the relationship between fundamental and clinical research may yield novel hypotheses and foster subsequent research. There is a gap between benchwork and the bedside. Bridging this gap with ML and interpretable models could reveal novel relationships and could inspire research both ways. Overall, ML with a mechanistic interpretation can provide the next big step in medical data science and can help bring these models to the bedside.

A.2.2 Dexamethasone in COVID-19

Coronavirus disease 2019 (COVID-19) was an unknown disease to intensive care clinicians worldwide. Both the natural course of the disease as well as optimal treatment were unknown throughout the onset of the pandemic. Since inflammatory organ injury appeared to play an important role in the pathophysiology of COVID-19, glucocorticoids were proposed to mitigate the damaging effects of the immune system [54]. In particular, Dexamethasone treatment has been shown to reduce mortality in patients on invasive mechanical ventilation or oxygen alone in the RECOVERY trial [32]. Moreover, the CoDEX trial demonstrated an increase in the number of ventilator free days with Dexamethasone treatment in moderate to severe COVID-19 acute respiratory distress syndrome (ARDS)[80]. As a result, COVID-19 treatment guidelines recommend Dexamethasone treatment in these settings [57].

Although beneficial effects have been shown of Dexamethasone on a group level, individual response to treatment remains unknown. Knowing this response would help clinicians to anticipate complications, to improve individualized prognosis, and potentially determine beneficial treatments in these patients. Moreover, clinicians could identify patients in which Dexamethasone has a desired effect and in which patient it may not. For example, in the case of coinfection, Dexamethasone may be discontinued in selected patients. Lastly, these models can identify novel mechanistic pathways in COVID-19 patients that can inspire both fundamental and clinical research. Taken together, individualized disease progression in response to Dexamethasone treatment would bring about a large step forward in COVID-19 research.

A.2.3 Potential negative impact

Any decision support system could be used negatively if the user intentionally chooses to worsen the outcome. This is very unlikely in our case because the intended users of LHM are clinicians.

A.3 Optimization and gradient calculation

We optimize ELBO by stochastic gradient ascent using the ADAM optimizer [45]. The gradient calculation is enabled by the following two methods.

Reparameterization. To evaluate the ELBO, we need to take samples from the variational distribution $\mathbb{Q}_\phi$. Here we use the Gaussian reparameterization in all sampling steps to obtain the gradients for the encoder [46].

Gradient for ODE. We use the torchdiffeq library to calculate the gradient with respect to the ODE solutions [13]. A variety of ODE solvers are available in the library, we used the adams solver, which is an adaptive step size solver.

A.4 Simulation study

A.4.1 Data generation details

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 measurable physiological variables $\mathbf{x}$. For the pharmacological model, we used the model provided in Appendix A.1, which involves five inter-related variables. We set the coefficients $h_P = h_C = 2$ and the rest to be one.

For each dataset, we set the number of un-modeled states $\mathbf{z}^m$ according to the number of physiological variables 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 un-modeled states.) The un-modeled states $\mathbf{z}^m$ are governed by a nonlinear ODE

$$\dot{\mathbf{z}}_i^m = \tanh(\mathbf{W}_1 \mathbf{z}_i^m + \mathbf{W}_2 \mathbf{z}_i^e),$$

with the coefficient matrices $\mathbf{W}_1 \in \mathbb{R}^{M \times M}$, $\mathbf{W}_2 \in \mathbb{R}^{M \times E}$. For each dataset, we sampled the entries in these matrices independently from $N(0, 1)$.

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 $\lambda = 100$ (this distribution is also given to the algorithms as the prior distribution). We consider a time horizon of $T = 14$ days; this is the median length of stay in hospital for Covid-19 patients [66].

Each patient i will receive a one-time dexamethasone treatment with dosage d_i at some time s_i, where $d_i \sim \text{uniform}[0, 10]$ mg and $s_i \sim \text{uniform}[0, T]$.

The true physiological variables are generated by

$$\mathbf{x}_i = \mathbf{W}_3 \mathbf{z}_i + \mathbf{W}_4 \mathbf{a}_i,$$

with the coefficient matrices $\mathbf{W}_3 \in \mathbb{R}^{X \times (M+E)}$, $\mathbf{W}_4 \in \mathbb{R}^{X \times 1}$. For each dataset, each element in these matrices was drawn independently from $N(0, 1)$ and then multiplied by a Bernoulli variable with $p = 0.5$, so that approximately half of the elements in each of these matrices $\mathbf{W}_3$, $\mathbf{W}_4$ were 0. (We did this in order to reflect the idea that each physiological variable is only related to some of the latent variables.) The measurements are generated by

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

with the measurement noise $\epsilon_{it} \sim N(0, \sigma)$ for $\sigma = 0.2, 0.4$ or 0.8; Equation (1). We first simulate all the daily measurements at $t = 1, 2, \ldots, T$, and then randomly remove measurements with probability 0.5; this represents the fact that measurements of made irregularly.

A.4.2 Hyper-parameter settings

As a reminder, the number of measured clinical variables is D, the number of expert variables is E, the number of ML latent variables is M. The sample size is N_0

The following is the hyperparameter setting used in the simulation study:

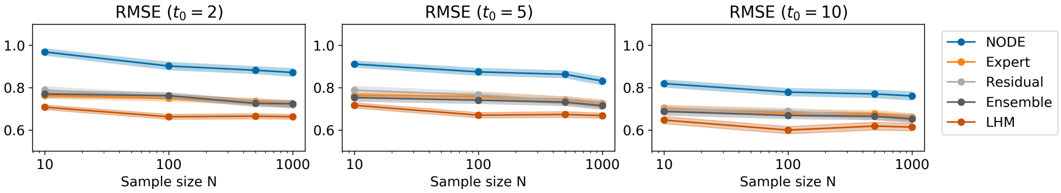

Figure 6: **Simulation results under different lengths of observed history** t_0. Prediction accuracy on future measurements $\mathcal{Y}[t_0 : T]$ given the observed history $\mathcal{Y}[0 : t_0]$ as measured by RMSE. The shaded areas represent 95% confidence intervals.

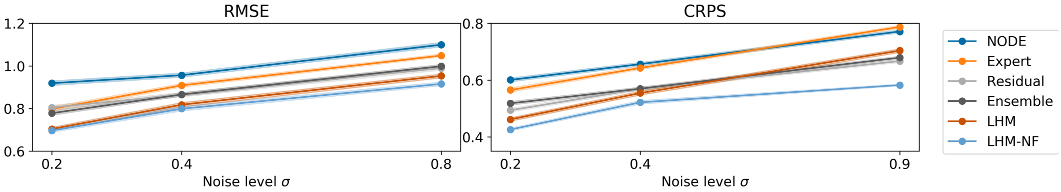

Figure 7: **Simulation results under different levels of measurement noise** σ. Prediction accuracy on future measurements $\mathcal{Y}[5 : T]$ given the observed history $\mathcal{Y}[0 : 5]$ as measured by RMSE and CRPS. The shaded areas represent 95% confidence intervals.

1. Learning rate: 0.01
2. Batch size: $\min(50, N_0)$
3. Early stopping tolerance: 10 epochs
4. Max iteration: 400
5. Number of latent variables in NODE: $Z = E + M$, i.e. the true value. (additional settings $E + M + 4$ and $E + M + 9$ are reported in the sensitivity analysis)
6. Number of ML latent variables in LHM: M, i.e. the true value.
7. Latent dimensionality in Encoder: $2D$
8. Number of layers in NODE: 2
9. ODE Solver: adams
10. ODE rtol: 1E-7 (library default)
11. ODE atol: 1E-8 (library default)

A.4.3 Performance under different lengths of observed history

As a reminder, we use the historical measurements $\mathcal{Y}[0 : t_0]$ up to some time t_0 to *predict* the future measurements $\mathcal{Y}[t_0 : T]$. To evaluate the performance under different lengths of observed history, we set $t_0 = 2, 5$ or 10 days and use the default setting $\sigma = 0.2$ and $M = 2$. The results are presented in Figure 6, where each panel corresponds to a different t_0. As expected, the predictive performance improved when longer observed history is used to make prediction. LHM outperforms the benchmarks for all t_0's we study.

A.4.4 Performance under different levels of measurement noise

In Figure 7, we show the model performance under different levels of measurement noise $\sigma = 0.2, 0.4, 0.8$ in a typical simulation with sample size $N_0 = 100$ and $M = 2$ un-modeled latent variables $\mathbf{z}^m$. In addition to the benchmarks introduced in Section 5.1, we compared with the LHM using normalizing flow as the variational distribution (LHM-NF) (detailed in the next section). Both LHM and LHM-NF outperform other benchmarks in RMSE. LHM-NF achieves the best CRPS with a fairly big improvement from the second best method (LHM). The result shows LHM's robustness to increased measurement noise.

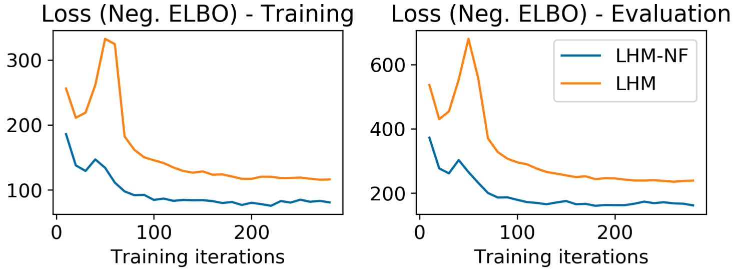

Figure 8: **Comparison between the standard LHM and the version with normalizing flow (LHM-NF)**. Loss on training and evaluation datasets are plotted over training iterations. The loss is the negative ELBO.

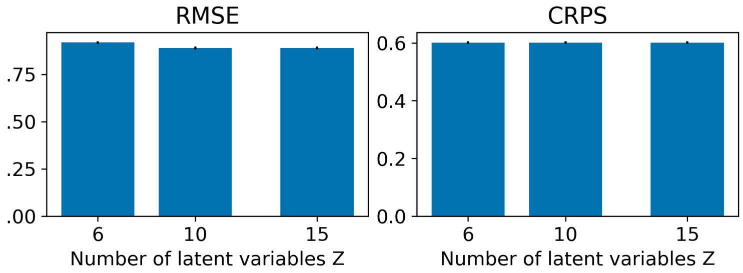

Figure 9: **NODE's performance under different numbers of latent variables** Z. The data are generated from six *true* latent variables. Prediction accuracy on future measurements $\mathcal{Y}[5:T]$ given the observed history $\mathcal{Y}[0:5]$ as measured by RMSE and CRPS.

A.4.5 Performance gain with Normalizing Flows

To ensure a fair comparison with existing methods, we use the diagonal Gaussian distribution in LHM as the variational distribution. However, diagonal Gaussian is a restrictive approximation because it does not capture any correlation structure between the latent variables.

Here we study if using a more flexible distribution will lead to further performance gain. We adopt the planar normalizing flow proposed in [67] with the number of flows set to 4. As is standard in the literature, we amortize the initial conditions $\mathbf{z}(0)$ as well as the flow parameters $\mathbf{u}$, $\mathbf{w}$ and $\mathbf{b}$. The following shows a typical simulation with $N_0 = 100$, $\sigma = 0.4$ and $M = 2$.

Figure 8 tracks the loss function (negative ELBO) during training on the training and the evaluation data respectively. As we expected, the version with normalizing flow (LHM-NF) consistently achieves smaller loss on the training data due to the increased flexibility of the variational distribution. The improvement persists when we turn to the evaluation data, and eventually translates into the performance gain illustrated in Figure 7. This suggests that using a more flexible variational distribution (e.g. normalizing flow) tends to improve accuracy as well as the uncertainty estimation.

A.4.6 Performance of NODE is not sensitive to adding more latent variables

In the simulations reported above, we set the number of latent variables in NODE to be the true value, i.e. $Z = M + E$. In practice, Z is a hyper-parameter that we do not know a priori. Here we study if the performance is sensitive to the exact choice of Z. We consider a setting where the data is generated from 6 latent variables (including both $\mathbf{z}^m$ and $\mathbf{z}^e$) and we vary $Z = 6, 10, 15$. We present the results in a typical simulation setting with $N_0 = 100$, $\sigma = 0.2$. As we show in Figure 9, the predictive performance does not significant change even when Z is more than doubled. Note that similar findings have been reported in prior research [21]. This supports our choice of setting $Z = M + E$ by default.

Table 2: Different methods to create hybrid ML models. We consider a static prediction problem with covariates $\mathbf{x} \in \mathbb{R}^D$ and target outcome $\mathbf{y} \in \mathbb{R}^K$ (notations differ from the rest of the paper). The $\mathbf{r} := \mathbf{y} - \hat{\mathbf{y}}$ denotes the residuals.

Method	Example	Expert model	ML model	Final output
Residual Model	[51, 82]	$\hat{\mathbf{y}} = f^e(\mathbf{x})$	$\hat{\mathbf{r}} = f^m(\mathbf{x})$	$\hat{\mathbf{y}} + \hat{\mathbf{r}}$
Ensemble	[89, 87]	$\hat{\mathbf{y}}_1 = f^e(\mathbf{x})$	$\hat{\mathbf{y}}_2 = f^m(\mathbf{x})$	$w_1\hat{\mathbf{y}}_1 + w_2\hat{\mathbf{y}}_2$
Feature Extraction	[40]	$\mathbf{z}^e = f^e(\mathbf{x})$	$\hat{\mathbf{y}} = f^m(\mathbf{z}^e)$	$\hat{\mathbf{y}}$
LHM	This work	Eq. 3	Eq. 4, 5	Eq. 4

A.4.7 Computational resources

The simulations were performed on a server with a Intel(R) Core(TM) i5-8600K CPU @ 3.60GHz and a Nvidia(R) GeForce(TM) RTX 2080 Ti GPU. All individual simulations were finished within 3 hours.

A.5 Extended related works

A.5.1 Hybrid models

Table 2 categorizes various hybrid modeling frameworks in terms the kind of expert model and the kind of machine learning that are used. For illustrative purposes, we consider a static prediction problem with measurements (covariates) $\mathbf{x} \in \mathbb{R}^D$ and target outcome $\mathbf{y} \in \mathbb{R}^K$.

A.5.2 Other research areas that involve ML and expert ODEs

There are several research areas that involve ML and expert ODEs but they are *unrelated* to hybrid model or LHM. We briefly describe them for clarification and completeness.

Reduced-Order Models. The expert model may involve a large number of variables, but not all of them are important to the system dynamics (e.g. in large, high-fidelity models of fluid dynamics [49]). Reduced-Order Models (ROMs) are compact representations of the more complex models [84]. They are often constructed using dimensionality reduction to retain only the most important dynamical characteristics of the original model. ROMs often achieve better estimation efficiency and lower the computational cost. Recently, ML has been applied to ROMs and achieved promising results [12, 86].

In ROMs, we start with an expert model that is *over-complete* and contains redundant variables. In contrast, in LHM, we are given a pharmacological model that is *incomplete*, i.e. it cannot fully explain the high-dimensional clinical measurements or provide the link between expert variables and the measurements. Hence, LHM is essentially solving the *opposite* problem of ROMs as we are introducing additional machine-learned latent variables into the system.

Using ML to solve expert equations. Some expert models involves ODEs or PDEs that are computationally challenging to solve (e.g. the quantum many-body problem [11]). ML has been used to speed up the solution process by making various approximations [35, 76]. However, the pharmacological models are generally well-behaved and the standard ODE solvers are able to find the solutions efficiently.

Learning unknown ODEs from data Step-wise regression is a general framework to discover unknown ODEs from data. It applies symbolic or sparse regression to the the observed time derivatives. When these time derivatives are not observed, they are first estimated from the (frequently-sampled) observations (e.g. by finite difference method) [9, 10, 71]. This approach is not applicable to our setting because the time derivatives of the expert variables $\dot{\mathbf{z}}^m$ are not observed or can be easily estimated from the data.

In addition to neural ODEs, Gaussian Processes (GP) have also been used to approximate unknown governing equations [5, 72]. However, most existing works focus on the discrete-time setting or use fixed step ODE solvers.

A.5.3 Using Pharmacology/Biology models in ML

Several other works have proposed to integrate pharmacological models into machine learning. But the problem settings they considered and the approach they took is different from LHM.

[38] introduces a pharmacological model (the log-cell kill model) to modulate the state transition dynamics of a state-space model. Their work considers discrete-time dynamical systems rather than the continuous-time systems we focus on. The authors recognize that the existing log-cell kill models are inadequate to model the disease dynamics (e.g. failure to capture relapses). To address this shortcoming, the authors designed a new set of expert equations to allow for more complex dynamics before integrating them with ML. Hence, this approach requires a deep understanding of the expert model, and a fair amount of mathematical knowledge and manual work to modify the expert model. Furthermore, this modification process has to be repeated for a different expert model. In contrast, LHM learns the missing dynamics by introducing the ML latent variables $\mathbf{z}^m$ and neural ODEs f^m.

[90] considers a problem with more expert variables than observable physiological variables, which is opposite to the setting we consider.[4] The problem setting is similar to the reduced-order models discussed above. The authors use a neural network with time t as input and outputs the system status at that time. In contrast, LHM uses neural ODEs to model the time derivatives and obtains the system status at time t by solving the ODEs. Finally, the authors evaluate the gradient of the neural network with respect to t by automatic differentiation and introduce an additional loss function to ensure the network gradient matches the expert ODEs. LHM does not involve any heuristic modification on the loss function and follows the standard practice in Bayesian inference.

A.5.4 Causal treatment effect estimation

Causal effect estimation is a diverse field with many different (and often incompatible) theories, notions and methods. Here we compare the approach taken by LHM with other well-known approaches in the literature.

As discussed in Section 4, LHM predicts the future health status given treatments using the governing equations (ODEs). This corresponds to the mechanistic (or physical) notion of the causality, which is recognized as the "gold standard" for modeling natural phenomena by [73].

The potential outcome framework widely used in Statistics is based on a different notion of causality [70]. It makes assumptions about the statistical properties of the unobservable potential outcomes (e.g. independence) to make inference about the (conditional) average treatment effect. Here, the focus is not on using or discovering the underlying governing equations, but on leveraging the statistical associations between the observed and the potential outcomes. Unlike the mechanistic framework, the potential outcome framework does not require the system to be observed over time, making it suitable for problems involving only static variables.

The causal graphical models [62] use another notion of causality. Causal graphical models describe the causal structure between variables as a graph (typically a directed acyclic graph, DAG). Various identification strategies have been developed to infer the causal effect given the graph (e.g. the backdoor criterion [63]). A closely related framework is the structural causal model [63], where a set of structural equations are given in addition to the causal graph. Typically, the structural equations are standard equations that link the (static or discretely sampled) variables, but they are not ODEs that describe the continuous-time dynamics. Some existing works attempt to establish the connection between the mechanistic ODEs with the structural equations [53].

A.6 Real data experiment

A.6.1 List of clinical variables

We use the measurements of the following temporal physiological variables. These variables are chosen by our clinical collaborators and reflect the information accessible and important to a clinician when deciding the treatment plan. They include vital signals, lung mechanics, and the biomarkers measured in blood tests.

[4] The Systems Biology models considered in [90] usually involve a large number of expert variables. This is not the case in the pharmacological models we consider.

1. P/F ratio
2. PEEP
3. SOFA
4. Temperature
5. Arterial blood pressure
6. Heart Rate
7. Bilirubin
8. Thrombocytes
9. Leukocytes
10. Creatinine
11. C Reactive Protein
12. Arterial lactate
13. Creatine kinase
14. Glucose
15. Alanine transaminase
16. Aspartate transaminase
17. Prone positioning
18. Tidal volume
19. Driving pressure
20. FiO2
21. Lung compliance (static)
22. Respiratory rate
23. Pressure above PEEP
24. Arterial PaCO2
25. Arterial PH
26. PaCO2 (unspecified)
27. PH (unspecified)

We used the following static covariates:

1. Age
2. Sex
3. Body Mass Index
4. Comorbidity: cirrhosis
5. Comorbidity: chronic dialysis
6. Comorbidity: chronic renal insufficiency
7. Comorbidity: diabetes
8. Comorbidity: cardiovascular insufficiency
9. Comorbidity: copd
10. Comorbidity: respiratory insufficiency
11. Comorbidity: immunodeficiency

A.6.2 Eligibility criterion

We selected all patients in DDW who stayed in the ICU for more than 2 days and less than 31 days (2097 out of 3464). Patients with a very short length of stay will not give us enough data points for training or evaluation.

Table 3: Prediction accuracy (RMSE) on $\mathcal{Y}[t_0 : t_0 + H]$ over different time horizons H (hours). The standard deviations are shown in the brackets.

Method \H=	6	12	24	72
Expert	0.734 (0.99)	0.724 (1.00)	0.713 (0.03)	0.993 (0.03)
Residual	0.555 (0.98)	0.575 (1.08)	0.607 (0.04)	0.983 (0.05)
Ensemble	0.556 (0.71)	0.573 (0.73)	0.599 (0.04)	**0.713 (0.05)**
NODE	0.661 (1.00)	0.654 (1.00)	0.650 (0.02)	0.996 (0.02)
ODE2VAE	0.627 (1.11)	0.616 (1.09)	0.619 (0.02)	1.113 (0.01)
GRU-ODE	0.549 (0.71)	0.571 (0.72)	0.601 (0.04)	**0.711 (0.05)**
Time LSTM	0.610 (0.81)	0.620 (0.82)	0.631 (0.04)	0.807 (0.05)
LHM	**0.517 (0.72)**	**0.511 (0.73)**	**0.511 (0.03)**	**0.691 (0.03)**

A.6.3 Hyper-parameter settings

The following is the hyperparameter setting used in the real-data study. They are decided based on a pilot study.

1. Learning rate: 0.01
2. Batch size: 100
3. Early stopping tolerance: 10 epochs
4. Max iteration: 1500
5. Number of latent variables in NODE: 20
6. Number of ML latent variables in LHM: 15 (this is to ensure the total number of latent variables is the same as NODE).
7. Latent dimensionality in Encoder: $1.2D$
8. Number of layers in NODE: 2
9. ODE Solver: adams
10. ODE rtol: 1E-7 (library default)
11. ODE atol: 1E-8 (library default)

A.6.4 Accuracy over different time horizons

Table 3 shows the performance over different prediction horizons H given $N_0 = 1000$ training samples. LHM achieves the best or equally the best performance in all cases.

A.6.5 License and anonymity

Access to the DDW is regulated. We have signed an end user license before access to the data was granted. All data were pseudonymized in DDW.

A.7 Practical extensions

Incorporating static covariates. Static covariates such as the demographics often impact disease progression. We can easily incorporate these variables in LHM by treating them as time-constant "treatments". This will allow the static covariates to impact the latent dynamics as well as the mapping between the latent and physiological variables (Equation 3 to 5).

Informative sampling. It is well known that the sampling frequency may carry information about the variables being measured (e.g. clinicians tend to take measurements more often if a patient is critically ill) [3]. One approach to incorporate informative sampling is to explicitly model it as a marked point process [69]. Another popular approach is to concatenate the measurements $\mathbf{x}$ with the masking vector that indicates which variable is measured, and train the model on the extended measurement vector [44]. Both approaches are compatible with LHM.

Correcting model mis-specification. *Equation Replacement* is a general approach that applies to any *misspecified* expert model and it can be combined with all the methods discussed above, and to LHM [34, 61, 94]. In this approach, one first identifies which equations in the expert model are misspecified, and then replaces these by flexible function approximators (such as neural networks), that will approximate the true equation after training. Equation replacement only attempts to correct the misspecifications in the original model, but does not introduce any new variables.

Efficient online inference. The inference method presented in Section 3.4 requires to re-process the entire history each time a new measurement is made. Instead, it may be desirable to incrementally update the posterior of $\mathbf{z}_i(0)$ based on the most recent measurement only. Fortunately, online Bayesian update (also known as Bayesian Filtering) is a well studied problem with many proven solutions (e.g. Kalman filter and extensions [68]). These inference methods can be applied when the efficiency of online inference is of concern.

Improving encoder architecture. For a fair comparison with related works, we used the reversed time-aware LSTM encoder proposed in [13]. Essentially, it is a LSTM with the observation time as an additional input channel and running backward through time. To further improve performance, one may explore other architectures. Essentially, any architecture that takes irregularly sampled data as input is applicable. Examples include the Neural Controlled Ordinary Differential Equation [44] and the Neural ODE Processes [58].

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] Cedric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, pages 1–16. PMLR, 2007.

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

[7] 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.

[8] 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.

[9] Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.

[10] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[11] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[12] Gang Chen, Yingtao Zuo, Jian Sun, and Yueming Li. Support-vector-machine-based reduced-order model for limit cycle oscillation prediction of nonlinear aeroelastic system. *Mathematical problems in engineering*, 2012, 2012.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

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

[32] RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19. *New England Journal of Medicine*, 384(8):693–704, 2021.

[33] 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.

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

[35] Jiequn Han, Arnulf Jentzen, and E Weinan. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[36] 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.

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

[38] Zeshan Hussain, Rahul G Krishnan, and David Sontag. Neural pharmacodynamic state space modeling. *arXiv preprint arXiv:2102.11218*, 2021.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[47] 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.

[48] 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.

[49] Toni Lassila, Andrea Manzoni, Alfio Quarteroni, and Gianluigi Rozza. Model order reduction in fluid dynamics: challenges and perspectives. *Reduced Order Methods for modeling and computational reduction*, pages 235–273, 2014.

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

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

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

[53] Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. *arXiv preprint arXiv:1304.7920*, 2013.

[54] John B Moore and Carl H June. Cytokine release syndrome in severe covid-19. *Science*, 368(6490):473–474, 2020.

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

[56] 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.

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

[58] Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, and Pietro Lio. Neural ode processes. *ICLR*, 2021.

[59] 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.

[60] 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.

[61] 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.

[62] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[63] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

[65] 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.

[66] 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.

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

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

[69] 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.

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

[71] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

[72] Andreas Ruttor, Philipp Batz, and Manfred Opper. Approximate gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems*, pages 2040–2048. Citeseer, 2013.

[73] 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.

[74] 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.

[75] 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.

[76] Justin Sirignano and Konstantinos Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364, 2018.

[77] 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.

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

[79] 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.

[80] Bruno M Tomazini, Israel S Maia, Alexandre B Cavalcanti, Otavio Berwanger, Regis G Rosa, Viviane C Veiga, Alvaro Avezum, Renato D Lopes, Flavia R Bueno, Maria Vitoria AO Silva, et al. Effect of dexamethasone on days alive and ventilator-free in patients with moderate or severe acute respiratory distress syndrome and covid-19: the codex randomized clinical trial. *Jama*, 324(13):1307–1316, 2020.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] D Xiao, CE Heaney, L Mottet, F Fang, W Lin, IM Navon, Y Guo, OK Matar, AG Robins, and CC Pain. A reduced order model for turbulent flows in the urban environment using machine learning. *Building and Environment*, 148:323–337, 2019.

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

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

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