Predicting the impact of treatments over time with uncertainty aware neural differential equations.

Edward De Brouwer, [1] **Javier González Hernández,** [2] **Stephanie Hyland,** [2]

[1] ESAT-STADIUS, KU Leuven, Belgium
[2] Microsoft Research, Cambridge, United Kingdom

Abstract

Predicting the impact of treatments from observational data only still represents a major challenge despite recent significant advances in time series modeling. Treatment assignments are usually correlated with the predictors of the response, resulting in a lack of data support for counterfactual predictions and therefore in poor quality estimates. Developments in causal inference have lead to methods addressing this confounding by requiring a minimum level of overlap. However, overlap is difficult to assess and usually not satisfied in practice. In this work, we propose Counterfactual ODE (CF-ODE), a novel method to predict the impact of treatments continuously over time using Neural Ordinary Differential Equations equipped with uncertainty estimates. This allows to specifically assess which treatment outcomes can be reliably predicted. We demonstrate over several longitudinal datasets that CF-ODE provides more accurate predictions and more reliable uncertainty estimates than previously available methods.

Introduction

Our main motivation in this work, consists of predicting the individual impact of medical interventions on patients based on available longitudinal information. As clinical trials are long and expensive, the ability to predict individual treatment effects from observational data alone is particularly attractive.

Based on observational clinical data, where treatments have been assigned to patients based on clinicians' judgement, we aim to predict the longitudinal progression of the disease course of individual patients depending on the intended treatment scenario. However, confounding, defined as the dependence of the treatment assignments on predictors of the future outcomes, can lead to biased estimates if not properly addressed (Bica et al. 2021). In this case, the cohorts of treated and non-treated patients are expected to be statistically different in the distribution of the predictors of the outcomes (Jesson et al. 2020). Crucially, this implies a distribution shift between the *factual* (the observed treatment-outcome pairs) and the *counterfactual* (when the treatment assignment is different than the one observed) distributions. A common way to address this distribution shift is to require a significant level of positive overlap between treated and non-treated distributions, which is both difficult to test and seldom realized in practice (Oberst et al. 2020; D'Amour et al. 2021).

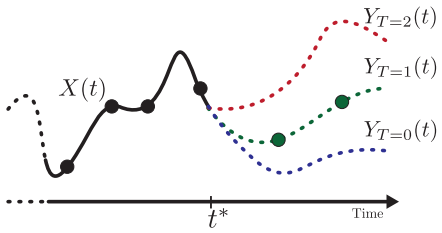


Figure 1: Based on available trajectory information $X(t)$, we aim at predicting in continuous time the potential outcomes of applying treatment regime T at time t^* (dotted lines). A single outcome is available for each instance in the dataset (solid green dots).

In contrast, rather than trying to improve the estimation of individual treatment effects (ITE) with a restrictive set of assumptions, we propose to learn a model to predict the different potential treatment outcomes over time with embedded uncertainties, reflecting the lack of data support in some regions of the predictors-treatment space (Jesson et al. 2020). We show that uncertainties in the prediction are crucial when it comes to the implementation of a treatment assignment recommendation system in (clinical) practice, where trust is of paramount importance. It informs the decision maker as to which specific treatment outcomes can be reliably estimated.

We propose a latent neural ODE model equipped with uncertainty estimates to predict the individual impact of treatment assignments. We show that uncertainties in the ODE parameters can be encoded by reformulating the problem as a latent stochastic differential equation (SDE) model, relying on recently developed techniques (Li et al. 2020; Xu et al. 2021). This formulation allows a flexible and efficient parametrization of the weights posterior probability distribution. Using datasets from cardiovascular system modeling and pharmacodynamics, we show that the uncertainties estimates reliably encode the error on the treatment effect estimator and provide a efficient way to detect patients for which an accurate estimation of treatment effect can be given.

Problem Setup

Temporal potential outcomes framework

We aim at predicting longitudinal treatment outcomes based on historical observational data. We consider a set of N multivariate time series $\mathcal{X} = \{X_i(t) \in \mathbb{R}^D; i \in 1, ..., N; t \in \mathbf{t}^i\}$, sampled at arbitrary and potentially irregular time points $\mathbf{t}^i = \{t_0, ..., t_{k_i}\}$ where k_i stands for the number of observations of time series $X_i(t)$. We refer to $\mathcal{S}_{t'}(X_i) = \{X_i(t) : t < t'\}$ as the set of available observations of $X_i(t)$ before t'. Some of those time series include a treatment assignment $T_i^* \in \{0, 1\}$, where the star indicates the observed (*i.e* factual) treatment assignment. In this case, we write the time of treatment assignment as t_i^*.

We formulate our problem within the potential outcomes framework (Rubin 1974). For each time series, we want to predict the potential outcomes $Y_{i,T}(t) : t \geq t_i^*$, the time series that would be observed when treatment T is applied at time $t = t_i^*$. However, for each time series, only one potential outcome is observed, corresponding to the treatment that was actually given. In this case ($T = T_i^*$), $Y_{i,T}(t) = X_i(t) \quad \forall t > t_i^*$. In our motivational patient trajectories example, the data would consist of N patients for whom we observe D-dimensional time series. A treatment T_i^* is then given at some time t_i^* and the resulting treatment effect is observed over time ($Y_{i,T_i^*}(t)$). Based on this information, we want to be able to predict all potential outcomes on a new patient based on the available longitudinal data before treatment assignment ($\mathcal{S}_{t_i^*}(X_i)$). An example of time series (with $D = 1$) and 3 potential treatment regimes is shown on Figure 1. For sake of readability, we drop the i subscript in the remaining of this text as the different time series are considered to be independent.

Dynamical System

We consider the available temporal data $X(t)$ is modelled by a latent continuous time process $h(t)$ whose dynamics are characterized by an ordinary differential equation (ODE) and an emmision function $g(\cdot)$ as defined in Equation 1.

The impact of interventions on the dynamics are assumed to come from an external exogenous continuous temporal input $u_T(t)$ where T indexes the treatment assignment.This leads to our first assumption about the data generating process.

Assumption 0.1. All observations $X(t)$ and potential outcomes $Y(t)$ are driven by a common dynamical system characterized by an unknown ODE:

$$\begin{aligned} \frac{dh_T(t)}{dt} &= f(h_T(t), u_T(t - t^*)) \\ X(t) &\sim g(h_{T^*}(t)) \\ Y_T(t) &\sim g(h_T(t)), \forall t \geq t^* \end{aligned} \tag{1}$$

where $f(\cdot)$, $g(\cdot)$ and $u_T(\cdot)$ are unknown functions. Note that before the treatment assignment ($t < t^*$), all potential outcomes trivially coincide with the observed process $X(t)$.

Treatment assignment and confounding

Without loss of generality, we consider that the treatment assignment at time $t = t^*$ depends on the latent process $h(t^*)$, giving the following propensity model:

Assumption 0.2. Conditioned on a treatment assignment time t^*, the probability of treatment assignment is generated as $T(t^*) \sim \tau(h(t^*))$.

As the treatment assignment depends on the *unobserved* latent process $h(t)$, we require another assumption to ensure that we can control for all possible confounders, corresponding to the classical strong ignorability condition:

Assumption 0.3. There exists a map $\phi(\cdot)$ between the set of available measurements at the time of treatment ($\mathcal{S}_{t^*}(X)$) and the latent process at time of treatment ($h(t^*)$) such that, for all observed times series $X(t)$: $\phi(\mathcal{S}_{t^*}(X)) = h(t^*)$.

Overlap of treated distributions

On top of the strong ignorability assumption, positive overlap between the distributions of the different treated groups is usually required in the potential outcomes framework. Unfortunately, this assumption is rarely satisfied, limiting the applicability of methods relying on it (Oberst et al. 2020). Therefore, in this work, we do not assume positive overlap between the various treated groups but rather model the lack of overlap by making our model uncertainty aware.

Model

Characterizing the lack of overlap with uncertainties

As stated above, we do not explicitly assume sufficient overlap. Nevertheless, the absence of overlap can be characterized as it would result in more *epistemic* uncertainty of the estimators in the region of poor coverage (Jesson et al. 2020). In particular, for the same observed time series $X(t)$, the uncertainty about the predicted potential outcomes can vary based on the treatment assignment, depending of how often particular treatments are observed for similar time series in the dataset. The resulting uncertainty can then be used to inform where reliable treatment effect predictions can be made. Indeed, in the limit of a dataset where there is no overlap between treated on non-treated, nothing can be expected from counterfactual prediction as literally no datapoints are available to train the counterfactual model. Therefore, we want to equip our model with epistemic uncertainty estimates, to reflect the lack of data coverage in specific observation-treatment regions.

Learning the dynamics with Neural ODEs

Building upon the formulation of the data generating process (Eq. 1), we propose to model the dynamics of the observed time series in two steps. We first encode the available measurements up until the treatment assignment time (t^*), aiming to recover the hidden process $h(t^*)$. We then integrate forward the hidden process $h(t)$ with the treatment-specific exogenous inputs ($u(t)$) using a Neural ODE to predict the potential outcomes $Y(t)$.

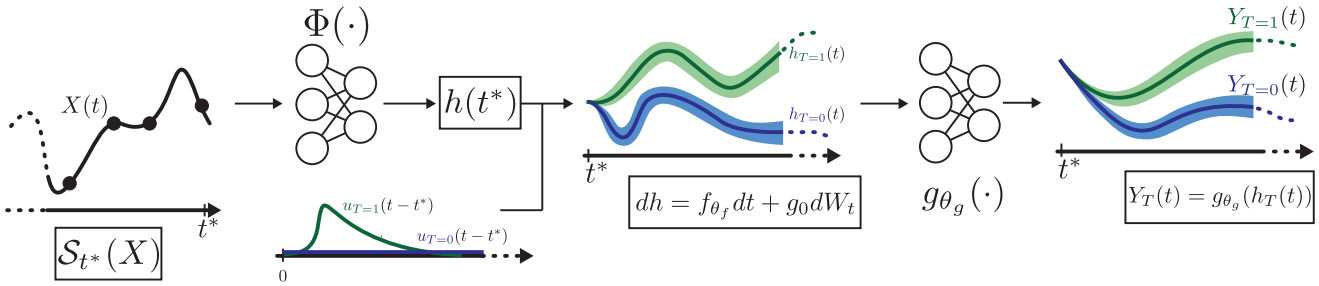


Figure 2: Overview of the CF-ODE architecture. Available temporal information ($\mathcal{S}_t(X)$) is mapped to the latent vector $h(t^*) = \Phi(\mathcal{S}_t^*(X))$. The latent process $h(t)$ is then integrated over time using a learnt Neural SDE. The impact of treatment is taken into account in the dynamics by learning an exogeneous treatment process $u_T(t)$. Finally, predictions of treatment effects in the observation space are obtained by applying a pointwise mapping $g_{\theta_g}(\cdot)$ on the hidden process.

The encoder mapping the observed data points to the latent space is a neural network Φ with parameters ϕ: $h(t^*) = \Phi(\mathcal{S}_{t^*}(X))$.

The Neural ODE used to compute predictions of treatment outcomes $Y(t)$ follows the same structure as Equation 1 where we parametrize unknown functions with neural networks $f_{\theta_f}(\cdot)$, $g_{\theta_g}(\cdot)$ and $u_{T,\theta_u}(\cdot)$:

$$\begin{aligned} \frac{dh_T(t)}{dt} &= f_{\theta_f}(h_T(t), u_{T,\theta_u}(t-t^*)) \\ Y_T(t) &\sim g_{\theta_g}(h_T(t)), \forall t \geq t^*. \end{aligned} \tag{2}$$

Incorporating uncertainties in ODEs using stochastic differential equations

We propose to embed uncertainties in the parameters of the ODE. Adopting the formalism from Bayesian neural networks, we posit a prior on the weights of the neural ODE function : $\theta_f \sim P(\theta_f)$ and aim to estimate the posterior weight distribution given the available data $P(\theta_f \mid \mathcal{D})$ where $\mathcal{D} = \{\mathcal{S}_{t^*}, Y\}$. Building upon recent results (Boué, Dupuis et al. 1998; Tzen and Raginsky 2019; Xu et al. 2021), we can show that the following bound holds:

$$\begin{aligned} &\log(p(\mathcal{D})) = \\ &\log \int_{\mathcal{H}} \int_{\theta_f} p(Y \mid \mathcal{H}) \cdot p(\mathcal{H} \mid \mathcal{S}_{t^*}, \theta_f) \cdot p(\theta_f) \cdot d\theta_f \cdot d\mathcal{H}. \\ &\qquad \leq \mathbb{E}_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}[\log p(Y \mid \mathcal{H})] \\ &\qquad - KL_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}(q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*}) \mid\mid p_0(\mathcal{H} \mid \mathcal{S}_{t^*})) \end{aligned} \tag{3}$$

where $p_0(\mathcal{H} \mid \mathcal{S}_{t^*})$ stands for the prior on the process $\mathcal{H} = h_T(t) \mid h_T(t^*)$. The process $\mathcal{H}$ being stochastic, we parametrize the variational approximation of the posterior distribution $q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*})$ with a neural *stochastic* differential equation (SDE) (in particular, a diffusion process),

$$\begin{aligned} &q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*}) \sim \\ &\qquad dh(t) = f_{\theta_f}(h(t), u_{T,\theta_u}(t-t^*))dt + g_{\phi_g}(h(t))dW_t \end{aligned} \tag{4}$$

Where dW_t stands for the multi-dimensional Wiener process. Similarly as in Neural ODEs, a Neural SDE is simply an SDE parametrized by neural networks.

CF-ODE

We are now ready to formalize the final form of our model. A graphical illustration is shown on Figure 2. We first map the observed time series to the hidden space $h(t^*)$. We then reconstruct the potential outcomes by integrating the SDE from Equation 4 and mapping from latent space to observation space using $\mu_{Y_T}(h_T(t)), \sigma_{Y_T}(h_T(t)) = g_{\theta_g}(h_T(t))$, where μ and σ are the parameters of the predicted distribution of $Y_T(t)$ (assumed Gaussian in our experiments). The loss we optimize is then

$$\begin{aligned} &\mathcal{L}(\mathcal{D}, \theta, \phi) = \\ &\mathbb{E}_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}[\log p_\theta(Y \mid \mathcal{H})] \\ &-KL_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}(q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*}, \phi) \mid\mid p_0(\mathcal{H} \mid \mathcal{S}_{t^*}, \phi)) \end{aligned}$$

which, remarkably, can be learned end to end using a differentiable SDE solver (Kidger et al. 2021; Li et al. 2020). In our experiments, we found that fixing $\sigma_X(h(t))$ as a hyperparameter helped the training and resulted in better performance. We provide an in-depth discussion of this effect and a corresponding ablation study in the Appendix. More details about the variational bound derivations are given in the Appendix.

Experiments

We evaluate our method on two datasets: a dataset from cardiovascular modeling and one from pharmacodynamics. The level of confounding in these datasets is controlled with a parameter γ. Full details about these datasets are given in the Appendix. For each of them, we report the root mean-squared error (RMSE) obtained on a left-out test set for the factual distribution (ID), the counterfactual (off-policy) distribution (OOD) and the treatment effect (PEHE (Hill 2011)). To assess the relevance of the uncertainty estimates, we also report the same metrics with only half of the test set samples as introduced in Jesson et al. (2020). They describe different strategies for picking the retained samples. The first one is to keep samples with lowest predicted uncertainty. The second one is to keep sample according to the propensity score

Table 1: Test RMSE for in-distribution (ID), out-distribution (OOD) and PEHE for the different methods. Rows with '50%' refer to trimming the test set to those with either lowest uncertainty, or propensity scores closest to 0.5 (propensity). Best results and the ones that do not significantly differ from best are in bold.

METHOD	ID RMSE	OOD RMSE	PEHE
Cardio-vascular Dataset			
CF-ODE	0.21±0.02	0.59±0.03	0.69±0.03
CFR	0.26±0.02	0.90±0.04	0.98±0.03
IMODE	**0.16±0.02**	**0.38±0.02**	**0.46±0.02**
GP	0.41±0.02	0.52±0.02	0.72±0.01
CF-ODE 50% (uncert.)	**0.11±0.02**	**0.39±0.03**	0.69±0.06
CF-ODE 50% (prop.)	0.20±0.02	**0.33±0.06**	0.45±0.06
CFR 50% (prop.)	0.24±0.01	0.55±0.04	0.61±0.08
IMODE 50% (prop.)	0.16±0.02	**0.36±0.02**	**0.35±0.02**
GP 50% (uncert.)	0.39±0.01	0.51±0.03	0.69±0.04
Dexamethasone Dataset			
CF-ODE	**0.027±0.002**	**0.049±0.005**	**0.050±0.003**
CFR	0.038±0.003	0.066±0.001	0.084±0.001
IMODE	0.051±0.024	**0.051±0.024**	0.094±0.043
GP	0.635±0.090	0.619±0.086	0.086±0.012
CF-ODE 50% ((uncert.)	**0.016±0.003**	**0.025±0.004**	**0.018±0.001**
CF-ODE 50% (prop.)	0.026±0.005	0.038±0.005	0.044±0.006
CFR 50% (prop.)	0.026±0.002	0.057±0.006	0.088±0.006
IMODE 50% (prop.)	0.052±0.024	0.052±0.024	0.094±0.043
GP 50% (uncert.)	0.620±0.062	0.607±0.059	0.083±0.011

(unlikely predictors-treatment assignment pairs are discarded first). The last one is a random sample of the test set.

Baselines

We compare our method against a set of state-of-the-art methods for the prediction of individual treatment effects. In particular, we compare against the counterfactual recurrent network (CRN) (Bica et al. 2020), Gaussian processes (GP) (Schulam and Saria 2017) and IMODE (Gwak et al. 2020).

Results

Table 1 summarizes the results from CF-ODE against the considered baselines in terms of RMSE on factual, counterfactual and treatment effect estimation. We see that our approach is competitive in all considered metrics and outperforms competitors on the dexamethasone datasets.

Using ODE uncertainties to assess when counterfactuals can be predicted. As stated above, we use the uncertainty of the CF-ODE to assess when potential outcomes predictions are reliable. Lower uncertainties should correspond to higher probability of accurate reconstruction and hence lower error (RMSE). We assess the value of the uncertainty estimates as in Jesson et al. (2020). In Figure 3, we show the relative improvement in PEHE that can be obtained with CF-ODE if we filter out datapoints according the predicted uncertainty of the model. We compare this approach against the random and propensity-based strategy. As observed on Figure 3 (Left), the overall PEHE is significantly reduced using the uncertainty of the model, highlighting the usefulness of the uncertainty estimates. Table 1 also reports the results of the trimming operation of our method for all datasets and compares it against baselines. We observe that the uncertainty-based strategy is almost always the most beneficial, leading to lower RMSE in all metrics.

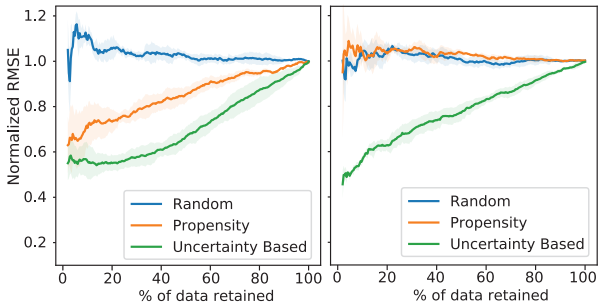


Figure 3: Evolution of normalized PEHE with the number of datapoints retained in the test set for different data pruning strategies on the harmonic oscillator dataset. Left : $\gamma = 8$. Right : $\gamma = 0$ (no counfounding).

Conclusion

The capacity to predict the effect of treatments at the individual level is an considerable challenge, underlying precision medicine. CF-ODE, an uncertainty-aware neural differential equation model to predict the impacts of treatments over time, opens up a novel promising perspective in that direction.

We demonstrated the improved performance of CF-ODE on various datasets with respect to the current state of the art methodologies. Importantly, we showed that incorporating uncertainties in the prediction of potential outcomes was crucial for allowing informed treatment decision. In the context of our recurring clinical example, the uncertainties of CF-ODE can guide healthcare professionals about when to trust the model to recommend treatments, fostering a synergistic collaboration between clinicians and machine learning models in the clinical practice.

Modeling the failures of individual treatment effects predictions with uncertainty is an exciting perspective as it relaxes many of the common assumptions made in this context. Yet, several challenges remain before leveraging this type of models into (clinical) practice. In particular, the best ways to endow neural networks with uncertainty is still a active research area. Our model uses the latest advances in this field but significant improvements (especially in out-of-distribution detection) of Bayesian neural networks are still needed. For simplicity, we made some restrictive assumptions such as binary treatment assignments and fixed time of treatment. While our approach does not preclude the more general case (as detailed in the Appendix) - the backbone of CF-ODE can be adapted to a broad range of practical scenarios - we leave the details of these specific adaptations, as well as the aforementioned open questions as future work.

References

Bica, I.; Alaa, A. M.; Jordon, J.; and van der Schaar, M. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*.

Bica, I.; Alaa, A. M.; Lambert, C.; and Van Der Schaar, M. 2021. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1): 87–100.

Boué, M.; Dupuis, P.; et al. 1998. A variational representation for certain functionals of Brownian motion. *The Annals of Probability*, 26(4): 1641–1659.

Dai, W.; Rao, R.; Sher, A.; Tania, N.; Musante, C. J.; and Allen, R. 2021. A Prototype QSP Model of the Immune Response to SARS-CoV-2 for Community Development. *CPT: pharmacometrics & systems pharmacology*, 10(1): 18–29.

D'Amour, A.; Ding, P.; Feller, A.; Lei, L.; and Sekhon, J. 2021. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654.

Falcon, W.; and The PyTorch Lightning team. 2019. PyTorch Lightning.

Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; and Wilson, A. G. 2018. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*.

Gwak, D.; Sim, G.; Poli, M.; Massaroli, S.; Choo, J.; and Choi, E. 2020. Neural Ordinary Differential Equations for Intervention Modeling. *arXiv preprint arXiv:2010.08304*.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Jesson, A.; Mindermann, S.; Shalit, U.; and Gal, Y. 2020. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33.

Kidger, P.; Foster, J.; Li, X.; Oberhauser, H.; and Lyons, T. 2021. Neural SDEs as Infinite-Dimensional GANs. *International Conference on Machine Learning*.

Li, X.; Wong, T.-K. L.; Chen, R. T.; and Duvenaud, D. 2020. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, 3870–3882. PMLR.

Linial, O.; Ravid, N.; Eytan, D.; and Shalit, U. 2021. Generative ODE modeling with known unknowns. In *Proceedings of the Conference on Health, Inference, and Learning*, 79–94.

Oberst, M.; Johansson, F.; Wei, D.; Gao, T.; Brat, G.; Sontag, D.; and Varshney, K. 2020. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, 788–798. PMLR.

Qian, Z.; Zame, W. R.; van der Schaar, M.; Fleuren, L. M.; and Elbers, P. 2021. Integrating Expert ODEs into Neural ODEs: Pharmacology and Disease Progression. *arXiv preprint arXiv:2106.02875*.

Rickles, D.; Hawe, P.; and Shiell, A. 2007. A simple guide to chaos and complexity. *Journal of Epidemiology & Community Health*, 61(11): 933–937.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Schulam, P.; and Saria, S. 2017. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems*, 30: 1697–1708.

Takens, F. 1981. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, 366–381. Springer.

Tzen, B.; and Raginsky, M. 2019. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*.

Xu, W.; Chen, R. T.; Li, X.; and Duvenaud, D. 2021. Infinitely Deep Bayesian Neural Networks with Stochastic Differential Equations. *arXiv preprint arXiv:2102.06559*.

Zenker, S.; Rubin, J.; and Clermont, G. 2007. From inverse problems in mathematical physiology to quantitative differential diagnoses. *PLoS Comput Biol*, 3(11): e204.

Baselines implementation details

In this section, we give additional detail regarding the implementation of the baselines used in the experiments.

GP

The GP baseline is inspired by the model proposed by Schulam and Saria (2017). Because the code of this paper is not publicly available, we reimplemented as closely as possible to the model described in the original paper. We make our implementation available in our codebase.

We use a a composition of a radial basis function kernel, a periodic kernel and a white noise kernel. Based on the available time series $X(t)$ and the observed potential outcomes, we train the parameters of the kernels that are shared for all time series. As in Schulam and Saria (2017), we model the treatment effect as a learnable additive term in the mean of the gaussian process. In contrast to the original paper, that uses simple exponential functions, we parametrize the impact on the mean with a fully connected neural network with 3 layers of 50 units each. We train the resulting model using GPytorch (Gardner et al. 2018).

The additive term in the mean being very restrictive, this model therefore does not allow for enough flexibility to fit the complex trajectories we investigate in this paper, resulting in poor performance as shown on Table 1.

CFR

We reused the code and hyperparameters made available by the authors and translated into a pytorch lightning module (Falcon and The PyTorch Lightning team 2019), to ease reproducibility and comparison between the different models. The resulting implementation is also available in our codebase.

IMODE

We used the code made available by the authors and added some modifications to evaluate counterfactuals. The changes we brought to the original implementations are also available in our codebase.

Dealing with time series without treatment assignment

For sake of simplicity, we assume in our experiments that the time of treatment t^* is constant. For instance, this would fit an example where patients in the ICU are given (or not given) a treatment a specific amount of time after admission. If this simplified assumption can fit various real-world cases, it might not be always be realistic. In this case, we propose a more general training scheme that acommodates arbitrary treatment times.

For trajectories for which a treatment is observed, the training remains unchanged. For each of those trajectory, we embed the available information $\mathcal{S}_{t^*}(X)$ to the hidden space $(h(t^*))$ and then predict the future trajectory using a neural differential equation.

For trajectories for which we do not observe a treatment however, we lack a time of treatment t^* and it's therefore unclear when the embedding function Φ should be applied. For the non-treated instances, we therefore propose to sample a treatment time from the observed empirical distribution of treatment times in the set of trajectories for which a treatment time is available. Abusing notations, let's denote $\mathcal{T}_{T^*\neq 0}$ the empirical distribution of treatment times, we then minimize

$$\begin{aligned}&\mathcal{L}(\mathcal{D},\theta,\phi) = \\ &\mathbb{E}_{t^*,q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}[\log p_\theta(Y \mid \mathcal{H})] \\ &-KL_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}(q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*},\phi) \mid\mid p_0(\mathcal{H} \mid \mathcal{S}_{t^*},\phi))\end{aligned}$$

where

$$t^* = \begin{cases}\sim \mathcal{T} & \text{if } T^* = 0 \\ t^* & \text{if } T^* \neq 0\end{cases}$$

A remaining interrogation is in the dependence of treatment time on the hidden process. Without loss of generality, we consider the general case of a point process with intensity function modulated by the latent process : $\lambda(h(t)) > 0$ such that $P(t^* \in [t_-, t_+]) = 1 - e^{-\int_{t_-}^{t_+} \lambda(h(s))ds}$. In this case, the confounding is addressed the same way as in Lemma **??**, because the same assumptions holds regarding the time of treatment.

About Assumption 0.3

As mentioned in the main text, Assumption 0.3 is less restrictive that it might appear. Let us first briefly recall Takens theorem (Takens 1981).

Let $X[t] \in \mathbb{R}^{d_X}$ be generated from a chaotic dynamical system that has a strange attractor $\mathcal{M}$ with box-counting dimension d_M, where we define an attractor as the manifold towards which the state of a chaotic dynamical system tends to evolve. The dynamics of this system are specified by a flow on $\mathcal{M}$, $\phi_{(\cdot)}(\cdot) : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$, where $\phi_\tau(\mathcal{M}_t) = \mathcal{M}_{t+\tau}$ and $\mathcal{M}_t$ stands for the point on the manifold at time index t. This flow is encoded in the ODE of the system. The observed time series $X[t]$ is then obtained through an observation function $f_{obs}(\cdot) : X[t] = f_{obs}(\mathcal{M}_t)$. Takens' theorem then states that a delay embedding Φ with delay τ and embedding dimension k

$$\begin{aligned}&\Phi^{k,\tau}_{\phi,\alpha}(\mathcal{M}_t) = \\ &\qquad(\alpha(\phi_0(\mathcal{M}_t)), \alpha(\phi_{-\tau}(\mathcal{M}_t)), \ldots, \alpha(\phi_{-k\tau}(\mathcal{M}_t)))\end{aligned}$$

is almost surely an embedding of the strange attractor $\mathcal{M}$ if $k > 2d_M$ and $\alpha : \mathbb{R}^{d_M} \to \mathbb{R}$ is a twice-differentiable observation function. More specifically, the embedding map Φ is a diffeomorphism between the original strange attractor manifold $\mathcal{M}$ and a shadow attractor manifold $\mathcal{M}'$ generated by the delay embeddings. Under these assumptions, one can then theoretically reconstruct the original time series from the delay embedding.

Using this key result, we can conclude that if t^* is large enough (if we have enough history on the time series before prediction), then any twice-differentiable observation function would almost surely provide us with an injective map between the filtration $\mathcal{F}^*_t(X)$ and $h(t^*)$ and Assumption 0.3 would then be satisfied.

The above provides supporting evidence for as why Assumption 0.3 might be often be satisfied in practice. However, two difficulties may still be apparent to the reader. First, the fact that the dynamical system is supposed to be chaotic. We note that all dynamical systems are either periodic, quasi-periodic or chaotic. And that many dynamical systems of interests are actually chaotic (Rickles, Hawe, and Shiell 2007). Second, because of the irregular sampling nature of the available time series, an appropriate twice-differentiable observation function α might not be available. This is a topic we are currently investigating, but, in the absence of formal proof, we claim that if enough information is contained in $\mathcal{S}_t^*(X)$, that is, the sampling rate is high enough, then Takens holds as well.

Using uncertainties to improve treatment assignment strategies

As shown in Section **??**, uncertainties can be used to derive probabilities of positive response to treatment that can then inform when to treat a specific patient. In this section, we highlight another approach of using uncertainties to guide treatment decision.

Similarly as in Section **??**, we use the oscillator dataset and aim at having a positive treatment effect at some arbitrary time after the giving the treatment. By making a decision of treatment on patients for whom the uncertainty is low, we can decrease the probability of mistakenly assigning a treatment to a patient who will not benefit from it. We frame this error as a false discovery rate (FDR) :

$$FDR = \frac{\sum_i^N (ITE_i < 0) \cdot (\hat{T}_i)}{\sum_i^N \hat{T}_i}$$

where ITE_i is the true individual treatment effect of patient i and $\hat{T}_i \in \{0, 1\}$ is the recommendation of treatment that we provide about this patient. We can decide to treat a patient whenever $ITE_i > 0$ (*i.e*, when the average predicted ITE is higher than 0). Using uncertainties, we can also decide to treat only patient for whom the uncertainty is not higher than a certain threshold. On Figure 4, we display the average predictions of the model for each patient against the true treatment effects in function of the proportion of data used for predictions. As in Section **??**, we remove data points with higher uncertainty first. We see that if we focus on the patients with the lowest uncertainties, the predictions get more accurate: they converge towards the optimal identity line (*i.e* the prediction is equal to the true value). Importantly, this also results in lower FDR, thus in a lower proportion of patients being treated and who would eventually not benefit from the treatment. On Figure 5, we show the evolution of the FDR in function of the proportion of data kept over the different folds. This rate decreases monotonically as patients with lowest uncertainties are kept.

Ablation Studies

To better understand how the building block of our model contribute individually to the overall performance, we performed an ablation study where we switch off the diffusion part of the model, that would therefore result in a standard Neural ODE. This is referred as the *no diffusion* variant in Table 2.

We observe that the performance degrades significantly, both in the full and the 50% dataset. This can be explained by the advantageous averaging/ensembling effect of the Bayesian model. What is more, the KL term in the loss function acts as a regularizer of the differential equation, which can contribute to the improvement of performance.

Secondly, we investigate a variant of our model where we $g(h(t))$ outputs both a mean and a standard deviation estimate of the observations $X(t)$ as described in Section . We refer to this variant as *with learnable std*. We observe an higher MSE than with the standard version of CF-ODE. However, this gap tends to close when the percentage of data points kept decreases. Importantly, despite more appealing theoretical properties, we found that learning the standard deviation of the output distribution made the model harder to train, with instabilities caused by the standard deviation being in the denominator of the log likelihood.

Details for the derivation of the variational lower bound

Incorporating uncertainties in ODEs using stochastic differential equations

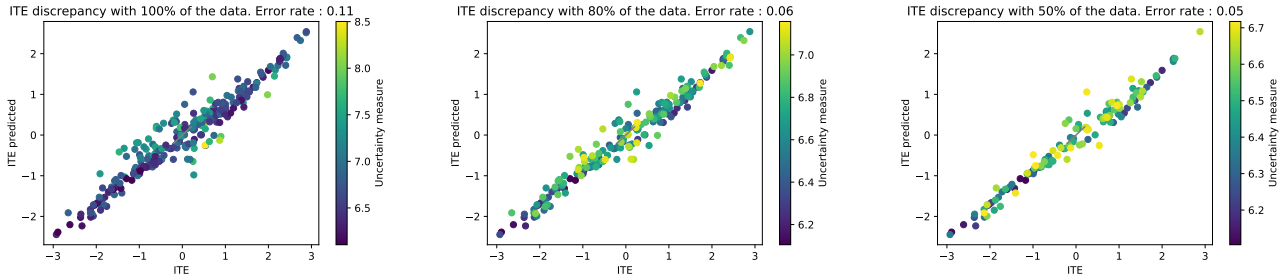
In order to provide uncertainties estimates for the predictions, we embed uncertainties in the parameters of the ODE. Adopting the formalism from Bayesian neural networks, we posit a prior on the weights of the neural ODE function $f_{\theta_f}(\cdot)$: $\theta_f \sim P(\theta_f)$ and aim at estimating the posterior of the distribution of the ODE parameters conditionned on the available data $P(\theta_f \mid \mathcal{D})$ where $\mathcal{D} = \{\mathcal{S}_{t^*}, Y\}$. The weights being probabilistic, the ODE therefore effectively becomes a *random* differential equation whose generating process goes as

$$\begin{aligned} &\theta_f \sim P(\theta_f) \\ &h_T(t) = h_T(t^*) + \int_{t^*}^{t} f(h_T(s), u_{T,\theta_u}(s - t^*), \theta) \cdot ds \end{aligned}$$

where we wrote the dependence on the weigths θ_f explicitly (f is then only the structure of the neural network, without the weigths). Crucially, this means that the prior on the weights θ_f parametrizes a prior distribution on the process $h_T(t) \mid h_T(t^*)$. For brevity of the notations, we refer to this process as $\mathcal{H}$. Using the natural decomposition of our model, we can derive a variational bound for the marginal probability of the available data $\mathcal{D}$, where we make the dependence on the treatment implicit:

$$\begin{aligned} &\log(p(\mathcal{D})) = \\ &log \int_{\mathcal{H}} \int_{\theta_f} p(Y \mid \mathcal{H}) \cdot p(\mathcal{H} \mid \mathcal{S}_{t^*}, \theta_f) \cdot p(\theta_f) \cdot d\theta_f \cdot d\mathcal{H}. \end{aligned}$$

The quantity $p(\mathcal{H} \mid \mathcal{S}_{t^*}, \theta_f)$ is a dirac function (as every realization of θ_f will lead to a single realization of $\mathcal{H}$) and



(a) 100% retained (b) 80% retained (c) 50% retained

Figure 4: Evolution of false discovery rate with the percentage of patients used for predictions. Patients with higher uncertainty are discarded first. The x-axis represents the exact individual treatment effect while the y-axis represents the average predictions for each patient. Predictions are getting more accurate (converging on the identity line) and the false discovery rate is decreasing.

Table 2: Test RMSE for in-distribution, out-distribution and PEHE for the different ablated versions of CF-ODE on the harmonic oscillator dataset.

	METHOD	In-distribution RMSE	Out-distribution RMSE	PEHE
$\gamma = 8$	CF-ODE	**0.04±0.01**	**0.06±0.01**	**0.07±0.01**
	CF-ODE no diffusion	0.17±0.10	0.28±0.17	0.38±0.25
	CF-ODE with learnable std	0.08±0.04	0.12±0.05	0.16±0.08
	CF-ODE 50 %	**0.03±0.01**	**0.03±0.01**	**0.05±0.01**
	CF-ODE 50 % no diffusion	0.11±0.14	0.24±0.28	0.34±0.39
	CF-ODE with learnable std 50%	**0.03±0.01**	0.05±0.01	0.07±0.03

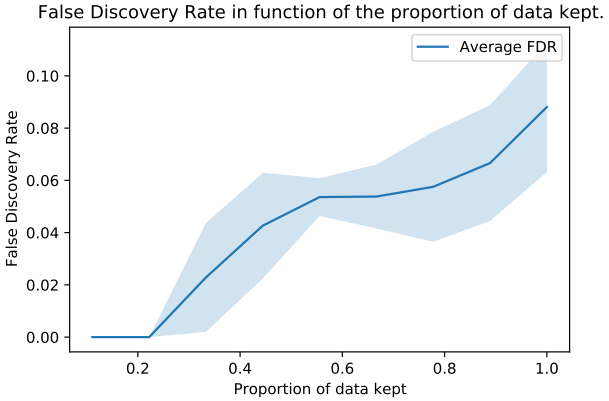


Figure 5: Evolution of the FDR in function of the proportion of data points (patients) kept. Patients with highest uncertainty are discarded first. Shaded are is 2 standard deviations wide.

the quantity $\int_{\theta_f} p(\mathcal{H} \mid \mathcal{S}_{t^*}, \theta_f) \cdot p(\theta_f) d\theta_f = p_0(\mathcal{H} \mid \mathcal{S}_{t^*})$ corresponds to a prior of the latent process generated by the prior on θ_f. We can then further write

$$\begin{aligned}
\log(p(\mathcal{D})) &= \log \int_h p(Y \mid \mathcal{H}) \cdot p_0(\mathcal{H} \mid \mathcal{S}_{t^*}) \cdot d\mathcal{H} \\
&= \log \int_h p(Y \mid \mathcal{H}) p_0(\mathcal{H} \mid \mathcal{S}_{t^*}) \frac{q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*})}{q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*})} \cdot d\mathcal{H} \\
&\leq \mathbb{E}_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})} \left[\frac{\log p(Y \mid \mathcal{H}) \cdot p_0(\mathcal{H} \mid \mathcal{S}_{t^*})}{q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*})} \right] \\
&= \mathbb{E}_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})} [\log p(Y \mid \mathcal{H})] \\
&\quad - KL_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})}(q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*}) \mid\mid p_0(\mathcal{H} \mid \mathcal{S}_{t^*}))
\end{aligned} \tag{5}$$

The process $\mathcal{H}$ being stochastic, we parametrize the posterior distribution $q_\theta(\mathcal{H} \mid X)$ with a stochastic differential equation, in particular a diffusion process:

$$q_\theta(\mathcal{H} \mid \mathcal{S}_{t^*}) \sim d\mathcal{H} = f_{\theta_f}(h(t), u_T(t - t^*))dt + g_\phi(h(t))dW_t \tag{6}$$

In order to have a tractable KL divergence term in equation 3, we choose the prior of process h(t) to be a diffusion process with same diffusion parameter g_θ. This requirement appears a priori very restrictive but any posterior can be approximated arbitrarily closely by such a functional form given a sufficiently expressive drift process (Boué, Dupuis et al. 1998; Tzen and Raginsky 2019; Xu et al. 2021). The prior on the process $h(t)$ then writes :

$$p_0(\mathcal{H} \mid X) \sim dh(t) = f_\phi(h(t))dt + g_\phi(h(t))dW_t,$$

The KL divergence is then tractable and given by

$$KL_{q_\theta(h|X)}(q_\theta(h \mid X) \mid\mid p_0(\mathcal{H} \mid X)) = \mathbb{E}_{q_\gamma(\mathcal{H}|X)}[\int_0^T \mid\mid u(t,\gamma) \mid\mid_2^2 dt]$$
$$\text{where} \quad u(t,\gamma) = g_\phi(h(t))^{-1}[f_\gamma(h(t)) - f_\phi(h(t))]$$

Datasets

Harmonic Oscillator

We first demonstrate the performance of our approach on a synthetic dynamical system. We model the dynamics of a pendulum where the intervention consists in injecting energy in the system over time by increasing the velocity of the pendulum weight. The structural equations are given in Equation 7. We further consider that only the angle of the pendulum is observed (the complete state is thus never fully observed) and that observations are made at random and irregularly over time. Different trajectories are generated by sampling different initial angles θ_0 and lengths $l : \theta_0 \sim \mathcal{U}(0.5, 1.5)$ and $l \sim \mathcal{U}(0.5, 4.5)$, resulting in different trajectories amplitudes and frequencies. Treatment assignments are coupled with the trajectories by setting $P(T = 0 \mid \theta_0) = \sigma(\gamma(\theta_0 - 1))$ where $\sigma(\cdot)$ is the *sigmoid* function and γ is a parameter tuning the degree of confounding. The dosage of the treatment A is set as a function of the the amplitude as well.

$$\begin{aligned} \frac{d\theta(t)}{dt} &= v \\ \frac{dv(t)}{dt} &= (1 + u(t)) \left(\frac{-g}{l}\right) \sin(\theta(t)) \\ u(t) &= A \sin(\phi t) \cdot e^{-\delta t} \end{aligned} \tag{7}$$

CardioVascular Model : assessing the impact of fluids intake

We use a model of the cardiovascular system as proposed in (Zenker, Rubin, and Clermont 2007; Linial et al. 2021) and use to study the capacity of our model to learn the impact of fluids intake. Fluid intake is commonly used for treating severe hypotension. However, the response of patient to fluids intake is difficult to assess beforehand. In particular, it depends on the patients cardiac contractility factor and the blood pressure at time of injection. If blood pressure is commonly and easily measured in standard clinical practice, assessing the cardiac contractility level of a patient requires imaging techinques such as echocardiography to measure the stroke volume. Yet, the injection of significant volume of fluids in an irreponsive patients cardiovascular system can lead to severe damage. This lead some clinicians to advocate for fluid challenges, or limited amount of fluid injection to test the responsiveness. This technique is still contested in the medical community and legs raising challenge, much less damaging but also less effective at assessing a patients response has been encouraged. This lack of availability of clear guidelines for fluids intake makes it a perfect case study for counterfactual prediction. Indeed, we'll try to address the question of if a clinician should administer fluids to particular patient based on his clinical history and therefore help informing clinical practice. The system of ODE used to generate the data is the following :

$$\begin{aligned} \frac{dSV(t)}{dt} &= I_{\text{external }(t)} \\ \frac{dP_a(t)}{dt} &= \frac{1}{C_a}\left(\frac{P_a(t) - P_v(t)}{R_{TPR}(S)} - SV \cdot f_{HR}(S)\right) \\ \frac{dP_v(t)}{dt} &= \frac{1}{C_v}\left(-C_a \frac{dP_a(t)}{dt} + I_{\text{external }(t)}\right) \\ \frac{dS(t)}{dt} &= \frac{1}{\tau_{\text{Baro}}}\left(1 - \frac{1}{1 + e^{-k_{\text{width}}\left(P_a(t) - P_{a_{\text{set}}}\right)}} - S\right) \end{aligned}$$

where

$$\begin{aligned} R_{TPR}(S) &= S(t)\left(R_{TPR_{Max}} - R_{TPR_{Min}}\right) \\ &\quad + R_{TPR_{Min}} + R_{TPR_{Mod}} \\ f_{HR}(S) &= S(t)\left(f_{HR_{Max}} - f_{HR_{Min}}\right) + f_{HR_{Min}}. \end{aligned}$$

In the above dynamical system, $P_a, P_v, SandSV$ stand for arterial blood pressure, venous blood pressure, autonomic baroreflex tone and cardiac stroke volume respectively. $I_{\text{external }(t)}$ is the amount of fluids given the patient over time and corresponds to the exogeneous input $u_T(t)$ in our model. In the data generation, we model it as

$$I_{\text{external }(t)} = 5 * e^{-(\frac{t-5}{5})^2}$$

The treatment assignment is confounded with the history of the patient and we make it explicity depends on the value of the arterial blood pressure at the time of treatment :

$$P(T = 1) = \sigma(\gamma \cdot (\frac{P_a(t^*) - P_{a,min}}{P_{a,width}} - 0.5))$$

where σ is the sigmoid function and $P_{a,min}$ and $P_{a,width}$ are defined parameters (75 and 10)).

Pharmacodynamics Model On top of the previous datasets, we also consider the pharmacodynamics of a dexamethasone, a glucocorticoid drug that has been used in treatment against COVID19. The dynamical system, presented in Equation 8 is adapted from Dai et al. (2021) and Qian et al. (2021). Variables z_1 and z_5 represebt the innate and adapative immune response, z_2 and z_3 the concentration of dexamethasone in the lung tissue and plasma and z_4 represents the viral load.

$$\begin{aligned} \dot{z}_1 &= k_{IR} \cdot z_4 + k_{PF} \cdot z_4 \cdot z_1 - k_O \cdot z_1 \\ &\quad + \frac{E_{\max} \cdot z_1^{h_P}}{EC_{50}^{h_P} + z_1^{h_P}} - k_{Dex} \cdot z_1 \cdot z_2 \qquad (8) \\ \dot{z}_2 &= -k_2 \cdot z_2 + k_3 \cdot z_3 \qquad (9) \\ \dot{z}_3 &= -k_3 \cdot z_3 \qquad (10) \\ \dot{z}_4 &= k_{DP} \cdot z_4 - k_{IIR} \cdot z_4 \cdot z_1 - k_{DC} \cdot z_4 \cdot z_5^{h_C} \qquad (11) \\ \dot{z}_5 &= k_1 \cdot z_1 \qquad (12) \end{aligned}$$

Only variables z_1 and z_5 can be realistically measured in the lab, through Type I IFNs and Cytotoxic T Cells respectively. Therefore we only use those two variables in $X(t)$. We set the variable z_1 as the variable of interest for the treatment effect. We model the intervention by simulating a constant injection of dexamethasone in the plasma ($\dot{z_3} = 10$). We introduce confounding by modeling a dependence of the treatment assignment on the factor k_{Dex}, that modulates the impact of dexamethasone on the immune response. We set $P(T = 1) = \sigma(\frac{k_{Dex}-1}{15} - 0.5)$ with $K_{Dex} \sim \mathcal{U}(1, 16)$.

Supplementary Experiments

To better understand how the building block of our model contribute individually to the overall performance, we performed an ablation study where we switch off the diffusion part of the model, that would therefore result in a standard Neural ODE. Results are presented in Table 2 in Appendix and show higher MSE than when an SDE is learnt. In an attempt to capture multiple sources of variances independently, we also report results when the networks outputs a mean estimate and a standard deviation (as discussed in Section). This resulted in higher MSE but the gaps resolves when 50% of the data is considered. In practice, we found this setup to be difficult to train, as discussed in more details in Appendix .