We thank the reviewers for thoughtful reviews and appreciation of our work. We expanded the "Related work" section to add more citations on the adjoint sensitivity method, which indeed dates back to 1962 [1]. We also added the very recent work by Ayed et al (2019). As reviewers noticed, detailed description of the experiments is provided in the supplement. We added more experimental details to the main text. We also clarified how the toy dataset was created and added a more detailed caption for Figure 3.

We updated our results on Physionet and Human Activity to include error bars, shown in the tables below. We also switched to a finer discretization of 1 minute on Physionet dataset, instead of 6 minutes. This resulted in the better performance for all the models on the interpolation and extrapolation tasks. Our conclusions remain consistent with the original submission.

Reviewer 1. 2) We agree that Latent ODEs may suffer from issues similar to the "shooting method". However, since the learned ODE is in a latent space, it is possible to move model complexity into the decoder, and regularize the learned ODE and latent landscape to be sufficiently smooth.

4) We agree on the value of including non-deep-learning baselines. Previous works [2,3] have conducted an extensive comparison with non-deep-learning methods, including powerful methods like MICE and ImputeTS. In turn, we perform comparison to the GRU-D model on the same datasets.

5) The Physionet dataset is typically discretized by one hour [2,3], which leads to the loss of 98% of all time points. In contrast, we discretized using 1 minute intervals, as this better showcases the benefits of continuous time-series modeling. Our approach can be straightforwardly applied without any discretization, but this makes batched training more expensive, with little gain in performance.

6) We performed hyperparameter search over a grid with respect to the baseline model (RNN GRU-D for autoregressive models, Latent ODE (RNN enc.) for encoder-decoder models). Our proposed models then used the same hyperparameters as the corresponding baselines where possible. We added the hyperparameter ranges used.

Reviewer 2. Our toy experiments are meant to illustrate the unique properties of these models. Quantitative comparisons and experiment details were placed in the supplement. Figure 3a demonstrates that Latent ODEs can condition on an arbitrary number of observations at test time. Figure 3b shows that samples from the trained prior resemble the data.

We argue our ODE-augmented model has a better inductive bias than vanilla RNNs. Modeling state-interactions via an ODE allows better generalization outside the training interval compared to directly modeling a function of time. Furthermore, this poor generalization can also occur for the RNN encoder, as shown in supplementary Figure 2.

Finally, we agree with the criticism about the absense of error bars. We re-ran experiments on Physionet and Human Activity datasets with different random seeds to provide error bars. All updated results are shown in the tables below. We report mean and standard deviation over three runs. We bolded significant results based on t-tests.

Reviewer 3. We implemented baseline RNNs by concatenating Δ_t to the input to account for the irregular time gaps. We will clarify this in the text. Concerning poor the performance of RNN encoders, one possible explanation is that RNN encoder is not robust to variable time gaps between points.

In Algorithm 2, z_0' is the evolving hidden state of the encoder, which is then fed into a fully connected network to output the means and variances of z_0. We will update this notation and refer to z_0' as h_0 (last hidden state of encoder).

Using an ODE-RNN as a decoder is a possible extension. However, we conjectured that this model will likely ignore the latent variable as observed in VAE-RNN language models, and so did not investigate this model.

We visualized the hidden states of an ODE-RNN encoder in between observations in Suppl. figure 5(b). Encoding is performed backwards in time.

Table 1: Test MSE (mean $\pm$ std) on PhysioNet. **Autoregressive** models.

Model	Interp ($\times 10^{-3}$)
RNN Δ_t	3.520 ± 0.276
RNN-Impute	3.243 ± 0.275
RNN-Decay	3.215 ± 0.276
RNN GRU-D	3.384 ± 0.274
ODE-RNN (Ours)	$\mathbf{2.361 \pm 0.086}$

[1] Pontryagin et al. The mathematical theory of optimal processes. 1962.
[2] Cao et al. Bidirectional recurrent imputation for time series, 2018. arxiv/1805.10572
[3] Che et al. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports, 8(1):6085, 2018.

Table 2: AUC on Physionet.

Method	AUC
RNN Δ_t	0.787 ± 0.014
RNN-Impute	0.764 ± 0.016
RNN-Decay	0.807 ± 0.003
RNN GRU-D	$\mathbf{0.818 \pm 0.008}$
RNN-VAE	0.515 ± 0.040
Latent ODE (RNN enc.)	0.781 ± 0.018
ODE-RNN	$\mathbf{0.833 \pm 0.009}$
Latent ODE (ODE enc)	$\mathbf{0.829 \pm 0.004}$
Latent ODE + Poisson	$\mathbf{0.826 \pm 0.007}$

Table 3: Human Activity.

Method	Accuracy
RNN Δ_t	0.797 ± 0.003
RNN-Impute	0.795 ± 0.008
RNN-Decay	0.800 ± 0.010
RNN GRU-D	0.806 ± 0.007
RNN-VAE	0.343 ± 0.040
Latent ODE (RNN enc.)	0.835 ± 0.010
ODE-RNN	0.829 ± 0.016
Latent ODE (ODE enc)	$\mathbf{0.846 \pm 0.013}$

Table 4: Test MSE (mean $\pm$ std) on PhysioNet. **Encoder-decoder** models.

Model	Interp ($\times 10^{-3}$)	Extrap ($\times 10^{-3}$)
RNN-VAE	5.930 ± 0.249	3.055 ± 0.145
Latent ODE (RNN enc.)	3.907 ± 0.252	3.162 ± 0.052
Latent ODE (ODE enc)	$\mathbf{2.118 \pm 0.271}$	$\mathbf{2.231 \pm 0.029}$
Latent ODE + Poisson	2.789 ± 0.771	$\mathbf{2.208 \pm 0.050}$