We thank the reviewers for valuable and thoughtful feedback, and for acknowledging the importance of this work.

Scope: Our main objective has been to provide insights on current limitations of the Neural ODE framework, and to design novel solutions backed by theory. Although evaluations for specific use cases e.g. CNF was not in within our primary scope, we include additional results to address requests and provide further evidence of practical usefulness.

Further augmentation evaluations [R2, R3]: We obtained improved results on all variants by means of a similar architecture equipped with a pooling layer (closer to standard image class. approaches) and tolerances 10^4. The neural ODEs learn richer dynamics (higher NFEs), boosting performance across all models and clarifying the relative ranking of aug. approaches. This also leads to improved parameter efficiency (8x less parameters). Tab. 1 includes results on both MNIST as well as CIFAR10 **(R1)**. The parameter efficiency of 2^{nd}–order models **(R2, R3)** is now more pronounced. We also report that they converge faster, often several epochs ahead of the alternatives.

Discussion on CNFs [R1, R2, R4]: We further showcase *data-control* (DC) strategies in the context CNFs as a more complex task. Compared to regular CNFs, DC–CNFs do not require changes to the formulation and converge faster with simpler flows as shown in Fig. 1, effectively reducing NFEs. Adaptive–depth models are also compatible with CNFs and would allow the model to allocate more *depth* to data further away from their target destination.

Depth-variance techniques [R2]: We agree that the choice of basis in Galërkin neural ODEs is important and worthy of several standalone investigations. However, regarding the *sinusoids* example, periodicity of the weights (inferred by the choice of the Fourier eigenbasis) does not imply periodicity of the Neural ODE and hence does not constitute a strong inductive bias. To confirm this, we tested with different signals and eigenbasis (Chebychev poly., RBFs). Fig. 2 shows a more complex experiment for time–varying nonlinear system.

Signal tracking [R2]: Here, depth–variance is not needed to actually learn the trajectory, generated by $\ddot{x} = x; [x_0; \dot{x}_0] = [1; 0]$ which does not contain any depth-varying harmonics. Rather, it ensures that for any initial condition of the neural ODE, sampled from $N([1; 0]; \ )$, the solution converges to the signal. The same result can be obtained for nonlinear systems whose solution does not admit a finite spectral decomposition as shown above.

Related work [R3]: We agree that these important references belong to Section 6 and have made the suggested changes. The approach of latent neural SDEs (and ODEs) is different compared to data-control, which does not require variational inference. It is correct to state that, however, both approaches condition the vector field on data.

Relation to PMP [R3]: We agree that Th.m 1 is directly derived via classic optimal control theory (PMP) and should be more appropriately referred to as "Proposition". However, including it was necessary for two main reasons: to extend vanilla adjoints to integral loss functions (used in practice for CNFs or signal tracking but not yet implemented) and to set the stage for Th.m 2. We note that Th.m 2 is a non–trivial generalization to infinite dimensional spaces.

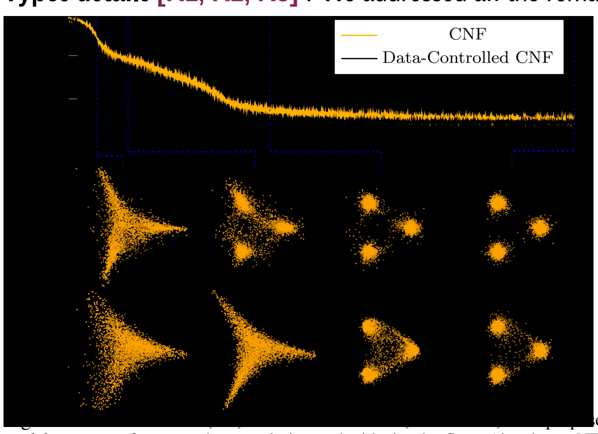
Relation to COD [R3]: We agree on COD and clarified the statements in Sec. 5. The phenomenon we want to highlight is that dimension of the state–space also drastically affects the behavior of dynamical systems in general.

Guidelines on choosing correct variants [R3]: In general, we observe data–control to be beneficial in all settings. We agree that additional guidelines on model choice could be useful to the reader; we will add more information.

Given Sec. 5.2, state–space crossing might be possible if each traj. could travel for different amounts of time. [R2]: In 5.2 we argue that adaptive–depth models can learn the reflection map **without** crossing flows (as they still cannot cross), consistently to what is stated in rest of Sec. 5. This is in fact the main *leitmotiv* of adaptive-depth models.

Clarifications: Figure 1 [R2]: The blue curves are learned flows of test init. cond., which converge to the signal to track (hence the decreasing variance across *depth*). **Figure 2 [R2, R3]:** Each traj. represents the evolution of a single parameter. **Training details [R2]** On the signal tracking task, we train on 10^2 initial conditions, full batch. The GalNODE architecture has a hidden layer of 64. On depth-varying classification, the architectures have two hidden layers of 32 with a dataset of 10^4 points (dense, to approx. *connected annuli*). **Why** s **instead of** t**? [R1]:** We chose s against t as a more general formulation for the (continuous) depth, to avoid confusion in static settings or whenever time is not directly related to depth–propagation dimension. **"This approach" refers to hypernetworks? [R2]:** Yes.

Typos details [R1, R2, R3] : We addressed all the remaining typos.



...ed model converges faster to a better solution and with simpler flows (thus lower NFEs).

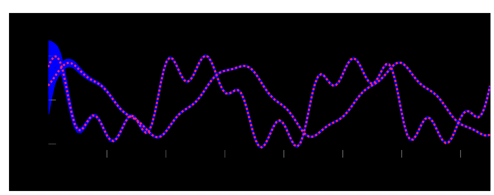


Figure 2: Galërkin Neural ODEs with Chebychev eigenbasis trained with integral loss to track the nonlinear time-varying *Duffing oscillator* $\ddot{x} = -\alpha x(1 + x^2) + \beta \cos(\omega s)$ ($\alpha = .5, \beta = 3, \omega = 5$). All traj.s, starting from rand. ICs, converge to the desired signal.

	NODE		ANODE		IL-NODE		2nd–Ord.	
	MNIST	CIFAR	MNIST	CIFAR	MNIST	CIFAR	MNIST	CIFAR
Test Acc.	98.3	59.1	99.1	68.7	99.4	70.7	99.5	71.8
NFE	130	152	124	153	112	150	106	142
Param.[K]	6.4	42.1	6.4	41.4	6.4	41.9	5.8	37.4

Table 1: Test results, 5 runs on MNIST, 3 on CIFAR10* (final results might vary slightly.)