We thank the four reviewers for their excitement about our work and detailed feedback! Overall, reviewers enjoyed the problem and approach: **"potential to impact the wide array of applications" "appealingly simple," "novel approach", "very extensive and promising" "better than existing techniques"**; criticism centered around description of prior work, and clarity of problem formulation and details. We address these concerns here

(a) Regularizer design and d_E (R2, R1). We thank R2, and we hope that they will reconsider. We've revised the text in order to clarify the problem framing and fix notational ambiguities and some typos, which we believe clarify R2's major concerns (and a related point about d_E by R1). d is the number of independent dynamical variables (number of ODEs); $d_F \leq d$ is attractor manifold dimension (non-integer for chaotic systems with fractal attractors; $d_F = 1$ for limit cycles); d_E is embedding dimension. The hyperparameter L sets the maximum d_E expressible by the autoencoder (AE). However, the AE does not pick an integer d_E; rather, d_E is estimated continuously post-training via the relative variance of each latent variable averaged across train data (similar to PCA weights). Dimensionality score thus compares explained variance (a function of latent index) between reconstruction and full dimensional system. Our experiments on systems with known d seek and find that $d_E \approx d$. Our regularizer takes a batch of latent activations, and estimates $\bar{F}_m$, the proportion of new false neighbors indexed by number of latent dimensions. Since $\bar{F}_m$ is intensive, it only weakly depends on batch size (we have added new physics references related to this observation).

(b) Dependence on L (R2, R1). We have added new experiments showing that L does not affect learned d_E of latent embedding as long as L is larger than d (see Fig H1); thus for unknown systems any large L can be used.

(c) Prior Work (R1, R3). We performed new experiments and extended the main text discussion of Ref. 35, which (to our knowledge) is the primary prior application of AE to attractor reconstruction. Ref. 35 embeds datasets via a one-layer AE with $\tanh$ activation and MSE loss, similar to our existing baseline unregularized MLP (see appendix). We performed a new set of experiments with a model exactly matching Ref. 35 (Fig H1), and find it is comparable to our other baselines for the noisy prediction task (same task as ref. 35). We have added these results to the paper.

(d) Higher dimensions (R4) Thank you! We have clarified that the ecosystem results are 10D; pendulum is 4D. We've added discussion and references to physics papers about mathematical limitations of embedding in the high-d limit.

(e) Existing work on state space modelling (R3): Thank you, and we hope that you will reconsider. Our paper doesn't claim to be the first AE applied to embedding (see (c) above). Indeed, 30% of our submission is a literature review of state space modelling (SSM); there we demonstrate several clear areas of novelty: (1) Our paper focuses on a **fully novel loss function and regularizer** rooted in the classical theory of dynamical systems, and we shows that this regularizer strongly constrains and improves AE representations, in contrast to prior AEs used for SSM. (2) Our paper uses a variety of **novel measures of attractor fidelity**—e.g. topology, neighbor coverage, fractal dimension—which go beyond few-timestep RMSE forecasting errors (the primary metric in previous works). (3) We show strong results for embedding **consistency across replicates, and robustness to Brownian stochasticity** (a more complex noise source than the measurement errors studied in prior works), two desirable embedding properties not explored previously.

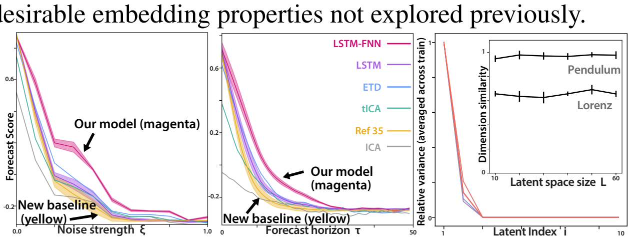

Figure H1: (A, B) Updated Fig 4 with ref. 35 baseline (yellow). (C) Similar activity patterns in first 10 latent units for fnn-AEs trained with $L = 10, 20, 30, 40, 50$ (blue to red). (Inset) dimension accuracy vs L also shows no dependence.

R1 & R2 additional comments: Thank you so much for detailed feedback; we've fixed all wording, framing, and added suggested references; we regret that space limits us to major concerns not covered above: **R1 Misc: 8.1.2a,b** We revised hyperparameter discussion to add nuance: we mean that our experiments achieve strong results only by varying learning rate and regularizer strength, and the former is only tuned to ensure that train loss plateaus. Rather than pre-select embedding timescale, we favor fixing large T and batch size, and letting the AE learn how much to weigh different timepoints. **8.1.2c,d,e** See (a) above. **8.3** We've moved details from appendix 5 into main text. There are few widely-used definitions of attractor similarity (many SSM papers from ML authors focus on prediction, not verisimilitude, and many physics papers are qualitative), and so we report multiple established and novel metrics in order to give a holistic view. We've added the caveat that FNN-AE is more expensive than ETD/tICA, especially on small datasets, but only marginally more expensive than unregularized AE. **8.5** We revised to clarify that time series are Fourier-resampled only to ensure consistency across datasets; preprocessing/filtering otherwise has little effect (hence noise results) **8.6** We re-ran experiments and confirmed with pendulum data; we will add this to the appendix.

R2 Misc: 3.2, 8.8, 8.12, 8.14-16 See (a, b); we've also moved Appendix 5 details to main text to clarify scoring metrics. We always refer to size L latent space as embedding space: T time delays only serve as a featurization of input to the AE, which seeks (and we find achieves) $d_E \approx d$ (not T). $d_E \leq L$ is computed *post-hoc* from relative latent activations by finding variance of the L latent variables across train set, giving continuous measure of dimensionality (thus unaffected by zero-padding). Thus d_E is neither a hyperparameter nor a direct AE output. **R2.8.2** The method and the code we're releasing now works for multivariate time series; we will highlight this. **R2.8.11** No leakage; when available, we use 2 different datasets or initial conditions; otherwise we use first N and last N points of a time series with length $\gg 2N$. **R2.8.16** We scale lower bound to mean, not theoretical min (see Nassar et al ICLR 2019, Eq. 25).