Effects of Changes in Time on Latent Variable Models of a Learning Brain

Arnold Cheskis[1], Kabir Dabholkar[2], & Omri Barak[3]
[1]*Electrical and Computer Engineering Faculty, Technion - Israel Institute of Technology*
[2]*Department of Mathematics, Technion - Israel Institute of Technology*
[3]*Rappaport Faculty of Medicine and Network Biology Research Laboratory, Technion - Israel Institute of Technology*

Commented [MR1]: Author names should be Arial 12 point, bold, left aligned.

List author names as First Name, Middle Initial, Last name, separate by commas, with superscript numbers indicating institution affiliation.

Latent variable models offer a powerful framework for uncovering hidden dynamics from observed data. However, in the absence of ground truth, model evaluation often relies on emission likelihood—a metric that can obscure discrepancies in latent representations. In this study, we introduce a structured approach for analyzing models trained via Hidden Markov Models (HMMs) within a teacher-student framework. By organizing evaluation across a three-dimensional space of students, teachers, and training epochs, we quantify both likelihood and decoding performance over time. Through this lens, we expose differences in generalization behavior and convergence dynamics among students trained on varying teacher sequences.
These findings reveal that decoding asymmetry, co-smoothing patterns, and adaptation rates offer richer signals about model quality than likelihood alone. Training on emissions from different teachers across epochs acts as a form of implicit regularization, surfacing distinctions between models that might otherwise appear equally effective. This method not only illuminates how training regimes shape learned representations but also identifies when larger or smoother models reflect true underlying dynamics versus overfitted emission patterns. Our framework contributes a principled method for assessing model reliability and representational fidelity in the broader context of latent variable modeling.

Keywords: Latent Variable Models; LVMs; Hidden Markov Models; HMMs; Teacher; Student; Likelihood; Decoding; Curriculum Learning; Perturbation

Introduction

Latent variable models (LVMs) are widely used to infer the hidden structure of dynamical systems based on observed data. In neuroscience, they help capture the underlying dynamics of neural circuits from recorded brain signals (Macke et al., 2011; Durstewitz et al., 2023). Traditional LVMs often assume a fixed dynamical system, yet real neural systems—such as the brain—continuously evolve due to learning (Brenner et al., 2022). This presents a challenge: standard modeling approaches fail to account for these gradual transformations.

In real-world applications, we cannot directly observe the latent variables governing neural activity—only their emissions are available. Thus, we must infer the latent structure using the emissions. Multiple models may achieve a high likelihood on these emissions, but **which models truly reflect the brain's latent parameters**? High likelihood alone is insufficient, as models with similar likelihood scores can still encode vastly different underlying dynamics (Pei et al., 2021; Morcos et al., 2018). Our goal is to gain a better understanding of latent models through their performance on data from a changing "ground truth," i.e., the brain, and develop a structured framework to evaluate student models in terms of their latent similarity to the ground-truth teacher model.

Hidden Markov Models (HMMs)

A common LVM used to model brain activity is the Hidden Markov Model (HMM), which is defined by the tuple: $(p(h_k \mid h_j), \pi, \{\mu, \Sigma\})$. Where: π is the initial state probability vector, i.e. what is the probability to start from each state, h_k represents a latent state, with transitions governed by the transition matrix:

$$A_{kj} = p(h_k|h_j),\ 1 \le \forall\, k,j \le Number\ of\ states.$$

$\{\mu, \Sigma\}$ represents the emissions mean vector and covariance matrix of the emissions, which together define the nature of the emissions, i.e. the LVM's output.
HMMs allow us to approximate the dynamics of neural circuits at a given time. However, they assume static parameters and do not account for changes that occur as the brain learns (Durstewitz et al., 2023). Capturing time-dependent latent structures requires a more flexible approach (Brenner et al., 2022).

Our research explores whether different training procedures lead to fundamentally different internal representations, even when models achieve similar likelihoods. This raises key questions: Do different curricula shape models in distinct ways? How well do trained models adapt to new, unseen teacher dynamics? Can we visualize the space of learned models to uncover hidden structural differences? By systematically analyzing trained HMMs, we seek to identify which models reflect meaningful aspects of the brain's underlying processes and which simply fit the observed emissions (Sorscher et al., 2022; Maheswaranathan et al., 2019).

The goal is to collect as much relevant data as possible across different training methodologies over time to understand latent

behavior and draw generalizable conclusions about the structure and adaptability of these models.

RELATED WORK

Our work lies at the intersection of interpretability and generalization in latent variable models (LVMs), with a particular focus on structured evaluation beyond likelihood-based metrics. Traditional approaches to evaluating LVMs, particularly in neuroscience, often rely on co-smoothing techniques—where latent representations are used to predict held-out neural signals given held-in ones from the same trial. This approach, first introduced in Macke et al. (2011) and Pandarinath et al. (2018), has become a popular metric for assessing predictive utility in neural datasets. Pei et al. (2021) formalized this approach by curating multiple primate datasets and establishing a benchmark that inspired advances in predictive modeling, including ensemble methods (Keshtkaran et al., 2022) and models with geometrically constrained latent spaces (Perkins et al., 2023).

While these methods show strong empirical performance, they rely heavily on prediction accuracy, which can obscure differences in the underlying latent structure. In response, a parallel body of work has focused on validating LVMs using synthetic data where ground truth is known (Brenner et al., 2022; Durstewitz et al., 2023; Sedler et al., 2022; Koppe et al., 2019). This enables direct evaluation of how well learned latent dynamics match the true generative process. Our work contributes to this latter line of research, using a synthetic student-teacher HMM setup to probe model interpretability through decoding asymmetries and co-smoothing patterns.

Most relevant to our framework is the recent study in Dabholkar and Barak (2024), which critiques the standard co-smoothing metric for failing to disambiguate models with extraneous latent dynamics. Their proposed few-shot prediction approach uses limited data to test whether latent representations generalize well to held-out channels. Our framework complements and extends this idea: instead of few-shot prediction within a single model, we evaluate cross-decoding performance across a population of student models trained on different perturbations of teacher emissions. This allows us to disentangle the role of training dynamics and model capacity in shaping learned representations—particularly in distinguishing models that reflect true dynamics from those that merely overfit emissions.

Moreover, our approach echoes recent work in representation comparison (Maheswaranathan et al., 2019; Morcos et al., 2018), where latent similarity across models is assessed using symmetric metrics like Canonical Correlation Analysis (CCA). In contrast, we apply asymmetric decoding-based comparisons, using directional decoding success (e.g., T→S vs. S→T) as a proxy for model interpretability and reliability. This provides insights into which models encode information that is both minimal and generalizable.

Finally, by training students on emissions from multiple teachers across epochs, our method introduces a form of implicit regularization. This strategy, though conceptually related to ideas in few-shot learning and meta-learning (e.g., Sorscher et al., 2022), is adapted here for structured temporal models and offers a new angle for evaluating LVMs in time-series contexts.

MATERIALS AND METHODS

Training a Hidden Markov Model (HMM) involves learning parameters that maximize the probability of observed sequences $v_{1:T}$. The process begins with sampling a hidden state $h_1 \sim \pi$, $P(h_1) = \pi$, and generating transitions via the transition matrix $A = p(h_2|h_1)$, and so on for $h_3, h_4, \ldots h_T$. Observations are drawn from $p(v_t|h_t)$, and governed by emission parameters $\{\mu, \Sigma\}$. Optimization algorithms, such as gradient descent, are used to find parameters $(A_{student},\ \pi_{student},\ \{\mu, \Sigma\}_{student})$ that maximize the joint likelihood of the observed emissions, $p(v_{1:T} = (v'_1, v'_2, \ldots v'_n))$.

To explore the behavior of the models, we adopt a student-teacher paradigm in which a known HMM teacher model, T_1, represents the "ground truth" that one cannot access, i.e. the brain, and the students are randomly or artificially initialized, and trained on the emissions of the teacher, to see which ones perform better, and observe whether or not their behavior could be distinguished. We repeat training with various random initializations to generate diverse student populations per curriculum.

However, as mentioned in the introduction, a core challenge with latent variable models is that high likelihood on emissions does not imply that the learned latent dynamics reflect the true generative structure. Instead, as shown in Figure 1, certain transition matrices can decode another model's emissions well if their dynamics encompass those of the source. If both models can decode each other well, we can infer a similar latent structure. To further illustrate the limits of likelihood-based evaluation, we also construct artificial students with high performance, identical to the ground truth, on emissions by adding a "ring" structure that expands the hidden space while preserving emission likelihood, as seen in the example in Figure 2. The students then decode the ground truth well, since, it contains the inner dynamics of the ground truth, but the ground truth is missing information to decode the student, which leads to only a 1 sided low decoding error, similar to the example in Figure 1. In addition, the way the student is constructed leaves the emissions unchanged, leaving to the same likelihood as the teacher. This allows us to artificially create "bad models" (Dabholkar and Barak, 2024) that are undetectable by merely looking at their emissions likelihood..

So far, we discussed the initialization of teacher and student models, but we did not factor in the changes over time. To do so, we apply random perturbations to the initialized teacher, i.e. the ground truth T_1. The perturbation changes the latent dynamic of the ground truth, in addition to the changes in the emissions that occur during each time step, i.e. when the model transitions from

h_k to h_{k+1}, representing the long-term changes that occur in a learning brain. In this study, $T_2, T_3 \dots T_k$ will represent the ground truth at different time steps, Such that $T_2 = T_1 + \epsilon$, $T_3 = T_2 + \epsilon$, and so on… Respectively, $S_1, S_2, \dots$ will represent students trained on data from $T_1, T_2, T_3 \dots T_k$. The students are then continued to be trained with several different curricula. $S_1, S_{11}, S_{12}, S_2, S_{22}$ will represent *trained* student models that were trained again respectively on $T_1, T_1 \to T_1, T_1 \to T_2$ and so on... This can be generalized to more complex curricula, $S_1, S_{11}, S_{12}, S_2, S_{22} \dots S_{12\dots k}$, which we did not experiment with in this study.

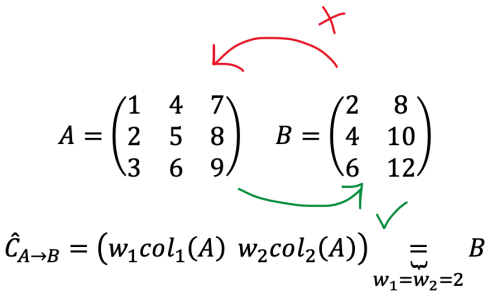

Figure 1: Example of how inner dynamics can yield a 1-sided good performance on decoding. These particular matrices don't necessarily represent any real dynamic and are used only for the sake of demonstrating the functionality of decoding. Here we can see that A's 2 left columns scaled by 2, $w_1 = w_2 = 2$, equal to B. Meaning, that A's dynamic encompasses B, so it will decode it well, but not vice versa, since B doesn't hold within it any information about A's right column.

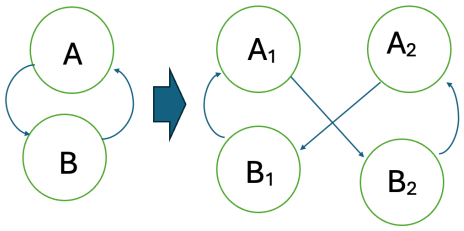

Figure 2: An example of an artificially created student, with a wrong latent structure but with a high likelihood score. The added ring,

In this study, the training is done using expectation maximization fitting. The teachers have 5 possible outputs, i.e. emission dimension of size 5, and 10 latent states. Initially, the number of states was decided based on related work (Kabir and Barak, 2024) but later adjusted to achieve conditions where it would be harder to detect a co-smoothing difference between the students and the teachers, in addition to choosing parts that will allow for a faster compute with less resources. In each of the initial runs, 7 models were used – 2 randomly initialized with the same number of states as the teachers, 2 artificially created by adding 1-2 rings (one with 20 states and the other with 30), and 3 deep copies of the teachers, which we will consider as the "good" models. To check the performance of the models, a normalized co-smoothing score was used, i.e. the log-likelihood of the models to output the same emissions as the ground truth. The normalization uses a baseline, an HMM model with one state that returns the average of the emissions. This value will be named B. In the same manner, the value for the likelihood of the student will be named S, and for the teacher T. The normalized likelihood should then be:

$$L_T(S) = \text{ReLU}\left\{\frac{S - B_T}{T - B_T}\right\}$$

Equation 1: A normalized likelihood (co-smoothing) score that provides a score between 0 and 1. T is the teacher's likelihood on its own data, which is guaranteed to give the highest likelihood, B is the minimum likelihood we are willing to accept, below that the models are of no interest to us, and S is the likelihood of the student model we check.

RESULTS

Using the methods mentioned in the previous section, we start with 3 teachers, training the students in a student-teacher framework. First, the models are trained using emissions from T0, Figures 3-4. As can be seen in Figures 3 and 4A, the randomly initialized students (models 1 and 3) begin with a low likelihood which rises during the training until convergence. The copies of the ground truth, i.e., the "good" models which are copies of T0, T1, and T2, start with the highest possible co-smoothing score on the test emissions, Figures 3 and 4B, and overfit immediately with the train co-smoothing score rising despite the test emissions score falling a bit down. The same behavior, as expected, is observed by the artificially made models, models 0 and 2 in Figure 3 and 4A. The results confirm the explanation in the previous section, regarding the fact that these artificially made models would possess a high T→S decoding error but a high co-smoothing score and low S→T decoding error.

From Figures 4A and 4B one can notice that the models overfit almost immediately to T0, resulting in a rise in training data likelihood but a lowering in the test dataset. In addition, due to the training parameters which use a relatively large number of time stamps, trials, and iteration, convergence happens right after 1 epoch.

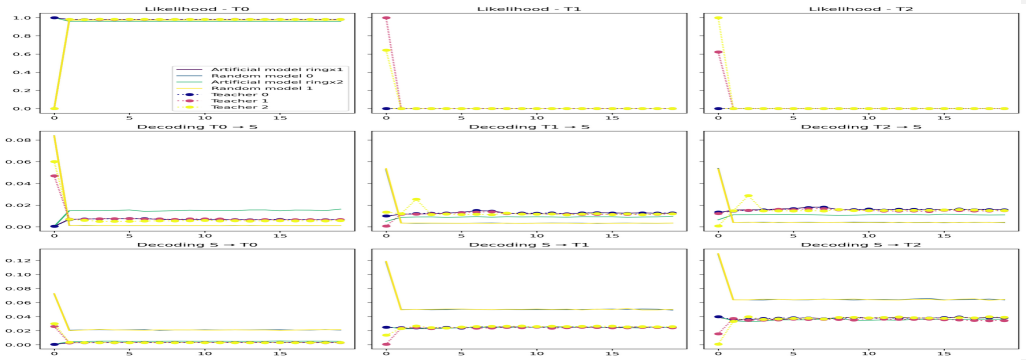

Figure 3: The figure shows the students' and the teachers' (ground truth) performance during the fitting to the emissions (train data) of T0. In each figure, the left column shows performance on T0's data, the middle T1, and the right T2. The first row shows the likelihood over time, the middle shows the trained model decoding score of the ground truth, and the lower shows the ground truth's decoding of the trained models. This was run using the following parameters: ***20 Epochs 250 Iterations per epoch 100 Timesteps 1000 Trials and fitted with expectation maximization.***

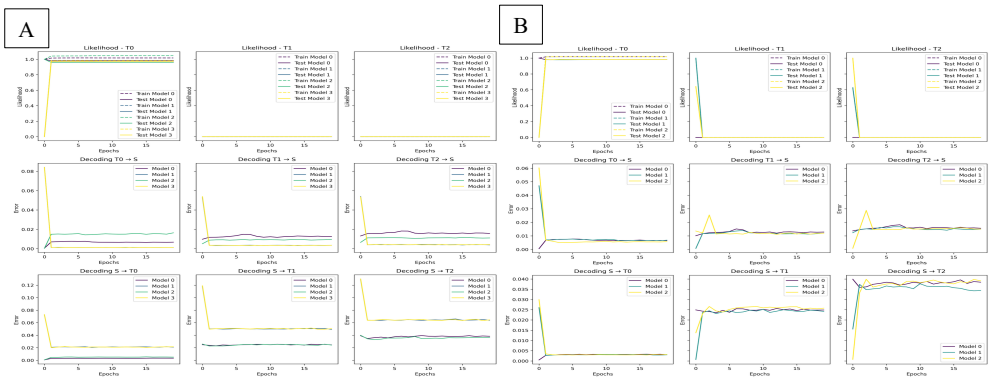

Figure 4. Same as Figure 3 only here we separated the artificially created and randomly initialized students from the teachers' copies to see the graphs more clearly, and see how models that are similar to the ground truth perform in comparison to each other. Same for the other models. On the right, we have the teachers' copies, and on the left the rest.

After the models converged, both the true params' copies and the student models, they are trained again on T1 and T2 *separately*, as shown in Figures 5-8. This allows us to compare the performance of models before and after converging.

In Figure 5-6 we can see that fitting to T1 creates a little variability, showing a difference between the randomly initialized students and the rest of the models.

As can be seen in Figure 5 the students have the same likelihood as the teachers on T1 and T2, except for one – student 3, which has a lower likelihood. This was not revealed in the initial training on T0, as can be seen in Figure 3.

The figures show more similarity between T2 and T1, perhaps due to smaller perturbation. This can be inferred from Figures 5 and 7 where fitting the models to T1 leads to a non-zero likelihood on T2, and vice versa. This indicates similarities between the 2 models, and a bigger difference between them and T0, where the left column in the figures shows the likelihood score being nullified, according to equation 1, meaning less similarity to T0.

TODO – maybe add the graphs for the run with less data? Final ppt ~21st slide, **20Epochs_30Iter_50Timesteps_100Trials.**

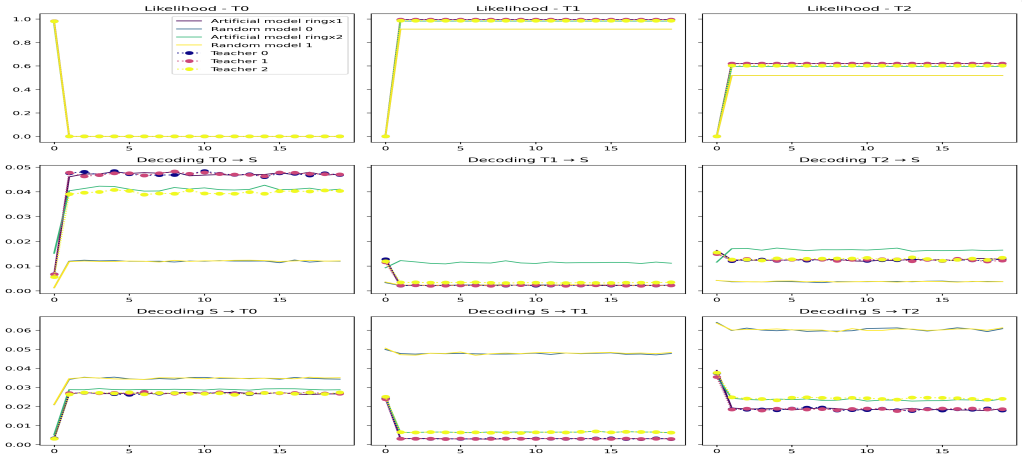

Figure 5: The figure shows the students' and the teachers' (ground truth) performance during the fitting to the emissions (train data) of T1 **after** being trained on T0. In each figure, the left column shows performance on T0's data, the middle T1, and the right T2. The first row shows the likelihood over time, the middle shows the trained model decoding score of the ground truth, and the lower shows the ground truth's decoding of the trained models. *This was run using the following parameters:* ***20 Epochs 250 Iterations per epoch 100 Timesteps 1000 Trials and fitted with expectation maximization.***

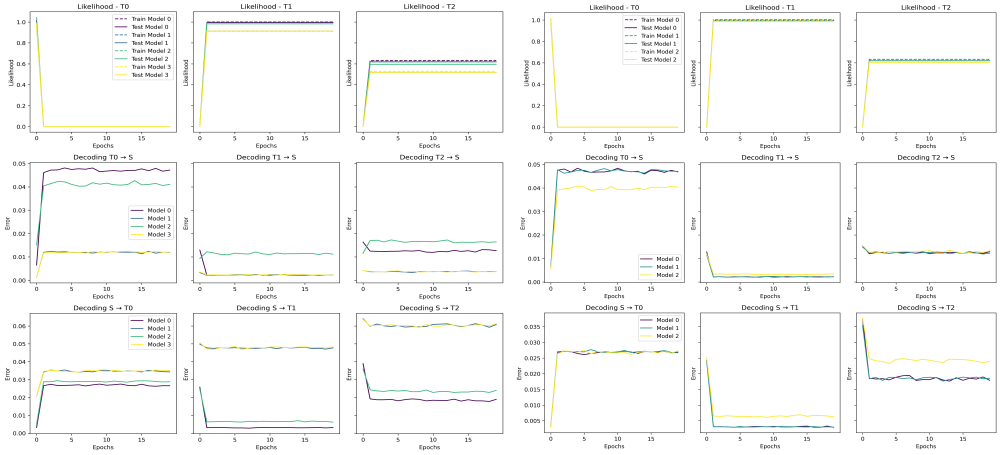

Figure 6. Same as Figure 5 only here we separated the artificially created and randomly initialized students from the teachers' copies to see the graphs more clearly and see how models that are similar to the ground truth perform in comparison to each other. Same for the other models. On the right, we have the teachers' copies, and on the left the rest.

<Trained on T2> - Run hmmST with first if == false. Set filename with something like ***20 Epochs 250 Iterations per epoch 100 Timesteps 1000 Trials and fitted with expectation maximization***

Figure 8. Exactly like Figure 5, with the same parameters,, except this time the models were trained on T2 instead of T1

Since these results yielded somewhat expected behavior, without much that could be concluded, another approach was explored in this study – to train each epoch on data from a differently perturbed teacher, Figures 9-10, and see whether some models can be "thrown off" from the indistinguishable converging pattern. This approach was attempted since some signs of variability were seen in Figures 5 and 8.

This study experimented with different parameters, some making the models transition too fast between the teachers with smaller datasets and fewer iterations, leading to a bad likelihood as can be seen in Figure 11, or too many iterations and a large dataset like in Figures 9-10, leading to the same co-smoothing score for all models.

With balanced parameters, the models were able to be trained to see a rise in likelihood, but not a convergence, before being fitted to a different teacher model. This leads to an additional variability that allows us to distinguish between the copies of similar teachers, T1 and T2, and the rest, as can be seen in Figure 12. At the first fitting to T1 and T2, the copies of T1 and T2 achieve the highest likelihood.

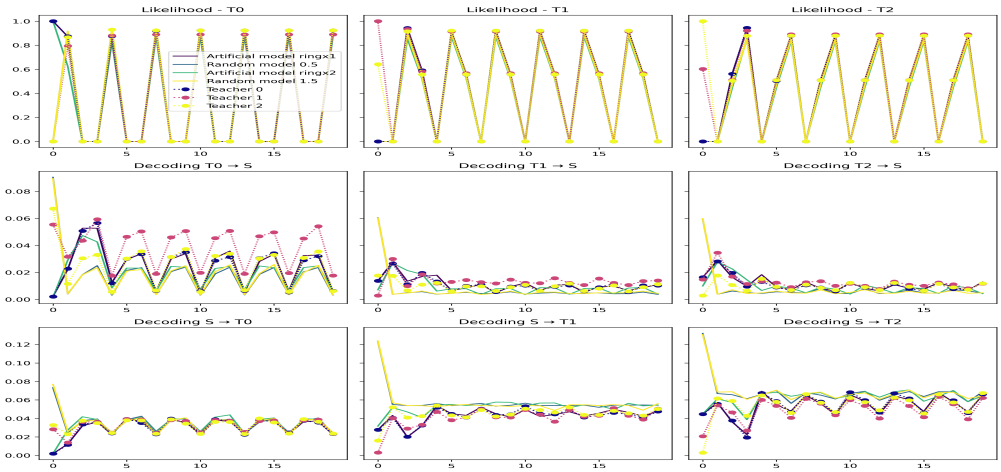

Figure 9: Each epoch, the models were fitted to a different teacher- T0, T1, or T2. This was run using the following parameters: ***20 Epochs 250 Iterations 100 Timesteps 100 Trials***

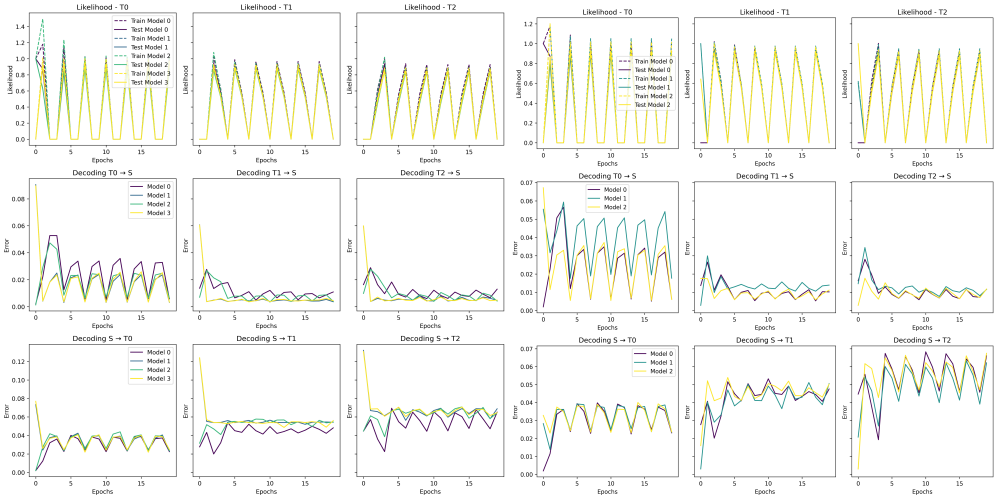

Figure 10: Same as Figure 9 only here we separated the artificially created and randomly initialized students from the teachers' copies to see the graphs more clearly and see how models that are similar to the ground truth perform in comparison to each other. Same for the other models. On the right, we have the teachers' copies, and on the left the rest.

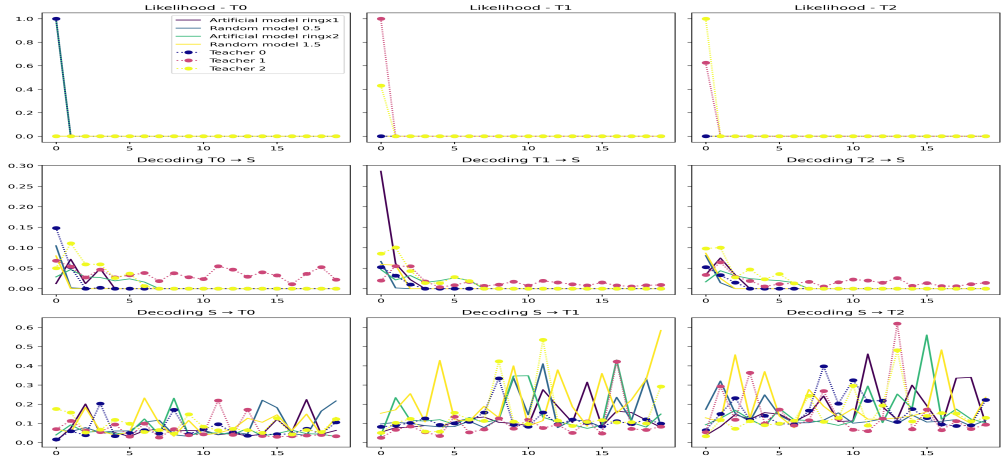

Figure 11: Each epoch, the models were fitted to a different teacher- T0, T1, or T2. This was run using the following parameters: ***20Epochs 10Iter 10Timesteps 10Trials***

Figure 11: Each epoch, the models were fitted to a different teacher- T0, T1, or T2. This was run using the following parameters: ***20Epochs 30Iter 50Timesteps 10Trials***

DISCUSSION

By training the student models on data from teachers at different times, i.e. different perturbations, this study reveals whether students trained on smoother transitions generalize better to unseen dynamics, or whether their decoding success is narrowly tuned to their training distribution. This framework allows us to distinguish between models that capture underlying dynamics and those that merely overfit emission patterns. In doing so, we move beyond likelihood as a metric and toward a deeper understanding of how training procedures shape learned representations.

As previously mentioned, high co-smoothing can be achieved by different models with different latent structures, similar to decoding from the trained model to the ground truth. This introduces ambiguity in evaluation metrics based solely on output matching, reinforcing the need for a deeper analysis of the latent structure and decoding asymmetry.

One of the hypotheses of this study was that the ground truth parameters are more likely to adjust quicker to the emissions of perturbed teachers in comparison to other models. This was confirmed in Figure 9, where with the correct parameters a clear distinction can be seen between the students and the teachers. Since achieving high co-smoothing can result from multiple configurations, the ability of the ground truth model to recover teacher dynamics more rapidly becomes a useful criterion for evaluation.

Another hypothesis is that bigger models are less likely to achieve higher T→S scores. They might contain the dynamics of the teachers, as seen in the case of the artificial ring students where high T→S is achieved, but not the other way around—since the teacher is less likely to have access to all the information encoded in the larger model's latents.

This study highlights the nuances of decoding performance in student-teacher Hidden Markov Model (HMM) settings, with a particular focus on how varying training regimes impact generalization and model interpretability. Training each epoch on emissions from differently perturbed teachers acts as a form of implicit regularization, encouraging the student to internalize stable dynamics rather than overfit to specific emission patterns. When the perturbations are balanced—neither too similar nor too divergent—this setup enables clearer distinctions between models that may otherwise appear equally performant under traditional metrics like likelihood. Larger models, while slower to adapt, often demonstrate stronger filtering behavior, suggesting a closer alignment with the underlying generative process. Although some students remain indistinguishable within this framework, the results overall support the hypothesis that carefully controlled teacher variability can surface meaningful differences between similarly "good" models. This work offers a principled approach to dissecting learned representations in time-series models and advances our understanding of how training procedures shape latent inference.

REFERENCES

Bauwens, L., & Veredas, D. (2004). The stochastic conditional duration model: A latent variable model for the analysis of financial durations. *Journal of Econometrics*, *119*(2), 381–412.

Brenner, M., Koppe, G., & Durstewitz, D. (2022). Multimodal teacher forcing for reconstructing nonlinear dynamical systems. *arXiv preprint* arXiv:2212.07892.

Commented [MR2]: References should follow Journal of Neuroscience style.

Use PubMed abbreviations for journals where available (please check PubMed for journal titles before submitting article).

List volume(issue number):page numbers for journal titles.

Author list construction:
Author A, Author B (Year) Source title…
-Single comma following author initials. No additional commas, periods, "&", "and" in the author list or separating author list from year and year from source title.

Dabholkar, K., & Barak, O. (2024). *When predict can also explain: Few-shot prediction to select better neural latents* (arXiv:2405.14425). arXiv. http://arxiv.org/abs/2405.14425

Durstewitz, D., Koppe, G., & Thurm, M. I. (2023). Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nature Reviews Neuroscience*, *24*(11), 693–710.

Maheswaranathan, N., Williams, A., Golub, M., Ganguli, S., & Sussillo, D. (2019). Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in Neural Information Processing Systems*, *32*.

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, *31*.

Pei, F., Ye, J., Zoltowski, D., Wu, A., Chowdhury, R., Sohn, H., O'Doherty, J., Shenoy, K. V., Kaufman, M., Churchland, M., Jazayeri, M., Miller, L., Pillow, J., Park, I. M., Dyer, E., & Pandarinath, C. (2021). Neural latents benchmark '21: Evaluating latent variable models of neural population activity. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, *1*. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/979d472a84804b9f647bc185a877a8b5-Paper-round2.pdf

Perkins, S. M., Cunningham, J. P., Wang, Q., & Churchland, M. M. (2023). Simple decoding of behavior from a complicated neural manifold. *bioRxiv*, 2023–04.

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, *119*(43), e2200800119. https://doi.org/10.1073/pnas.2200800119

Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., & Sahani, M. (2011). Empirical models of spiking in neural populations. *Advances in Neural Information Processing Systems*, *24*. https://proceedings.neurips.cc/paper_files/paper/2011/file/7143d7fbadfa4693b9eec507d9d37443-Paper.pdf

---

Not used but maybe useful?

Duncker, L., & Sahani, M. (2021). Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings. *Current Opinion in Neurobiology*, *70*, 163–170. https://doi.org/10.1016/j.conb.2021.10.014

Maybe:

Barak O. Recurrent neural networks as versatile tools of neuroscience research. Curr Opin Neurobiol. 2017 Oct;46:1-6. doi: 10.1016/j.conb.2017.06.003. Epub 2017 Jun 29. PMID: 28668365.

Mathematical models of learning and what can be learned from them—PubMed. (n.d.). Retrieved August 4, 2024, from https://pubmed.ncbi.nlm.nih.gov/37043892/

https://neurallatents.github.io/
Stanford document about HMM

,