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

**Arnold Cheskis**1**, Kabir Dabholkar**2**, & Omri Barak**3

*1Electrical and Computer Engineering Faculty, Technion - Israel Institute of Technology*

*2**Department of Mathematics, Technion - Israel Institute of Technology*

*3Rappaport Faculty of Medicine and Network Biology Research Laboratory,* *Technion - Israel Institute of Technology*

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 ....

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. Traditional LVMs assume a fixed dynamical system, yet real neural systems, such as the brain, continuously evolve due to learning. 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 (Dabholkar & Barak, 2024). 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: . Where: π is the initial state probability vector, i.e. what is the probability to start from each state, represents a latent state, with transitions governed by the transition matrix

.

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, which do not account for changes occurring as the brain learns. Capturing *time-dependent latent structures*\*\* requires a more flexible approach.

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.

The goal is to get as much relevant data as possible with different methods of training the models over time, to understand latent behavior, draw conclusions,

**RELATED WORK**

Mention few shots?

Talk about krelated work, also this:  
Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed Chowdhury, Hansem Sohn, Joseph

O' Doherty, Krishna V Shenoy, Matthew Kaufman, Mark Churchland, Mehrdad Jazayeri,

Lee Miller, Jonathan Pillow, Il Memming Park, Eva Dyer, and Chethan Pandarinath.

Neural latents benchmark ‘21: Evaluating latent variable models of neural population

activity. In J. Vanschoren and S. Yeung, editors, Proceedings of the Neural Information

Processing Systems Track on Datasets and Benchmarks, volume 1. Curran, 2021. URL

https://datasets-benchmarks-proceedings.neurips.cc/paper\_files/paper/

2021/file/979d472a84804b9f647bc185a877a8b5-Paper-round2.pdf.

**MATERIALS AND METHODS**

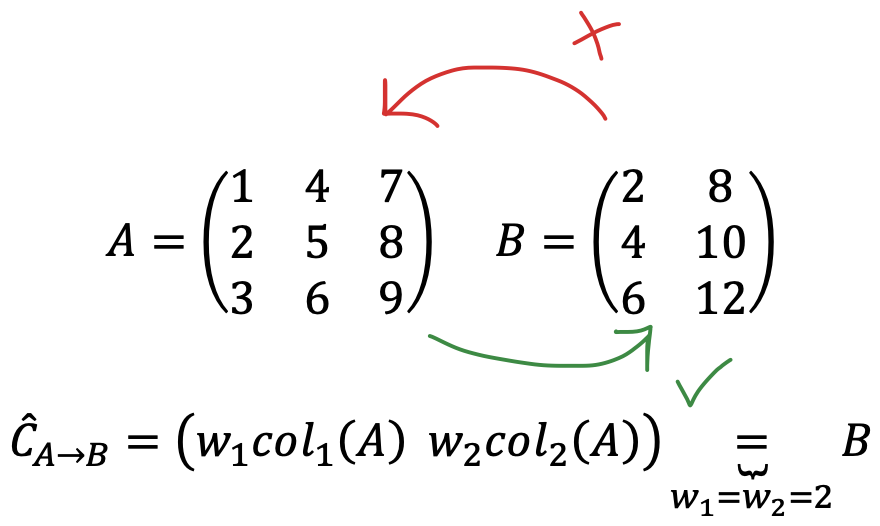
Training a Hidden Markov Model (HMM) involves learning parameters that maximize the probability of observed sequences ​. The process begins with sampling a hidden state , and generating transitions via the transition matrix and so on for. Observations are drawn from , and governed by emission parameters . Optimization algorithms, such as gradient descent, are used to find parameters (, , ) that maximize the joint likelihood of the observed emissions, .

To explore the behavior of the models, we adopt a student-teacher paradigm in which a known HMM teacher model, 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.

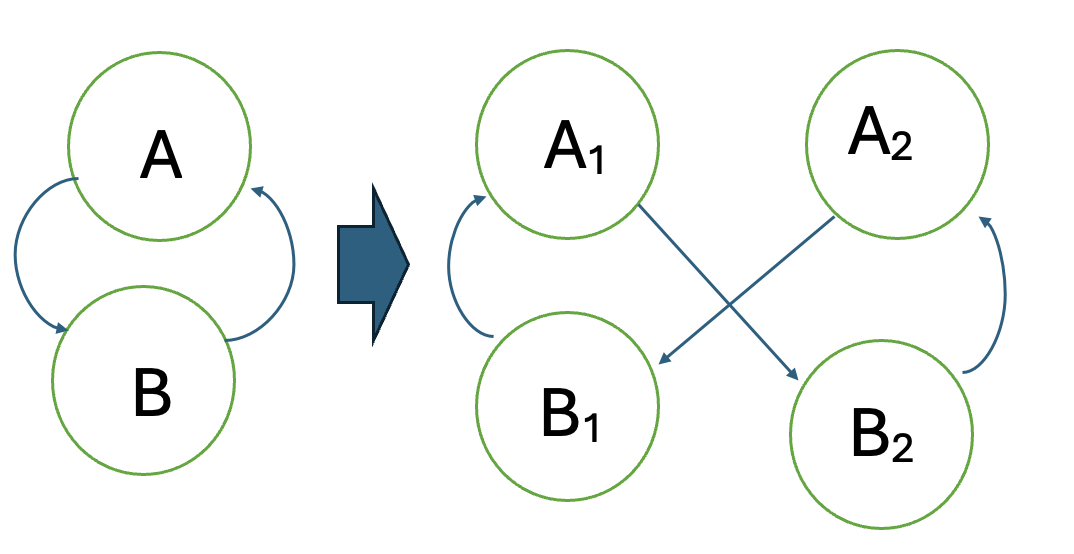
However, as mentioned in the introduction, a core challenge with latent variable models is that high likelihood on emissions does not imply that the learned latent dynamics reflect the true generative structure. Instead, as shown in Figure 1, certain transition matrices can decode another model’s emissions 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” (\*) 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 . 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 to , representing the long-term changes that occur in a learning brain. In this study,  will represent the ground truth at different time steps, Such that = + ,  
 = , and so on… Respectively, will represent students trained on data from . The students are then continued to be trained with several different curricula. will represent *trained* student models that were trained again respectively on and so on... This can be generalized to more complex curricula, which we did not experiment with in this study.

The fitting is done using EM



*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, , 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,*

## 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 posses a high T🡪S decoding error but a high co-smoothing score and low S🡪T decoding error.

**A graph of students and teachers

AI-generated content may be incorrect.**

*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.*

After the models converged, both the true params and the student models, they were trained again on T1 and T2 *separately*, as shown in Figures 5-6. This allows us to compare the performance of models before and after converging.

…

Since these results yielded somewhat expected behavior, without much that could be concluded, another approach was explored in this study – each epoch train on data from a differently perturbed teacher, Figures 7-9, and see whether some models can be “thrown off” from the indistinguishable converging pattern.

…

From Figures 3-4 one can notice that the models overfit almost immediately to T0

In Figure 5-6 teachers T1\* we can see that fitting to T1 causes T2 itself and T1 in a worse way, but T0 decodes T2 better.

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

A group of graphs with numbers

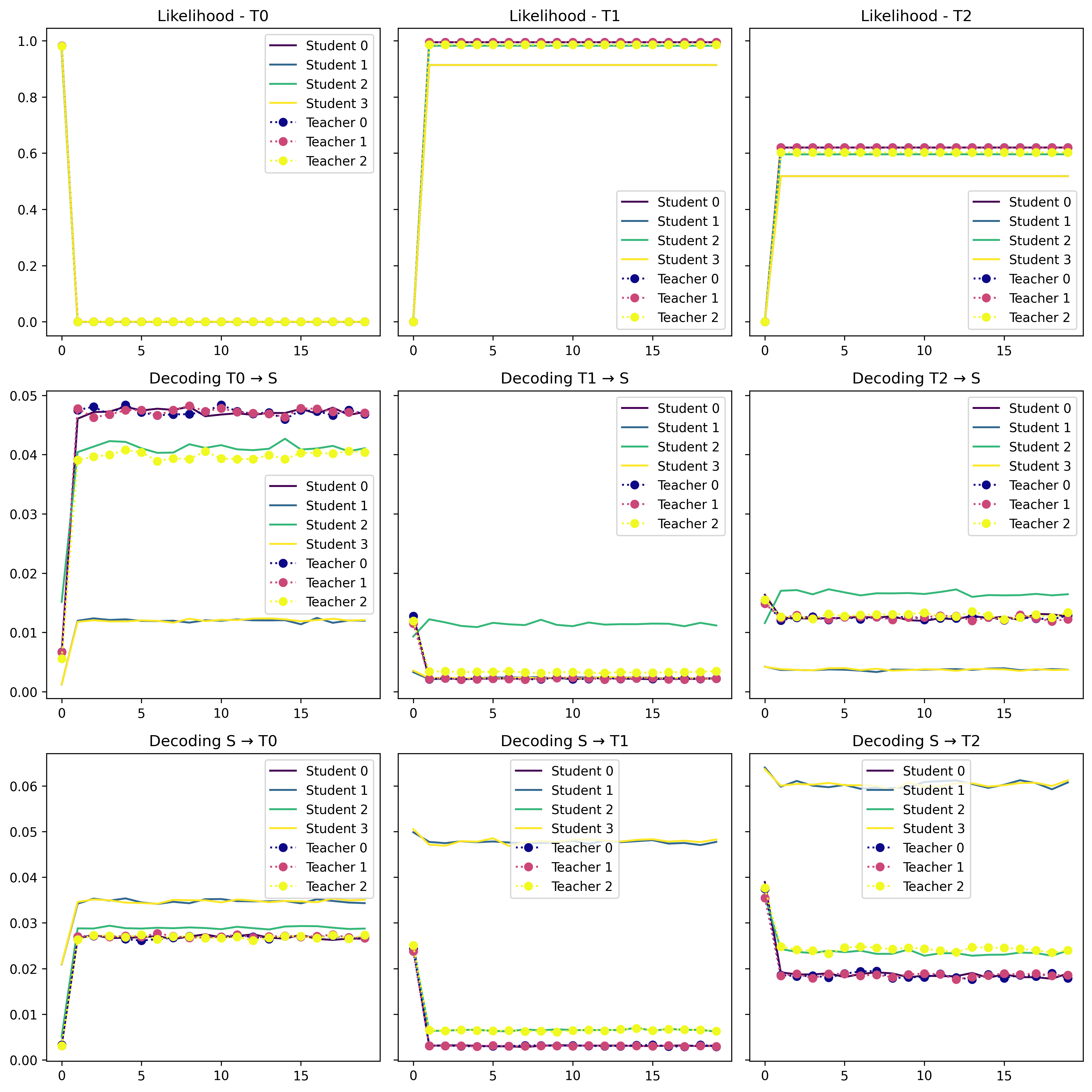
AI-generated content may be incorrect.A group of graphs with text and numbers

AI-generated content may be incorrect.

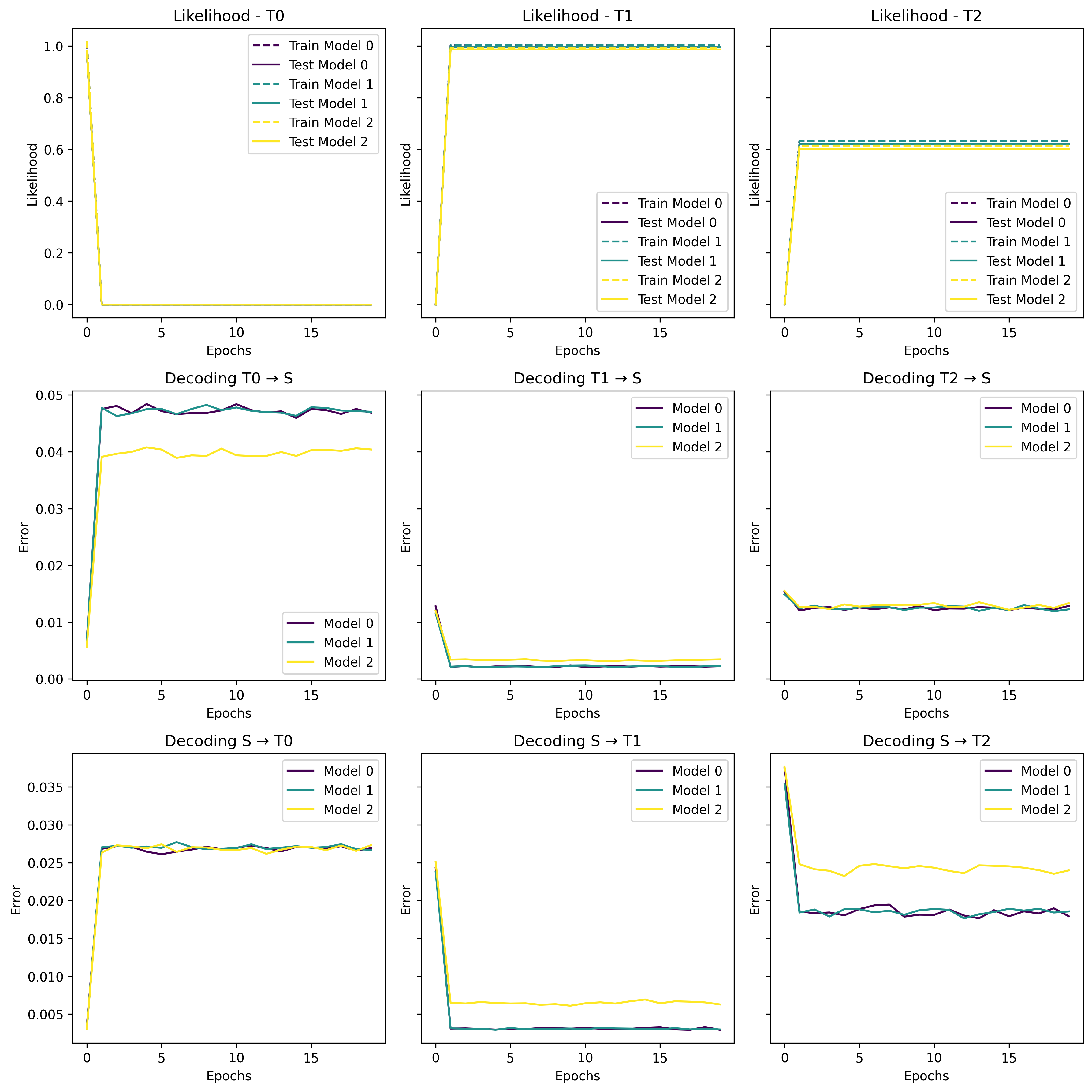
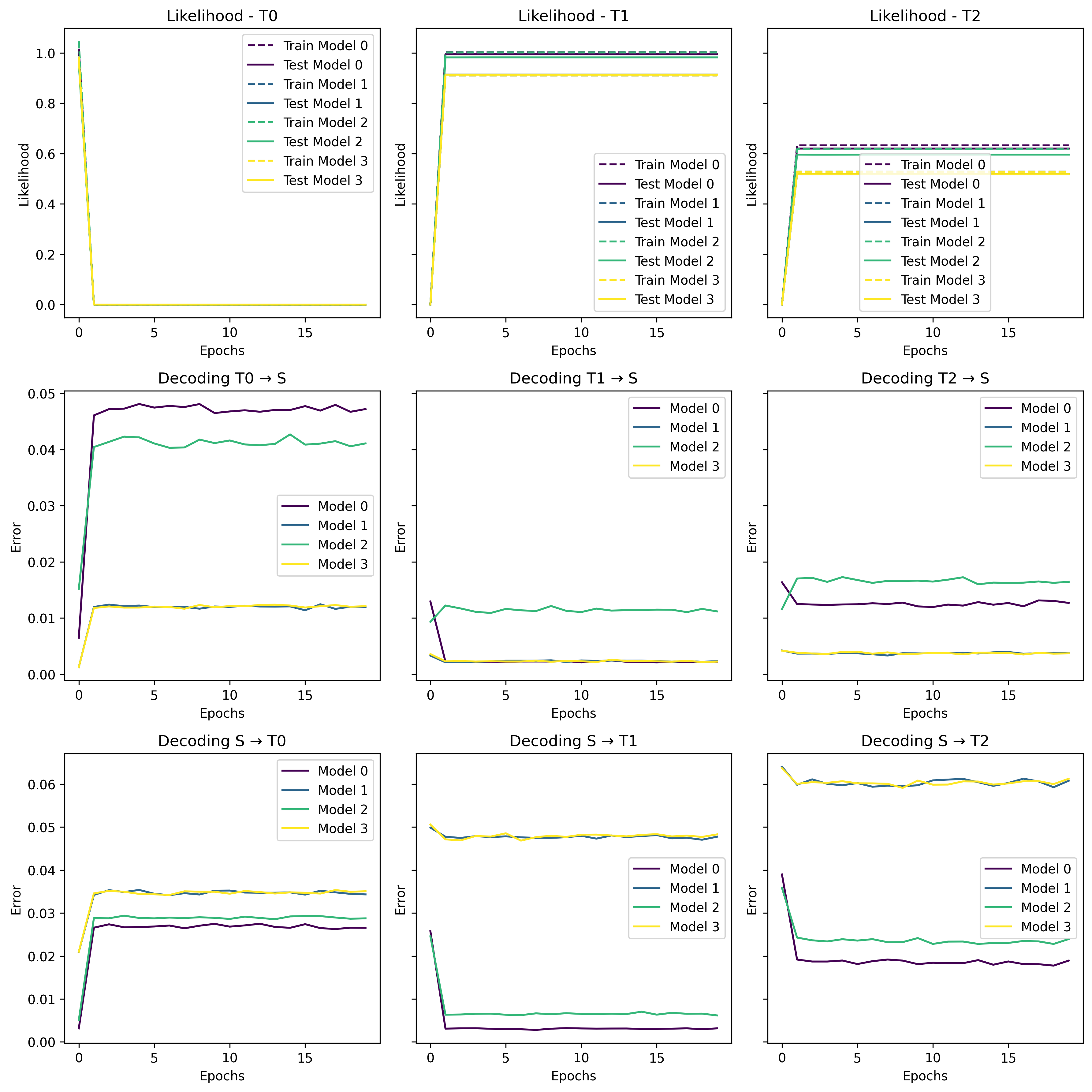
A

B

*Figure 4. Same as figure x-1 only here we separated the students and the teachers to see the graphs more clearly. On the left, we have the students and on the right, we have the teachers*



*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.



*Figure 6. Same as figure x-1 only here we separated the students and the teachers to see the graphs more clearly. On the left, we have the students and on the right, we have the teachers*

Experimenting with different parameters, eventually, the models were able to be trained to see a rise in likelihood, but not a convergence before being fitted to the next teacher model.

## 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, similarly, to decoding from the trained model to the ground truth. Decoding f

One of the hypotheses of this study was that the ground trurth 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 …

###### REFERENCES

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>

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>

*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/>

Pei, F., Ye, J., Zoltowski, D., Wu, A., Chowdhury, R. H., Sohn, H., O’Doherty, J. E., Shenoy, K. V., Kaufman, M. T., Churchland, M., Jazayeri, M., Miller, L. E., Pillow, J., Park, I. M., Dyer, E. L., & Pandarinath, C. (2022). *Neural Latents Benchmark ’21: Evaluating latent variable models of neural population activity* (arXiv:2109.04463). arXiv. <https://doi.org/10.48550/arXiv.2109.04463>

<https://neurallatents.github.io/>

Stanford document about HMM  
  
*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.

,