Predictive variational autoencoder for learning robust representations of time-series data

Julia H. Wang
Cold Spring Harbor Laboratory School of Biological Sciences
Cold Spring Harbor Laboratory
Cold Spring Harbor, New York, USA
julwang@cshl.edu

Dexter Tsin
Princeton Neuroscience Institute
Prineton University
Princeton, New Jersey, USA
dexter.tsin@princeton.edu

Tatiana A. Engel
Princeton Neuroscience Institute
Prineton University
Princeton, New Jersey, USA
tatiana.engel@princeton.edu

Abstract

Variational autoencoders (VAEs) have been used extensively to discover low-dimensional latent factors governing neural activity and animal behavior. However, without careful model selection, the uncovered latent factors may reflect noise in the data rather than true underlying features, rendering such representations unsuitable for scientific interpretation. Existing solutions to this problem involve introducing additional measured variables or data augmentations specific to a particular data type. We propose a VAE architecture that predicts the next point in time and show that it mitigates the learning of spurious features. In addition, we introduce a model selection metric based on smoothness over time in the latent space. We show that together these two constraints on VAEs to be smooth over time produce robust latent representations and faithfully recover latent factors on synthetic datasets.

1 Introduction

One central goal of neuroscience is to understand the link between neural activity in the brain and behavior. Both neural activity and behavior are often measured via high-dimensional observables but are governed by low-dimensional latent factors. Identifying these factors from data can reveal the correspondence between the neural activity and behavior leading to new scientific hypotheses (Cunningham and Yu [2014]).

This latent to observable mapping can be modeled as a generative process in variational autoencoders (VAE), and many have leveraged VAEs to uncover latent factors from animal behavior and neural recordings (Pandarinath et al. [2018], Zhou and Wei [2020], Pei et al. [2021], Peterson et al. [2023], Luxem et al. [2022]). However, to reliably interpret latent representations, we need to ensure that they capture some true underlying structure of the data and do not contain spurious features resulting

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

from noise. For example, if different instances of a model produced different representations from the same data, it would be impossible to meaningfully interpret these representations. Thus, robustness and reproducibility are prerequisites for interpretability.

However, standard VAEs do not produce robust representations (Locatello et al. [2019], Keshtkaran and Pandarinath [2019], Radhakrishnan et al. [2019]), indicating that these models can learn spurious features unrelated to underlying structure of the data. There are two possible causes of how VAEs may learn spurious features. First, the VAE might learn a functional mapping that is irrelevant to the low-dimensional structure but sufficient for reconstructing the data (Keshtkaran and Pandarinath [2019], Radhakrishnan et al. [2019]). The reconstruction goal of VAEs only requires the model to separate dissimilar points but does not incentivize the model to keep similar points close in the latent space. Thus, there may be several local optima in the loss landscape at which the VAE can map points in the latent space to points in the original space without learning the correct low-dimensional structure.

Second, VAE can overfit to noise in the data by trying to separate data points based on the noise features. Traditional practice to avoid overfitting is regularization, which involves selection of hyperparameters by optimizing the model's performance on a held-out validation dataset. However, flexible machine learning models with many parameters can generalize well on unseen data despite learning to interpolate perfectly through noise in the training data (Belkin et al. [2018], Bartlett et al. [2020]). In these instances, many models can generalize equally well but have different learned features and therefore interpretations (Genkin and Engel [2020]). This type of overfitting is not an issue if generalization performance is the only goal, but it undermines the interpretation of the latent representations learned by the model.

Solutions proposed to avoid the learning of spurious features involve 1) adding inductive bias to the model architecture or through regularization (Locatello et al. [2019], Khemakhem et al. [2020]) and 2) using appropriate model selection metrics to choose robust features that are reproducible across different model instances (Duan et al. [2016], Genkin and Engel [2020]). Here we extend these ideas to learning robust low-dimensional representations of time series data with VAEs. We first show that standard VAEs are prone to learning spurious features. We then introduce an inductive bias in the VAE architecture and a model selection metric which both promote smoothness of latent factors over time. The temporal smoothness of latent factors naturally arises from the continuous dynamics inherent in the vast majority of physical and biological processes. Therefore, it serves as a natural prior for time-series data generated by these processes. We show that with these changes, VAEs learn robust representations on synthetic and biological datasets.

2 Related Work

2.1 Learning interpretable representations

Uncovering robust low-dimensional representations of biological data with deep learning models allows for downstream analysis and interpretation to derive new scientific hypotheses. The goal of achieving robust representations directly aligns with efforts in developing identifiable models. A model is *identifiable* if its parameterization of a certain set of observations is unique (Khemakhem et al. [2020]). The standard VAE is not identifiable itself, but some modifications of the VAE do achieve identifiability, for example, through conditioning on an additionally observed auxillary variable (Khemakhem et al. [2020], Zhou and Wei [2020], Mita et al. [2021]).

Learning robust or identifiable representations is distinct from achieving disentanglement in VAEs (Khemakhem et al. [2020]). Disentanglement seeks to create a representation that separates meaningful features of the high-dimensional data into independent generative factors in the latent space. However, the goals of identifiability in representation learning and disentanglement are related, as many of the modifications to VAE architecture and learning paradigms that produce robustness or identifiability are also useful in achieving disentanglement (Mita et al. [2021]).

Various modifications to VAEs have been explored, both in and outside of neuroscience, to achieve more robust and interpetable representations. Some of these variations promote smoothness over time or spatial dimensions (Gulrajani et al. [2016], Klindt et al. [2021]). Weak supervision and self-supervision are also often explored in pursuit of robust representation learning and identifiability. Weak supervision often leverages multiple data modalities for regularization by introducing an

auxillary variable, such as activity of another neuron or behavioral measurements (Mita et al. [2021], Zhou and Wei [2020], Liu et al. [2021], Azabou et al. [2021]). However, not all datasets have associated auxillary variables, so such a method is unsuitable in those cases. Self-supervised learning involves training the model to predict an augmented version of the original data, called a self-supervised label. Doing so ensures that the model learns to preserve features that are shared between the original and augmented data. For example, data augmentation can involve using a semantic-preserving transformation of an image (such as a rotation) (Sinha and Dieng [2021]) or dropping and swapping spikes in a neural activity dataset (Liu et al. [2021]). These kinds of data augmentation are specific to the dataset at hand, and it is generally not obvious which transformations preserve the true latent structure for every data type.

Some non-variational model architectures also seek to learn robust representations of time-series data, such as contrastive predictive coding (Oord et al. [2019]) or prediction of neighboring words (Mikolov et al. [2013]). However, representation learning with predictive coding models has not been studied in terms of identifiability and robustness of representation, nor have these models been systematically compared to VAEs in this area.

2.2 Model selection metrics for interpretability

Even with inductive biases added to promote robustness or identifiability, there is still a need for model selection metrics to avoid overfitting to noise in the training and validation data. A standard metric for model selection in unsupervised settings is cross-validated performance, such as reconstruction loss for VAEs. However, since overparametrized models can achieve high validation performance despite learning spurious features (Belkin et al. [2018], Genkin and Engel [2020]), there is a need for validation metrics that directly assess the robustness of features learned by the model. One approach is to evaluate the model performance in predicting an external variable (Higgins et al. [2017], Pei et al. [2021]); however, this relies on some external label rather than evaluating the quality of latent representations directly. Another approach measures the similarity of latent representations learned by different model instances, as true features should be similar, whereas spurious features are not reproducible (Duan et al. [2020], Genkin and Engel [2020]).

3 Methods

3.1 Datasets

To study how standard VAEs learn spurious features and to test how our approach overcomes this problem, we used synthetic data with known low-dimensional structure, as well as biological neural recording data. Synthetic datasets are useful because we know the ground-truth latent structure used to generate the data, although they may not capture the entire complexity of real biological data.

Hidden Markov Model (HMM) and Gaussian clusters: We generated synthetic data points from a three-state HMM with emissions sampled from 3 clusters mimicking wake, rapid eye movement (REM), and slow wave sleep (SWS) brain states. Each cluster was modeled as a Gaussian distribution in 31-dimensional space, where dimensions represent local field potential (LFP) power in 30 frequency bands and total power of the electromyography (EMG) signal with mean and variance of the Gaussian distributions matched to the biological data. The transition dynamics were also matched to wake, REM and SWS states in biological data. In Fig. 2, we use this dataset without considering time information, essentially treating data-points as independent samples from the 3 high-dimensional Gaussian clusters.

Spiral: We sampled data points uniformly along a two-dimensional spiral, with time index increasing monotonically along the spiral starting from the innermost point. We then nonlinearly embedded these points into 31-dimensional space and added Gaussian white noise.

Visual cortex LFP data: We used combined EMG and LFP data recorded from the visual cortex of a mouse during continuous 24-hour recordings over 12 days (Soltani et al. [2019]). Each data point represents a 2 second time window. For this 2-second time window, we extracted the LFP power in 30 frequency bands and total EMG power to form a 31-dimensional feature vector representing the signals. A subset of data points had expert-provided labels of wake, REM and SWS brain states.

3.2 Time-Neighbor VAE architecture

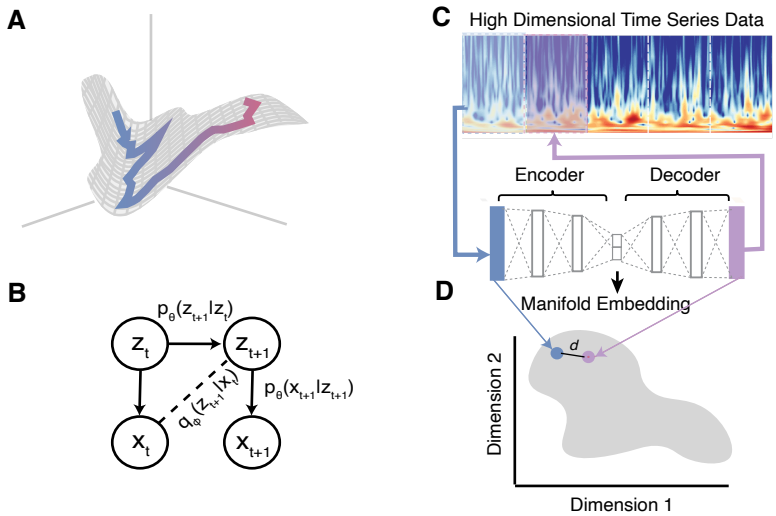

Figure 1: **A**, Generative model. The data arises from latent dynamics (illustrated as a blue-to-pink trajectory) evolving on a low-dimensional manifold (2-D grey surface). This manifold is nonlinearly embedded in a high-dimensional space of observations, in which each axis represent one feature in the data. **B**, Directed graphical model. Solid lines indicate the generative model and dashed lines indicate the variational approximation. **C**, Neighbor VAE Architecture. High-dimensional time series data is binned to desired time windows and represented as a high-dimensional vector. This vector is passed into a VAE that predicts the vector for the next point in time. **D**, The distance between neighboring points in time on the manifold embedding discovered by the VAE is used as a metric for model selection.

We added an inductive bias to the VAE architecture that promotes learning of underlying low-dimensional structure in time-series data. We assume that latent factors evolve smoothly in time as a Markov process over the latent space (Fig. 1A,B). Thus, we constrained the VAE architecture to be autoregressive by modifying the objective from reconstructing the original data point to predicting the next point in time (Fig. 1B,C). In this way, the function the model parameters learn encompasses both the generative function that maps low-dimensional latent factors to observables and the transition probability over time between points in the latent space. Thus, we minimize the following loss function (derivation in Appendix: A.1):

$$\mathscr{L} = -\mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)}[\log p_\theta(x_{t+1}|z_{t+1})] + \beta \mathrm{KL}(q(z_{t+1}|x_t)\|p(z_{t+1})). \tag{1}$$

Here x_t is the data point at time t and z_t is its latent representation. $p_\theta(x_{t+1}|z_{t+1})$ is the likelihood of x_{t+1} conditioned on z_{t+1} under generative model p_θ, and $q_\phi(z_{t+1}|x_t)$ is the approximated posterior of the future latent state z_{t+1} given the observation x_t. We allowed for a weight on the KL divergence term β, which was proposed as the β-VAE (Higgins et al. [2017]). We call our autoregressive VAE the Time-Neighbor VAE (TN-VAE).

3.3 Model selection metric: Neighbor Loss

To prevent learning spurious features due to overfitting to noise in the training and validation data, we defined a model selection metric based on our assumption that latent factors evolve smoothly over time in the latent space. For each model instance, we calculated the distance in the latent space between the representations of each data point and the next point in time, normalized by the overall size of the latent manifold (Fig. 1D). This metric which we termed Neighbor Loss (NL) was defined as:

$$\mathrm{NL} = \sum_{t=0}^{N-1} \frac{\|z_{t+1} - z_t\|}{\bar{z}}, \tag{2}$$

where $\bar{z} = \sum_{t=0}^{N} \|z_t\|/N$ is the average distance from the origin across latent representations of all N points in the dataset. Theoretically, if transitions between points in the latent space follow a Gaussian random walk, then minimizing the absolute distance between latent representations of neighboring points in time is equivalent to maximizing their log-likelihood (Appendix: A.2). Calculating NL loss has a small computational cost scaling linearly with the number of data points, thus calculating NL throughout training adds only a negligible computational overhead.

3.4 Training details

We trained large ensembles of VAE and TN-VAE models across a comprehensive range of hyperparameters. We compiled a range of values across 5 hyperparameters: number of layers in the model ($2-4$), dimensions of each layer ($50-400$), KL-divergence weight ($10^{-4}-10^{-3}$), batch size (spiral dataset: $128-1024$; HMM and LFP datasets: $2048-8192$), and learning rate ($10^{-5}-10^{-3}$). We used multiple unique seeds (spiral dataset: 5; HMM and LFP datasets: 3) that determined the partition of data into training and validation splits and model parameters at initialization. We fitted all models for 500 epochs, at which point the validation loss curves fully plateaued. In total, we trained multiple model instances per dataset (spiral dataset: 1440; HMM and LFP datasets: 648) for subsequent analysis.

3.5 Metrics for the quality of learned representations

We used silhouette score with respect to the ground-truth clusters to quantify how well the latent representations recover the true low-dimensional structure in the data. The silhouette score for a single sample is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}, \tag{3}$$

where $a(i)$ is the mean distance between point i and all other points in the same cluster, and $b(i)$ is the mean distance between point i and all points from the other nearest cluster. The silhouette score of the whole dataset is the mean over silhouette score of all samples. We use the ground-truth clusters for the HMM dataset, and the expert labels of wake, SWS, and REM states for the LFP dataset. For the spiral dataset, we define ground-truth clusters as groups of 100 consecutive data points along the spiral.

We quantified the robustness of representations learned by different model instances using procrustes distance between their encodings of the same test data. Generalized procrustes distance is the minimum Euclidean distance between two data matrices $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times d}$ (n: number of data points, d: number of dimensions) subject to any rotation, translation, scaling and reflection (Gower [1975]).

4 Results

4.1 VAE learns spurious features

To illustrate that standard VAEs are prone to learning spurious features, we trained VAEs on a synthetic dataset of points drawn from 3 high-dimensional Gaussian clusters. Both the training and validation loss decreased monotonically throughout 500 training epochs (Fig. 2A). Thus, the standard approach would be to select the model at the last training epoch that has the lowest validation loss. However, despite good reconstruction performance on validation data, the VAE not only failed to correctly separate the 3 clusters in the learned latent representation, but also learned spurious features. Over training epochs, the latent representation gradually lost the smooth Gaussian shape and feathered into streak-like patterns. (Fig. 2B). To quantify how well a model separates the three clusters, we computed the silhouette score of the latent encodings using the ground-truth cluster labels. Across 288 VAEs trained with different hyperparameter combinations, there was not a single model that separated clusters well, and the silhouette score did not correlate with the validation loss (Fig. 2C). In addition, models with lower validation loss had higher per-cluster skew and kurtosis of the latent encodings, indicating deviation from Gaussian distribution (Fig. 2C). These findings show that lower validation loss does not imply a more accurate representation of the ground-truth structure in the latent encodings.

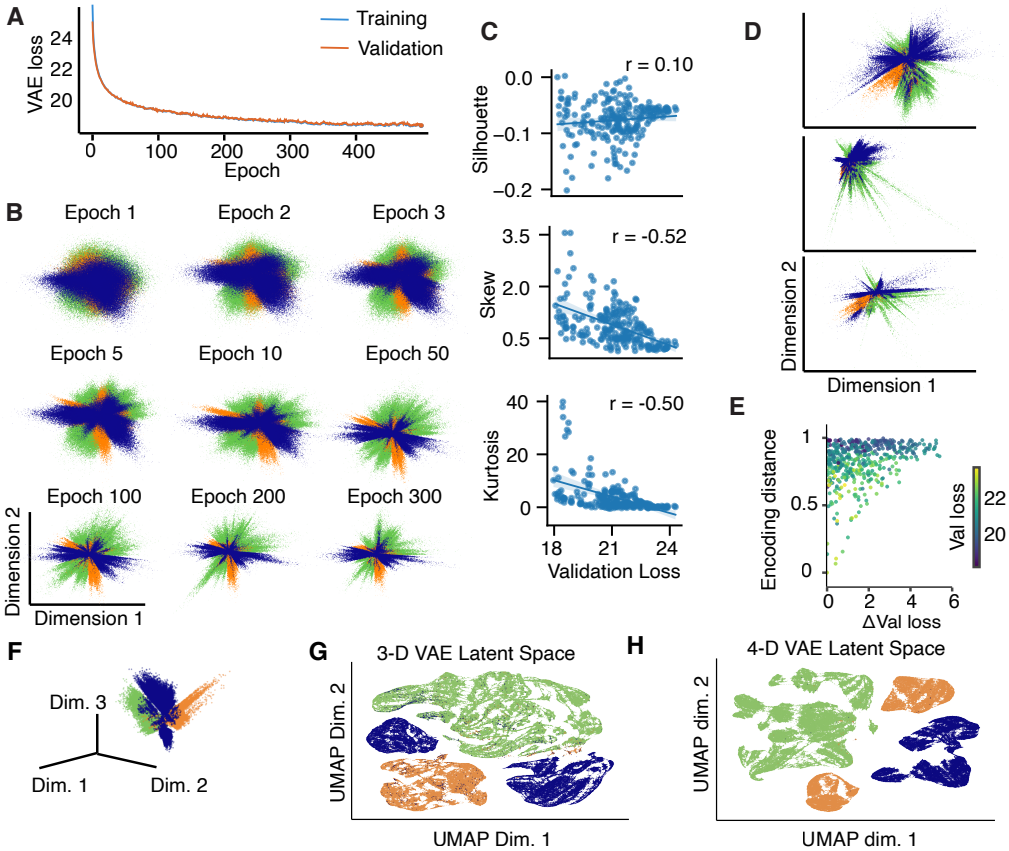

Figure 2: **A**, Training and validation loss for an example VAE trained on the Gaussian clusters dataset. **B**, Latent representations learned by this VAE at different training epochs. **C**, Validation loss versus silhouette score (upper panel), skew (middle panel) and kurtosis (lower panel) of latent encodings of the ground-truth clusters across 288 VAEs with unique hyperparameters combinations. **D**, Latent representations after 500 epochs in 3 VAEs with different hyperparameters with similar validation loss show lack of robustness. **E**, Models with low, more similar validation loss did not necessarily have more similar latent encodings as measured by procrustes distance (each dot is a model pair, a random sample of 500 model pairs is shown). Color indicates average validation loss between two models. **F**, Latent representations in a three-dimensional VAE with the best validation loss across 288 models with different hyperparameters trained on the same data. **G**, UMAP-projection of the three-dimensional latent space from F. **H**, Same as G for the VAE with four latent dimensions that had the best validation loss.

Furthermore, separate model instances trained on different splits of the data achieved similar validation loss but uncover different latent representations (Fig. 2D), indicating that the VAE learns spurious features that are not robust. We quantified similarity of latent representations between a pair of models using the encoding distance, defined as procrustes distance (Gower [1975]) between their latent encodings of the same test data. The encoding distance measures the disparity between two latent representations of the data allowing for arbitrary scaling, dilation, rotation, and reflection of the latent space. One might expect that models with lower and more similar validation loss produce more similar encoding spaces. However, we observed the opposite behavior: models with more similar, low validation loss have more dissimilar encoding spaces as indicated by higher encoding distance (Fig. 2E).

One concern may be that VAEs learn spurious features because two latent dimensions are not sufficient to faithfully represent the ground-truth structure in the data. Therefore, it is common to fit VAEs with a high-dimensional latent space and then project the latent encodings to two dimensions for visualization using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. [2020])

or t-distributed Stochastic Neighbor Embedding (tSNE) (Maaten and Hinton [2008]). However, we found that VAEs with three latent dimensions learn similar spurious features as two-dimensional VAEs (Fig. 2F). Moreover, UMAP-projection of higher-dimensional latent spaces produced spurious structures with more clusters than the ground truth (Fig. 2G,H), consistent with the fact that UMAP and tSNE methods are not guaranteed to preserve local or global distances, are sensitive to the choice of hyperparameters, and thus can produce spurious features (Chari and Pachter [2023]). These results demonstrate that variational autoencoders are not suitable for faithfully identifying latent structure in high-dimensional data, and validation loss is not a suitable metric for selecting interpretable models with robust features.

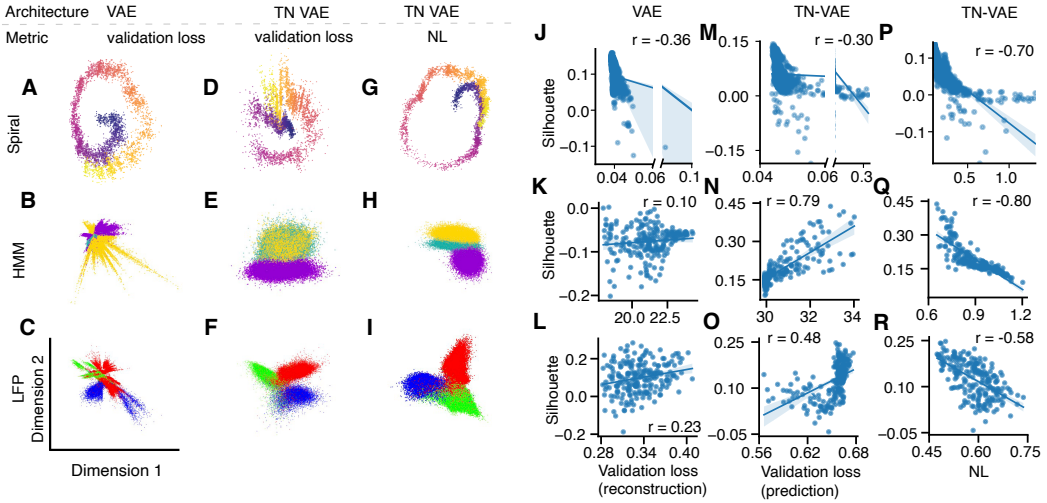

Figure 3: Latent representations on a held-out test set learned by the standard VAE with validation-loss model selection (**A-C**), TN-VAE with validation-loss model selection (**D-F**), and TN-VAE with NL model selection (**G-I**) for three datasets: synthetic spiral (**A, D, G**, points colored by the position on the original spiral), synthetic HMM (**B, E, H**, points colored by the 3 HMM states), and biological LFP data (**C, F, I**, points colored by expert-provided labels of wake, REM, and SWS brain states). We quantify the correspondence between latent encodings and the ground truth using the silhouette score of encodings with respect to clusters formed by each 100 consecutive data points for the spiral dataset (**J, M, P**), silhouette score of encodings with respect to the ground-truth clusters for the HMM dataset (**K, N, Q**), and silhouette score of encodings with respect to the expert-labeled wake, REM, and SWS states for the LFP dataset (**L, O, R**). For the VAE, lower validation loss does not correlate with the higher silhouette score (**J, K, L**). For the TN-VAE, validation loss is only slightly negatively correlated with silhouette score for spiral (**M**) and positively correlated with silhouette score for the HMM and LFP datasets (**N, O**). In contrast, lower NL corresponded to higher silhouette score for all datasets (**P, Q, R**), indicating better correspondence of latent encodings with the ground truth.

4.2 Priors on smoothness over time promote learning of true and not spurious features

We introduce a modified model architecture TN-VAE and a NL model selection metric which promote smoothness of latent factors over time (Methods). We used two synthetic datasets with known low-dimensional structure (Fig. 3A-H) to test how these modifications affect the quality of learned latent representations in comparison to the standard VAE and standard model selection based on validation loss. First, the standard VAE with the validation-loss model selection did not necessarily recover the correct low-dimensional structure without learning spurious features (Fig. 3A,B, Figs. A1A, A2A). The level and pattern of overfitting (to noise or by memorization) depends on the noise level in the data observations (Fig. A4). TN-VAE introduces an inductive bias that pushes the model to learn meaningful features, but can still overfit to noise in the data or fail to select the best model (Fig. 3D,E, Figs. A1B, A2B). Using TN-VAE in combination with NL model selection metric resulted in the most faithful reconstruction of the original low-dimensional structure in the data without learning spurious features (Fig. 3G,H, Figs. A1C, A2C).

We then applied each approach to biological LFP data recorded in the visual cortex of a mouse over normal day and night activity. While the ground-truth structure in these data is unknown, the domain knowledge suggests the existence of 3 major clusters in these data corresponding to wake, REM, and SWS brain states. The standard VAE with best validation loss failed to separate the clusters at all and instead learns spurious features (Fig. 3C, Fig. A3A). The TN-VAE with validation-loss model selection separates the clusters (Fig. 3F) but does not reliably achieve consistency between latent encodings of the models with the best validation loss (Fig. A3B). Finally, TN-VAE with the NL model selection metric produced a smooth latent representation with 3 clusters that matched the human-expert labeling of wake, REM, and SWS brain states (Fig. 3I), which was highly consistent across the models with the best NL (Fig. A3C).

Across all models trained on these three datasets, lower validation loss did not necessarily indicate better correspondence between latent encodings and the ground truth. We quantified this relationship using silhouette score of encodings with respect to clusters formed by each 100 consecutive data points for the spiral dataset (Fig. 3J,M), the silhouette score of encodings with respect to the ground-truth clusters for the HMM dataset (Fig. 3K,N), and the silhouette score of encodings with respect to expert-labeled wake, REM, and SWS states in the LFP dataset (Fig. 3L,O). In contrast, lower NL in the TN-VAE model was systematically correlated with better correspondence between the latent encodings and the ground truth, indicated by higher silhouette score (Fig. 3P,Q,R). Thus, the TN-VAE model architecture and NL model selection metric together are most effective in recovering latent encodings representing ground truth structure.

4.3 Neighbor Loss metric selects models with robust representations

Finally, we tested how the Neighbor Loss (NL) model selection metric corresponds to the robustness of the learned representations. For all three datasets, we chose a variety of different hyperparameter configurations, and for each configuration, we trained multiple model instances with unique training and validation data splits. Each model instance produces a latent representation of the same held-out test dataset. We then measured the average encoding distance between latent representations of the test set for each pair of model instances across a set of unique seeds. Lower encoding distance indicates more consistency in representations across model instances and thus, can be used to assess the robustness of learned latent representations. We found that NL correlates well with encoding distance, whereas validation loss was not correlated with how consistent the latent representations are across model instances (Fig. 4). Intuitively, if models underfit, there would be no correlation between the structure of the latent representation and time information. Vice versa, overfitted models separate neighboring points in time based on noise features that are random across training instances. As a result, minimizing the NL metric as a criterion for model selection leads to the selection of robust latent representations.

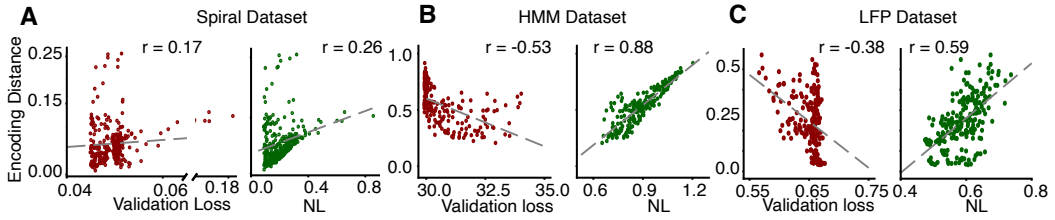

Figure 4: For each hyperparameter combination, we trained multiple model instances (spiral dataset: 1440, HMM and LFP datasets: 648) on different training and validation splits. We measured the encoding distances between latent representations learned by each pair of model instances on a test set. We also calculated the average validation loss and average validation NL over the model instances for each hyperparameter combination. NL correlated more with encoding distance than validation loss for synthetic spiral (**A**) HMM (**B**) and LFP (**C**) datasets.

5 Discussion

Achieving similar representations across different models trained on the same biological data is a prerequisite to scientific interpretation of such representations. We tested how and when this similarity

can fail and degenerate representations can arise. Specifically, we seek to understand overfitting to noise in VAE latent representations of high-dimensional neural data and to uncover strategies to mitigate such overfitting. We show that the representations produced by the standard VAE with validation-loss model selection often lack robustness and thus cannot be reliably interpreted. We leverage smoothness of latent representations over time to introduce both an inductive bias on model architecture and a model selection metric, which together prevent learning of spurious features and lead to robust representations.

The introduction of time as a regularization in neural networks has been used previously, both in variational autoencoder architectures (Schneider et al. [2023], Liu et al. [2021], Sedler et al. [2023]) and more broadly in feedforward networks (Toosi and Issa [2023]) and transformer architectures (Devlin et al. [2019]). Here, we extend the use of time to mitigate overfitting and learning spurious features in VAEs. In addition, we propose its use as a model selection metric, which unsupervised methods such as VAEs often lack. In contrast to current model selection metrics for VAEs (Higgins et al. [2017], Pei et al. [2021], Duan et al. [2020]), our proposed NL metric is intrinsic to the latent representations themselves, not requiring external labels, and also can be used during training unlike methods that need to compare large ensembles of trained models. Thus, our contributions are to 1) show that standard VAEs are prone to learning spurious features; 2) show how using time in VAE architecture is an effective inductive bias and regularizer and 3) introduce a novel model selection metric based on smoothness over time of latent representations.

One way to recover meaningful and interpretable representations of data by VAEs is through achieving identifiability. It is important to note that true identifiability cannot be achieved without an external label (Locatello et al. [2019]), and our methods does not necessarily achieve identifiability. Instead, we focus on robustness of latent representations as a prerequisite for interpretability.

Acknowledgments and Disclosure of Funding

This work was supported by the Cold Spring Harbor Laboratory School of Biological Sciences (J.H.W.), the Jenny and Jeff Kelter Neuroscience Scholarship (J.H.W.), NIH grant RF1DA055666 (T.A.E.), Starr Foundation award I15-0037 (J.H.W. and T.A.E.), and NIH grant RF1NS128901 (J.H.W. and T.A.E). We thank S. Chauvette and I. Timofeev for sharing of electrophysiological data, which are presented in Soltani et al. [2019].

References

M. Azabou, M. G. Azar, R. Liu, C.-H. Lin, E. C. Johnson, K. Bhaskaran-Nair, M. Dabagia, B. Avila-Pires, L. Kitchell, K. B. Hengen, W. Gray-Roncal, M. Valko, and E. L. Dyer. Mine Your Own vieW: Self-Supervised Learning Through Across-Sample Prediction, Dec. 2021. URL `http://arxiv.org/abs/2102.10106`. arXiv:2102.10106 [cs].

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, Dec. 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907378117. URL `https://pnas.org/doi/full/10.1073/pnas.1907378117`.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. page 23, 2018.

T. Chari and L. Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, Aug. 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011288. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288`. Publisher: Public Library of Science.

J. P. Cunningham and B. M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, Nov. 2014. ISSN 1546-1726. doi: 10.1038/nn.3776. URL `https://www.nature.com/articles/nn.3776`. Number: 11 Publisher: Nature Publishing Group.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL `http://arxiv.org/abs/1810.04805`. arXiv:1810.04805 [cs].

S. Duan, L. Matthey, A. Saraiva, N. Watters, C. P. Burgess, A. Lerchner, and I. Higgins. Unsupervised Model Selection for Variational Disentangled Representation Learning, Feb. 2020. URL http://arxiv.org/abs/1905.12614. arXiv:1905.12614 [cs, stat].

Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv:1611.02779 [cs, stat]*, Nov. 2016. URL http://arxiv.org/abs/1611.02779. arXiv: 1611.02779.

M. Genkin and T. A. Engel. Moving beyond generalization to accurate interpretation of flexible models. *Nature Machine Intelligence*, 2(11):674–683, Nov. 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00242-6. URL https://www.nature.com/articles/s42256-020-00242-6. Number: 11 Publisher: Nature Publishing Group.

J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. ISSN 00333123. doi: 10.1007/bf02291478. URL https://doi.org/10.1007/bf02291478. Number: 1 Publisher: Springer.

I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville. PixelVAE: A Latent Variable Model for Natural Images, Nov. 2016. URL http://arxiv.org/abs/1611.05013. arXiv:1611.05013 [cs].

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. -VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK. *International Conference of Learning Representations*, 2017.

M. R. Keshtkaran and C. Pandarinath. Enabling hyperparameter optimization in sequential autoencoders for spiking neural data. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6948bd44c91acd2b54ecdd1b132f10fb-Abstract.html.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, June 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html. ISSN: 2640-3498.

D. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding, Mar. 2021. URL http://arxiv.org/abs/2007.10930. arXiv:2007.10930 [cs, stat].

R. Liu, M. Azabou, M. Dabagia, C.-H. Lin, M. G. Azar, K. B. Hengen, M. Valko, and E. L. Dyer. Drop, Swap, and Generate: A Self-Supervised Approach for Generating Neural Activity. *Advances in neural information processing systems*, 34:10587–10599, Dec. 2021. ISSN 1049-5258. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9713686/.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, June 2019. URL http://arxiv.org/abs/1811.12359. arXiv:1811.12359 [cs, stat].

K. Luxem, P. Mocellin, F. Fuhrmann, J. Kürsch, S. R. Miller, J. J. Palop, S. Remy, and P. Bauer. Identifying behavioral structure from deep variational embeddings of animal motion. *Communications Biology*, 5(1):1–15, Nov. 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-04080-7. URL https://www.nature.com/articles/s42003-022-04080-7. Number: 1 Publisher: Nature Publishing Group.

L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Sept. 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space, Sept. 2013. URL http://arxiv.org/abs/1301.3781. arXiv:1301.3781 [cs].

G. Mita, M. Filippone, and P. Michiardi. An Identifiable Double VAE For Disentangled Representations, Feb. 2021. URL http://arxiv.org/abs/2010.09360. arXiv:2010.09360 [cs, stat].

A. v. d. Oord, Y. Li, and O. Vinyals. Representation Learning with Contrastive Predictive Coding, Jan. 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs, stat].

C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, and D. Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, Oct. 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0109-9. URL http://www.nature.com/articles/s41592-018-0109-9.

F. Pei, J. Ye, D. Zoltowski, A. Wu, R. H. Chowdhury, H. Sohn, J. E. O'Doherty, K. V. Shenoy, M. T. Kaufman, M. Churchland, M. Jazayeri, L. E. Miller, J. Pillow, I. M. Park, E. L. Dyer, and C. Pandarinath. Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity. *arXiv*, Sept. 2021. doi: 10.48550/arXiv.2109.04463. URL https://arxiv.org/abs/2109.04463v4.

R. E. Peterson, A. Choudhri, C. Mitelut, A. Tanelus, A. Capo-Battaglia, A. H. Williams, D. M. Schneider, and D. H. Sanes. Unsupervised discovery of family specific vocal usage in the Mongolian gerbil. *eLife*, 12, Aug. 2023. doi: 10.7554/eLife.89892.

A. Radhakrishnan, K. Yang, M. Belkin, and C. Uhler. Memorization in Overparameterized Autoencoders, Sept. 2019. URL http://arxiv.org/abs/1810.10333. arXiv:1810.10333 [cs, stat].

S. Schneider, J. H. Lee, and M. W. Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://www.nature.com/articles/s41586-023-06031-6. Number: 7960 Publisher: Nature Publishing Group.

A. R. Sedler, C. Versteeg, and C. Pandarinath. Expressive architectures enhance interpretability of dynamics-based neural population models, Feb. 2023. URL http://arxiv.org/abs/2212.03771. arXiv:2212.03771 [cs, q-bio].

S. Sinha and A. B. Dieng. Consistency Regularization for Variational Auto-Encoders. *arXiv:2105.14859 [cs]*, May 2021. URL http://arxiv.org/abs/2105.14859. arXiv:2105.14859.

S. Soltani, S. Chauvette, O. Bukhtiyarova, J.-M. Lina, J. Dubé, J. Seigneur, J. Carrier, and I. Timofeev. Sleep–Wake Cycle in Young and Older Mice. *Frontiers in Systems Neuroscience*, 13, 2019. ISSN 1662-5137. URL https://www.frontiersin.org/article/10.3389/fnsys.2019.00051.

T. Toosi and E. B. Issa. Brain-like representational straightening of natural movies in robust feedforward neural networks, Aug. 2023. URL http://arxiv.org/abs/2308.13870. arXiv:2308.13870 [cs, q-bio].

D. Zhou and X.-X. Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. In *Advances in Neural Information Processing Systems*, volume 33, pages 7234–7247. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc//paper/2020/hash/510f2318f324cf07fce24c3a4b89c771-Abstract.html.

A Appendix

A.1 TN-VAE Loss derivation

Here, we approximate the relationship between x_t and the latent at the next time point, z_{t+1} with q_ϕ (encoder). We seek to minimize the KL divergence between this approximate posterior q and $p_\theta(z_{t+1}, x_{t+1})$.

$$
\begin{aligned}
\mathrm{KL}(q_\phi(x_{t+1}|x_t)||p_\theta(z_{t+1}|x_{t+1})) &= \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log q_\phi(z_{t+1}|x_t) - \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log p_\theta(z_{t+1}|x_{t+1}) \\
&= \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log q_\phi(z_{t+1}|x_t) - \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log p_\theta(x_{t+1}|z_{t+1}) \\
&- \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log p_\theta(z_{t+1}) + \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log p_\theta(x_{t+1})
\end{aligned}
$$

Minimizing this quantity is equivalent to maximizing the Evidence Lower Bound ($\mathscr{L}$).

$$
\begin{aligned}
\mathscr{L} &= \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log p_\theta(x_{t+1}|z_{t+1}) + \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} p_\theta(z_{t+1}) - \mathrm{E}_{z\sim z_{t+1}\ q_\phi(z_{t+1}|x_t)} \log q_\phi(z_{t+1}|x_t) \\
&= \mathrm{E}_{z\sim q_\phi(z_{t+1}|x_t)} \log p_\theta(x_{t+1}|z_{t+1}) - \mathrm{KL}(q_\phi(z_{t+1}|x_t)||p_\theta(z_{t+1}))
\end{aligned}
$$

A.2 Neighbor Loss as the likelihood of a random walk in the latent space

Assume data-points traverse a low-dimensional latent space where $\mathbf{x}_{t+1}$ is drawn from a circular Gaussian distribution centered at $\mathbf{x}_t$ with variance σ. The log-likelihood of this function is:

$$
\begin{aligned}
\log(L) &= \log\left(\prod_{t=1}^{n} \frac{1}{2\pi\sigma} \exp(-\frac{(\mathbf{x}_{t+1}-\mathbf{x}_t)(\mathbf{x}_{t+1}-\mathbf{x}_t)^T}{2\sigma})\right) \\
&= \sum_{t=1}^{n} \log\left(\frac{1}{2\pi\sigma} \exp(-\frac{(\mathbf{x}_{t+1}-\mathbf{x}_t)(\mathbf{x}_{t+1}-\mathbf{x}_t)^T}{2\sigma})\right) \\
&= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma}\sum[(\mathbf{x}_{t+1}-\mathbf{x}_t)(\mathbf{x}_{t+1}-\mathbf{x}_t)^T].
\end{aligned}
$$

Thus, maximizing the log-likelihood is equivalent to minimizing sum of the absolute distance between each point and its neighboring point in time.

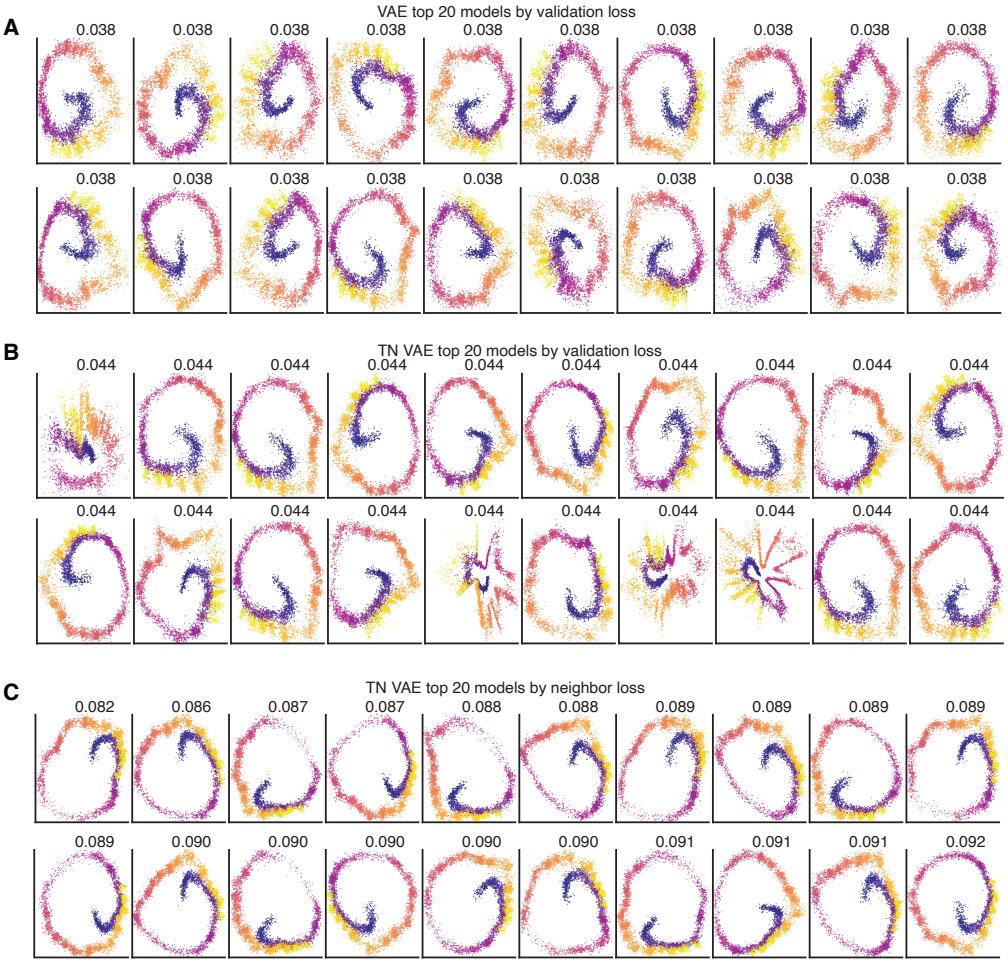

Fig. A1: Latent representations of the spiral dataset on a held-out test set from the top 20 models learned by (A) VAE and selected based on the validation loss, (B) TN-VAE and selected based on the validation loss, (C) TN-VAE and selected based on the neighbor loss.

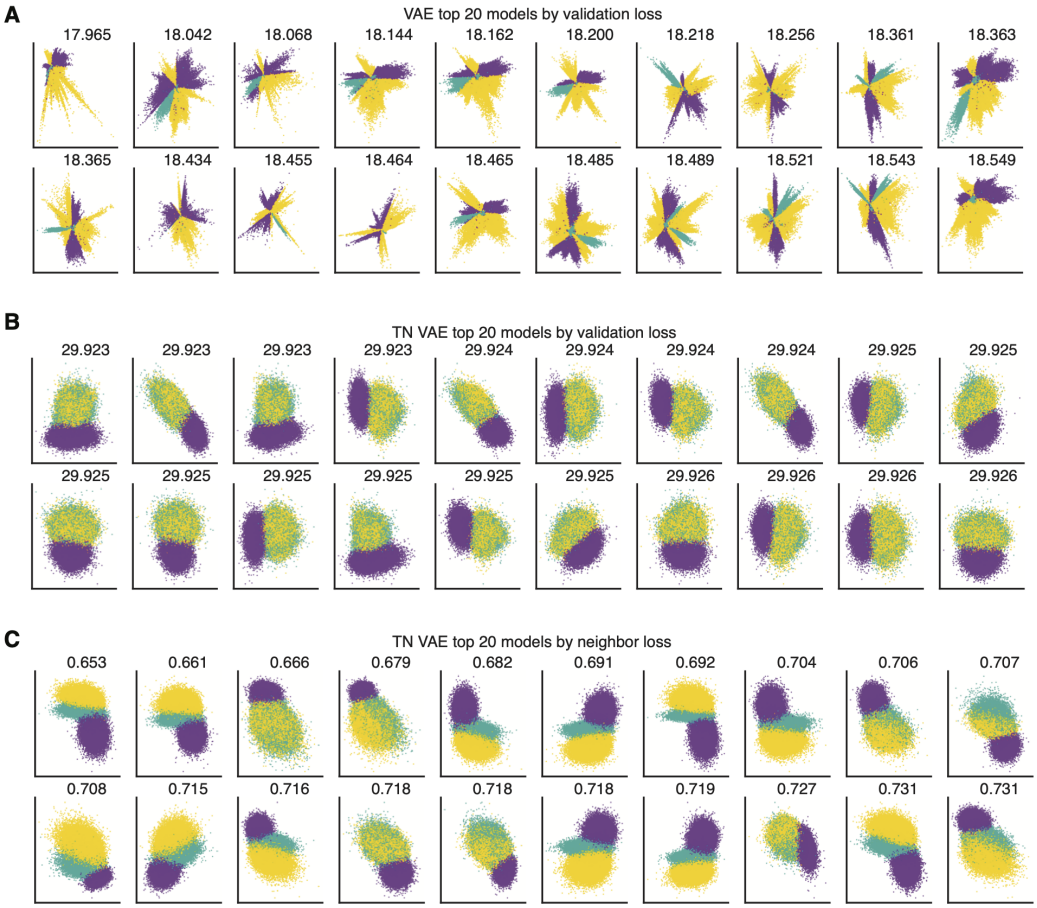

Fig. A2: Latent representations of the HMM dataset on a held-out test set from the top 20 models learned by (A) VAE and selected based on the validation loss, (B) TN-VAE and selected based on the validation loss, and (C) TN-VAE and selected based on the neighbor loss.

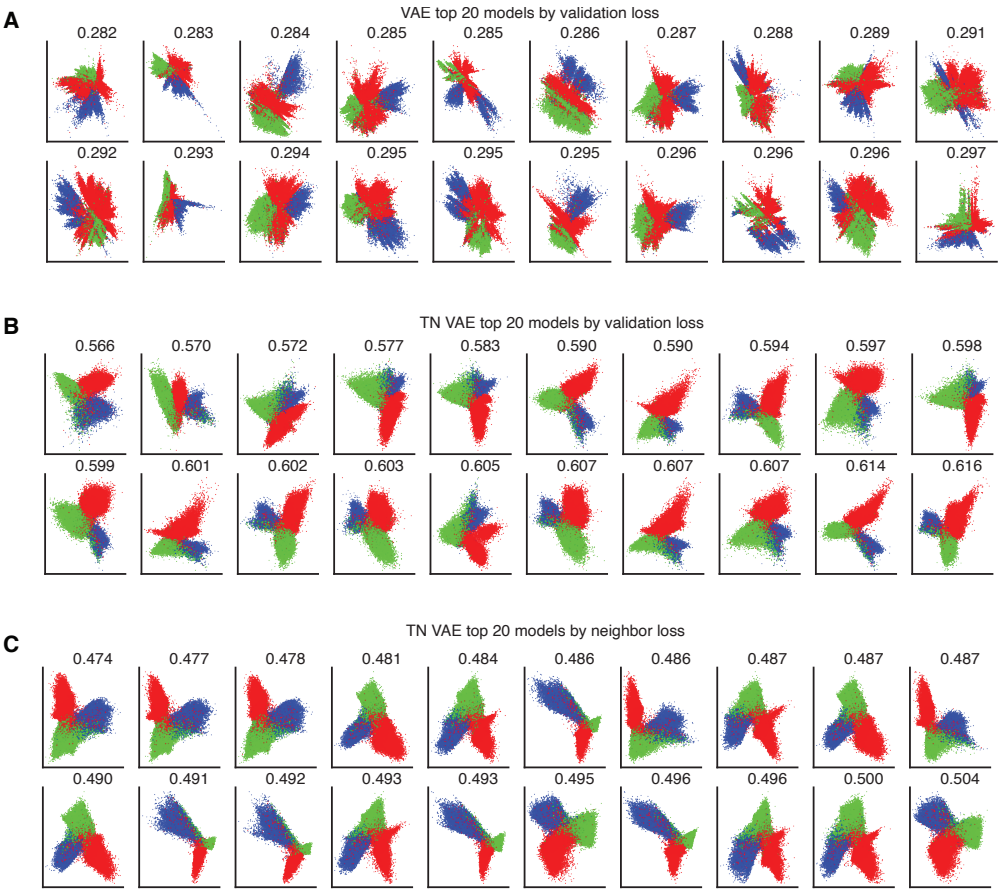

Fig. A3: Latent representations of the LFP dataset on a held-out test set from the top 20 models learned by (A) VAE and selected based on the validation loss, (B) TN-VAE and selected based on the validation loss, and (C) TN-VAE and selected based on the neighbor loss.

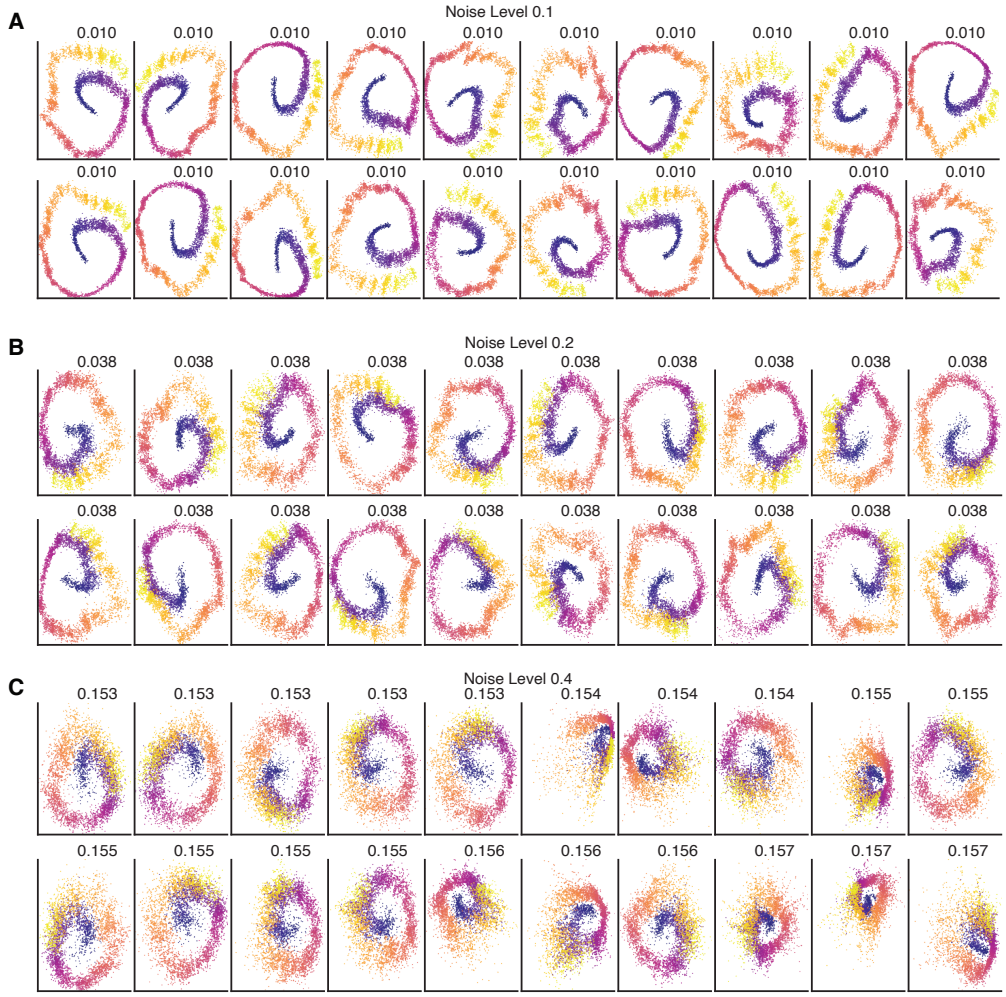

Fig. A4: The level and pattern of overfitting in the VAE depends on the magnitude of the high-dimensional noise in the observations. Latent representations of the spiral dataset on a held-out test set from the top 20 models learned by VAE and selected based on the validation loss, for training and test datasets with noise level (A) 0.1, (B) 0.2, and (C) 0.4.