LATENT REPRESENTATION AND SIMULATION OF MARKOV PROCESSES VIA TIME-LAGGED INFORMATION BOTTLENECK

Marco Federici[*†]
AMLab
University of Amsterdam
m.federici@uva.nl

Patrick Forré
AI4Science Lab, AMLab
University of Amsterdam
p.d.forre@uva.nl

Ryota Tomioka
Microsoft Research AI4Science
ryoto@microsoft.com

Bastiaan S. Veeling[*]
Microsoft Research AI4Science
basveeling@microsoft.com

ABSTRACT

Markov processes are widely used mathematical models for describing dynamic systems in various fields. However, accurately simulating large-scale systems at long time scales is computationally expensive due to the short time steps required for accurate integration. In this paper, we introduce an inference process that maps complex systems into a simplified representational space and models large jumps in time. To achieve this, we propose Time-lagged Information Bottleneck (T-IB), a principled objective rooted in information theory, which aims to capture relevant temporal features while discarding high-frequency information to simplify the simulation task and minimize the inference error. Our experiments demonstrate that T-IB learns information-optimal representations for accurately modeling the statistical properties and dynamics of the original process at a selected time lag, outperforming existing time-lagged dimensionality reduction methods.

1 INTRODUCTION

Markov processes have long been studied in the literature (Norris, 1997; Ethier & Kurtz, 2009), as they describe relevant processes in nature such as weather, particle physics, and molecular dynamics. Despite being well-understood, simulating large systems over extensive timescales remains a challenging task. In molecular systems, analyzing meta-stable molecular configurations requires unfolding simulations over several milliseconds ($\tau \approx 10^{-3}s$), while accurate simulation necessitates integration steps on the order of femtoseconds ($\tau_0 \approx 10^{-15}s$). The time required to simulate 10^{12} steps is determined by the time of a single matrix multiplication, which takes on the order of milliseconds on modern hardware, resulting in a simulation time of multiple years.

Deep learning-based approximations have shown promising results in the context of time series forecasting (Staudemeyer & Morris, 2019; Lim & Zohren, 2021), including applications in weather forecasting (Veillette et al., 2020), sea surface temperature prediction (Ham et al., 2019; Gao et al., 2022), and molecular dynamics (Sidky et al., 2020; Klein et al., 2023; Schreiner et al., 2023). Mapping observations into lower-dimensional spaces has proven to be an effective method for reducing computational costs. Successful examples in molecular dynamics include learning system dynamics through coarse-grained molecular representations (Wang et al., 2019a; Köhler et al., 2023; Arts et al., 2023), or linear (Koopman, 1931; Molgedey & Schuster, 1994) and non-linear (Wehmeyer & Noé, 2018; Mardt et al., 2018; Sidky et al., 2020) projections of molecular features.

Modern deep representation learning methods have proven effective in creating representations for high-dimensional structured data, including images (Hjelm et al., 2019; Chen et al., 2020), audio

[*]Corresponding author.
[†]Work partially done during an internship at Microsoft Research, AI4Science.

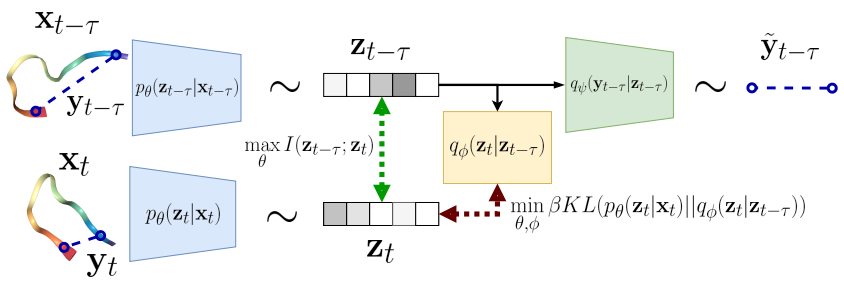


Figure 1: The Time-lagged Information Bottleneck objective aims to maximize the mutual information between sampled representations $\mathbf{z}_{t-\tau}, \mathbf{z}_t$ at temporal distance τ while minimizing mismatch between the encoding distribution $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ and the learned variational transitional distribution $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau})$. This results in minimal representations capturing dynamics at timescale τ or larger, which can be used to predict properties of interest $\mathbf{y}_t$, such as inter-atomic distances, over time.

(van den Oord et al., 2018; Saeed et al., 2021), text (Devlin et al., 2018; Radford et al., 2018), and graphs (Veličković et al., 2018; Wang et al., 2022). These methods often aim to capture relevant information while reducing the complexity of the data. In this context, information theory provides a compelling direction for further analysis (Wennekers & Ay, 2003; Gao et al., 2022; Lozano-Durán & Arranz, 2022). In particular, the information bottleneck principle (Tishby et al., 2000; Tishby & Zaslavsky, 2015) suggests that an optimal representation should retain relevant information while discarding unnecessary features. Applying this principle to the context of Markov process simulations has the potential to simplify the modeling task, reduce computational complexity, and aid in identifying the salient characteristics that define the relevant dynamics.

In this paper, we make the following contributions: (i) we introduce a probabilistic inference scheme for Markov processes, *Latent Simulation* (LS), and characterize the inference error by defining *Time-lagged InfoMax* (T-InfoMax) as a general family of principled training objectives. (ii) We propose *Time-lagged Information Bottleneck* (T-IB, Figure 1), a novel objective that follows the T-InfoMax principle to preserve system dynamics while discarding superfluous information to simplify modeling tasks. (iii) We empirically compare the performance of models trained using the T-InfoMax and T-IB objectives on synthetic trajectories and molecular simulations, showcasing the importance of the T-InfoMax principle and the advantages of the proposed T-IB method for both representation learning and latent simulation inference compared to other models in the literature.

2 Method

We delve into the problem of efficiently representing and simulating Markov processes starting by defining *Latent Simulation* as an inference procedure and characterizing the corresponding error (section 2.1). Next, in section 2.2, we analyze the problem of capturing system dynamics from an information-theoretic perspective, defining and motivating *Time-Lagged InfoMax*: a family of objectives that minimizes the latent simulation error. Finally, we introduce *Time-lagged Information Bottleneck* (section 2.3) as an extension of T-InfoMax that aims to simplify the representation space. A schematic representation of our proposed model is visualized in Figure 1.

2.1 Latent Simulation

Consider a sequence of T random variables, denoted as $[\mathbf{x}_t]_{t=0}^T$, which form a homogeneous Markov Chain. This chain models a dynamical process of interest, such as molecular dynamics, global climate systems, or particle interactions. Let $\mathbf{y}_t$ represent a specific (noisy) property of $\mathbf{x}_t$ that we aim to model over time. Formally, we define $\mathbf{y}_t = f(\mathbf{x}_t, \epsilon_t)$, where $f : \mathbb{X} \times \mathcal{E} \to \mathbb{Y}$ is some function and ϵ_t is temporally uncorrelated noise. Examples of such properties could include the energy or momentum of a particle, the meta-stable state of a molecular structure, and the amount of rainfall. Each of these properties $\mathbf{y}_t$ can be derived from a more comprehensive high-dimensional state description $\mathbf{x}_t$.

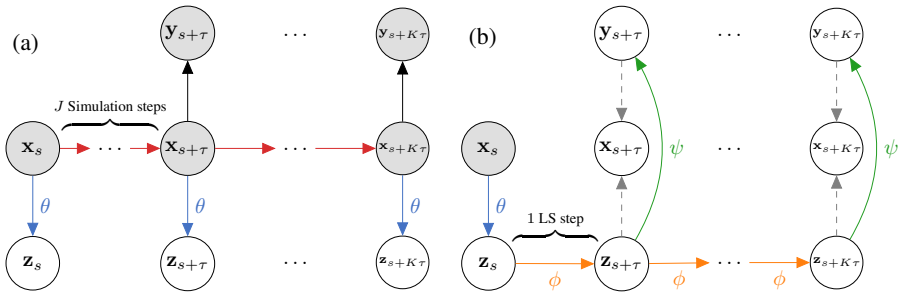


Figure 2: Graphical models for the joint distribution of a sequence, targets, and representations. 2a shows the data-generating process in which red arrows denote computationally expensive simulation steps. 2b represents the corresponding Variational Latent Simulation, in which the transitions are modeled in the latent space. Gray dashed lines indicate distributions that are not used for inference.

Given an initial observation $\mathbf{x}_s$, the joint distribution of the sequence of K future targets $[\mathbf{y}_{s+k\tau}]_{k=1}^K$ at some *lag time* $\tau > 0$ can be expressed as:

$$p([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s) = \int \dots \int \prod_{k=1}^K \underbrace{p(\mathbf{x}_{s+k\tau}|\mathbf{x}_{s+(k-1)\tau})}_{\text{Transition}} \underbrace{p(\mathbf{y}_{s+k\tau}|\mathbf{x}_{s+k\tau})}_{\text{Prediction}} d\mathbf{x}_{s+\tau} \dots d\mathbf{x}_{s+K\tau}. \quad (1)$$

Each transition distribution $p(\mathbf{x}_{t+\tau}|\mathbf{x}_t)$ may necessitate J integration steps at a finer timescale $\tau_0 < \tau$. Given the sequential nature of simulation, generating trajectories over extended time horizons may require substantial computational resources. To mitigate the challenges of simulating large-scale system dynamics, we adopt two modeling strategies: (i) rather than modeling the transition distribution in the original space $\mathbb{X}$, we learn a time-independent *encoder* $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ that maps into a simpler representation space $\mathbb{Z}$; and (ii) we directly model the dynamics for larger jumps $\tau > \tau_0$. We refer to the process of unfolding simulations in the latent representation space as *Latent Simulation* (LS). The joint distribution for trajectories of targets unfolded using LS starting from $\mathbf{x}_s$ is defined as:

$$p^{LS}([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s) := \int \dots \int \underbrace{p_\theta(\mathbf{z}_s|\mathbf{x}_s)}_{\text{Encoding}} \prod_{k=1}^K \underbrace{p(\mathbf{z}_{s+k\tau}|\mathbf{z}_{s+(k-1)\tau})}_{\text{Latent transition}} \underbrace{p(\mathbf{y}_{s+k\tau}|\mathbf{z}_{s+k\tau})}_{\text{Latent prediction}} d\mathbf{z}_s \dots \mathbf{z}_{s+K\tau}. \quad (2)$$

Unfolding LS requires access to the *latent transition* $p(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$ and *predictive* $p(\mathbf{y}_t|\mathbf{z}_t)$ distributions, which are generally intractable for an arbitrary choice of encoding distribution $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$. To circumvent this intractability, we introduce *variational transition* and *variational target predictive* distributions, denoted as $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau})$ and $q_\psi(\mathbf{y}_t|\mathbf{z}_t)$, respectively. The resulting joint inference distribution for the future targets $[\mathbf{y}_{s+k\tau}]_{k=1}^K$, unfolding from the initial observation $\mathbf{x}_s$, is referred to as the *Variational Latent Simulation* distribution $q^{LS}([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s)$ and visualized together with the graphical model for the data-generating process in Figure 2.

The Kullback-Leibler (KL) divergence, which quantifies the discrepancy between the ground truth and the variational latent simulation distributions, can be upper-bounded as follows:

$$\underbrace{\mathrm{KL}(p([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s)||q^{LS}([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s))}_{\text{Variational Latent Simulation error}} \leq \underbrace{\mathrm{KL}(p([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s)||p^{LS}([\mathbf{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s))}_{\text{Latent Simulation error}}$$
$$+ \sum_{k=1}^K \underbrace{\mathrm{KL}(p(\mathbf{z}_{s+k\tau}|\mathbf{z}_{s+(k-1)\tau})||q_\phi(\mathbf{z}_{s+k\tau}|\mathbf{z}_{s+(k-1)\tau}))}_{\text{Variational latent transition gap}} + \underbrace{\mathrm{KL}(p(\mathbf{y}_{s+k\tau}|\mathbf{z}_{s+k\tau})||q_\psi(\mathbf{y}_{s+k\tau}|\mathbf{z}_{s+k\tau}))}_{\text{Variational latent prediction gap}}. \quad (3)$$

The upper bound consists of the latent simulation error and the sum of the variational gaps for both the latent transition and target predictive distributions. Unfortunately, terms on the right side of equation 3 are intractable. To address this, we propose a two-step optimization procedure: (i) we first learn an encoding distribution $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ that minimizes the latent simulation error, effectively capturing the dynamical properties of the system in the representation; then (ii), assuming a fixed (optimal) encoding distribution, we optimize the variational latent transition $q_\phi(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$ and

predictive $q_\psi(\mathbf{y}_t|\mathbf{z}_t)$ distributions using maximum likelihood to minimize their respective variational gaps. In the following sections, we will focus on step (i), analyzing the problem of learning representations that preserve dynamical properties from an information theoretical perspective. Additional details about this two-step procedure are available in Appendix C.1.

2.2 Temporal Information on Markov Chains

A crucial prerequisite for ensuring that the latent simulation process does not introduce any error is to guarantee that each representation $\mathbf{z}_t$ is as informative as the original data $\mathbf{x}_t$ for the prediction of any future target of interest $\mathbf{y}_{t+\tau}$. If $\mathbf{z}_t$ is less predictive than $\mathbf{x}_t$ for $\mathbf{y}_{t+\tau}$, the statistics for the corresponding predictive distribution $p(\mathbf{y}_{t+\tau}|\mathbf{z}_t)$ would deviate from those based on the original data $p(\mathbf{y}_{t+\tau}|\mathbf{x}_t)$. This first requirement can be expressed by equating *mutual information*[1] that $\mathbf{x}_t$ and $\mathbf{z}_t$ share with $\mathbf{y}_{t+\tau}$: $I(\mathbf{x}_t;\mathbf{y}_{t+\tau}) = I(\mathbf{z}_t;\mathbf{y}_{t+\tau})$. We will refer to this requirement as *sufficiency* of $\mathbf{z}_t$ for $\mathbf{y}_{t+\tau}$. Sufficiency is achieved only when $\mathbf{x}_t$ and $\mathbf{z}_t$ yield identical predictive distributions for the future target, i.e., $p(\mathbf{y}_{t+\tau}|\mathbf{x}_t) = p(\mathbf{y}_{t+\tau}|\mathbf{z}_t)$.

Secondly, we introduce the concept of *autoinformation*. Autoinformation at a given lag time τ is defined as the mutual information between the current observation $\mathbf{x}_t$ and its corresponding future $\mathbf{x}_{t+\tau}$. Formally, $AI(\mathbf{x}_t;\tau) := I(\mathbf{x}_t;\mathbf{x}_{t+\tau})$. This concept extends the statistical notion of autocorrelation, which measures the linear relationship between values of a variable at different times (Brockwell & Davis, 2002), to include nonlinear relationships (Chapeau-Blondeau, 2007; von Wegner et al., 2017).

Since $\mathbf{z}_t$ is derived from $\mathbf{x}_t$, the autoinformation for $\mathbf{x}_t$ sets an upper-bound for the autoinformation for $\mathbf{z}_t$: $AI(\mathbf{x}_t;\tau) \geq AI(\mathbf{z}_t;\tau)$ (proof in Appendix B.3). We refer to the difference between the two values as the *autoinformation gap* $AIG(\mathbf{z}_t;\tau) := AI(\mathbf{x}_t;\tau) - AI(\mathbf{z}_t;\tau)$ and we say that $\mathbf{z}_t$ *preserves autoinformation* whenever autoinformation gap is zero.

Lemma 1. *Autoinformation and Sufficiency (proof in Appendix B.5)*
A representation $\mathbf{z}_t$ preserves autoinformation at lag time τ if and only if it is sufficient for any target $\mathbf{y}_{t+\tau}$. Conversely, whenever $\mathbf{z}_t$ does not preserve autoinformation for a lag time τ is always possible to find a target $\mathbf{y}_{t+\tau}$ for which $\mathbf{z}_t$ is not sufficient:

$$AIG(\mathbf{z}_t;\tau) = 0 \iff I(\mathbf{x}_t;\mathbf{y}_{t+\tau}) = I(\mathbf{z}_t;\mathbf{y}_{t+\tau}) \quad \forall \mathbf{y}_{t+\tau} := f(\mathbf{x}_{t+\tau}, \epsilon).$$

In simpler terms, a representation that preserves autoinformation encapsulates all dynamic properties of the original data for the temporal scale τ. As a result, the representation $\mathbf{z}_t$ can replace $\mathbf{x}_t$ in predicting any future properties at time $t+\tau$.

For a temporal sequence $[\mathbf{x}_t]_{t=s}^{T}$, we define the autoinformation at lag time τ as the average autoinformation between all pairs of elements in the sequence that are τ time-steps apart: $AI([\mathbf{x}_t]_{t=s}^{T};\tau) := \mathbb{E}_{t\sim U(s,T-\tau)}\left[AI(\mathbf{x}_t;\tau)\right]$, where $U(s, T-\tau)$ refers to a uniform distribution. If $p(\mathbf{x}_s)$ is stationary, the amount of autoinformation for a sequence $[\mathbf{x}_t]_{t=s}^{T}$ is equivalent to autoinformation at any point $\mathbf{x}_t$. Using this definition, we can show:

Lemma 2. *Autoinformation and Markov Property (proof in Appendix B.6)*
If a sequence of representations $[\mathbf{z}_t]_{t=s}^{T}$ of a homogeneous Markov chain $[\mathbf{x}_t]_{t=s}^{T}$ preserves autoinformation at lag time τ, then any of its sub-sequences of elements separated by τ time-steps must also form a homogeneous Markov chain:

$$AIG([\mathbf{z}_t]_{t=s}^{T};\tau) = 0 \implies [\mathbf{z}_{s'+k\tau}]_{k=0}^{K} \text{ is a homogeneous Markov Chain},$$

for every $s' \in [s, T-\tau]$, $K \in [0, \lfloor (T-s')/\tau \rfloor]$.

Building on this, we further establish that dynamics at a predefined timescale τ also encode information relevant to larger timescales:

Lemma 3. *Slower Information Preservation (proof in Appendix B.8)*
Any sequence of representations $[\mathbf{z}_t]_{t=s}^{T}$ that preserves autoinformation at lag time τ must also preserve autoinformation at any larger timescale τ':

$$AIG([\mathbf{z}_t]_{t=s}^{T};\tau) = 0 \implies AIG([\mathbf{z}_t]_{t=s}^{T};\tau') = 0 \quad \forall \tau' \geq \tau.$$

[1] We refer the reader to Appendix A for further details on the notation.

By synthesizing the insights from Lemma 1, 2, and 3, we can infer that any representation preserving autoinformation at lag time τ captures the dynamical properties of the system across timescales τ' that are equal or larger than τ. Specifically, we conclude that: (i) $\mathbf{z}_t$ can replace $\mathbf{x}_t$ in predicting any $\mathbf{y}_{t+\tau'}$ (Lemma 1 + Lemma 3); (ii) any sequence of representations $[\mathbf{z}_{s+k\tau'}]_{k=0}^{K}$ will form a homogeneous Markov Chain (Lemma 2 + Lemma 3). Furthermore, we establish an upper bound for the expected Latent Simulation error in equation 3 using the autoinformation gap:

$$\underbrace{\mathbb{E}_t\left[\mathrm{KL}(p([\mathbf{y}_{t+k\tau}]_{k=1}^{K}\,|\mathbf{x}_t)||p^{LS}([\mathbf{y}_{t+k\tau}]_{k=1}^{K}\,|\mathbf{x}_t))\right]}_{\text{Latent Simulation error for } K \text{ simulations steps with lag time } \tau} \leq K \quad \underbrace{AIG([\mathbf{z}_t]_{t=s}^{T};\tau)}_{\text{Autoinformation gap for lag time } \tau} \quad , \tag{4}$$

with $t \sim U(s, s+\tau-1)$ and $T := s + (K+1)\tau - 1$. In words, the latent simulation error is upper-bounded by the product of the number of simulation steps and the autoinformation gap. A full derivation is reported in Appendix B.9.

Given that the autoinformation between elements of the original sequence is fixed, we can train representations that minimize the autoinformation gap at resolution τ by maximizing the autoinformation between the corresponding representations at the same or higher temporal resolution. We refer to this training objective as *Time-lagged InfoMax* (T-InfoMax):

$$\mathcal{L}^{\text{T-InfoMax}}([\mathbf{x}_t]_{t=s}^{T}, \tau; \theta) := AIG([\mathbf{z}_t]_{t=s}^{T};\tau) = -\mathbb{E}_{t\sim U(s,T-\tau)}[I(\mathbf{z}_t;\mathbf{z}_{t+\tau})]. \tag{5}$$

Among the various differentiable methods for maximizing mutual information in the literature (Poole et al., 2019; Hjelm et al., 2019; Song & Ermon, 2020), we focus on noise contrastive methods (InfoNCE) due to their flexibility and computational efficiency (van den Oord et al., 2018; Chen et al., 2020). Therefore, we introduce an additional *critic* architecture $F_\xi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ with parameters ξ to define an upper-bound on the T-InfoMax loss:

$$\mathcal{L}^{\text{T-InfoMax}}([\mathbf{x}_t]_{t=s}^{T}, \tau; \theta) \leq \mathcal{L}^{\text{T-InfoMax}}_{\text{InfoNCE}}([\mathbf{x}_t]_{t=s}^{T}, \tau; \theta, \xi) \approx -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{F_\xi(\boldsymbol{z}_{t_i}, \boldsymbol{z}_{t_i-\tau})}}{\frac{1}{B}\sum_{j=1}^{B} e^{F_\xi(\boldsymbol{z}_{t_j}, \boldsymbol{z}_{t_i-\tau})}}. \tag{6}$$

In this equation, t_i is sampled uniformly in the interval $(s, T-\tau)$, $\boldsymbol{z}_{t_i}$ and $\boldsymbol{z}_{t_i-\tau}$ are the representations of $\boldsymbol{x}_{t_i}$ and $\boldsymbol{x}_{t_i-\tau}$ encoded via $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$, and B denotes the mini-batch size. We refer the reader to Appendix C.2 for additional discussion regarding the proposed approximations.

2.3 From Time-lagged InfoMax to Time-lagged Information Bottleneck

In the previous section, we emphasized the importance of maximizing autoinformation for accurate latent simulation. However, it is also critical to design representations that discard as much irrelevant information as possible. This principle, known as *Information Bottleneck* (Tishby et al., 2000), aims to simplify the implied transition $p(\mathbf{z}_t|\mathbf{z}_{t-\tau})$ and predictive $p(\mathbf{y}_t|\mathbf{z}_t)$ distributions to ease the variational fitting tasks, decreasing their sample complexity. In dynamical systems, the information that $\mathbf{z}_t$ retains about $\mathbf{x}_t$ can be decomposed into the autoinformation at the lag time τ and superfluous information:

$$\underbrace{I(\mathbf{x}_t;\mathbf{z}_t)}_{\text{Total Information}} = \underbrace{AI(\mathbf{z}_{t-\tau};\tau)}_{\text{Autoinformation at lag time } \tau} + \underbrace{I(\mathbf{x}_t;\mathbf{z}_t|\mathbf{z}_{t-\tau})}_{\text{Superfluous information}} . \tag{7}$$

As shown in Appendix B.11, superfluous information consists of time-independent features and dynamic information for temporal scales smaller than τ.

Incorporating sufficiency from equation 4 with the minimality of superfluous information we obtain a family of objectives that we denote as *Time-lagged Information Bottleneck* (T-IB):

$$\mathcal{L}^{\text{T-IB}}([\mathbf{x}_t]_{t=s}^{T}, \tau, \beta; \theta) = \mathcal{L}^{\text{T-InfoMax}}([\mathbf{x}_t]_{t=s}^{T}, \tau; \theta) + \beta\, \mathbb{E}_t\left[I(\mathbf{x}_t;\mathbf{z}_t|\mathbf{z}_{t-\tau})\right]. \tag{8}$$

Here, β is a hyperparameter that trades off sufficiency (maximal autoinformation, $\beta \to 0$) and minimality (minimal superfluous information, $\beta \to +\infty$). Given that superfluous information can not be computed directly, we provide a tractable upper bound based on the variational latent transition distribution $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau})$. Together with equation 6, this defines a tractable T-IB InfoNCE objective:

$$\mathcal{L}^{\text{T-IB}}_{\text{InfoNCE}}([\mathbf{x}_t]_{t=s}^{T}, \tau, \beta; \theta, \phi, \xi) \approx \frac{1}{B}\sum_{i=1}^{B} -\log \frac{e^{F_\xi(\boldsymbol{z}_{t_i}, \boldsymbol{z}_{t_i-\tau})}}{\frac{1}{B}\sum_{j=1}^{B} e^{F_\xi(\boldsymbol{z}_{t_j}, \boldsymbol{z}_{t_i-\tau})}} + \beta \log \frac{p_\theta(\boldsymbol{z}_{t_i}|\boldsymbol{x}_{t_i})}{q_\phi(\boldsymbol{z}_{t_i}|\boldsymbol{z}_{t_i-\tau})}, \tag{9}$$

in which the encoder $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ is parametrized using a Normal distribution with learnable mean and standard deviation as in Alemi et al. (2016); Federici et al. (2020). Details on the upper bound in equation 9 are reported in Appendix C.3.

3 Related Work

Information-theoretic methods have gained traction in fluid mechanics, offering valuable insights into energy transfer mechanisms (Betchov, 1964; Cerbus & Goldburg, 2013; Lozano-Durán & Arranz, 2022). Measures like *Transfer Entropy* (Schreiber, 2000) and *Delayed Mutual Information* (Materassi et al., 2014) closely align with the concept of *Autoinformation*, which is central in this work. However, previous literature predominantly focused on designing localized reduced-order models (Lozano-Durán & Arranz, 2022) by factorizing spatial scales and independent sub-system components, rather than learning flexible representations that capture dynamics at the desired temporal scale. Moreover, the theory and application of these principles have largely been confined to discrete-state systems (Kaiser & Schreiber, 2002) and model selection tasks (Akaike, 1974; Burnham & Anderson, 2004).

A widely used approach in dynamical system representation involves measuring and maximizing linear autocorrelation (Calhoun et al., 2001; Pérez-Hernández et al., 2013; Wiskott & Sejnowski, 2002). In particular, Sidky et al. (2020) proposes a latent simulation inference that leverages linear correlation maximization, coupled with a mixture distribution for latent transitions. As shown in Appendix D.1, autocorrelation maximization can be also interpreted as autoinformation maximization constrained to jointly Normal random variables (Borga, 2001). However, the linear restriction requires high-dimensional embeddings (Kantz & Schreiber, 2003; von Wegner et al., 2017), and may introduce training instabilities for non-linear encoders (Mardt et al., 2018; Wu & Noé, 2020; Lyu et al., 2022). In this work, we prove that the requirement of linear transitions is not necessary to capture slow-varying signals, demonstrating the benefits of modern non-linear mutual information maximization strategies.

The proposed T-InfoMax family also generalizes existing models based on the reconstruction of future states (Wehmeyer & Noé, 2018; Hernández et al., 2018). On one hand, these approaches are proven to maximize mutual information (Barber & Agakov, 2003; Poole et al., 2019), on the other their effectiveness and training costs are contingent on the flexibility of the decoder architectures (Chen et al., 2019). For this reason, we chose to maximize autoinformation using contrastive methods, which rely on a more flexible critic architecture (van den Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020) instead of a decoder[2]. While contrastive methods have already been applied to temporal series (van den Oord et al., 2018; Opolka et al., 2019; Gao & Shardt, 2022; Yang et al., 2023), our work additionally provides a formal characterization of InfoMax representations of Markov processes.

Another key contribution of our work lies in the introduction of an explicit bottleneck term to remove superfluous fast features. The proposed T-IB approach builds upon Wang et al. (2019b), which first proposes a reconstruction-based information bottleneck objective for molecular time series, utilizing a dimensionality-reducing linear encoder instead of a flexible deep neural architecture to implicitly reduce information. Wang & Tiwary (2021) later developed a related bottleneck objective, focusing on future target reconstruction instead of autoinformation maximization and using a marginal prior for compression. Although less reliant on the decoder architecture, this objective is not guaranteed to produce accurate simulation for arbitrary targets, as demonstrated in Appendix D.3.

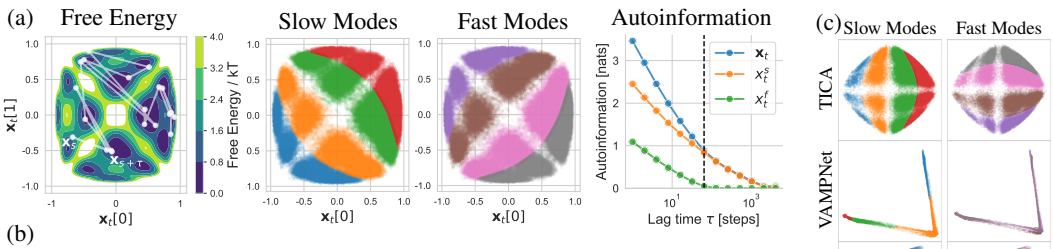
4 Experimental Results

We perform experiments on (i) a controlled dynamical system consisting of non-linear mixing of slow and fast processes, and (ii) molecular simulations of peptides. Our goal is, primarily, to examine the effect of the information maximization strategy (linear vs. contrastive) and the impact of the bottleneck regularization on the trajectories unfolded using LS. We further aim to validate our theory by estimating autoinformation and superfluous information for the models considered in this analysis.

Models We analyze representations obtained using correlation maximization methods based either on linear projections (TICA) (Molgedey & Schuster, 1994) or non-linear encoders (VAMPNet) (Mardt et al., 2018) against and non-linear autoinformation maximization (T-InfoMax) and corresponding bottleneck (T-IB) based on InfoNCE. The regularization strength β is selected based on the validation scores[3]. We use a conditional Flow++ architecture (Ho et al., 2019) to model the variational transition distribution $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau})$. This is because of the modeling flexibility, the tractability of the likelihood, and the possibility of directly sampling to unfold latent simulations. Multi-layer perceptrons (MLPs)

[2]We refer the reader to Appendix D.2 for further details.

[3]Ablation studies on the effect of β and the effect of a stochastic encoder can be found in Appendix F.1.



(b)

	$AIG([\mathbf{z}_t];64)\downarrow$	$I(\mathbf{z}_t;\mathrm{y}_t^s)\uparrow$	$I(\mathbf{z}_t;\mathrm{y}_t^f)\downarrow$	$MJS\downarrow$	$TJS@512\downarrow$
Original ($\mathbf{z}_t{=}\mathbf{x}_t$)	**0**	$\mathbf{1.31 \pm 0.00}$	1.30 ± 0.00	4.1 ± 4.7	4.0 ± 1.2
Constant ($\mathbf{z}_t{=}c$)	0.88 ± 0.00	0	**0**	13.6 ± 17.9	105.9 ± 17.2
TICA (1D)	0.10 ± 0.00	1.16 ± 0.00	0.06 ± 0.00	14.5 ± 11.2	17.9 ± 11.9
VAMPNet	0.05 ± 0.02	0.99 ± 0.03	$\mathbf{0.00 \pm 0.00}$	20.5 ± 21.0	10.4 ± 6.4
T-InfoMax	$\mathbf{0.03 \pm 0.00}$	$\mathbf{1.29 \pm 0.02}$	0.30 ± 0.07	13.0 ± 7.0	6.5 ± 2.5
T-IB ($\beta = 0.01$)	0.05 ± 0.00	$\mathbf{1.30 \pm 0.01}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{2.4 \pm 1.6}$	$\mathbf{2.4 \pm 0.3}$

Figure 3: Visualization of the results on the Prinz 2D dataset. 3a: free energy and short sample trajectory (left), samples colored by the slow and fast mode index (center), and autoinformation for the full process and its components at several lag times (right). 3b: measures of autoinformation gap, mutual information between the representation and the discrete fast and slow modes in nats, and value of marginal and transition JS divergence for unfolded sequences in milli-nats. 3c: trajectories encoded in the latent space $\mathbf{z}_t$ through various trained models. Quantitative and qualitative results confirm that T-IB uniquely captures relevant (slow) information while discarding irrelevant (fast) components. This results in more accurate LS as measured by the marginal and transition JS.

are used to model $q_\psi(\mathrm{y}_t|\mathbf{z}_t)$, mapping the representations $\mathbf{z}_t$ into the logits of a categorical distribution over the target y_t. For all objectives, we use the same encoder, transition, and predictive architectures.

Training We first train the parameters θ of the encoder $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ using each objective until convergence. Note that T-IB also optimizes the parameters of the transition model $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau})$ during this step (as shown in equation 9). Secondly, we fix the parameters θ and fit the variational transition $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau})$ and predictive $q_\psi(\mathrm{y}_t|\mathbf{z}_t)$ distributions. This second phase is identical across all the models, which are trained until convergence within a maximum computational budget (50 epochs) with the AdamW optimizer (Loshchilov & Hutter, 2019) and early stopping based on the validation score. Standard deviations are obtained by running 3 experiments for each tested configuration with different seeds. Additional details on architectures and optimization can be found in Appendix E.2.

Quantitative evaluation We estimate the autoinformation of the representations $AI([\mathbf{z}_t]_{t=s}^T;\tau)$ at several lag time τ using SMILE (Song & Ermon, 2020) and measure the amount of information that the representations contain about the targets of interest $I(\mathbf{z}_t;\mathrm{y}_t)$ using difference of discrete entropies: $H(\mathrm{y}_t) - H(\mathrm{y}_t|\mathbf{z}_t)$ (Poole et al., 2019; McAllester & Stratos, 2020). Given an initial system state $\mathbf{x}_s$ of a test trajectory $[\mathbf{x}_t]_{t=s}^T$ and the sequence of corresponding targets $[y_t]_{t=s}^T$, we use the trained encoder, transition, and prediction models to unfold trajectories $[\tilde{y}_{s+k\tau}]_{k=1}^K \sim q^{LS}([\mathrm{y}_{s+k\tau}]_{k=1}^K|\mathbf{x}_s)$ that cover the same temporal span as the test trajectory ($K = \lfloor (T-s)/\tau \rfloor$). Similarly to previous work (Arts et al., 2023), for evaluation purposes, we consider only discrete targets y_t so that we can estimate the marginal and transition probabilities for the ground truth and unfolded target trajectories by counting the frequency of each target state and the corresponding transition matrix (Figure 5a). We evaluate the fidelity of the unfolded simulation by considering the Jensen-Shannon divergence (JS) between the ground truth and unfolded target marginal (MJS) and target transition distribution for several $\tau' > \tau$ ($TJS@\tau'$). Further details on the evaluation procedures are reported in Appendix E.3.

2D Prinz Potential Inspired by previous work (Mardt et al., 2018; Wu et al., 2018) we design a 2D process consisting of a fast x_t^f and slow x_t^s components obtained from 2 independent simulations on the 1D Prinz potential (Prinz et al., 2011). This potential energy function consists of four interconnected low-energy regions, which serve as the discrete targets y_t^f and y_t^s. The two components are mixed through a linear projection and a $tanh$ non-linearity to produce a 2D process consisting of a total of 4 (fast) $\times$ 4 (slow) modes, visualized in Figure 3a. We generated separate training, validation, and test trajectories of 100K steps each. The encoders $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ consist of simple MLPs and $\mathbf{z}_t$ is fixed

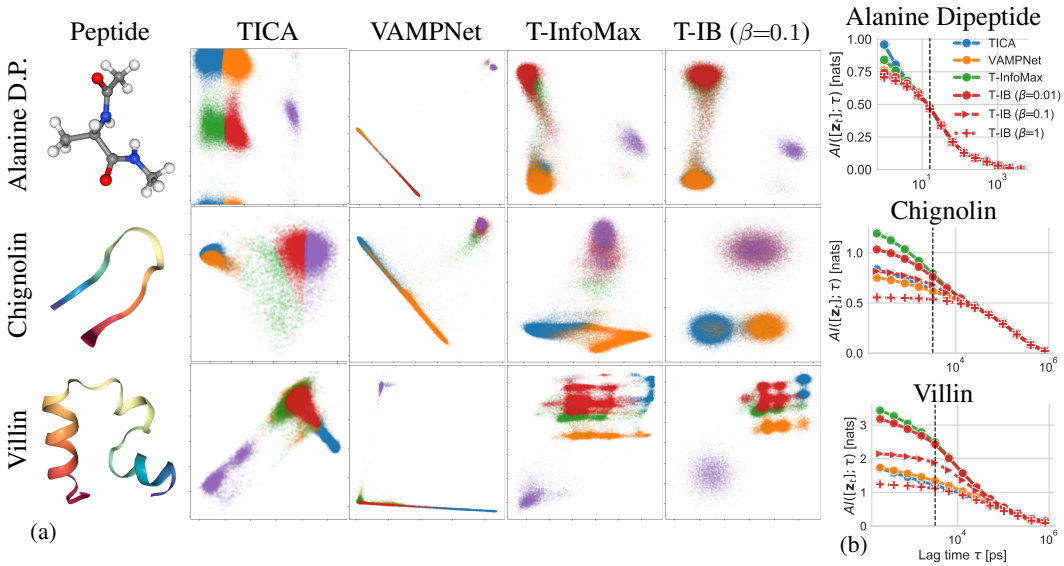


Figure 4: Comparison of 2D representations for Alanine Dipeptide, Chignolin, and Villin simulations. 4a: visualizations are colored by molecular configuration clusters $\boldsymbol{y}_t$ obtained by clustering torsion angles (Alanine Dipeptide) and TICA projections (Chignolin, Villin). 4b: corresponding values of autoinformation (y-axis) at multiple lag times (x-axis). An optimal representation should maximize autoinformation at the trained lag time τ (indicated by the dashed vertical line) while minimizing information on faster processes (to the left of the dashed line). Correlation maximization methods struggle to capture all relevant dynamics in larger systems, while T-IB regularization can effectively regulate the amount fast information in $\mathbf{z}_t$. Visually this results in simpler clustered regions.

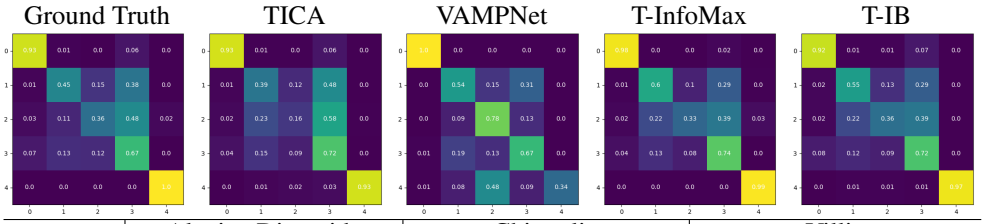
to be 2D. As shown in the autoinformation plot in Figure 3a (on the right), at the choesen train lag time ($\tau = 64$, vertical dashed line), the fast components are temporally independent, and all the relevant information is given by the slow process: $AI(\mathbf{x}_t; 64) \approx AI(\mathrm{x}_t^s; 64) > AI(\mathrm{x}_t^f; 64) \approx 0$. Therefore, information regarding x_t^f can be considered superfluous (equation 7), and should be discarded.

Figure 3c visualizes the representations obtained with several models colored by the slow (left column) and fast (right column) mode index y_t^s and y_t^f. We can visually observe that our proposed T-IB model preserves information regarding the slow process while removing all information regarding the irrelevant faster component. This is quantitatively supported by the measurements of mutual information reported in Table 3b, which also reports the values of marginal and transition JS divergence for the unfolded slow targets trajectories $[\tilde{y}_t^s]_{t=s}^T$. We observe that the latent simulations unfolded from the T-IB representations are statistically more accurate, improving even upon trajectories unfolded by fitting the transition distribution directly in the original space $\mathbf{x}_t$. We believe this improvement is due to the substantial simplification caused by the T-IB regularization.

Molecular Simulations We analyze trajectories obtained by simulating *Alanine Dipeptide* and two fast-folding mini-proteins, namely *Chignolin* and *Villin* (Lindorff-Larsen et al., 2011) in water solvent. We define disjoint *train*, *validation*, and *test* splits for each molecule by splitting trajectories into temporally distinct regions. Encoders $p_\theta(\mathbf{z}_t|\mathbf{x}_t)$ employ a TorchMD Equivariant Transformer architecture (Thölke & Fabritiis, 2022) for rotation, translation, and reflection invariance. Following previous work (Köhler et al., 2023), TICA representations are obtained by projecting invariant features such as inter-atomic distances and torsion angles. Following Arts et al. (2023), the targets y_t are created by clustering 32-dimensional TICA projections using K-means with 5 centroids. Further details on the data splits, features and targets can be found in Appendix E.1.2.

In Figure 4, we show 2D representations obtained by training the encoders on the molecular trajectories (Figure 4a), and the corresponding measure of autoinformation (Figure 4b) at several time scales (x-axis), while Figure 5 reports transition and marginal JS for trajectories unfolded on larger latent spaces (16D for Alanine simulations and 32D for Chignolin and Villin). While previous work demon-

(a)



(b)

	Alanine Dipeptide		Chignolin		Villin	
	$MJS \downarrow$	TJS@256 ps $\downarrow$	$MJS \downarrow$	TJS@51.2 ns $\downarrow$	$MJS \downarrow$	TJS@51.2ns $\downarrow$
TICA	2.3 ± 1.8	14.9 ± 2.8	8.1 ± 7.8	11.2 ± 12.1	12.1 ± 10.8	14.1 ± 1.7
VAMPNet	220 ± 350	150 ± 21	35 ± 30	106 ± 100	96 ± 64	130 ± 110
T-InfoMax	3.9 ± 1.6	13.7 ± 2.3	11.0 ± 4.0	3.3 ± 0.2	35.2 ± 37.8	9.1 ± 0.9
T-IB	0.8 ± 0.2	12.9 ± 1.2	5.5 ± 1.9	3.0 ± 0.6	22.5 ± 17.9	6.8 ± 0.9

Figure 5: Evaluation of the statistical fidelity of unfolded molecular trajectories. 5a: visualization of transition matrices for ground-truth and VLS target trajectories for different models on Villin at 51.2 ns. 5b: corresponding values of marginal and transition JS on Alanine Dipeptide, Chignolin and Villin. LS based on T-IB representations consistently results in lower simulation error, improving upon linear methods and unregularized T-InfoMax models.

strated that a linear operator can theoretically approximate expected system dynamics on large latent spaces (Koopman, 1931; Mezić, 2005), we note that models trained to maximize linear correlation (TICA, VAMPNet) face difficulties in extracting dynamic information in low dimensions even with non-linear encoders. Moreover, our empirical observations indicate that higher-dimensional representations obtained with VAMPNet yield transition and prediction distributions that are more difficult to fit (see Table 5 and Appendix F) resulting in less accurate unfolded target trajectories. Methods based on non-linear contrastive T-InfoMax produce more expressive representations in low dimensions. The addition of a bottleneck term aids in regulating the amount of information on faster processes (Figure 4b, left of the dashed line). As shown in Figure 5a and Table 5b, T-IB consistently improves the transition and marginal statistical accuracy when compared to the unregularized T-InfoMax counterpart. Results for additional targets and train lag times are reported in Appendix F. We estimated that training and unfolding Villin latent simulations of the same length of the training trajectory with T-IB take approximately 6 hours on a single GPU. In contrast, running molecular dynamics on the same hardware takes about 2-3 months. Further details on the run times can be found in Appendix G.

5 Conclusions

In this work, we propose an inference scheme designed to accelerate the simulation of Markov processes by mapping observations into a representation space where larger time steps can be modeled directly. We explore the problem of creating such a representation from an information-theoretic perspective, defining a novel objective aimed at preserving relevant dynamics while limiting superfluous information content through an Information Bottleneck. We demonstrate the effectiveness of our method from both representation learning and latent inference perspectives by comparing the information content and statistics of unfolded trajectories on synthetic data and molecular dynamics.

Limitations and Future work The primary focus of this work is characterizing and evaluating the dynamic properties of representations. Nevertheless, modeling accurate transition in the latent space remains a crucial aspect, and we believe that more flexible classes of transition models could result in higher statistical fidelity at the cost of slower sampling. Another challenging aspect involves creating representations of systems with large autoinformation content (e.g. chaotic and unstable systems). This is because the variance of modern mutual information lower bounds increases exponentially with the amount of information to extract (McAllester & Stratos, 2020). To mitigate this issue and validate the applicability of our method to other practical settings, future work will consider exploiting local similarity and studying the generalization capabilities of models trained on multiple systems and different simulation conditions. We further aim to evaluate the accuracy of long-range unfolded trajectories when only collections of shorter simulations are available during training time.

ACKNOWLEDGMENTS

We thank Frank Noé, Rianne van den Berg, Victor Garcia Satorras, Chin-Wei Huang, Marloes Arts, Wessel Bruinsma, Tian Xie, and Claudio Zeni for the insightful discussions and feedback provided throughout the project. This work was supported by the Microsoft Research PhD Scholarship Programme.

REFERENCES

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *CoRR*, abs/1612.00410, 2016. URL http://arxiv.org/abs/1612.00410.

Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 1247–1255. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/andrew13.html.

Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zügner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *CoRR*, abs/2302.00600, 2023. doi: 10.48550/arXiv.2302.00600. URL https://doi.org/10.48550/arXiv.2302.00600.

David Barber and Felix V. Agakov. The im algorithm: a variational approach to information maximization. In *NIPS 2003*, 2003.

R. Betchov. Measure of the Intricacy of Turbulence. *The Physics of Fluids*, 7(8):1160–1162, 08 1964. ISSN 0031-9171. doi: 10.1063/1.1711356. URL https://doi.org/10.1063/1.1711356.

Magnus Borga. Canonical correlation: A tutorial, Jan 2001.

Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer, 2002.

Kenneth Burnham and David Anderson. Model selection and multimodel inference. *A Practical Information-theoretic Approach*, 01 2004. doi: 10.1007/978-0-387-22456-5_5.

Vince Calhoun, Tülay Adali, Godfrey Pearlson, and J.J. Pekar. Spatial and temporal independent component analysis of functional mri data containing a pair of task-related waveforms. *Human brain mapping*, 13:43–53, 06 2001. doi: 10.1002/hbm.1024.

R. T. Cerbus and W. I. Goldburg. Information content of turbulence. *Phys. Rev. E*, 88:053012, Nov 2013. doi: 10.1103/PhysRevE.88.053012. URL https://link.aps.org/doi/10.1103/PhysRevE.88.053012.

François Chapeau-Blondeau. Autocorrelation versus entropy-based autoinformation for measuring dependence in random signal. *Physica A: Statistical Mechanics and its Applications*, 380:1–18, 2007. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2007.02.077. URL https://www.sciencedirect.com/science/article/pii/S0378437107002075.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Wei Chen, Hythem Sidky, and Andrew L. Ferguson. Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems. *CoRR*, abs/1906.00325, 2019. URL http://arxiv.org/abs/1906.00325.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

Marco Federici, David Ruhe, and Patrick Forré. On the Effectiveness of Hybrid Mutual Information Estimation. *arXiv e-prints*, art. arXiv:2306.00608, June 2023. doi: 10.48550/arXiv.2306.00608.

Ian S. Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020. doi: 10.3390/e22090999. URL https://doi.org/10.3390/e22090999.

Xinrui Gao and Yuri A.W. Shardt. Evolve·infomax: A new criterion for slow feature analysis of nonlinear dynamic system from an information-theoretical perspective. *IFAC-PapersOnLine*, 55(20):43–48, 2022. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2022.09.069. URL https://www.sciencedirect.com/science/article/pii/S2405896322012514. 10th Vienna International Conference on Mathematical Modelling MATHMOD 2022.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year enso forecasts. *Nature*, 573(7775):568–572, 2019.

Carlos X Hernández, Hannah K Wayment-Steele, Mohammad M Sultan, Brooke E Husic, and Vijay S Pande. Variational encoding of complex dynamics. *Physical Review E*, 97(6):062412, 2018.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. *CoRR*, abs/1902.00275, 2019. URL http://arxiv.org/abs/1902.00275.

Moritz Hoffmann, Martin Scherer, Tim Hempel, Andreas Mardt, Brian de Silva, Brooke E Husic, Stefan Klus, Hao Wu, Nathan Kutz, Steven L Brunton, et al. Deeptime: a python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology*, 3(1):015009, 2021.

A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1):43–62, 2002. ISSN 0167-2789. doi: https://doi.org/10.1016/S0167-2789(02)00432-3. URL https://www.sciencedirect.com/science/article/pii/S0167278902004323.

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2 edition, 2003. doi: 10.1017/CBO9780511755798.

Leon Klein, Andrew Y. K. Foong, Tor Erlend Fjelde, Bruno Mlodozeniec, Marc Brockschmidt, Sebastian Nowozin, Frank Noé, and Ryota Tomioka. Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. *CoRR*, abs/2302.01170, 2023. doi: 10.48550/arXiv.2302.01170. URL https://doi.org/10.48550/arXiv.2302.01170.

B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315. URL https://www.pnas.org/doi/abs/10.1073/pnas.17.5.315.

Jonas Köhler, Yaoyi Chen, Andreas Krämer, Cecilia Clementi, and Frank Noé. Flow-matching: Efficient coarse-graining of molecular dynamics without forces. *Journal of Chemical Theory and Computation*, 19(3):942–952, 2023. doi: 10.1021/acs.jctc.3c00016. URL https://doi.org/10.1021/acs.jctc.3c00016. PMID: 36668906.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011. doi: 10.1126/science.1208351. URL https://www.science.org/doi/abs/10.1126/science.1208351.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Adrián Lozano-Durán and Gonzalo Arranz. Information-theoretic formulation of dynamical systems: Causality, modeling, and control. *Phys. Rev. Res.*, 4:023195, Jun 2022. doi: 10.1103/PhysRevResearch.4.023195. URL https://link.aps.org/doi/10.1103/PhysRevResearch.4.023195.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=5FUq05QRc5b.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets: Deep learning of molecular kinetics. *Nature Communications*, 9, 01 2018. doi: 10.1038/s41467-017-02388-1.

Massimo Materassi, Giuseppe Consolini, Nathan Smith, and Rossana De Marco. Information theory analysis of cascading process in a synthetic model of fluid turbulence. *Entropy*, 16(3):1272–1286, 2014. doi: 10.3390/e16031272. URL https://doi.org/10.3390/e16031272.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 2020. URL http://proceedings.mlr.press/v108/mcallester20a.html.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41:309–325, 2005.

L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3637, Jun 1994. doi: 10.1103/PhysRevLett.72.3634. URL https://link.aps.org/doi/10.1103/PhysRevLett.72.3634.

Frank Noé and Feliks Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.*, 11(2):635–655, 2013. doi: 10.1137/110858616. URL https://doi.org/10.1137/110858616.

J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi: 10.1017/CBO9780511810633.

Felix L. Opolka, Aaron Solomon, Catalina Cangea, Petar Velickovic, Pietro Liò, and R. Devon Hjelm. Spatio-temporal deep graph infomax. *CoRR*, abs/1904.06316, 2019. URL http://arxiv.org/abs/1904.06316.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019. URL http://proceedings.mlr.press/v97/poole19a.html.

J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F Noé. Markov models of molecular kinetics: generation and validation. *The Journal of Chemical Physics*, 134(17):174105, 2011. doi: https://doi.org/10.1063/1.3565032.

Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1):015102, 07 2013. ISSN 0021-9606. doi: 10.1063/1.4811489. URL https://doi.org/10.1063/1.4811489.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3875–3879. IEEE, 2021.

Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000. doi: 10.1103/PhysRevLett.85.461. URL https://link.aps.org/doi/10.1103/PhysRevLett.85.461.

Mathias Schreiner, Ole Winther, and Simon Olsson. Implicit Transfer Operator Learning: Multiple Time-Resolution Surrogates for Molecular Dynamics. *arXiv e-prints*, 2023.

David E. Shaw, Peter J. Adams, Asaph Azaria, Joseph A. Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J. Adam Butts, Timothy Correia, Robert M. Dirks, Ron O. Dror, Michael P. Eastwood, Bruce Edwards, Amos Even, Peter Feldmann, Michael Fenn, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Maria Gorlatova, Brian Greskamp, J. P. Grossman, Justin Gullingsrud, Anissa Harper, William Hasenplaugh, Mark Heily, Benjamin Colin Heshmat, Jeremy Hunt, Douglas J. Ierardi, Lev Iserovich, Bryan L. Jackson, Nick P. Johnson, Mollie M. Kirk, John L. Klepeis, Jeffrey S. Kuskin, Kenneth M. Mackenzie, Roy J. Mader, Richard McGowen, Adam McLaughlin, Mark A. Moraes, Mohamed H. Nasr, Lawrence J. Nociolo, Lief O'Donnell, Andrew Parker, Jon L. Peticolas, Goran Pocina, Cristian Predescu, Terry Quan, John K. Salmon, Carl Schwink, Keun Sup Shim, Naseer Siddique, Jochen Spengler, Tamas Szalay, Raymond Tabladillo, Reinhard Tartler, Andrew G. Taube, Michael Theobald, Brian Towles, William Vick, Stanley C. Wang, Michael Wazlowski, Madeleine J. Weingarten, John M. Williams, and Kevin A. Yuh. Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In Bronis R. de Supinski, Mary W. Hall, and Todd Gamblin (eds.), *International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2021, St. Louis, Missouri, USA, November 14-19, 2021*, pp. 1. ACM, 2021. doi: 10.1145/3458817.3487397. URL https://doi.org/10.1145/3458817.3487397.

Hythem Sidky, Wei Chen, and Andrew L. Ferguson. Molecular latent space simulators. *Chem. Sci.*, 11:9459–9467, 2020. doi: 10.1039/D0SC03635H. URL http://dx.doi.org/10.1039/D0SC03635H.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1x62TNtDS.

Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zNHzqZ9wrRB.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000. URL http://arxiv.org/abs/physics/0004057.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019, 2020.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

F. von Wegner, E. Tagliazucchi, and H. Laufs. Information-theoretical analysis of resting state eeg microstate sequences - non-markovianity, non-stationarity and periodicities. *NeuroImage*, 158:99–111, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.06.062. URL https://www.sciencedirect.com/science/article/pii/S1053811917305347.

Jiří Vymětal and Jiří Vondrášek. Metadynamics as a tool for mapping the conformational and free-energy space of peptides — the alanine dipeptide case study. *The Journal of Physical Chemistry B*, 114(16):5632–5642, 2010. doi: 10.1021/jp100950w. URL https://doi.org/10.1021/jp100950w. PMID: 20361773.

Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154(13):134111, 04 2021. ISSN 0021-9606. doi: 10.1063/5.0038198. URL https://doi.org/10.1063/5.0038198.

Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E Charron, Gianni De Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS central science*, 5(5):755–767, 2019a.

Yihang Wang, João Marcelo Lamim Ribeiro, and Pratyush Tiwary. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications*, 10, 2019b. URL https://api.semanticscholar.org/CorpusID:199491659.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics*, 148(24):241703, 2018.

Thomas Wennekers and Nihat Ay. Temporal infomax on markov chains with input leads to finite state automata. *Neurocomputing*, 52-54:431–436, 2003. ISSN 0925-2312. doi: https://doi.org/10.1016/S0925-2312(02)00862-7. URL https://www.sciencedirect.com/science/article/pii/S0925231202008627. Computational Neuroscience: Trends in Research 2003.

Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.*, 14(4):715–770, 2002. doi: 10.1162/089976602317318938. URL https://doi.org/10.1162/089976602317318938.

Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *J. Nonlinear Sci.*, 30(1):23–66, 2020. doi: 10.1007/s00332-019-09567-y. URL https://doi.org/10.1007/s00332-019-09567-y.

Hao Wu, Andreas Mardt, Luca Pasquali, and Frank Noé. Deep generative markov state models. *CoRR*, abs/1805.07601, 2018. URL http://arxiv.org/abs/1805.07601.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Yuncong Yang, Jiawei Ma, Shiyuan Huang, Long Chen, Xudong Lin, Guangxing Han, and Shih-Fu Chang. Tempclr: Temporal alignment representation with contrastive learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=CIFOsnhZvON.