Deep Learning the slow modes for rare event sampling

Supervisors:

Dr Rakesh S Singh

Department of Physics

Semester Project submitted by:

Venkata Sai Sivam Mantha

Department of Physics

Semster:- Aug2025

1 Introduction

The configuration space is a $3N$-dimensional space in which each point corresponds to a unique configuration of the system. The Free Energy Surface (FES) is a function f defined in the configuration space:

$$f : \mathbb{R}^{3N} \to \mathbb{R} \tag{1}$$

By understanding the topography of the configuration surface, we can develop pathways for the reaction to reach the global minima of the energy without getting stuck at the local minima. Atomistic simulations, and in particular molecular dynamics (MD), play an important role in several fields of science, serving as a virtual microscope that greatly aids in the study of physical, chemical, and biological processes. However, whenever the free energy barrier between metastable states is large relative to the thermal energy, k_BT, transitions between states become rare events, occurring on timescales too long to be simulated by standard methods. Different sampling problems have been developed and a large amount of these depend upon identification of a small set of collective variables. The Collective variables are functions of system atomic Cooridnates $\mathbb{R}^{3n}$.
Now to develop good collective variables we need to have transition dynamics of the system and to simulate transition dynamics successfully, we need to have good collective variables and this has become a chicken and a egg problem. A solution to this conundrum is to use collective variable developed during biased simulations and model the behaviour of collective variable as bias in biased simulation changes with time.

This report has two parts:

1. **Mathematics**
   In this section, we discuss the main mathematical ideas from functional analysis and measure theory that will aid us in our pursuit.

2. **Physics**
   In this section, we discuss how the mathematical ideas discussed are applied to our work.

2 Mathematics

2.1 Markov Processes

Let $x(t)$ be an element of our configuration space, i.e. $x(t) \in \Omega \subset \mathbb{R}^{3N}$. Each $x(t + \Delta t)$ depends only on $x(t)$, meaning that the dynamics depend only on the position in the previous step and not on earlier states.

This property is known as **Markovianity**.

Let $x(t), y(t) \in \Omega$ and let dy be an infinitesimal region around y. As a result of Markovianity in Ω, the transition probability that a trajectory starting at time t from point x will be in region dy is given by

$$p(x, y, \tau)\, dy = \mathbb{P}[x(t + \Delta t) \in y + dy]. \tag{2}$$

Let $A \subset \Omega$. Then

$$p(x, A; \tau) = \mathbb{P}[x(t+\tau) \in A \mid x(t) = x] = \int_{y \in A} dy\, p(x, y, \tau). \tag{3}$$

2.2 ergodicity

The process $x(t)$ is said to be *ergodic*, meaning that the space does not contain two or more dynamically disconnected subsets, and for $t \to \infty$, each state x will be visited infinitely often. The fraction of time that the system spends in any of its states during an infinitely long trajectory is given by its unique stationary density (invariant measure) $\mu(x) : \Omega \to \mathbb{R}_0^+$, corresponding to the equilibrium probability density for a thermodynamic ensemble (e.g., NVT, NpT).

For molecular dynamics at constant temperature T, the dynamics yield a stationary density $\mu(x)$ that depends on T and is given by the Boltzmann distribution:

$$\mu(x) = Z(\beta)^{-1} \exp(-\beta H(x)), \tag{4}$$

where $H(x)$ is the Hamiltonian, $\beta = 1/(k_B T)$, with k_B the Boltzmann constant, and $k_B T$ the thermal energy.

The partition function is defined as

$$Z(\beta) = \int dx\, \exp(-\beta H(x)). \tag{5}$$

2.3 detailed balance

The process $x(t)$ is said to be *reversible* if the transition probability $p(x, y; \tau)$ satisfies the condition of detailed balance:

$$\mu(x)\, p(x, y; \tau) = \mu(y)\, p(y, x; \tau), \tag{6}$$

meaning that, in equilibrium, the fraction of systems transitioning from x to y per unit time is equal to the fraction transitioning from y to x.

2.4 Lebsegue integrable functions

Given a measure space $(X, \mathcal{F}, \mu)$, the space

$$L^1(X) = \left\{ f : X \to \mathbb{R} \;\middle|\; \int_X |f(x)|\, d\mu(x) < \infty \right\}.$$

So any function whose absolute value is integrable belongs to $L^1(X)$.

Let (X, F, μ) be a measure space (In our case $X = \mathbb{R}^{3n}$, volume element dx is the Lebesgue measure). A probability density function (PDF) $p : X \to [0, \infty)$ satisfies:

$$\int_X p(x)\, dx = 1.$$

$$\begin{array}{ccc} p_t & \xrightarrow{\mathcal{Q}(\tau)} & p_{t+\tau} \\ \downarrow \cdot\mu^{-1} & & \uparrow \cdot\mu \\ u_t & \xrightarrow{\mathcal{T}(\tau)} & u_{t+\tau} \end{array}$$

Figure 1: Diagram equating different relation between p_t and μ_t

So the set of all probability densities forms a subset of $L^1(X)$.
Since $L^1(x)$ is a Banach space, we can define linear transformations on $L^1(x)$ and hence on our probability densities.

2.5 Transfer Operator

The change of the probability density $p_t(x)$ from the time t to time $t+\tau$ can be described by the action of a continuous operator. Let us define propogator $\mathcal{Q}(\tau)$ as follows:

$$p_{t+\tau}(y) = \mathcal{Q}(\tau) \circ p_t(y) = \int dx\, p(x, y; \tau)\, p_t(x).$$

An equivalent description is provided by the transfer operator $\mathcal{T}(\tau)$, which has nicer properties from a mathematical point of view. $T(\tau)$ is defined as:

$$u_{t+\tau}(y) = \mathcal{T}(\tau) \circ u_t(y) = \frac{1}{\mu(y)} \int dx\, p(x, y; \tau)\, \mu(x)\, u_t(x).$$

where $\mu(x)$, $\mu(y)$ is our stationary probability density at x and y.

$T(\tau)$ does not propagate probability densities, but instead functions $u_t(x)$ that differ from probability densities by a factor of stationary density $\mu(x)$, i.e.,

$$p_t(x) = \mu(x)\, u_t(x).$$

2.6 dominant spectrum

Both the operators $\mathcal{T}$ and $\mathcal{Q}$ follow Chapman-Kolmogrov equations given by

$$u_{t+k\tau} = \mathcal{T}^k \circ\ u_t(x)$$

$$p_{t+k\tau} = \mathcal{Q}^k \circ\ p_t(x)$$

$\mathcal{Q}(\tau)$ has eigenfunctions $\phi_i(x)$ and associated eigenvalues λ_i [see Figs. 1(c) and 1(e)]:

$$\mathcal{Q}(\tau) \circ \phi_i(x) = \lambda_i\, \phi_i(x), \tag{7}$$

while $\mathcal{T}(\tau)$ has eigenfunctions $\psi_i(x)$ with the same corresponding eigenvalues:

$$\mathcal{T}(\tau) \circ \psi_i(x) = \lambda_i\, \psi_i(x). \tag{8}$$

Given any two state functions $u_1(x)$ and$u_2(x)$ can be written as

$$\begin{aligned}\langle u_1(x)\,\mathcal{T}\,|u_2(x)\rangle_\mu &= \int dx\, u_1(x) \int dy\, p(y|x)\, u_2(y)\, \mu(y) \\ &= \int dx\, u_1(x) \int dy\, p(x|y)\, u_2(y)\, \mu(x) \\ &= \int dy\, u_2(y) \int dx\, p(x|y)\, u_1(x)\, \mu(x) \\ &= \int dx\, u_2(x) \int dy\, p(y|x)\, u_1(y)\, \mu(y) \\ &= \langle \mathcal{T}\, u_1(x)\,|u_2(x)\rangle_\mu\end{aligned}$$

which demonstrates that $\mathcal{T}$ is self-adjoint with respect to the inner product

$$\langle a, b\rangle = \int a(x)\, b(x)\, d\mu(x).$$

Let $\psi_i(x)$ be the eigenfunctions of $\mathcal{T}$ corresponding to the eigenvalues λ_i in non-ascending order:

$$\mathcal{T}\,\psi_i(x) = \lambda_i\,\psi_i(x). \tag{9}$$

The self-adjoint nature of $\mathcal{T}$ implies that it possesses real eigenvalues λ_i and its eigenvectors $\psi_i(x)$ form a complete orthonormal basis, with orthonormality relations:

$$\langle \psi_i, \psi_j\rangle = \delta_{ij}. \tag{10}$$

Normalization of the transition density

$$\int dx\, p(x \mid y) = 1$$

together with the assumption of ergodicity implies that the eigenvalue spectrum is bounded from above by a unique unit eigenvalue such that

$$\lambda_0 = 1 > \lambda_1 \geq \lambda_2 \geq \ldots$$

. Any state $u_t(x)$ at a specific time t can be written as a linear expansion in this basis of ψ_i:

$$u_t(x) = \sum_i c_i(t)\,\psi_i(x),$$

where $c_i(t)$ are the expansion coefficients.

The evolution of $u_t(x)$ to $u_{t+k\tau}(x)$ after a time period $k\tau$ can be written as

$$\begin{aligned}u_{t+k\tau}(x) = \mathcal{T}^k \circ u_t(x) &= \sum_i \langle \psi_i|u_t\rangle_\pi \mathcal{T}^k \psi_i(x) \\ &= \sum_i \langle \psi_i|p_t\rangle_\pi \lambda_i^k \psi_i(x)\end{aligned}$$

$$= \sum_i \langle \psi_i | p_t \rangle_\pi \exp\left(-\frac{k\tau}{t_i}\right) \psi_i(x),$$

where t_i is the implied time scale corresponding to the eigenfunction i, given by

$$t_i = -\frac{\tau}{\log \lambda_i}.$$

We can decompose $\mathcal{T}$ into $\mathcal{T}_{slow} + \mathcal{T}_{fast}$
This will give us

$$u_{t+k\tau}(x) = T_{\text{slow}}(k\tau) \circ u_t(x) + T_{\text{fast}}(k\tau) \circ u_t(x), \tag{11}$$

$$= \sum_{i=1}^{m} \lambda_i^k \langle u_t, \phi_i \rangle \psi_i(x) + T_{\text{fast}}(k\tau) \circ u_t(x), \tag{12}$$

$$= \sum_{i=1}^{m} \lambda_i^k \langle u_t, \psi_i \rangle_\mu \mu\psi_i(x) + T_{\text{fast}}(k\tau) \circ u_t(x). \tag{13}$$

Which implies

$$u_{t+k\tau}(x) = 1 + \sum_{i=2}^{m} \exp\left(-\frac{k\tau}{t_i}\right) \langle u_t, \psi_i \rangle_\mu \mu\psi_i(x) + T_{\text{fast}}(k\tau) \circ u_t(x).$$

2.7 Variational approach to conformational dynamics

$\psi_0 = 1$ is our trivial eigenfunction with eignevalue λ=1 corresponding to equilibrium distribution at t$\to \infty$.

$$\mathcal{T}\,\psi_0(x) = \mathcal{T} \circ \mathbf{1} = \frac{1}{\mu(x)} \int dy\, p(y,x)\, \mu(y) = \frac{1}{\mu(x)} \int dy\, p(x,y)\, \mu(x) = \mathbf{1} = \psi_0(x).$$

To learn $\mathbf{1}(x)$, we note that any state function $u(x)$ which is orthogonal to $\psi_0(x)$ can be expressed as

$$u(x) = \sum_{i>1} \langle u | \psi_i \rangle_\mu \psi_i(x) = \sum_{i>1} c_i\, \psi_i(x).$$

where $c_i = \langle \psi_i, u \rangle$ are the expansion coefficients, and where $c_i = \langle \psi_i | u \rangle_\pi$ are expansion coefficients, and

$$\tilde{\lambda} = \frac{\langle u \mid \mathcal{T} \circ u \rangle_\pi}{\langle u \mid u \rangle_\pi} = \frac{\sum_{i\geq 1} c_i^2 \lambda_i}{\sum_{i\geq 1} |c_i^2|} \leq \frac{\sum_{i\geq 1} c_i^2 \lambda_1}{\sum_{i\geq 1} c_i^2} = \lambda_1.$$

$\tilde{\lambda}$ is upper bounded by λ

We can exploit this fact to approximate the first non-trivial eigenfunction ψ_1 by searching for a function u that maximizes an appropriate objective, subject to $u \perp 1$ i.e $\langle u | \psi_0 \rangle = 0$ and $u \neq 0$. The learned u is then an approximation to the first non-trivial eigenfunction ψ_1.

We can continue this procedure to approximate higher order eigenfunctions. In general we approximate $\psi_i(x)$ by maximizing

$$\tilde{\lambda}_i = \frac{\langle \tilde{\psi}_i \mid \mathcal{T} \circ \tilde{\psi}_i \rangle_\pi}{\langle \tilde{\psi}_i \mid \tilde{\psi}_i \rangle_\pi},$$

under the orthogonality constraints

$$\langle \tilde{\psi}_k \mid \tilde{\psi}_i \rangle_\pi = 0, \quad 0 \leq k < i.$$

Given an arbitrary input basis $\{\zeta_j\}$, the eigenfunction approximations may be written as linear expansions

$$\psi_i = \sum_j s_{ij}\, \zeta_j \tag{14}$$

Adopting this basis, the VAC can be shown to lead to a generalized eigenvalue problem analogous to the quantum mechanical Roothaan–Hall equations :

$$C\, s_i = \lambda_i\, Q\, s_i, \tag{15}$$

where

$$C_{jk} = \langle \phi_j(x),\, T\, \phi_k(x) \rangle, \tag{16}$$

$$Q_{jk} = \langle \phi_j(x),\, \phi_k(x) \rangle. \tag{17}$$

Here s_i is the (eigen)vector of linear expansion coefficients for the approximate eigenfunction ψ_i, and λ_i is the corresponding eigenvalue. Solving (15) gives the best linear estimates of the eigen functions within the basis $\mathcal{C}_j$

2.8 Estimates of VAC equations from trajectory data

Here we show how Eq. [12] and [13] can be estimated from empirical trajectory data [1][2][27]. The numerator of Eq. [12] becomes

$$\langle \bar{\psi}_i \mid \mathcal{T} \bar{\psi}_i \rangle_\mu = \int dx\, \mu(x) \bar{\psi}_i(x) \frac{1}{\mu(x)} \int dy\, p_\tau(y,x) \bar{\psi}_i(y) \mu(y) = \int dx\, dy\, \bar{\psi}_i(x) p_\tau(y,x) \bar{\psi}_i(y) \mu(y)$$

$$= \int dx\ dy\ \bar{\psi}_i(x) \mathbb{P}(x_{t+\tau} = x | x_t = y) \bar{\psi}_i(y) \mathbb{P}(x_t = y) \approx \mathrm{E}\left[\bar{\psi}_i(x_t) \bar{\psi}_i(x_{t+\tau})\right],$$

(18)

where $\mathrm{E}\left[\bar{\psi}_i(x_t) \bar{\psi}_i(x_{t+\tau})\right]$ can be estimated from a trajectory $\{x_t\}$. The denominator follows similarly as

$$\langle \bar{\psi}_i \mid \bar{\psi}_i \rangle_\mu = \int dx\ \mu(x) \bar{\psi}_i(x) \bar{\psi}_i(x) = \int dx\ \bar{\psi}_i(x) \bar{\psi}_i(x) \mathbb{P}(x_t = x) \approx \mathrm{E}\left[\bar{\psi}_i(x_t) \bar{\psi}_i(x_t)\right].$$

(19)

The full expression for Eq. [12] becomes

$$\tilde{\lambda}_i = \frac{\langle \bar{\psi}_i \mid \mathcal{T} \circ \bar{\psi}_i \rangle_\pi}{\langle \psi_i \mid \bar{\psi}_i \rangle_\pi} \approx \frac{\mathrm{E}\left[\bar{\psi}_i(x_t)\bar{\psi}_i(x_{t+\tau})\right]}{\mathrm{E}\left[\bar{\psi}_i(x_t)\bar{\psi}_i(x_t)\right]}.$$

(20)

Similarly, Eq. [13] becomes

$$\langle \bar{\psi}_k \mid \bar{\psi}_i \rangle_\pi \approx \mathrm{E}\left[\bar{\psi}_k(x_t)\bar{\psi}_i(x_t)\right] = 0,\ 0 \leq k < i.$$

(21)

Using the same reasoning, the components (Eq. [16] and [17]) of the generalized eigenvalue problem (Eq. [15]) are estimated as

$$C_{jk} = \langle \zeta_j(x) | \mathcal{T} \circ \zeta_k(x) \rangle_\pi \approx \mathrm{E}\left[\zeta_j(x_t)\zeta_k(x_{t+\tau})\right],$$

(22)

$$Q_{jk} = \langle \zeta_j(x) | \zeta_k(x) \rangle_\pi \approx \mathrm{E}\left[\zeta_j(x_t)\zeta_k(x_t)\right].$$

(23)

2.9 Time Independent Component analysis

In TICA, we represent $\tilde{\psi}_i(x)$ as a linear combination of molecular coordinates x, where a_i is a vector of linear expansion coefficients and C is an additive constant

$$\tilde{\psi}_i(x) = a_i \cdot x + C.$$

(24)

The orthogonality condition Eq. [13] of $\tilde{\psi}_i$ relative to $\tilde{\psi}_0(x) = 1$ becomes

$$0 = \int dx\ \mu(x)\tilde{\psi}_0(x)\tilde{\psi}_i(x) = \int dx\ \mu(x)\tilde{\psi}_i(x) = \mathrm{E}\left[\tilde{\psi}_i(x)\right]$$

(25)

It follows that

$$0 = \mathrm{E}\left[\tilde{\psi}_i(x)\right] = \mathrm{E}\left[a_i \cdot x + C\right] = a_i \cdot \mathrm{E}\left[x\right] + C$$

(26)

$$\Rightarrow C = -a_i \cdot \mathrm{E}\left[x\right],$$

(27)

and therefore Eq. [24] can be written as

$$\tilde{\psi}_i(x) = a_i \cdot x - a_i \cdot \mathrm{E}\left[x\right] = a_i \cdot \delta x,$$

(28)

where $\delta x = x - \mathrm{E}[x]$ is a mean-free coordinate. Under this specification for $\tilde{\psi}_i(x)$, Eq. [12] and Eq. [13] become

$$\tilde{\lambda}_i = \frac{\langle \tilde{\psi}_i \mid \mathcal{T} \circ \tilde{\psi}_i \rangle_\pi}{\langle \tilde{\psi}_i \mid \tilde{\psi}_i \rangle_\pi} = \frac{\mathrm{E}\left[(a_i \cdot \delta x_t)\,(a_i \cdot \delta x_{t+\tau})\right]}{\mathrm{E}\left[(a_i \cdot \delta x_t)^2\right]},$$

(29)

$$0 = \langle \tilde{\psi}_0 | \tilde{\psi}_1 \rangle_\mu = \mathbb{E}[(a_i.\delta x_t).(a_k.\delta x_t)]$$

which are exactly the objective function and orthogonality constraints of TICA.

2.10 State Free Reversible VAMP nets (SRV's)

In general, the eigenfunction approximations need not be a linear function within either the input coordinate space or the feature space.

$$\begin{aligned}\tilde{\lambda}_i &= \frac{\langle \tilde{\psi}_i \mid \mathcal{J} \circ \tilde{\psi}_i \rangle_\pi}{\langle \tilde{\psi}_i \mid \tilde{\psi}_i \rangle_\pi} \\ &= \frac{\mathbf{E}\left[\tilde{\psi}_i(x_t)\tilde{\psi}_i(x_{t+\tau})\right]}{\mathbf{E}\left[\tilde{\psi}_i^2(x_t)\right]},\end{aligned}$$

$$0 = \langle \tilde{\psi}_k \mid \tilde{\psi}_i \rangle_\pi = \mathrm{E}\left[\tilde{\psi}_i(x_t)\tilde{\psi}_k(x_t)\right].$$

We now introduce the SRV approach that employs a neural network f to learn nonlinear approximations to $\{\tilde{\psi}_i\}$ directly without requiring the kernel trick. The neural network f maps configuration x to a n-dimensional output $f_i(x)$ $(i = 1, ..., n)$, where n is the number of slow modes we want to learn. Then a linear variational method is applied to obtain the corresponding $\tilde{\psi}_i(x)$ such that $\{\tilde{\psi}_i(x)\}$ form an orthonormal basis set that minimizes the network loss function.

The method proceeds as follows. Given a neural network f with n-dimensional outputs, we feed a training set $X = \{x\}$ and train the network with loss function L

$$L = \sum_i g(\tilde{\lambda}_i) = \sum_i g\left(\frac{\mathrm{E}\left[\tilde{\psi}_i(x_t)\tilde{\psi}_i(x_{t+\tau})\right]}{\mathrm{E}\left[\tilde{\psi}_i^2(x_t)\right]}\right).$$

(38)

where g is a monotonically decreasing function. Minimizing L is equivalent to maximizing the sum over $\tilde{\lambda}_i$ and therefore maximizing the sum over t_i (Eq. 9). The $\{\tilde{\psi}_i\}$ correspond to linear combinations of the neural network outputs $\{f_i(x)\}$

$$\tilde{\psi}_i(x) = \sum_j s_{ij} f_j(x),$$

(39)

computed by applying the linear VAC to the neural network outputs . The linear VAC is equivalent to the generalized eigenvalue problem in Eq. [15] where $\zeta_j(x) = f_j(x)$.

Minimization of the loss function by gradient descent requires the derivative of L with respect to neural network parameters θ

$$\frac{\partial L}{\partial \theta} = \sum_i \frac{\partial L}{\partial g} \frac{\partial g}{\partial \lambda_i} \frac{\partial \tilde{\lambda}_i}{\partial \theta}.$$

(40)

The first two partial derivatives are straightforwardly calculated by automatic differentiation once a choice for g has been made. The third partial derivative requires a little more careful consideration. We first expand this final derivative using Eq. [15] to make explicit the dependence of λ_i on the matrices $\tilde{C}$ and $\tilde{Q}$

$$\frac{\partial \tilde{\lambda}_i}{\partial \theta} = \frac{\partial \tilde{\lambda}_i}{\partial C} \frac{\partial C}{\partial \theta} + \frac{\partial \tilde{\lambda}_i}{\partial Q} \frac{\partial Q}{\partial \theta}.$$

(41)

To our best knowledge, no existing computational graph frameworks provide gradients for generalized eigenvalue problems. Accordingly, we rewrite Eq. [15] as follows. We first apply a Cholesky decomposition to Q, such that

$$Cs_i = \tilde{\lambda}_i L L^T s_i,$$

(42)

where L is a lower triangular matrix. We then left multiply both sides by L^{-1} to obtain

$$\left(L^{-1} C (L^T)^{-1}\right) \left(L^T s_i\right) = \tilde{\lambda}_i \left(L^T s_i\right).$$

(43)

Defining $\tilde{C} = L^{-1} C (L^T)^{-1}$ and $\tilde{s}_i = L^T s_i$ we convert the generalized eigenvalue problem into a standard eigenvalue with a symmetric matrix $\tilde{C}$

$$\tilde{C} \tilde{s}_i = \tilde{\lambda}_i \tilde{s}_i,$$

(44)

where the Cholesky decomposition assures numerical stability. Now, Eq. [41] becomes

$$\frac{\partial \tilde{\lambda}_i}{\partial \theta} = \frac{\partial \tilde{\lambda}_i}{\partial \tilde{C}} \left(\frac{\partial \tilde{C}}{\partial C} \frac{\partial C}{\partial \theta} + \frac{\partial \tilde{C}}{\partial Q} \frac{\partial Q}{\partial \theta} \right)$$

(45)

where all terms are computable using automatic differentiation: $\frac{\partial \tilde{\lambda}_j}{\partial \tilde{C}}$ from routines for a symmetric matrix eigenvalue problem, $\frac{\partial \tilde{C}}{\partial C}$ and $\frac{\partial \tilde{C}}{\partial Q}$ from those for Cholesky decomposition, matrix inversion, and matrix multiplication, and $\frac{\partial C}{\partial \theta}$ and $\frac{\partial Q}{\partial \theta}$ by applying the chain rule to

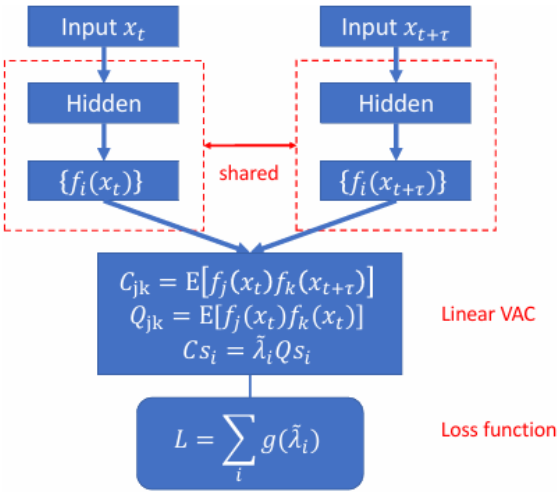

Figure 2: Schematic diagram of state-free reversible VAMPnets (SRVs). A pair of configurations $(x_t, x_{t+\tau})$ are fed into a Siamese neural network in which each subnet possesses the same architecture and weights. The Siamese net generates network outputs $f_i(x_t)$ and $f_i(x_{t+\tau})$. The mappings $f_i : x \mapsto f_i(x)$ evolve as network training proceeds and the network weights are updated. Optimal linear combinations of f_i produce estimates of the transfer operator eigenfunctions ψ_i by applying linear VAC, which can be formulated as a generalized eigenvalue problem. Network training proceeds by backpropagation and is terminated when the loss function $L = \sum_i g(\tilde{\lambda}_i)$ is minimized.

Eq. [22] and [23] with $\zeta_j(x) = f_j(x)$ and computing the derivatives $\frac{\partial f}{\partial \theta}$ through the neural network.

Training is performed by passing $\{x_i, x_{i+\tau}\}$ pairs to the SRV and updating the network parameters using mini-batch gradient descent using Adam and employing the automatic differentiation expression for $\partial L / \partial \theta$ to minimize the loss function. To prevent overfitting, we shuffle all $\{x_i, x_{i+\tau}\}$ pairs and reserve a small portion as validation set with which to implement early stopping. Training is terminated when validation loss no longer decreases for a pre-specified number of epochs.

3 Physics

3.1 OPES

The OPES method can be used to sample either kind of target distributions. It does so by adding to the potential energy of the system $U(x)$ a bias potential $V(x)$ such that the sampled distribution is not the equilibrium Boltzmann distribution, $P(x) \propto e^{-\beta U(x)}$, but the target one, $p_{\mathrm{tg}}(x)$. This bias potential is defined as

$$V(x) = -\frac{1}{\beta} \log \frac{p_{\mathrm{tg}}(x)}{P(x)}, \tag{14}$$

where β is the inverse Boltzmann temperature. The bias potential is not known a priori, but it is self-consistently learned during the simulation via an on-the-fly estimate of the probability distributions. Statistics of the unbiased system can be retrieved via a reweighting procedure, by assuming that the bias is updated in an adiabatic way. Given any observable $O(x)$, its ensemble average $\langle O \rangle$ over the unbiased system can be estimated via ensemble averages over the sampled biased system:

$$\langle O \rangle = \frac{\langle O e^{\beta V} \rangle_V}{\langle e^{\beta V} \rangle_V}. \tag{15}$$

At convergence, the free energy surface (FES) as a function of s is computed from $F(s) = -k_B T \log P(s)$.

3.2 Outline of the Procedure

We outline here the key steps of our recommended procedure (Fig. 1).

1. **Exploration.** Harness a number of reactive events using a CV-based OPES simulation with a trial CV s_0, multithermal sampling (in such a case, $s_0 = U(R)$), or even a combination of the two. Store the final bias potential $V^*(s_0)$ of this initial simulation.

2. **CV construction.** Select the descriptors to be used as inputs of the NN. Train the Deep-TICA CVs using the trajectories generated in step 1 by calculating the correlation functions in time t.

3. **Sampling.** Perform an OPES simulation using the leading Deep-TICA eigenfunction as CV on the Hamiltonian modified by the addition of the bias potential $V^*(s_0)$.

4 Future Direction

We aim to apply this technique to the polymorph selection problem.The "polymorph selection problem" is the challenge of reliably producing a specific crystal form (polymorph) of a substance from multiple possible forms, which is critical in industries like pharmaceuticals where different polymorphs have different properties. This problem is addressed by understanding and controlling the kinetics and thermodynamics of crystallization, with factors like temperature, solvent, shear stress, and pressure influencing the outcome.

-The End-