Data-driven discovery and enhanced sampling in slow collective variables

Acknowledgements

Collab: Prof. J.D. Tovar (JHU)  
Prof. Howard Katz (JHU)  
Prof. André Schleife (UIUC)  
Prof. Charles Schroeder (UIUC)  
Prof. Rama Ranganathan (UofC)  
Prof. Aaron Esser-Kahn (UofC)  
Prof. Juan de Pablo (UofC)  
Prof. Amish Patel (UPenn)

GS: Kirill Shmilovich  
Xinran Lian  
Nick Rego  
Shiqi Chen  
Walter Alvarado  
Max Topel  
Mike Jones  
Nick Herringer  
Kate Johnson  
Nikša Praljak  
Jianming Mao  
Alex Berlaga

UG: Smayan Khanna  
Diya Ghandi  
Junhee Lee

Post-Docs: Dr. Siva Dasetty

We are hiring for three postdocs

Ferguson Laboratory
www.ferglab.com

MICCoM
5J-30161-0021A

NSF
DMR-1841800
CHE-1841805
DMS-1841810
DMR-1841807
DMR-1844505
ACI-1547580
CHE-2152521

THE UNIVERSITY OF CHICAGO
PRITZKER SCHOOL OF MOLECULAR ENGINEERING
THE UNIVERSITY OF CHICAGO

U.S. DEPARTMENT OF ENERGY
DE-SC0019288

75N93019C00041
Data-driven protein engineering startup in Chicago
ALF is a co-founder, consultant, equity holder, and IP licensor
No funding or role of Evozyne in the reported work
Limitations of molecular simulation

Two limitations in existing simulations are the approximations in the potential energy functions and the lengths of the simulations. The first introduces systematic errors and the second statistical errors.


1. Accurate force fields (systematic errors)

2. Sampling configurational space (statistical errors)
### Enhanced sampling

<table>
<thead>
<tr>
<th>Tempering</th>
<th>Collective variable bias</th>
<th>Path sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated annealing</td>
<td>Umbrella sampling</td>
<td>Transition path sampling</td>
</tr>
<tr>
<td>Multicanonical algorithm</td>
<td>Hyperdynamics</td>
<td>Transition interface sampling</td>
</tr>
<tr>
<td>Replica exchange</td>
<td>Metadynamics</td>
<td>Milestoning</td>
</tr>
<tr>
<td>Hamiltonian exchange</td>
<td>Adaptive force biasing</td>
<td>Forward flux sampling</td>
</tr>
<tr>
<td>Parallel tempering</td>
<td>Wang-Landau</td>
<td>Weighted ensemble</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Tempering** modifies $T$ or Hamiltonian to accelerate barrier crossing
  - substantial CPU time expended on conditions not of direct interest
- **CV biasing** efficiently directs sampling along relevant order parameters
  - presupposes *a priori* availability of “good” CVs
- **Path sampling** efficiently simulates reactive paths between states
  - requires *a priori* knowledge of metastable states and connectivity
CV discovery + enhanced sampling

- Challenging to intuit CVs for all but the most trivial systems
- Data-driven CV discovery and enhanced sampling is typically iterative

Good CVs required to drive sampling of configurational space (chicken)
Trajectories with good sampling needed to discover good CVs (egg)
CV discovery + enhanced sampling

Simulation trajectory

Highest variance

Linear

Nonlinear

Slowest

PCA
MDS
...

LLE
Isomap
dMaps
iMapD
...

tICA
DMD
MSM
Ulam's method
...

VAC
SRVs
deep tICA
...

H. Sidky, W. Chen, and ALF Molecular Physics 118 5 e1737742 (2020)
Inference of slow CVs

TICA
approximates eigenvalues and eigenfunctions

\[ \psi(x) = x \]

DMD
approximates eigenvalues and modes

\[ \psi(x) = x \]

VAC
approximates eigenvalues and eigenfunctions

\[ \psi(x) = \begin{bmatrix} \mathbb{1}_{A_1}(x) \\ \vdots \\ \mathbb{1}_{A_k}(x) \end{bmatrix} \]

EDMD
approximates eigenvalues, eigenfunctions, and modes

\[ \psi(x) = \begin{bmatrix} \mathbb{1}_{A_1}(x) \\ \vdots \\ \mathbb{1}_{A_k}(x) \end{bmatrix} \]

MSM
approximates eigenvalues and eigenfunctions

Ulam’s method
approximates eigenvalues and eigenfunctions

Slow CVs: Transfer operator theory

- Dynamical evolution of a system is governed by the transfer operator that propagates probability densities through time in discrete steps of size $\tau$

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

transfer or Perron-Frobenius operator on densities (self-adjoint to Koopman operator on observables)

- For equilibrium systems $\mathcal{T}$ is self-adjoint $\Rightarrow$ possesses an orthonormal basis of real eigenfunctions and eigenvalues

$$\mathcal{T} \circ \psi_i(x) = \lambda_i \psi_i(x)$$

$$\langle \psi_i | \psi_j \rangle_\pi = \delta_{ij}$$

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

inner product wrt equilibrium distribution

References:

F. Noé and F. Nüske, Multiscale Modeling & Simulation 11, 635 (2013)
G. Andrew, R. Arora, J. Bilmes, and K. Livescu Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP (vol. 28)
Transfer operator

• Any state function $\chi(x)$ can be projected into this orthonormal basis

$$\chi_t(x) = \sum_i \langle \psi_i | \chi_t \rangle \pi \psi_i(x)$$

- expansion coefficients
- basis functions

• The time evolution of $\chi(x)$ after $k$ applications of $\mathcal{T}$ becomes

$$\chi_{t+k\tau}(x) = \mathcal{T}^k \circ \chi_t(x) = \sum_i \langle \psi_i | \chi_t \rangle \pi \mathcal{T}^k \psi_i(x)$$

$$= \sum_i \langle \psi_i | \chi_t \rangle \pi \lambda_i^k \psi_i(x)$$

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

- implied timescale of $(\psi_i, \lambda_i)$

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

Leading eigenfunctions of $\mathcal{T}$ are slowest dynamical modes
Time-lagged ICA (tICA)

\[ \zeta_j(x) \]

basis functions

\[ C_{jk} = \langle \zeta_j(x) \mid \mathcal{T} \circ \zeta_k(x) \rangle_{\pi} \]

\[ Q_{jk} = \langle \zeta_j(x) \mid \zeta_k(x) \rangle_{\pi} \]

correlation matrices

\[ C s_i = \tilde{\lambda}_i Q s_i \]

generalized eigenvalue problem

\[ \tilde{\psi}_i = \sum_j s_{ij} \zeta_j \]

linear basis expansion of tICs

estimate from (unbiased) simulation trajectories

Learned approximations to eigenfunctions of \( \mathcal{T} \)

Isomorphic to Roothan-Hall equations in Hartree-Fock theory

http://msmbuilder.org/3.3.0/_images/tica_vs_pca.png

F. Noé and F. Nüske, Multiscale Modeling & Simulation 11, 635 (2013)
State-free reversible VAMPnets (SRVs)

\[ x(t) \xrightarrow{\text{featurizer}} \{ \zeta_i(x(t)) \} \]

learned optimal basis functions \( \{ \zeta_i(x) \} \)

\[ x(t + \tau) \xrightarrow{\text{featurizer}} \{ \zeta_i(x(t + \tau)) \} \]

learned approximations to eigenfunctions of \( \mathcal{T} \)

\[
C_{jk} = \langle \zeta_j(x) | \mathcal{T} \circ \zeta_k(x) \rangle_{\pi} \\
Q_{jk} = \langle \zeta_j(x) | \zeta_k(x) \rangle_{\pi} \\
C s_i = \tilde{\lambda}_i Q s_i \\
\tilde{\psi}_i = \sum_j s_{ij} \zeta_j \\
\mathcal{L} = - \sum_i \tilde{\lambda}_i^r
\]

molecular configs  molecular features  fully connected twin-lobe ANN [k-100-100-(d≈5-10)]
Iterative slow CV discovery and enhanced sampling?

- Slow CVs are inherently dynamical variables estimated as path observables.
- **Thermodynamic reweighting** of individual configurations is not enough, must also **dynamically reweight** the paths.

\[
\tilde{V}(x) = V(x) + U(\xi(x))
\]

\[
W(\omega_{t\rightarrow t+\tau}) = g(x_t) \times M(\omega_{t\rightarrow t+\tau})
\]

\[
\xi = \text{ML} \left( \omega_{t\rightarrow t+\tau}, W(\omega_{t\rightarrow t+\tau}) \right)
\]
Girsanov reweighting

- Elegant framework for dynamical reweighting of phase space paths under stochastic integrators (e.g., Langevin dynamics)

\[
V_{\text{sim}} = V_{\text{target}} - U_{\text{bias}}
\]

\[
P_{\text{sim}}(\eta) = \frac{1}{(2\pi)^{n/2}} \exp \left( -\frac{1}{2} \sum_{k=0}^{n-1} \eta_k^2 \right)
\]

\[
P_{\text{target}}(\tilde{\eta}) = \frac{1}{(2\pi)^{n/2}} \exp \left( -\frac{1}{2} \sum_{k=0}^{n-1} \tilde{\eta}_k^2 \right)
\]

\[
M(\omega) = \frac{P_{\text{target}}(\tilde{\eta})}{P_{\text{sim}}(\eta)}
\]

\[
= \exp \left( -\frac{1}{2} \sum_{k=0}^{n-1} \eta_k \cdot \Delta \eta_k \right)
\]

\[
\times \exp \left( -\frac{1}{2} \sum_{k=0}^{n-1} \eta_k^2 \right)
\]

Girsanov reweighting

- Exact expressions known for
  - Brownian dynamics (overdamped Langevin) by Euler-Maruyama
  - (underdamped) Langevin dynamics by Izaguirre, Sweet, and Pande (ISP)
- Applied by Weber & Pande and Keller & co. to estimate MSMs from biased simulations collected under pre-defined CV

**Brownian dynamics (E-M)**

\[
M_0(\omega_o, \eta_o; \Delta t|x_0) = \exp\left(-\sum_{k=0}^{n-1} \sqrt{\frac{\Delta t}{2k_B T \xi_m}} \nabla U(x_k) \cdot \eta_{o,k}\right) \cdot \exp\left(-\frac{1}{2} \sum_{k=0}^{n-1} \frac{\Delta t}{2k_B T \xi_m} (\nabla U(x_k))^2\right).
\]

**Langevin dynamics (ISP)**

\[
M_L(\omega_L, \eta_L; \Delta t|x_0, v_0) = \exp\left(-\frac{1 - \exp(-\xi \Delta t)}{\sqrt{1 - \exp(-2\xi \Delta t)}} \cdot \frac{\sum_{k=0}^{n-1} \nabla U(x_k) \eta_{L,k}}{\sqrt{k_B T \xi^2 m}}\right) \times \exp\left(-\frac{(1 - \exp(-\xi \Delta t))^2}{1 - \exp(-2\xi \Delta t)} \cdot \frac{\sum_{k=0}^{n-1} \nabla^2 U(x_k)}{2k_B T \xi^2 m}\right).
\]


Girsanov reweighting within SRVs

\[
C_{jk} = \langle \zeta_j(x_t) \mid \zeta_k(x_{t+\tau}) \rangle_{V_{\text{target}}}
\]
\[
\approx \frac{1}{L} \sum_{l=1}^{L} W(\omega_{t\rightarrow t+\tau}) \zeta_j(x_t^{(l)}) \zeta_k(x_{t+\tau}^{(l)})
\]

\[
Q_{jk} = \langle \zeta_j(x_t) \mid \zeta_k(x_t) \rangle_{V_{\text{target}}}
\]
\[
\approx \frac{1}{L} \sum_{l=1}^{L} W(\omega_{t\rightarrow t+\tau}) \zeta_j(x_t^{(l)}) \zeta_k(x_t^{(l)})
\]

\[
C_s_i = \tilde{\lambda}_i Q_s_i
\]

\[
\tilde{\psi}_i = \sum_j s_{ij} \zeta_j
\]

\[
g(x_t) = \exp \left( -\beta U_{\text{bias}}(x_t) \right)
\]

thermodynamic weighting of initial state

\[
M(\omega_{t\rightarrow t+\tau})
\]

Girsanov reweighting of phase space path

learned approximations to eigenfunctions of \( \mathcal{F} \)

learned optimal basis functions \{\zeta_i(x)\}
GREST: Girsanov Reweighting Enhanced Sampling Technique

(1) training data: biased simulation trajectory

(2) CV discovery: Girsanov reweighted SRV

(3) enhanced sampling: attenuated metadynamics

(4) convergence: CV stabilization
1D 4-well landscape
1D 4-well landscape

\[ V(x) \text{ [kJ/mol]} \]

\[ V_{\text{sim}} \]

\[ V_{\text{target}} \]

\[ \phi_1(x) \]

\[ \phi_2(x) \]

\[ \phi_3(x) \]
1D 4-well landscape

![Diagram showing 1D 4-well landscape with implied timescale and eigenfunction index.]

- **Implied timescale [s]**
  - Values range from 0 to 35.

- **Eigenfunction index**
  - Labeled as 1, 2, and 3.

- **Graphs**
  - **$\phi_1(x)$**
    - Values range from -0.6 to 0.6.
  - **$\phi_2(x)$**
    - Values range from -0.6 to 0.6.
  - **$\phi_3(x)$**
    - Values range from -0.6 to 0.6.

- **Legend**
  - **Unbiased reference**
  - **Thermodynamic correction**
  - **Thermo + dynamical correction**
  - **No correction**

The graphs illustrate the implied timescale across different eigenfunction indices, demonstrating the impact of various corrections on the behavior of the system.
Alanine dipeptide
Alanine dipeptide

![Diagram of Alanine dipeptide with FES, Φ_1, Φ_2, and WTMetaD plots showing free energy and SRV CV](image)
Alanine dipeptide

FES

Reference
250 ns

Round 0
5 ns

Round 1
25 ns

$\Phi_1$

$\Phi_2$

WTMetaD

Sample density, rad$^{-2}$

Free energy / kT

SRV CV
Alanine dipeptide
**Alanine dipeptide**

**Figure:**
- **Top:** Graph showing the similarity of consecutive SRV CVs with Pearson correlation coefficients.
  - Differentiation between the 1st and 2nd SRV CVs.
- **Bottom:** Heatmaps with color scales indicating free energy in kT and SRV CV values.
  - Round 4 with data points and color gradients.

**Note:**
- 25 ns
WLALL pentapeptide

Trp-Leu-Ala-Leu-Leu (WLALL)  
250 ns

(a)  
(b)  
(c)  

Misfolded (3)  

Unfolded (1)  
Folded (2)  

$\phi_1$  
$\phi_2$  
$t_1 = 19.6$ ns  
$t_2 = 11.3$ ns
WLALL pentapeptide
WLALL pentapeptide

Round 3
25 ns

(d-i) [Graph]

(d-ii) [Graph]

(d-iii) [Graph]

(d-iv) [Graph]

Round 4
25 ns

(e-i) [Graph]

(e-ii) [Graph]

(e-iii) [Graph]

(e-iv) [Graph]

Round 5
25 ns

(f-i) [Graph]

(f-ii) [Graph]

(f-iii) [Graph]
WLALL pentapeptide

Similiarity of consecutive SRV CVs

Pearson correlation coefficient, $\rho$

$1^{st}$ SRV CV
$2^{nd}$ SRV CV

Round 5

25 ns

(f-i)

(f-ii)

(f-iii)
WLALL pentapeptide

2/5 wells

3/5 wells

4/5 wells

5/5 wells
## Conclusions & Outlook

Girsanov reweighting enables rigorous iterative learning of and enhanced sampling in slow molecular CVs

\[ W(\omega_{t \rightarrow t+t}) = g(x_t) \times M(\omega_{t \rightarrow t+t}) \]

Slow CVs reveal **important dynamical motions** and optimal for enhanced sampling & long-time kinetic models

\[ \tilde{\lambda}_k = \frac{\langle u_k | \mathcal{T}(\tau) \circ u_k \rangle_\pi}{\langle u_k | u_k \rangle_\pi} \]

Incorporation of **solvent coordinates** respecting permutational invariance

Application to **larger protein systems**