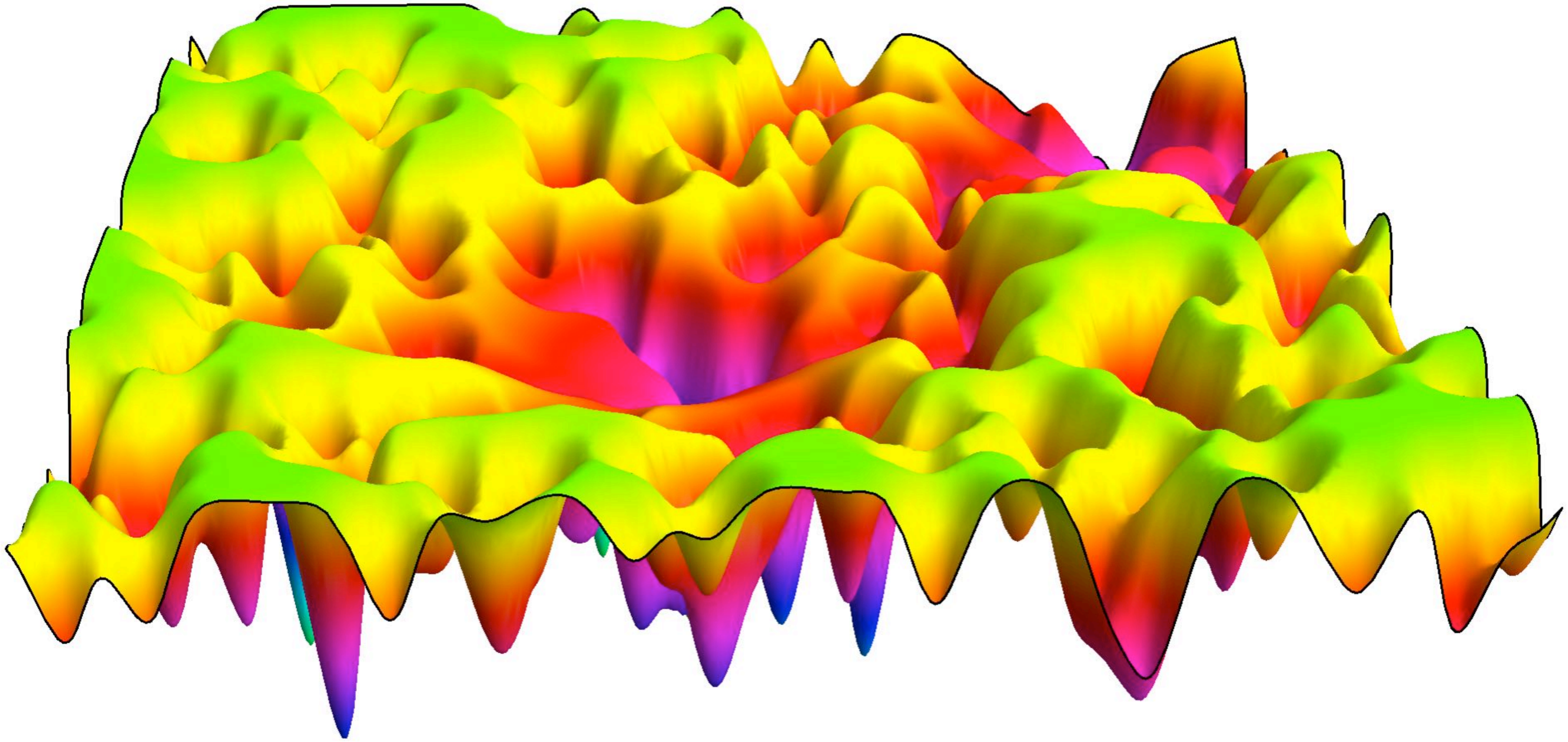
Data-driven discovery and enhanced sampling in slow collective variables



K. Shmilovich and A.L. Ferguson "Girsanov Reweighting Enhanced Sampling Technique (GREST): On-the-fly data-driven discovery of and enhanced sampling in slow collective variables" (submitted, 2023)

Brin Math Research Center: Rare Events, Analysis, Numerics, and Applications
27 February 2023
Andrew Ferguson, U. Chicago

THE UNIVERSITY OF CHICAGO

PRITZKER SCHOOL OF MOLECULAR ENGINEERING
THE UNIVERSITY OF CHICAGO

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)

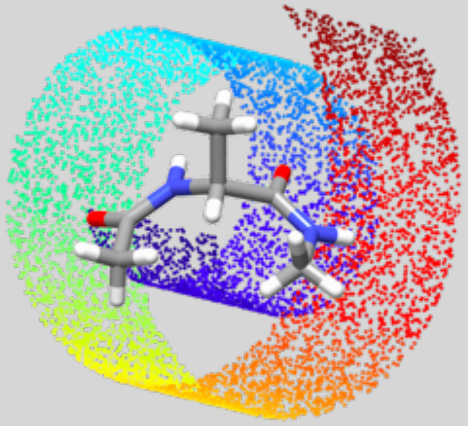
GS:
Kirill Shmilovich
Xinran Lian
Nick Rego
Shiqi Chen
Walter Alvarado
Max Topel
Mike Jones
Nick Herringer
Kate Johnson
Nikša Praljak
Jianming Mao
Alex Berlaga

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

NIH NIAID
75N93019C00041

The Camille & Henry Dreyfus Foundation, Inc.

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

U.S. DEPARTMENT OF ENERGY
DE-SC0019288

THE UNIVERSITY OF CHICAGO

PRITZKER SCHOOL OF MOLECULAR ENGINEERING
THE UNIVERSITY OF CHICAGO

Disclosure

Rama Ranganathan

Professor
Biochemistry and Molecular Biology
Pritzker School of Molecular Engineering
University of Chicago

evozyne

https://evozyne.com

Andrew Ferguson

Associate Professor
Pritzker School of Molecular Engineering
University of Chicago

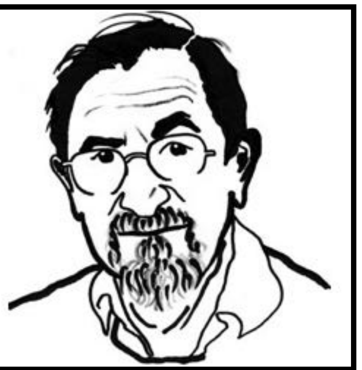
Data-driven protein engineering startup in Chicago

ALF is a co-founder, consultant, equity holder, and IP licensor

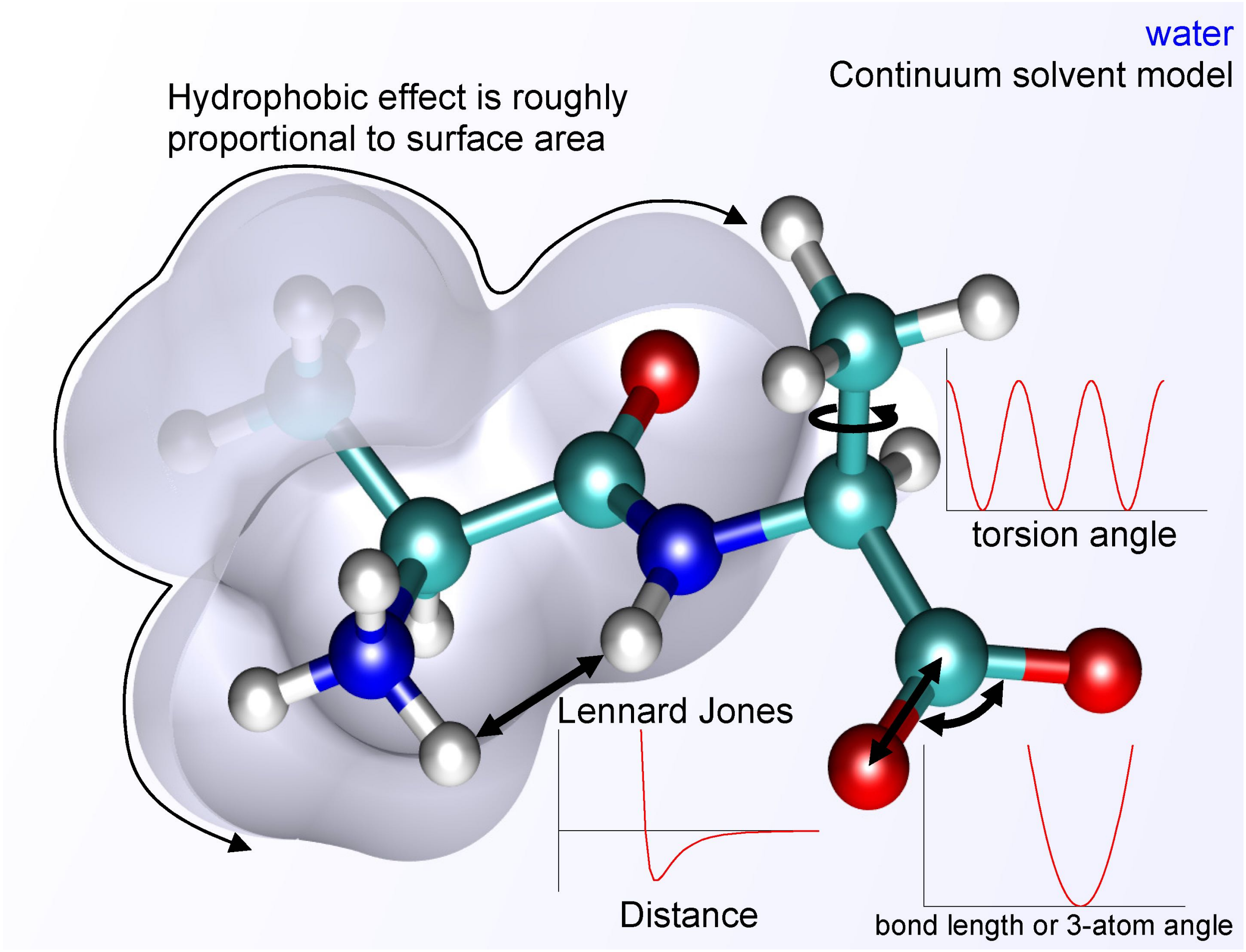
No funding or role of Evozyne in the reported work

Limitations of molecular simulation

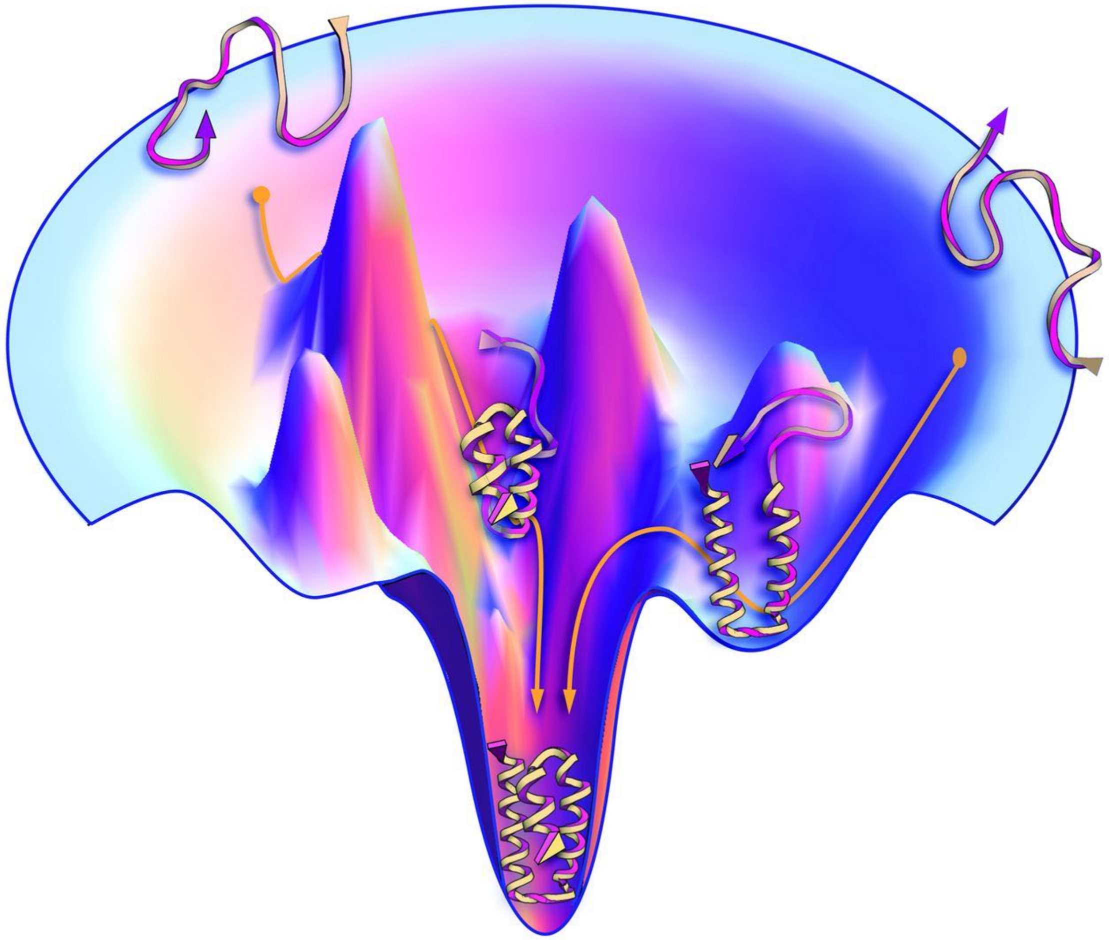
> Two limitations in existing simulations are the approximations in the potential energy functions and the lengths of the simulations. The first introduces systematic errors and the second statistical errors.
>
> — M. Karplus & G. Petsko *Nature* (1990)



1. Accurate force fields (systematic errors)



2. Sampling configurational space (statistical errors)



M. Karplus and G.A. Petsko Nature 347 631-639 (1990)

https://upload.wikimedia.org/wikipedia/commons/5/5c/MM_PEF.png http://jonlieffmd.com/wp-content/uploads/2012/12/F3.large_.jpg

Enhanced sampling

Tempering

Simulated annealing
Multicanonical algorithm
Replica exchange
Hamiltonian exchange
Parallel tempering
…

Collective variable bias

Umbrella sampling
Hyperdynamics
Metadynamics
Adaptive force biasing
Wang-Landau
…

Path sampling

Transition path sampling
Transition interface sampling
Milestoning
Forward flux sampling
Weighted ensemble
…

- **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

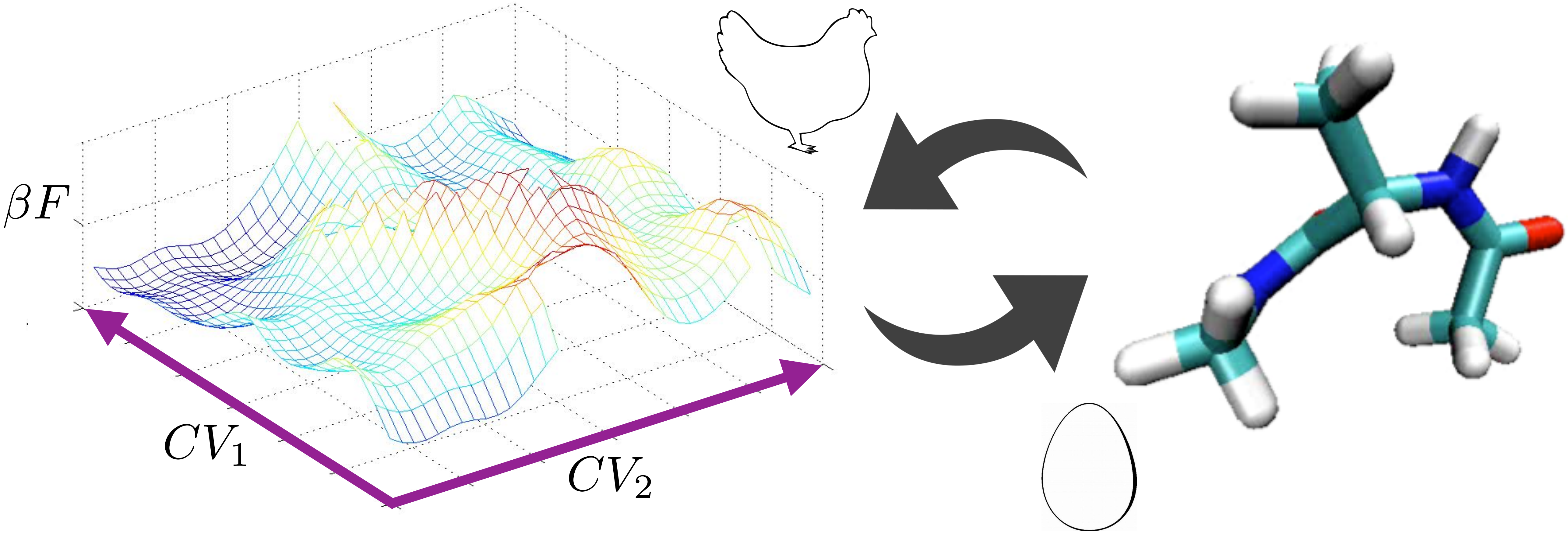
- **Path sampling** efficiently simulates reactive paths between states
  ➔ requires *a priori* knowledge of metastable states and connectivity

C. Abrams and G. Bussi Entropy 16 163–199 (2013)
P.G. Bolhuis and D.W.H. Swenson Adv. Theor. Sim. 4 2000237 (2021)
M.A. Rohrdanz, W. Zheng, and C. Clementi, C. Annu. Rev. Phys. Chem. 64 295–316 (2013)

CV discovery + enhanced sampling

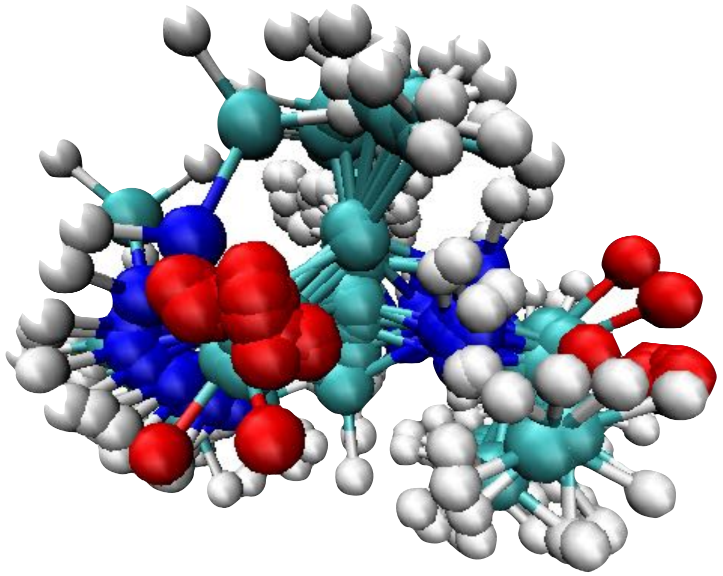
- Challenging to intuit CVs for all but the most trivial systems
- Data-driven CV discovery and enhanced sampling is typically **iterative**

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

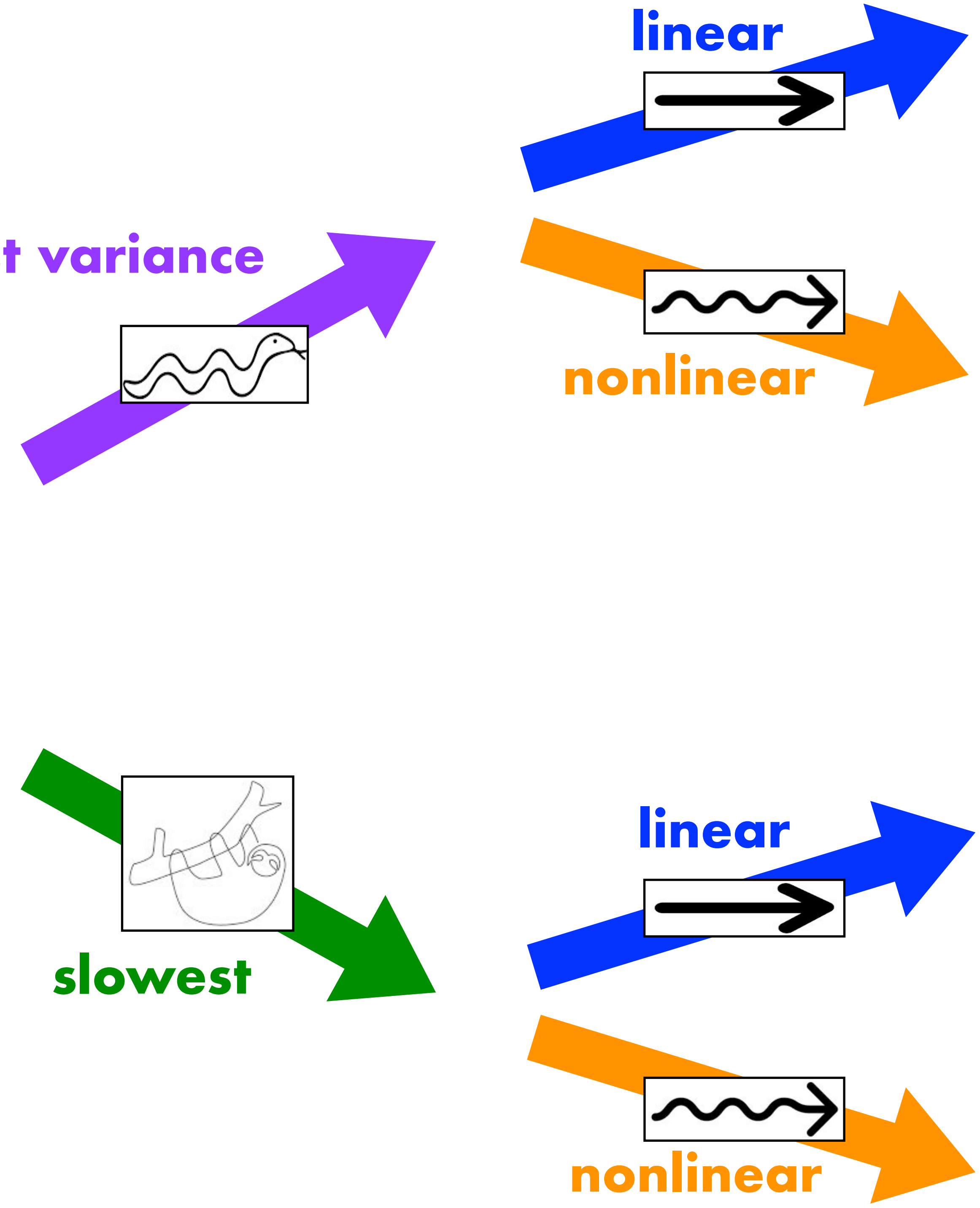


M.A. Rohrdanz, W. Zheng, and C. Clementi, C. Annu. Rev. Phys. Chem. 64 295–316 (2013)

CV discovery + enhanced sampling



simulation trajectory



PCA
MDS
…

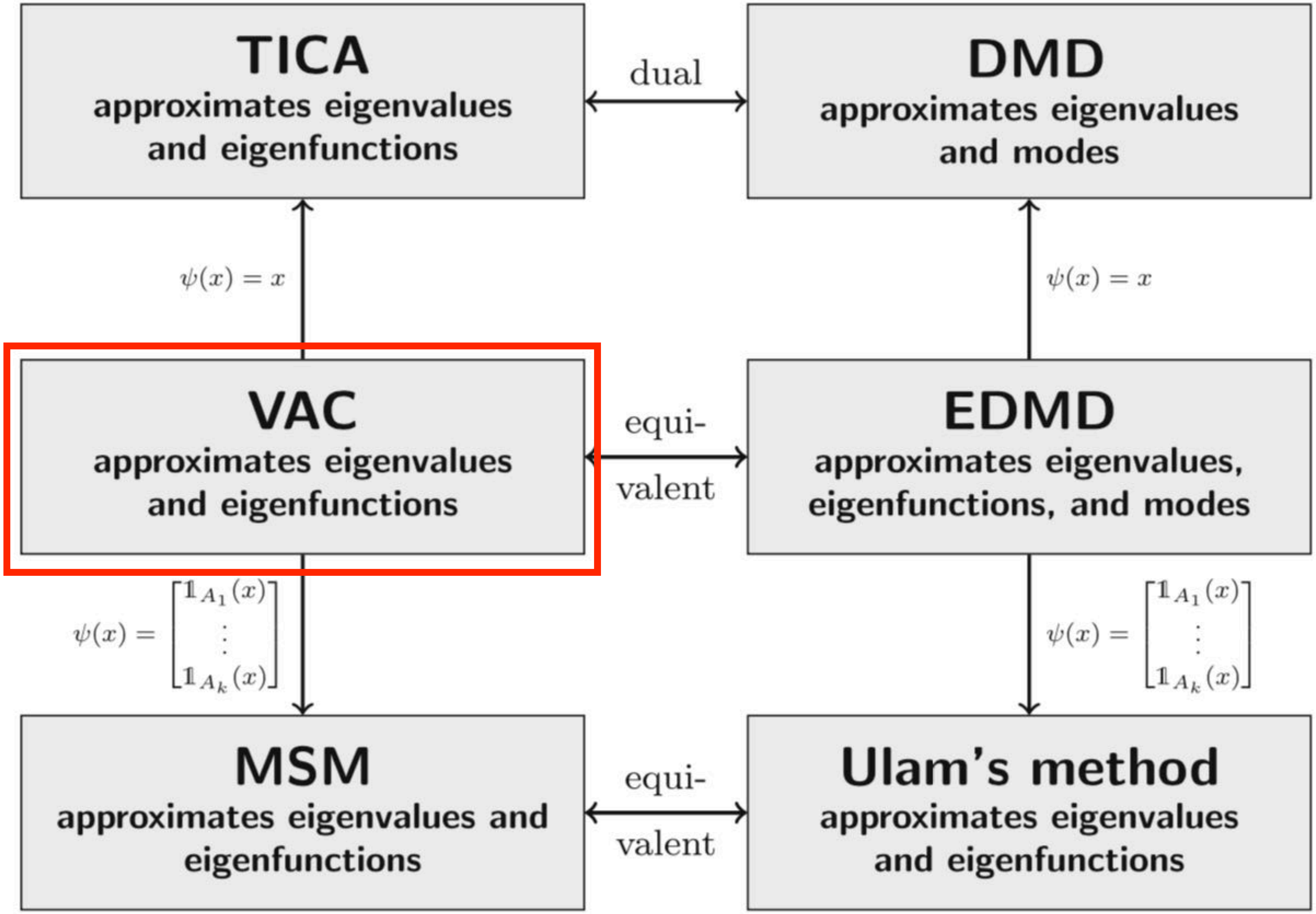
LLE
Isomap
dMaps
iMapD
…

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

dual

DMD
approximates eigenvalues and modes

$\psi(x) = x$

$\psi(x) = x$

VAC
approximates eigenvalues and eigenfunctions

equi-valent

EDMD
approximates eigenvalues, eigenfunctions, and modes

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

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

MSM
approximates eigenvalues and eigenfunctions

equi-valent

Ulam's method
approximates eigenvalues and eigenfunctions

S. Klus, F. Nüske, P. Koltai, H. Wu, I.G. Kevrekidis, C. Schütte, and F. Noé J Nonlinear Sci 28 985–1010 (2018)

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 τ

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

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

probability distribution over state space *x*
at time *t* scaled by equilibrium distribution

$$u_t(x) = p_t(x)/\pi(x)$$

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

$$\mathscr{T} \circ \psi_i(x) = \lambda_i \psi_i(x) \qquad \langle \psi_i | \psi_j \rangle_\pi = \delta_{ij} \qquad 1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \ldots$$

inner product wrt equilibrium distribution

F. Noé and F. Nüske, Multiscale Modeling & Simulation 11, 635 (2013)
F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, Journal of Chemical Theory and Computation 10, 1739 (2014)
F. Noé and C. Clementi, Current Opinion in Structural Biology 43, 141 (2017)
S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, Journal of Nonlinear Science 28, 985 (2018)
C. Schütte, W. Huisinga, and P. Deuflhard, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems (Springer, 2001) pp. 191–223.
J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, The Journal of Chemical Physics 134, 174105 (2011)
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 χ(x) can be projected into this orthonormal basis

$$\chi_t(x) = \sum_i \underset{\uparrow}{\langle \psi_i | \chi_t \rangle_\pi} \underset{\uparrow}{\psi_i(x)}$$

expansion coefficients

basis functions

- The time evolution of χ(x) after *k* applications of $\mathscr{T}$ becomes

$$\begin{aligned}\chi_{t+k\tau}(x) = \mathscr{T}^k \circ \chi_t(x) &= \sum_i \langle \psi_i | \chi_t \rangle_\pi \mathscr{T}^k \psi_i(x) \\ &= \sum_i \langle \psi_i | \chi_t \rangle_\pi \lambda_i^k \psi_i(x) \\ &= \sum_i \langle \psi_i | \chi_t \rangle_\pi \exp\left(-\frac{k\tau}{t_i}\right) \psi_i(x)\end{aligned}$$

implied timescale of (ψ_i, λ_i)

$$\downarrow$$

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

Leading eigenfunctions of $\mathscr{T}$ are slowest dynamical modes

Time-lagged ICA (tICA)

estimate from (unbiased) simulation trajectories

$\zeta_j(x)$ **basis functions**

$$C_{jk} = \langle \zeta_j(x) | \mathscr{T} \circ \zeta_k(x) \rangle_\pi$$
$$Q_{jk} = \langle \zeta_j(x) | \zeta_k(x) \rangle_\pi$$

correlation matrices

$$C s_i = \tilde{\lambda}_i Q s_i$$

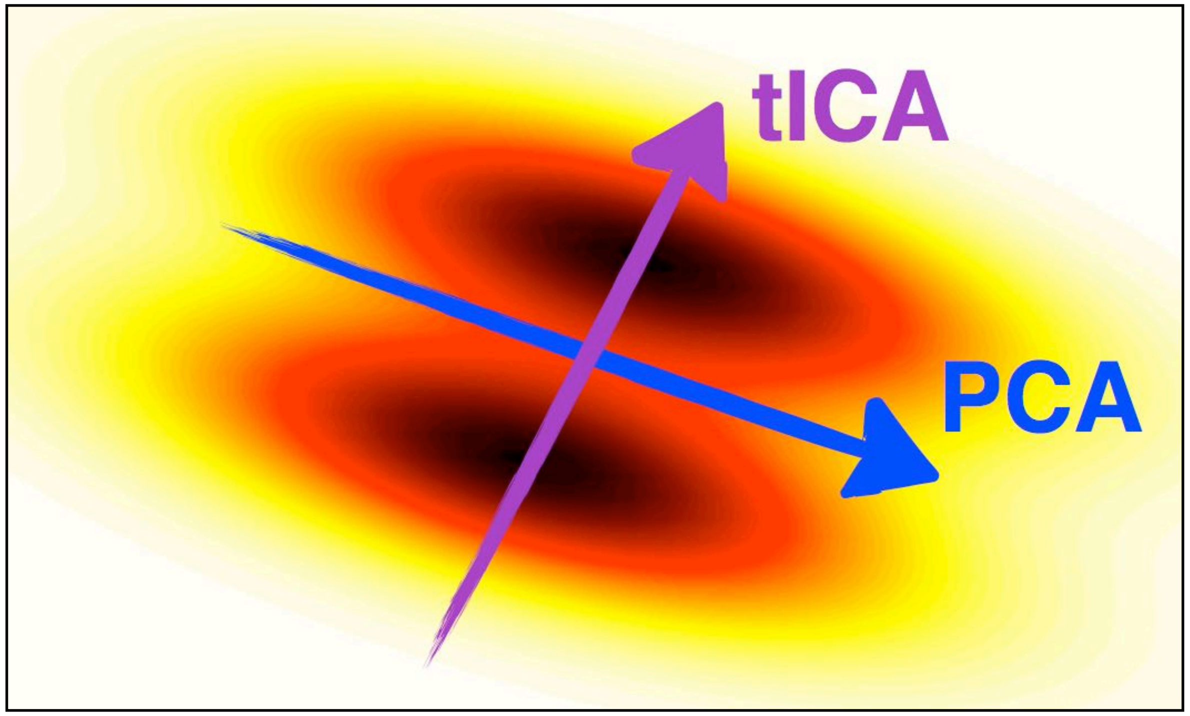
generalized eigenvalue problem

$$\tilde{\psi}_i = \sum_j s_{ij} \zeta_j$$

linear basis expansion of tICs

learned approximations
to eigenfunctions of $\mathscr{T}$

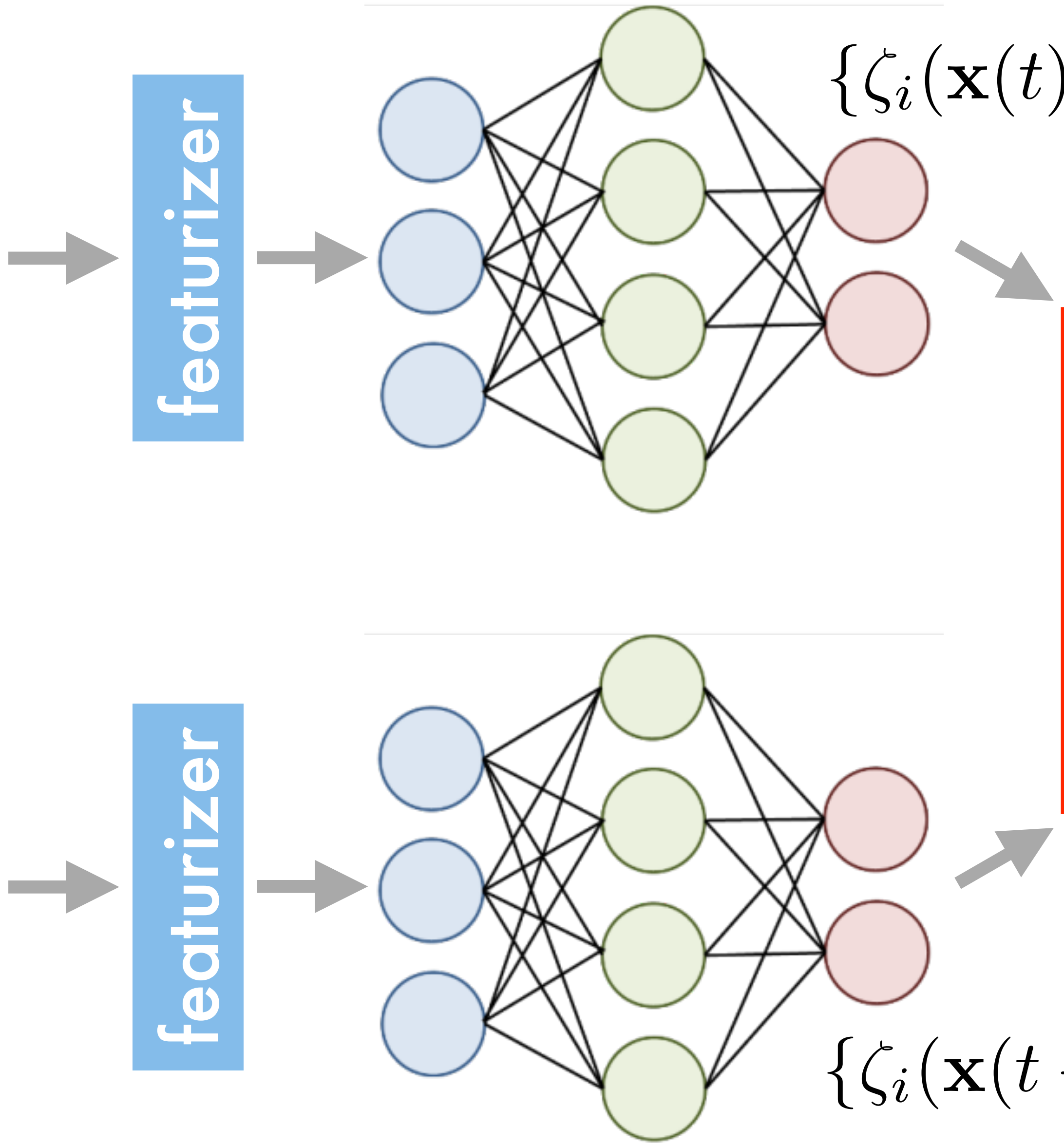
isomorphic to Roothan-Hall
equations in Hartree-Fock theory



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

L. Molgedey and H.G. Schuster Phys. Rev. Lett. 72 3634–3637 (1994)
F. Noé and F. Nüske, Multiscale Modeling & Simulation 11, 635 (2013)

State-free reversible VAMPnets (SRVs)



$\mathbf{x}(t)$

$\{\zeta_i(\mathbf{x}(t))\}$

learned optimal basis functions **{ζᵢ(x)}**

$\mathbf{x}(t+\tau)$

$\{\zeta_i(\mathbf{x}(t+\tau))\}$

$$C_{jk} = \langle \zeta_j(x) | \mathscr{T} \circ \zeta_k(x) \rangle_\pi$$
$$Q_{jk} = \langle \zeta_j(x) | \zeta_k(x) \rangle_\pi$$
$$C s_i = \tilde{\lambda}_i Q s_i$$
$$\tilde{\psi}_i = \sum_j s_{ij} \zeta_j$$

tICA

$$\mathcal{L} = -\sum_i \tilde{\lambda}_i^r$$

learned approximations to eigenfunctions of $\mathscr{T}$

molecular configs

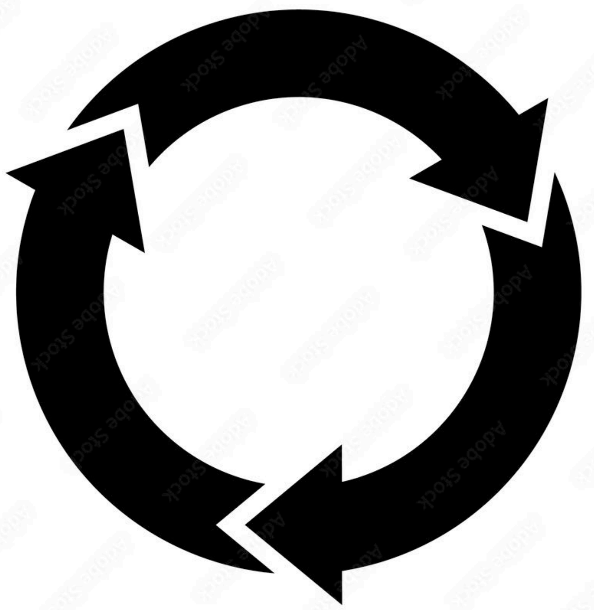
molecular features

fully connected twin-lobe ANN [k-100-100-(d≈5-10)]

W. Chen, H. Sidky, and ALF J. Chem. Phys. 150 214114 (2019)

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



enhanced sampling

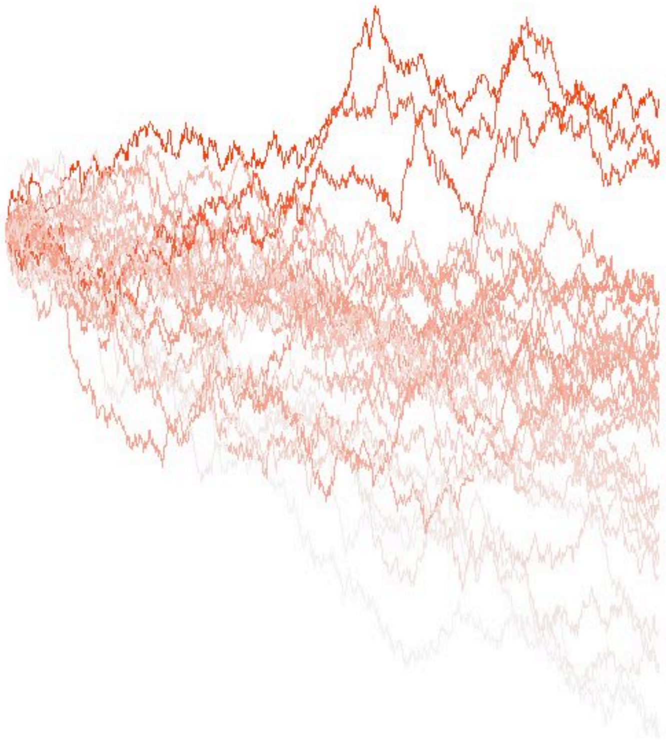
$\tilde{V}(x) = V(x) + U(\xi(x))$

thermodynamic and dynamical reweighting

$W(\boldsymbol{\omega}_{t\rightarrow t+\tau}) = g(\boldsymbol{x}_t) \times M(\boldsymbol{\omega}_{t\rightarrow t+\tau})$

slow CV learning

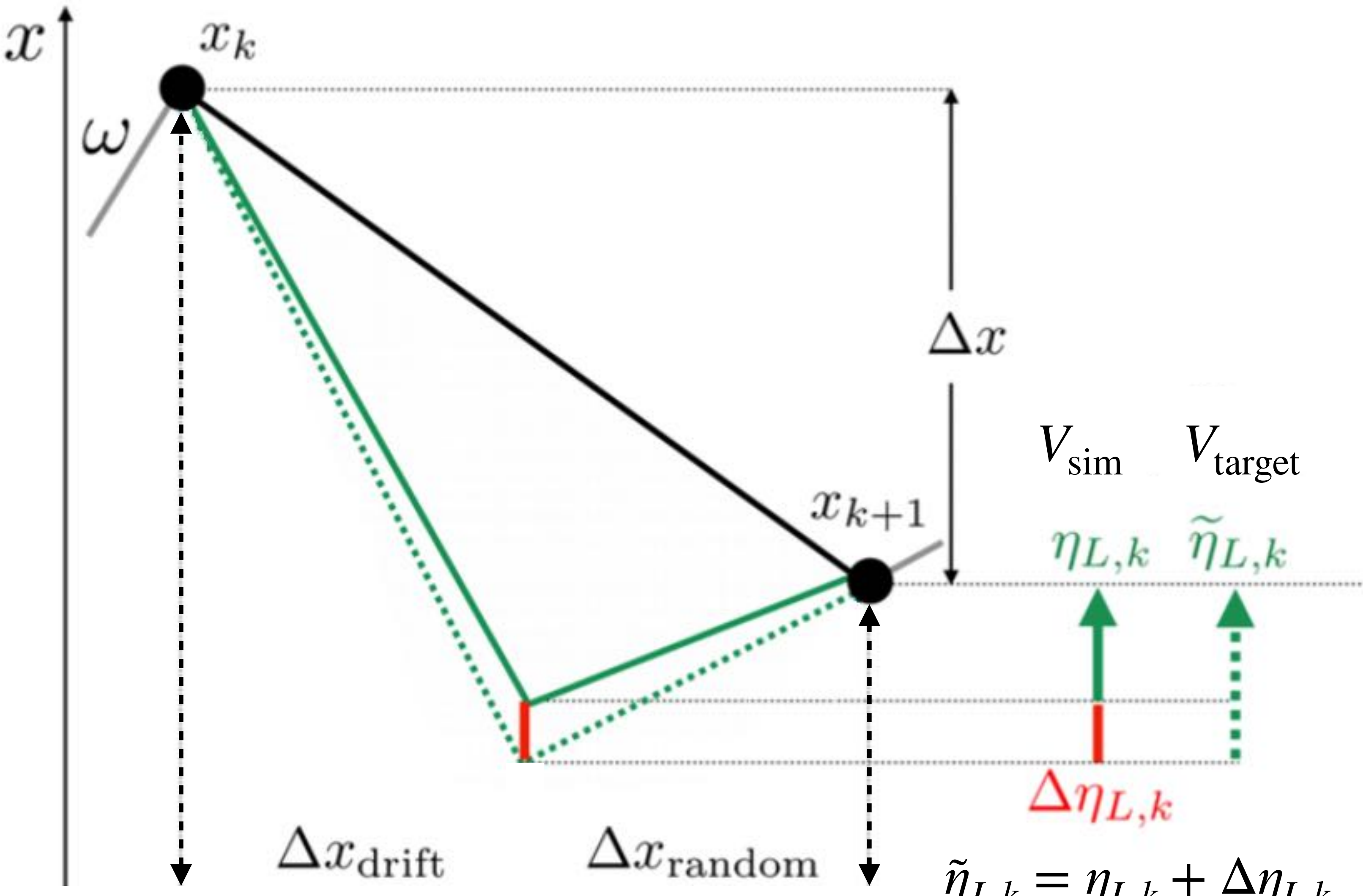
$\xi = \mathrm{ML}\left(\boldsymbol{\omega}_{t\rightarrow t+\tau}, W(\boldsymbol{\omega}_{t\rightarrow t+\tau})\right)$



S. Kieninger and B.G. Keller J. Chem. Phys. 154 094102 (2021)
Martin Keller-Ressel (uploaded by Thomas Steiner) - created with GNU R, see source below, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=928046

Girsanov reweighting

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



$$\tilde{\eta}_{L,k} = \eta_{L,k} + \Delta\eta_{L,k}$$

S. Kieninger and B.G. Keller J. Chem. Phys. 154 094102 (2021)

recorded during simulation

calculable from integration algorithm

$$V_{\text{sim}} = V_{\text{target}} - U_{\text{bias}}$$

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

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

$$M(\boldsymbol{\omega}) = \frac{P_{\text{target}}(\tilde{\eta})}{P_{\text{sim}}(\eta)}$$

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

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

J.K. Weber and V.S. Pande J. Chem. Theory Comput. 11 2412–2420 (2015)
L. Donati, C. Hartmann, and B.G. Keller J. Chem. Phys. 146 244112 (2017)
L. Donati and B.G. Keller J. Chem. Phys. 149 072335 (2018)
S. Kieninger and B.G. Keller J. Chem. Phys. 154 094102 (2021)

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)

$$M_o(\omega_o, \eta_o; \Delta t | x_0) = \exp\left(-\sum_{k=0}^{n-1}\sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k)\cdot\eta_{o,k}\right) \cdot \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\frac{\Delta t}{2k_BT\xi m}(\nabla U(x_k))^2\right).$$

Langevin dynamics (ISP)

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

S. Kieninger and B.G. Keller J. Chem. Phys. 154 094102 (2021)

J.K. Weber and V.S. Pande J. Chem. Theory Comput. 11 2412–2420 (2015)
L. Donati, C. Hartmann, and B.G. Keller J. Chem. Phys. 146 244112 (2017)
J.A. Izaguirre, C.R. Sweet, and V. Pande Pac. Symp. Biocomput. 15 240 (2010)
L. Donati and B.G. Keller J. Chem. Phys. 149 072335 (2018)
S. Kieninger and B.G. Keller J. Chem. Phys. 154 094102 (2021)

Girsanov reweighting within SRVs

$$C_{jk} = \langle \zeta_j(\boldsymbol{x}_t) \,|\, \zeta_k(\boldsymbol{x}_{t+\tau}) \rangle_{V_{\text{target}}}$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} W(\boldsymbol{\omega}_{t\rightarrow t+\tau}) \zeta_j(\boldsymbol{x}_t^{(l)}) \zeta_k(\boldsymbol{x}_{t+\tau}^{(l)})$$

$$Q_{jk} = \langle \zeta_j(\boldsymbol{x}_t) \,|\, \zeta_k(\boldsymbol{x}_t) \rangle_{V_{\text{target}}}$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} W(\boldsymbol{\omega}_{t\rightarrow t+\tau}) \zeta_j(\boldsymbol{x}_t^{(l)}) \zeta_k(\boldsymbol{x}_t^{(l)})$$

$$W(\boldsymbol{\omega}_{t\rightarrow t+\tau}) = g(\boldsymbol{x}_t) \times M(\boldsymbol{\omega}_{t\rightarrow t+\tau})$$

$$g(\boldsymbol{x}_t) = \exp\left(-\beta U_{\text{bias}}(\boldsymbol{x}_t)\right)$$

thermodynamic weighting of initial state

$$M(\boldsymbol{\omega}_{t\rightarrow t+\tau})$$

Girsanov reweighting of phase space path

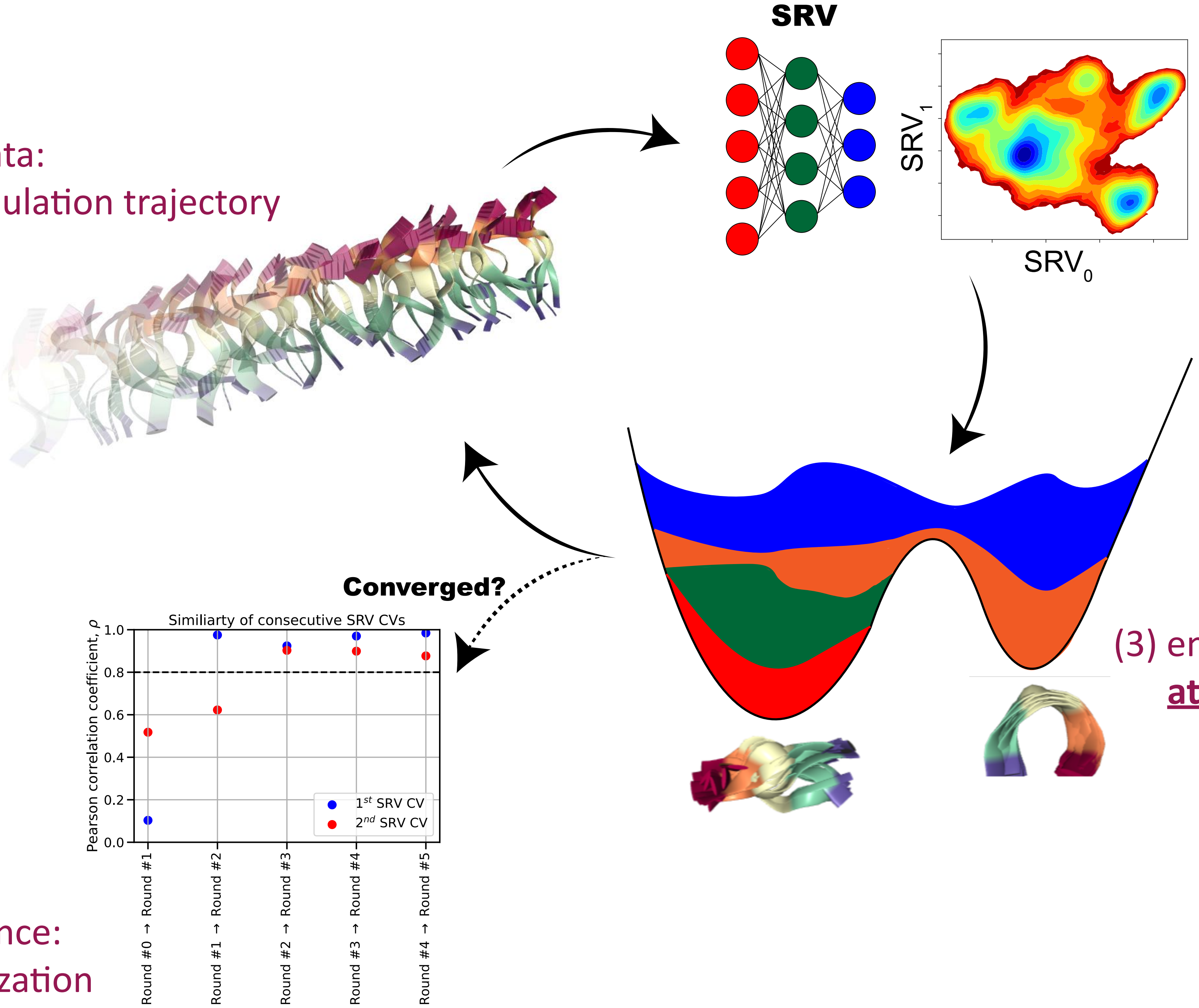
$$Cs_i = \tilde{\lambda}_i Q s_i$$

learned optimal basis functions **{ζᵢ(x)}**

$$\tilde{\psi}_i = \sum_j s_{ij} \zeta_j$$

learned approximations to eigenfunctions of $\mathscr{T}$

GREST: Girsanov Reweighting Enhanced Sampling Technique



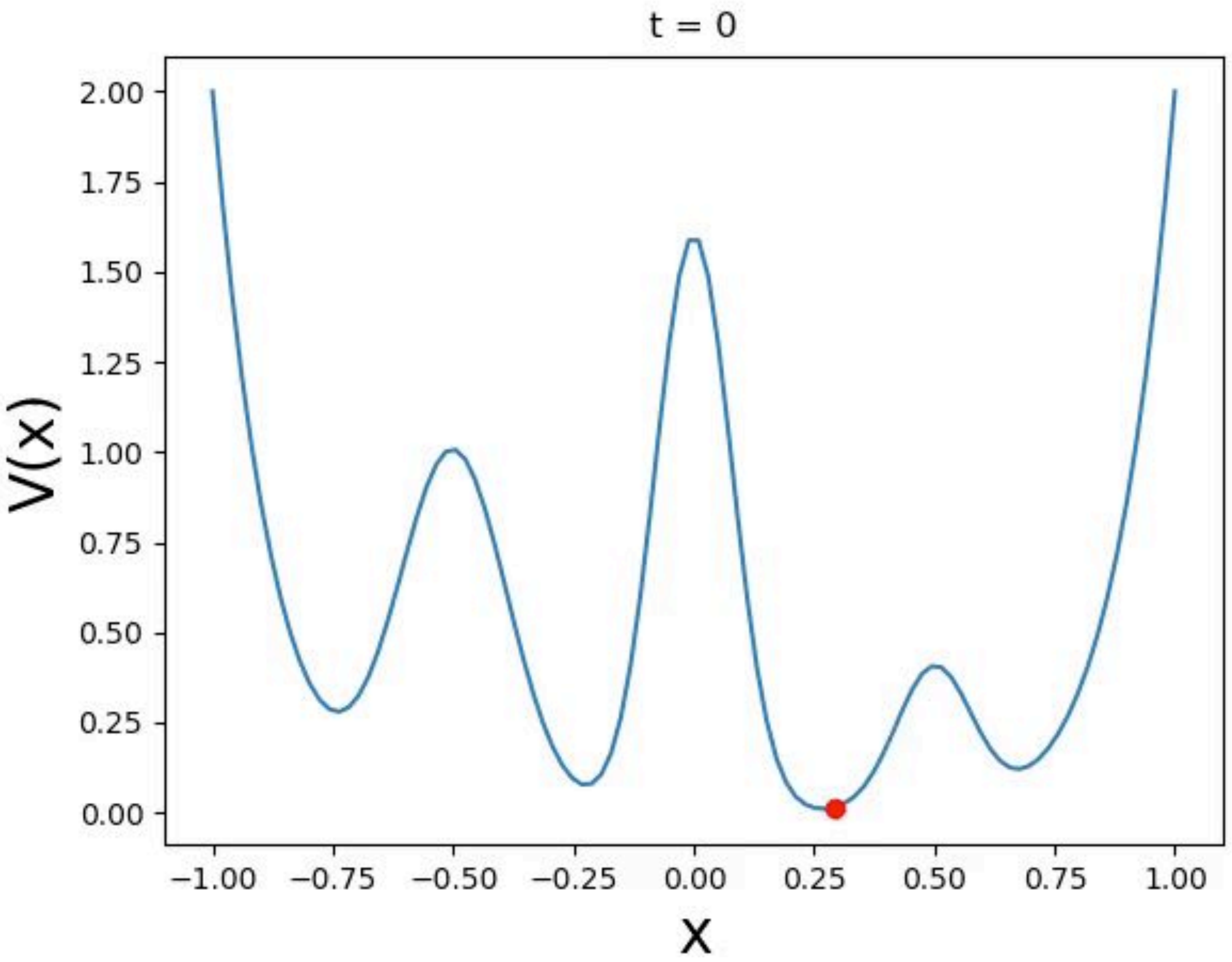
(1) training data:
biased simulation trajectory

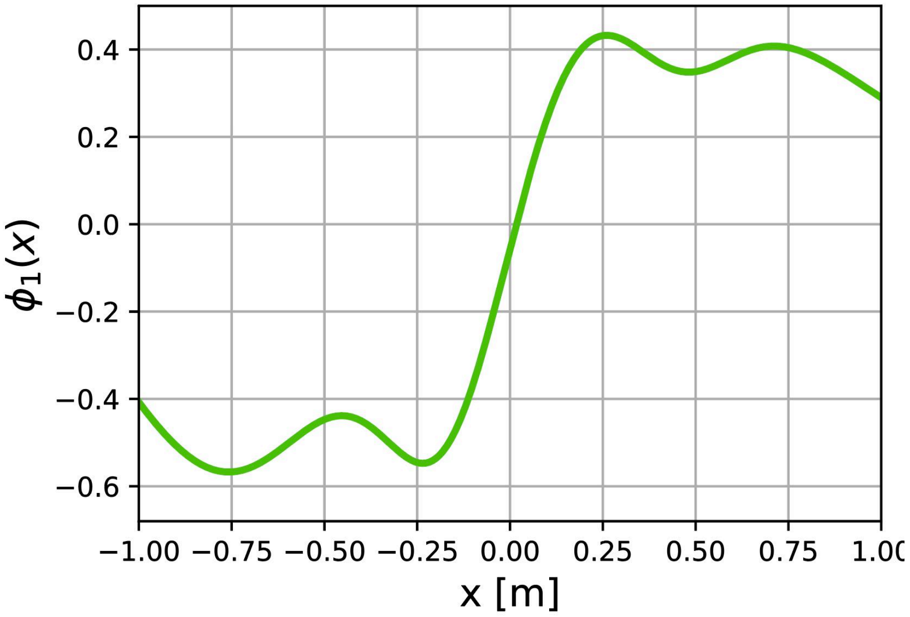
(2) CV discovery:
Girsanov reweighted SRV

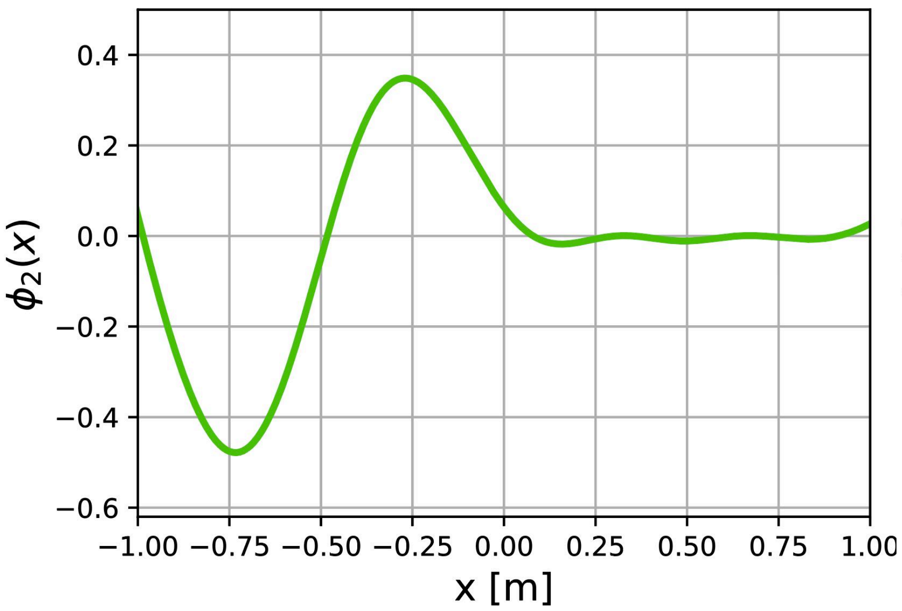
(3) enhanced sampling:
attenuated metadynamics

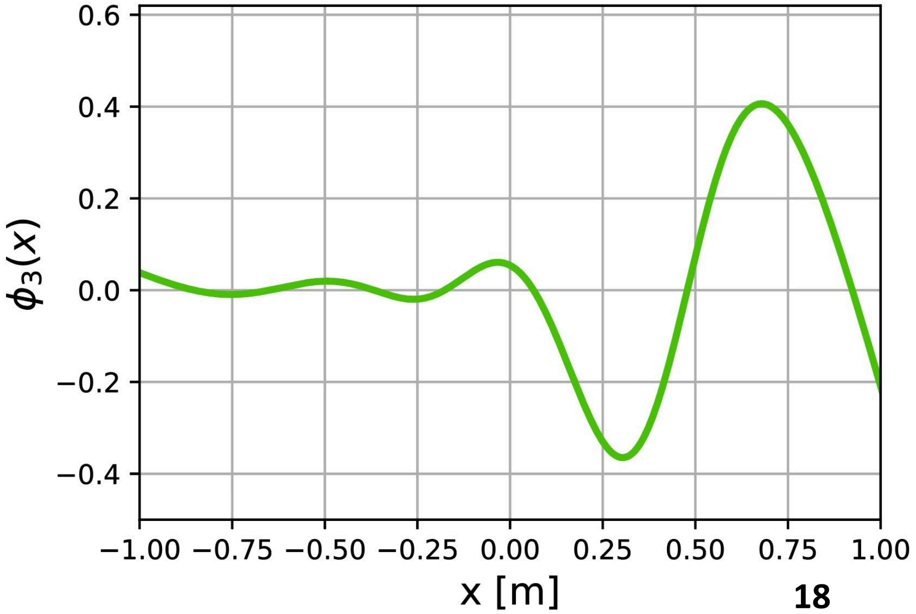
(4) convergence:
CV stabilization

1D 4-well landscape

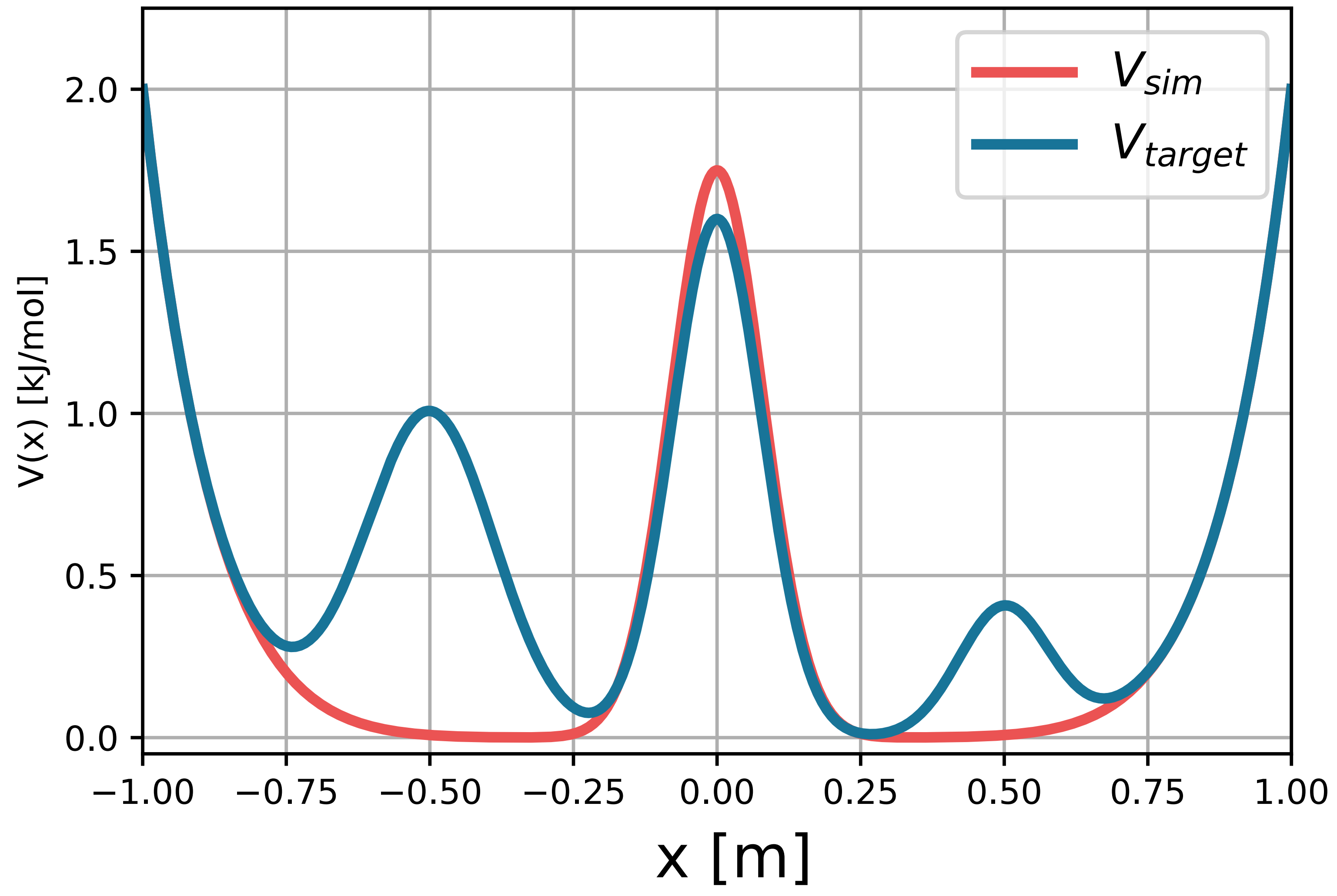


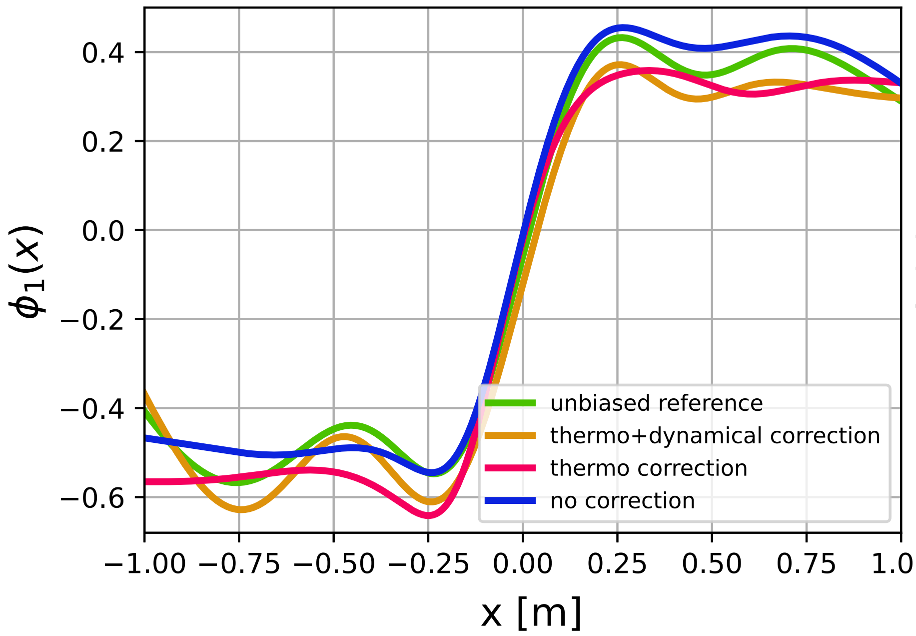


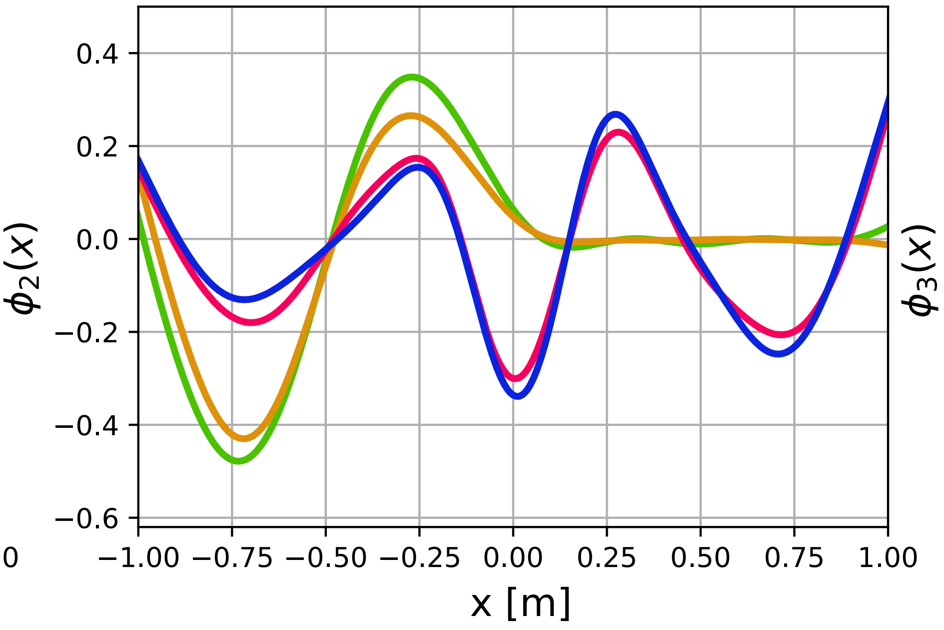


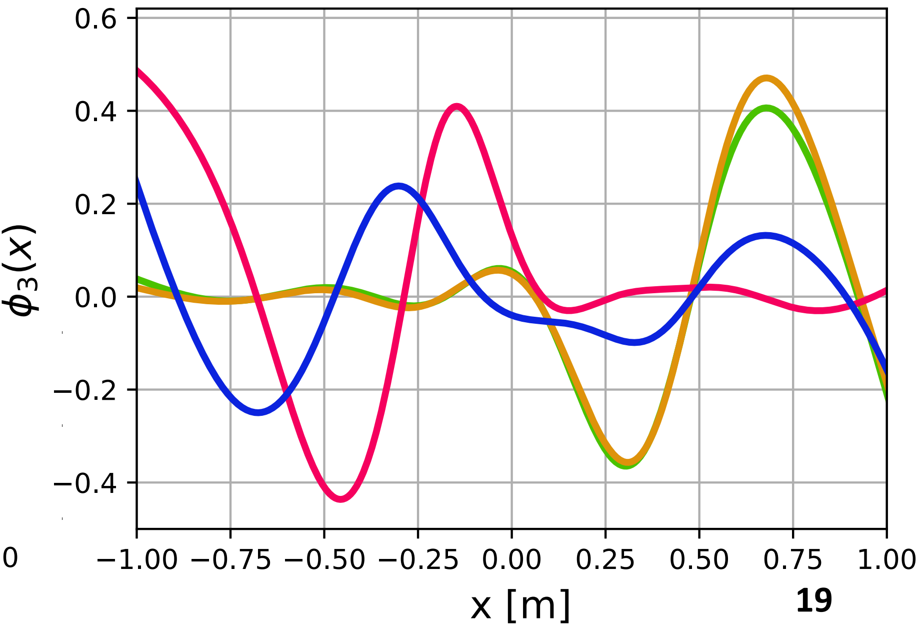


1D 4-well landscape

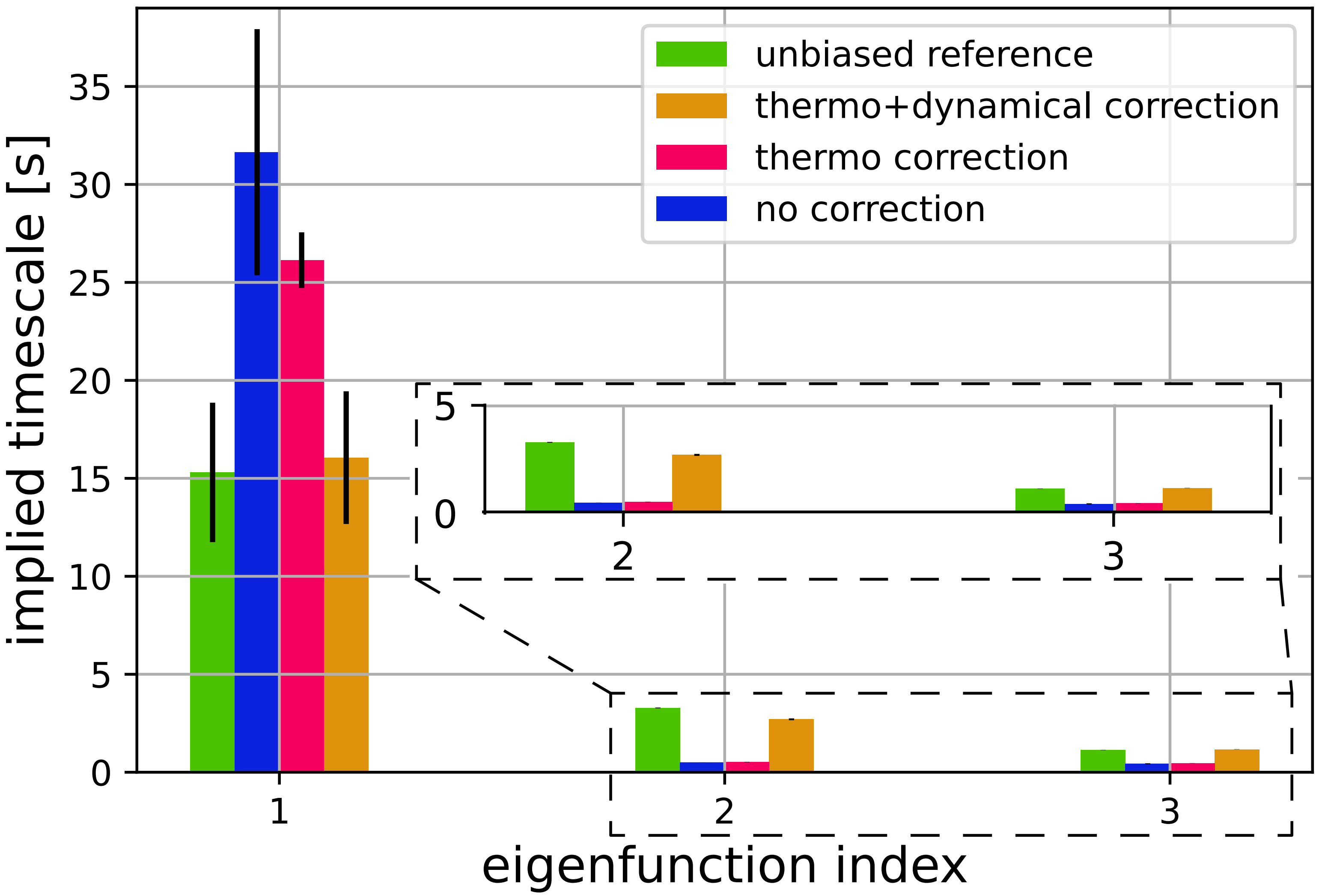


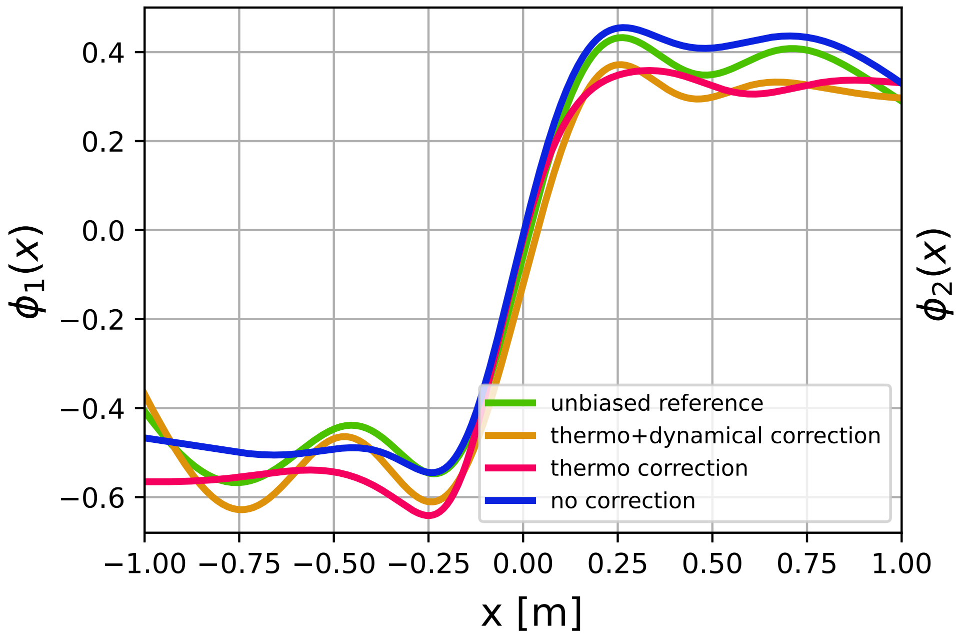


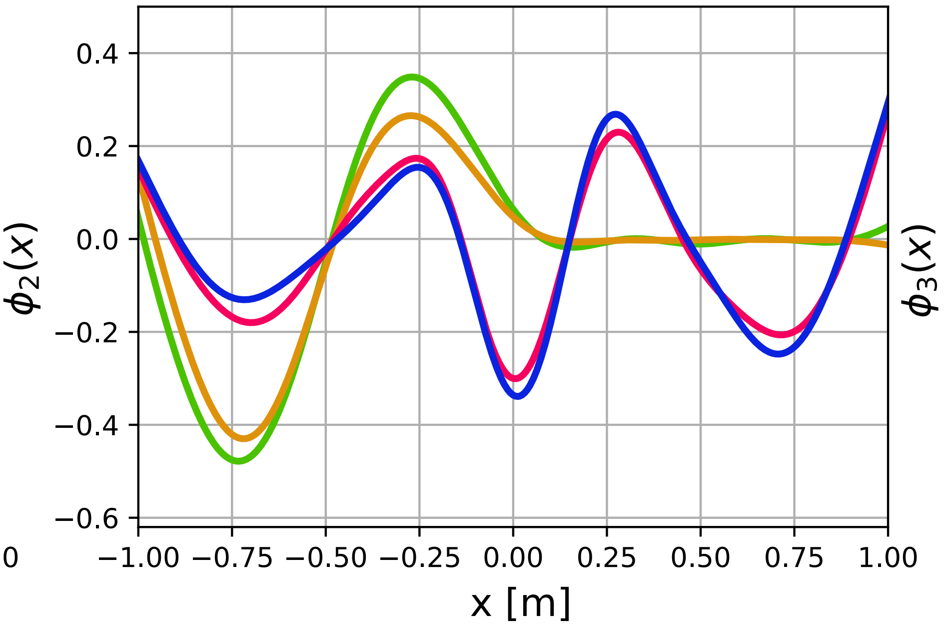


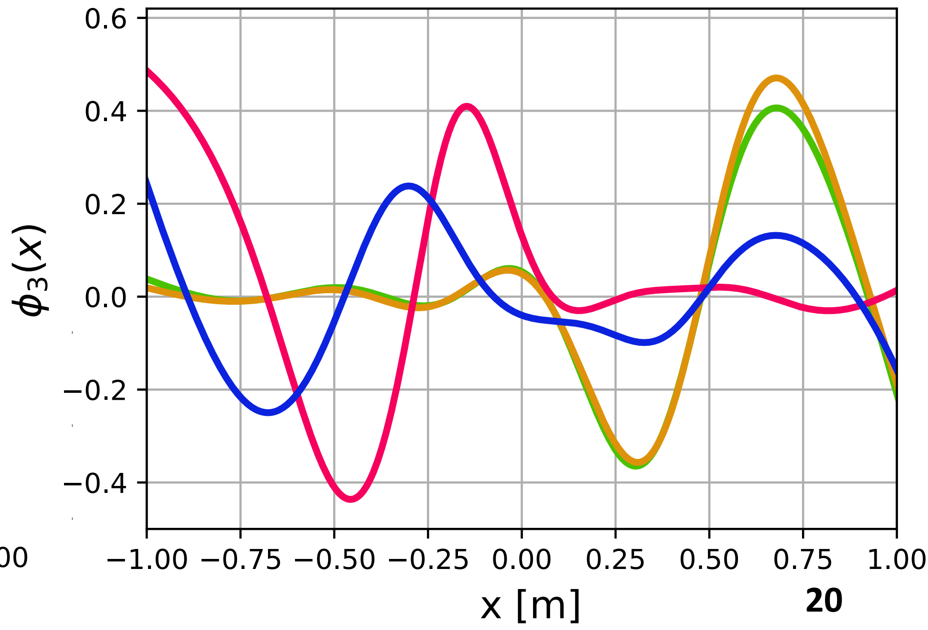


1D 4-well landscape

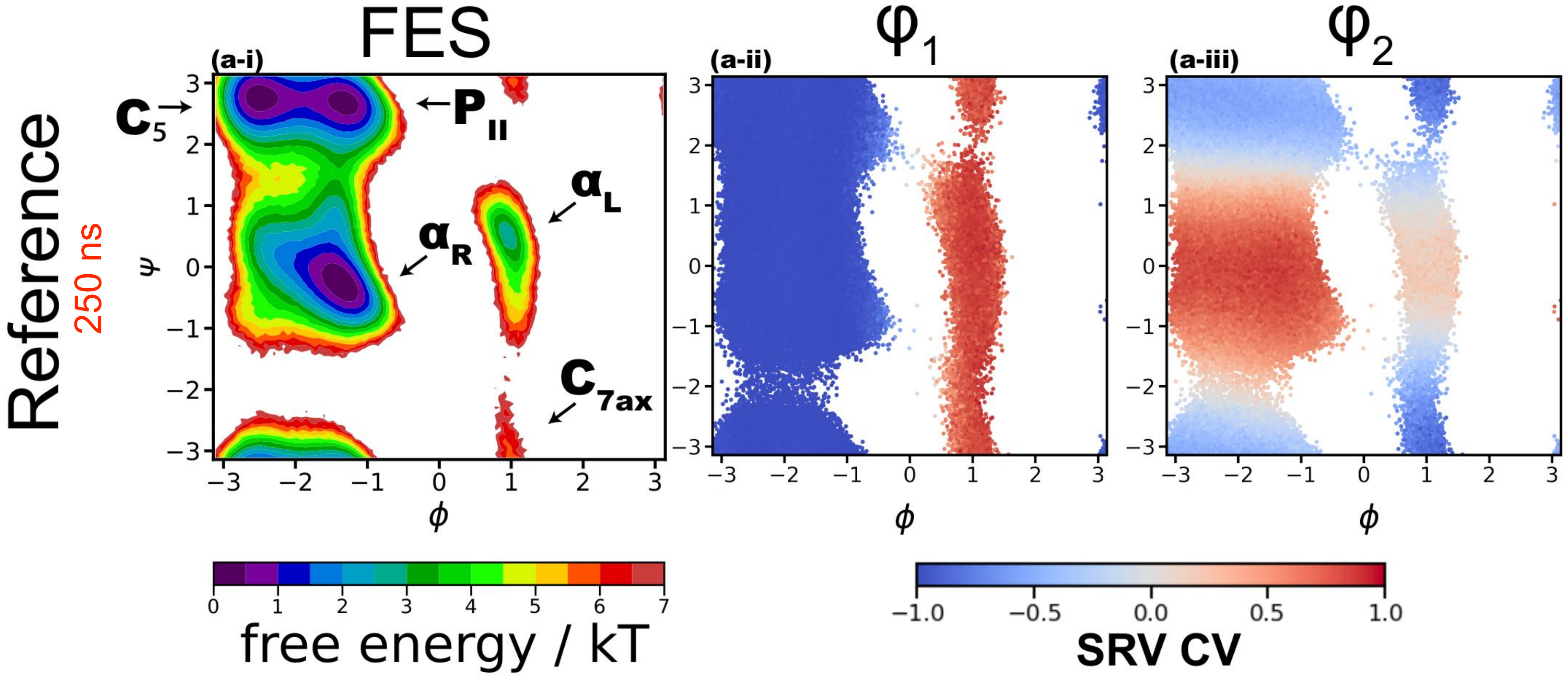


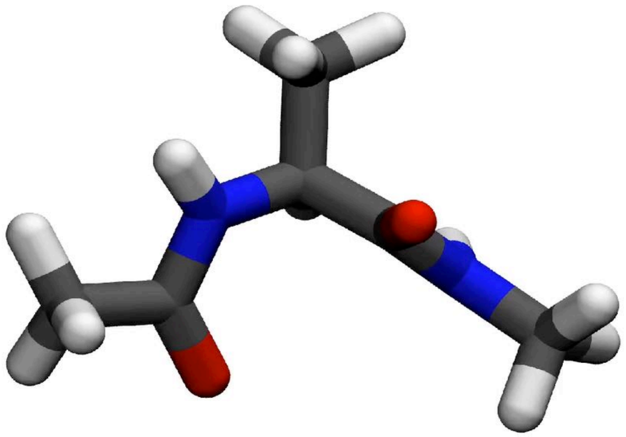




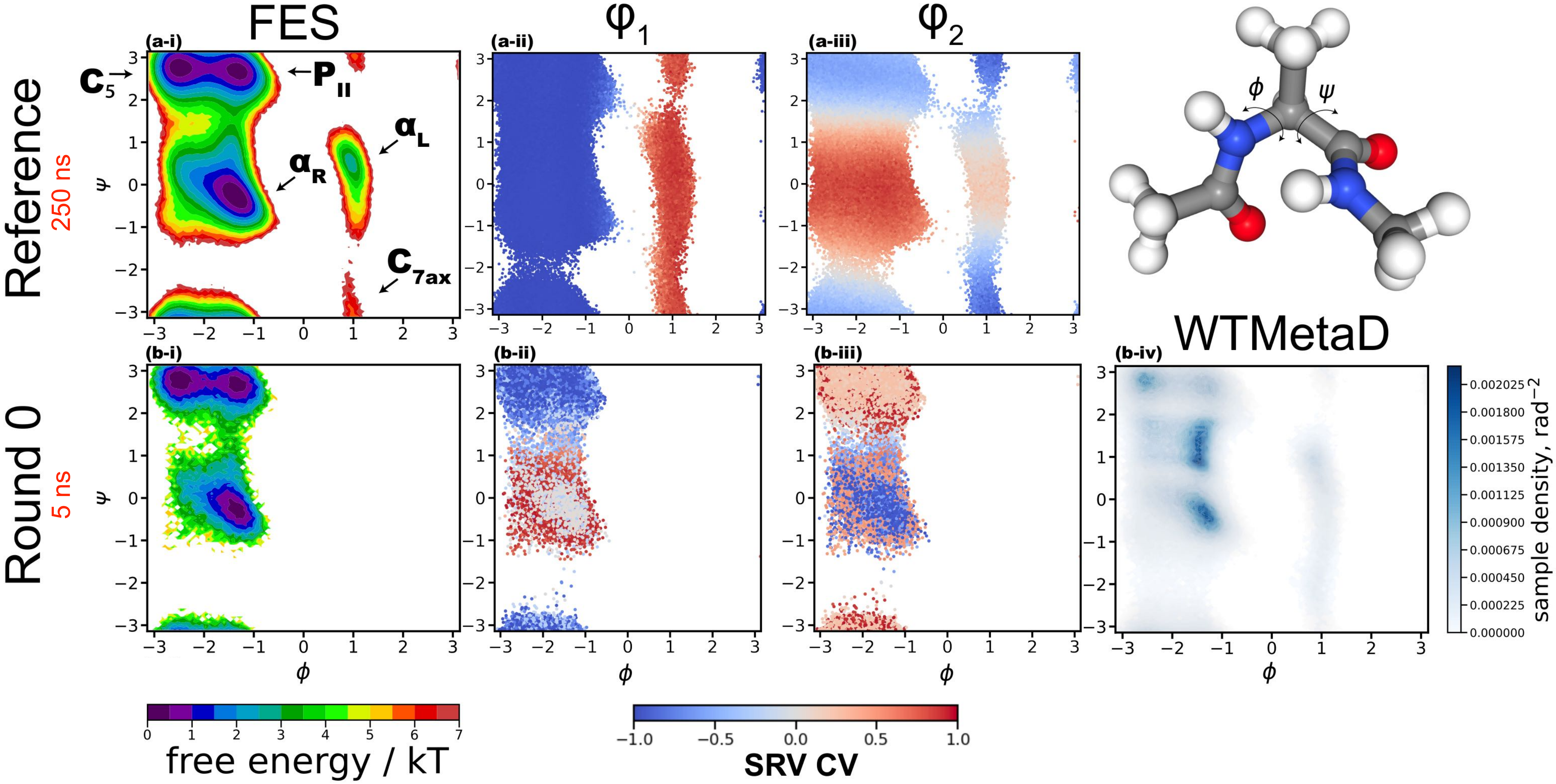


Alanine dipeptide

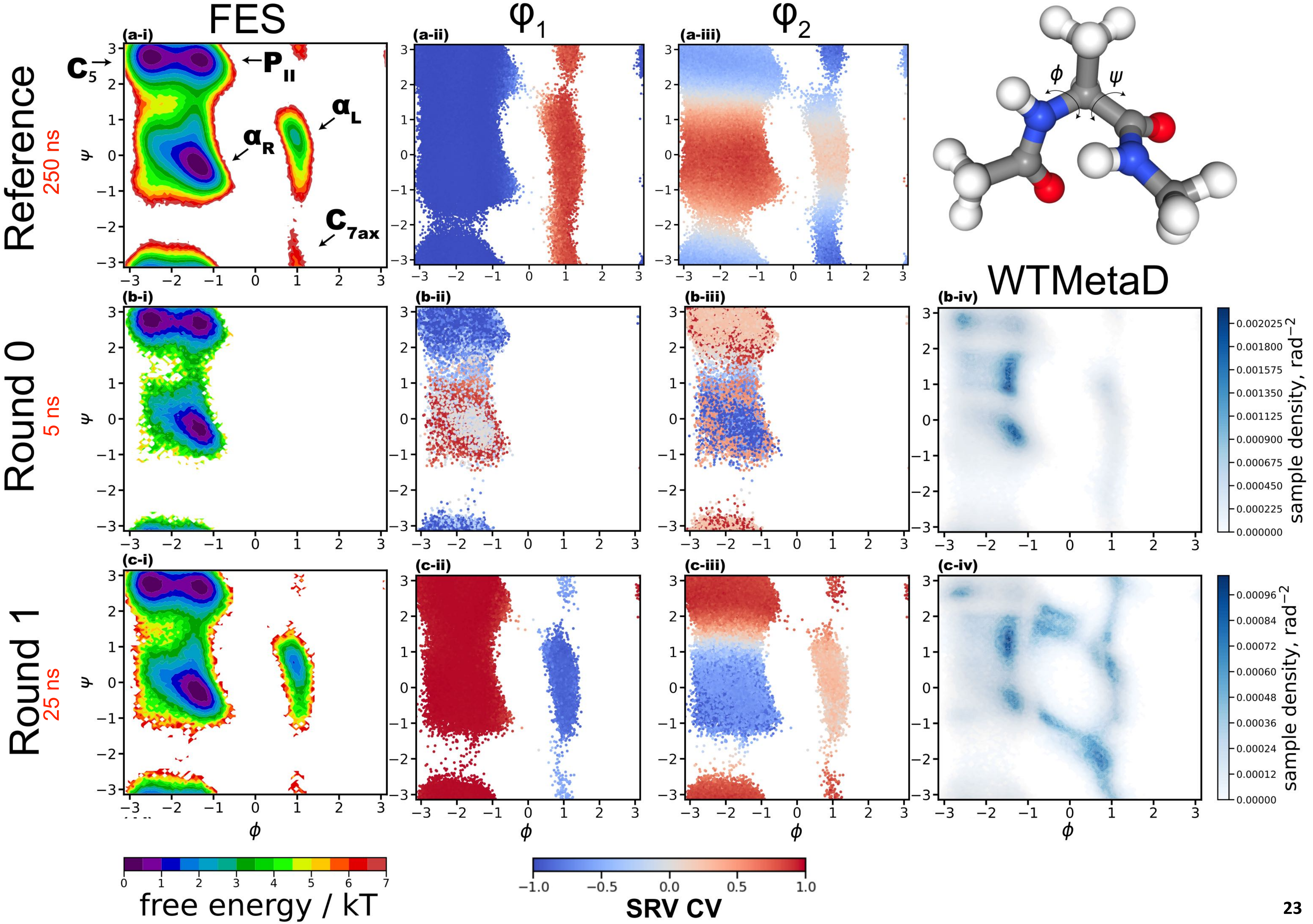




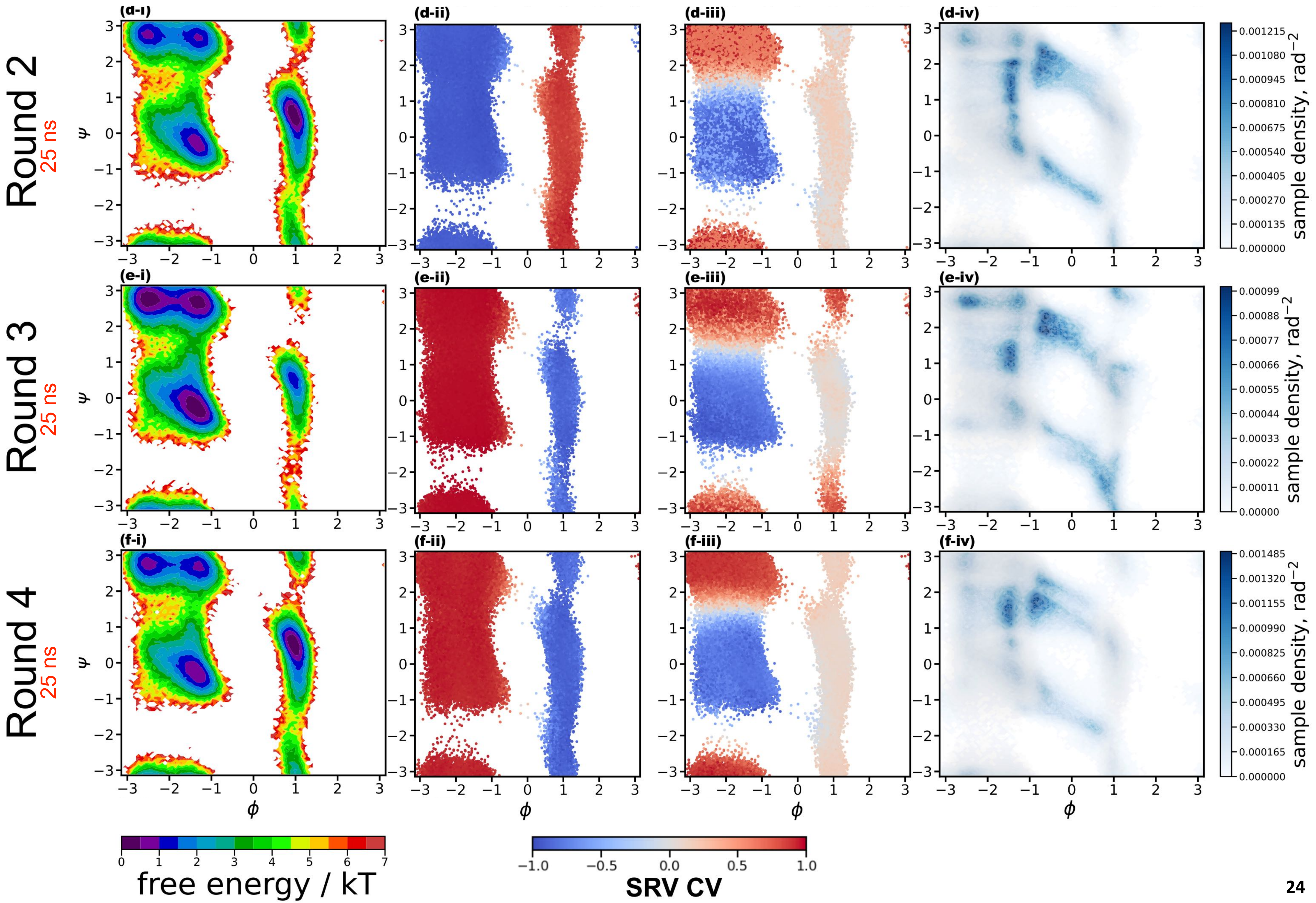
Alanine dipeptide



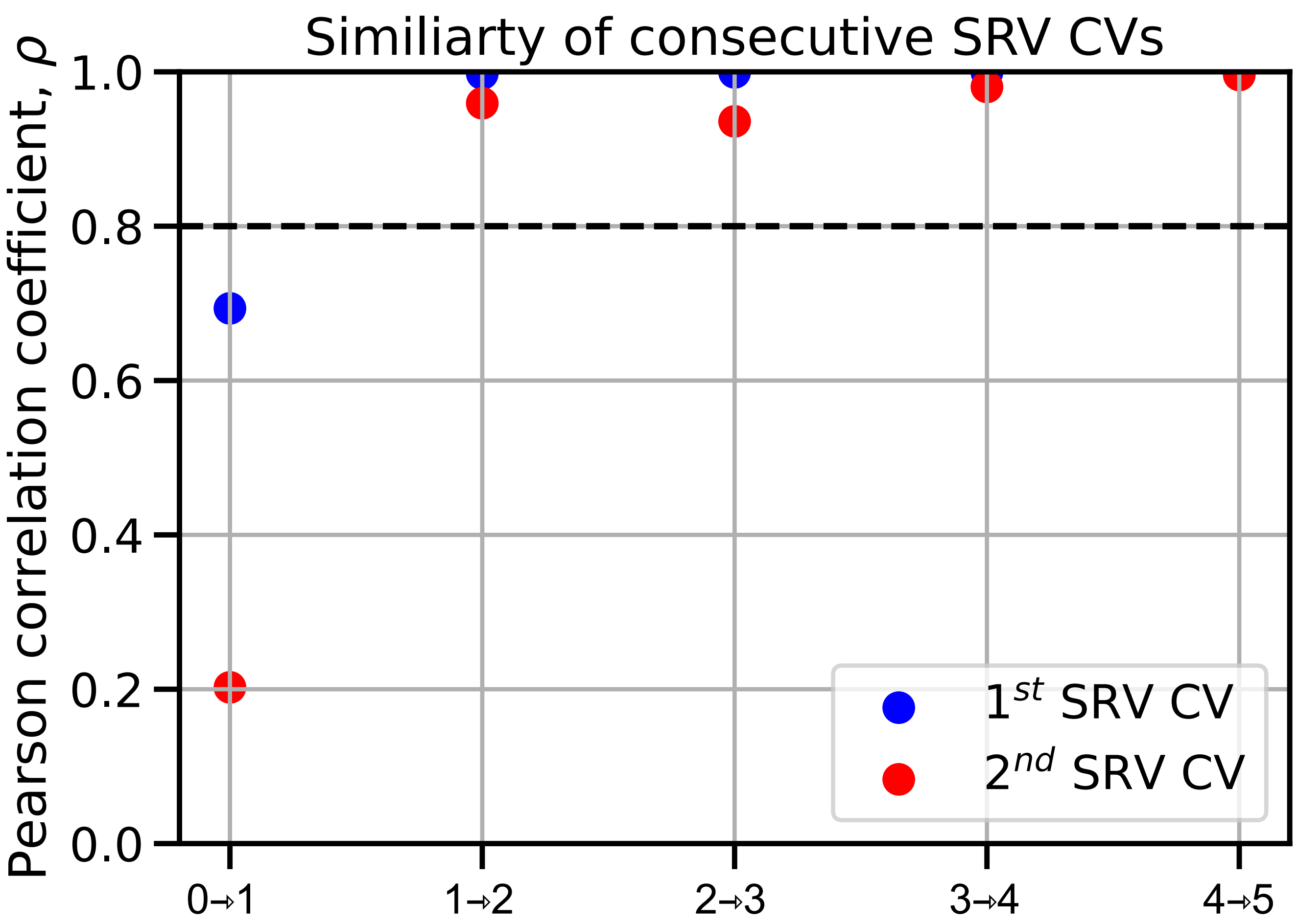
Alanine dipeptide

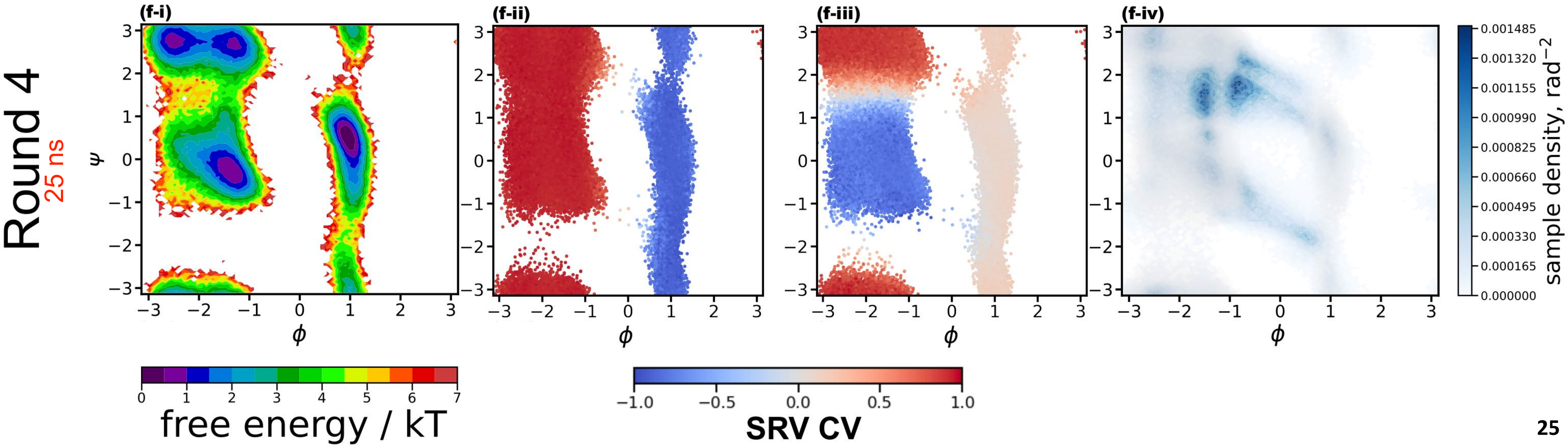


Alanine dipeptide

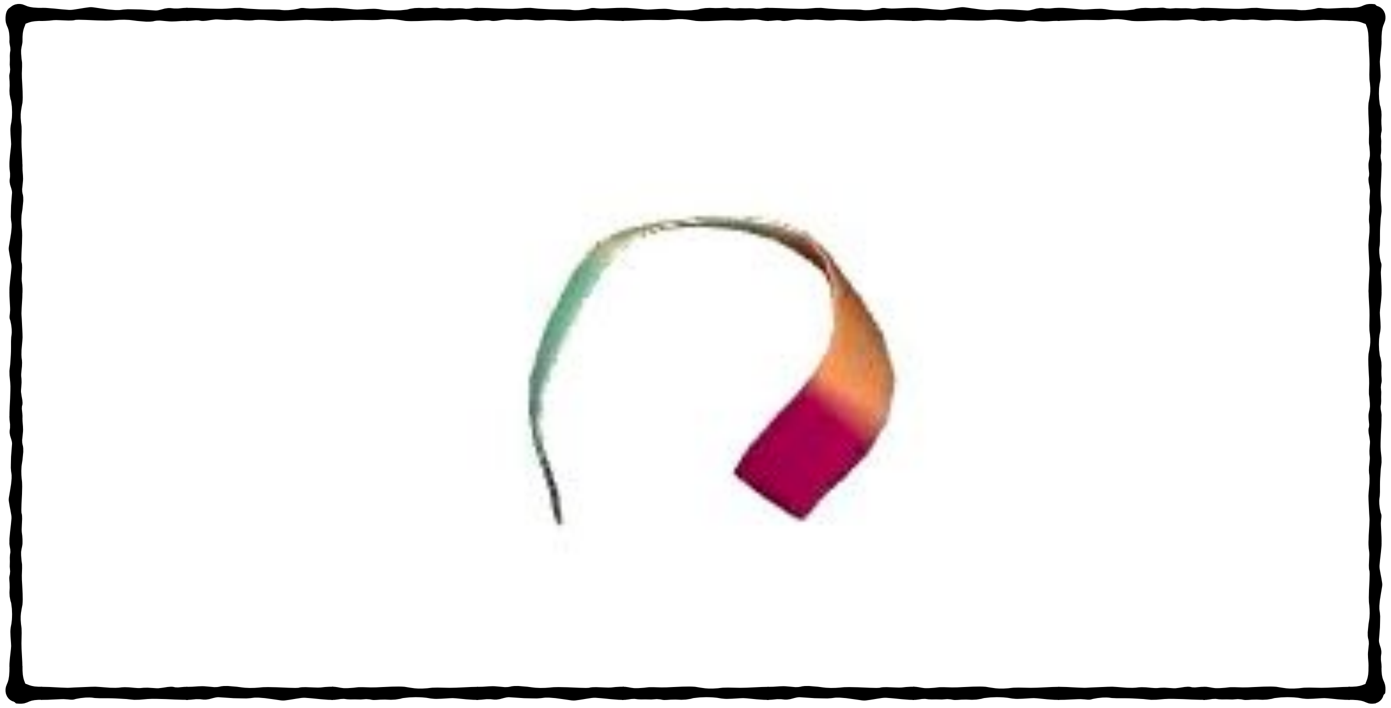


Alanine dipeptide



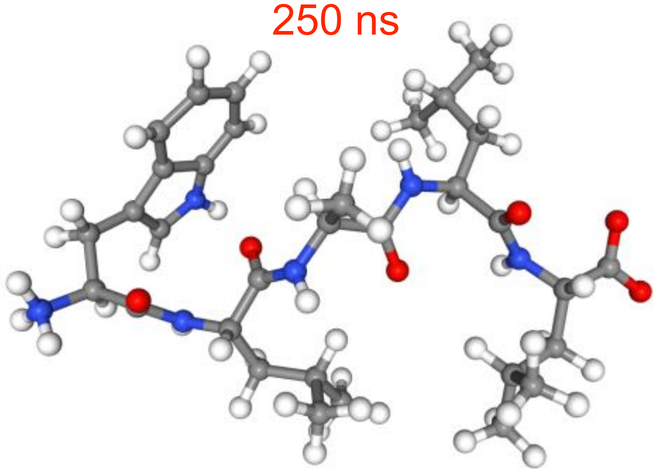


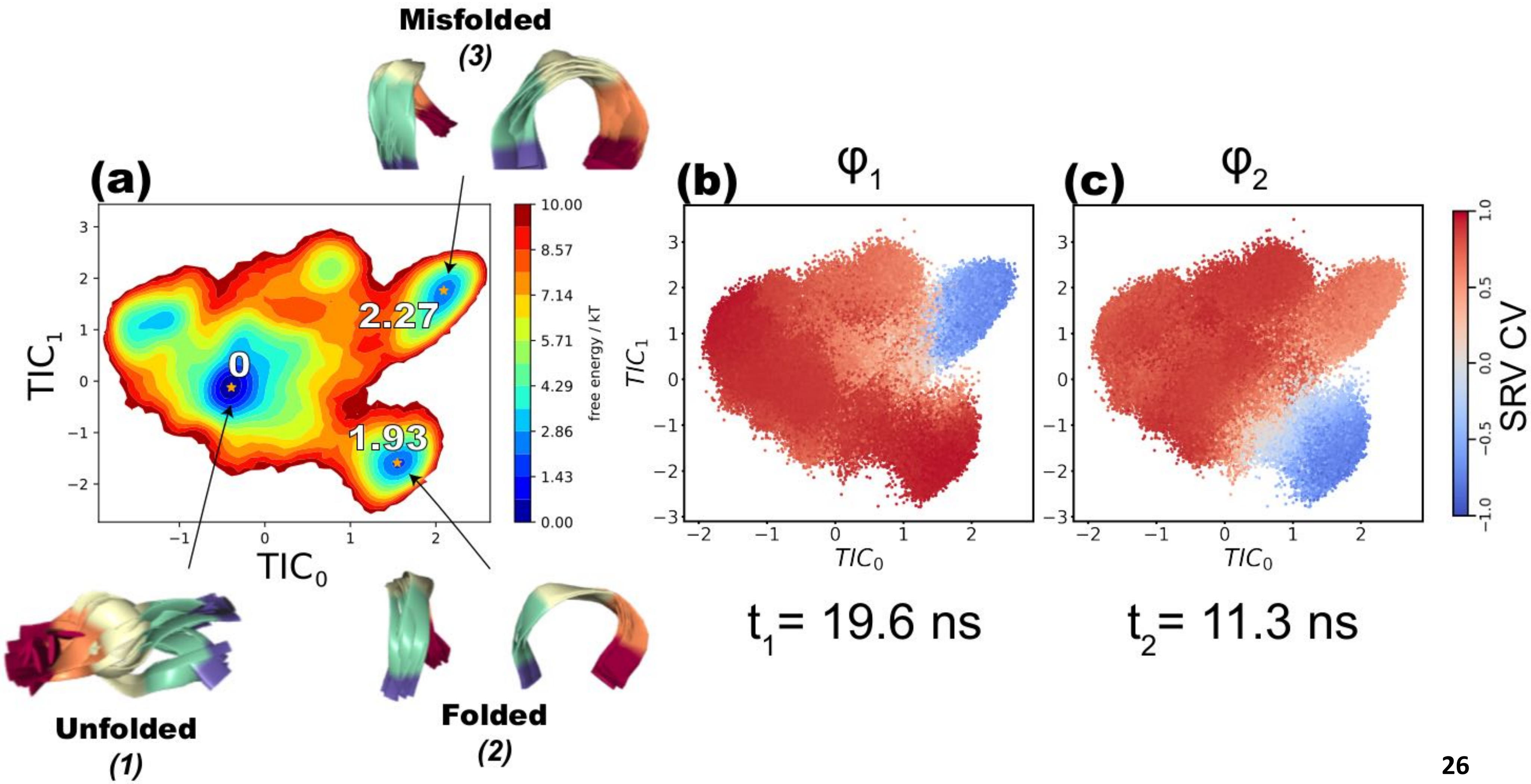
WLALL pentapeptide



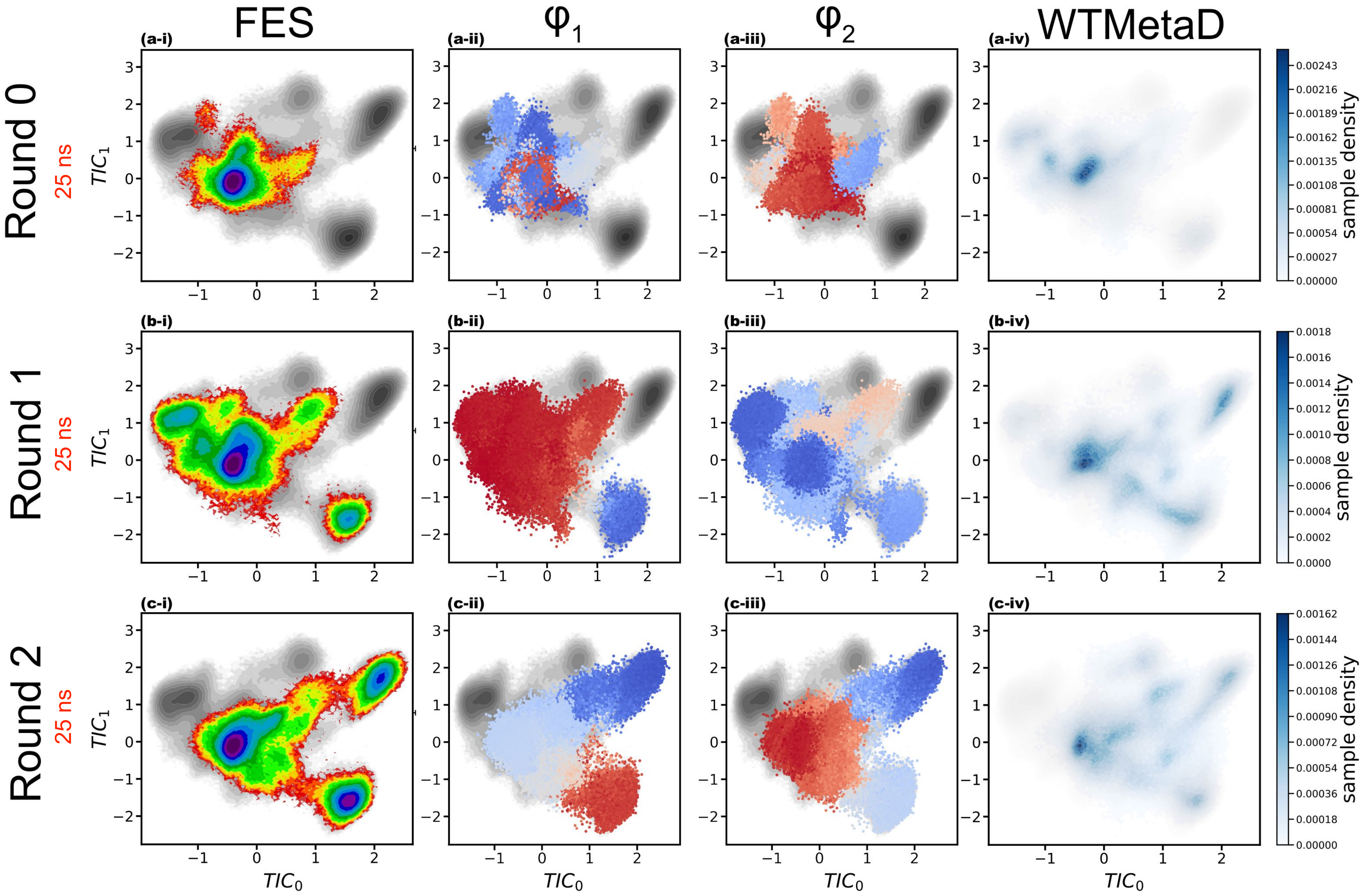
Trp-Leu-Ala-Leu-Leu (WLALL)

250 ns

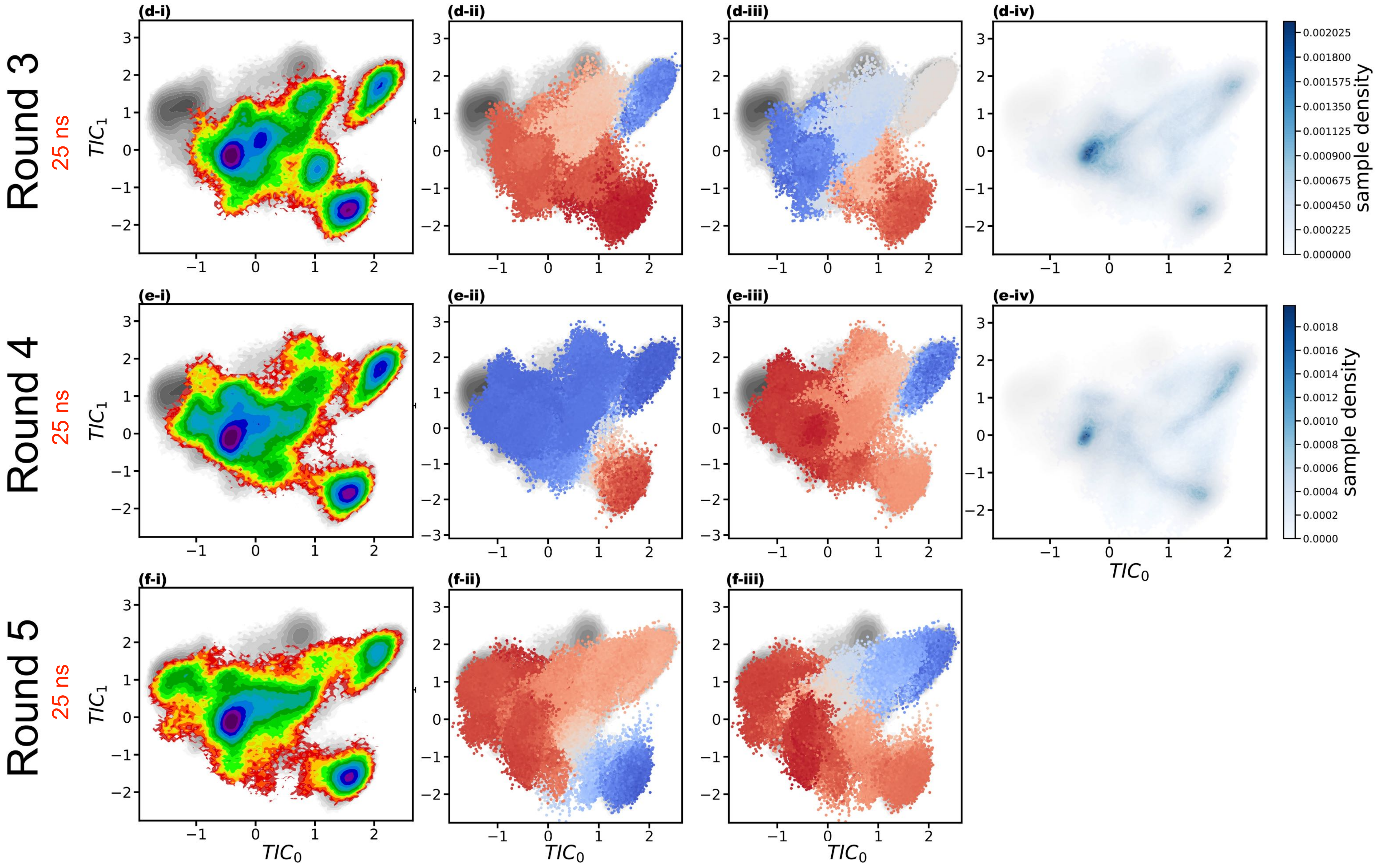




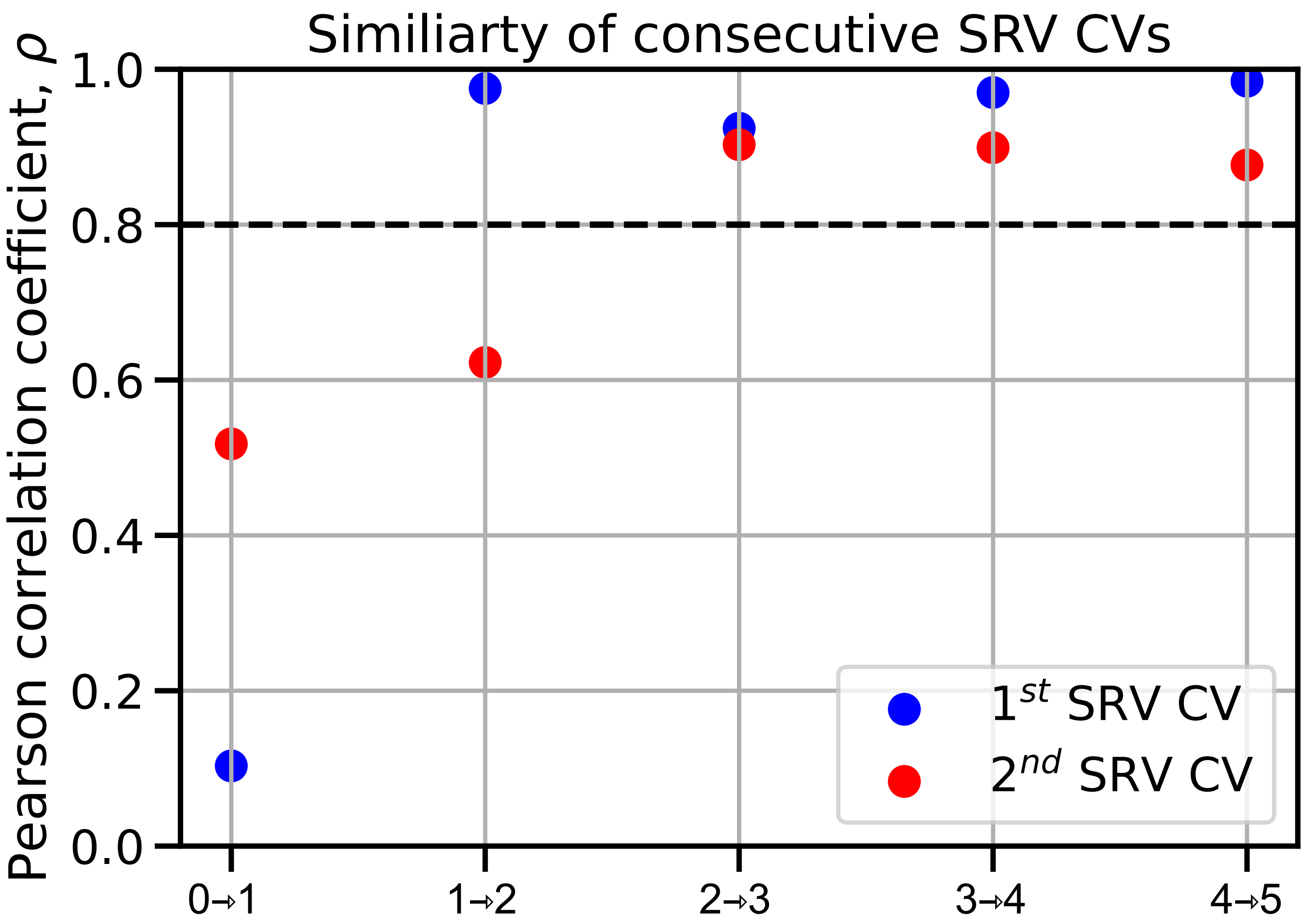
WLALL pentapeptide



WLALL pentapeptide

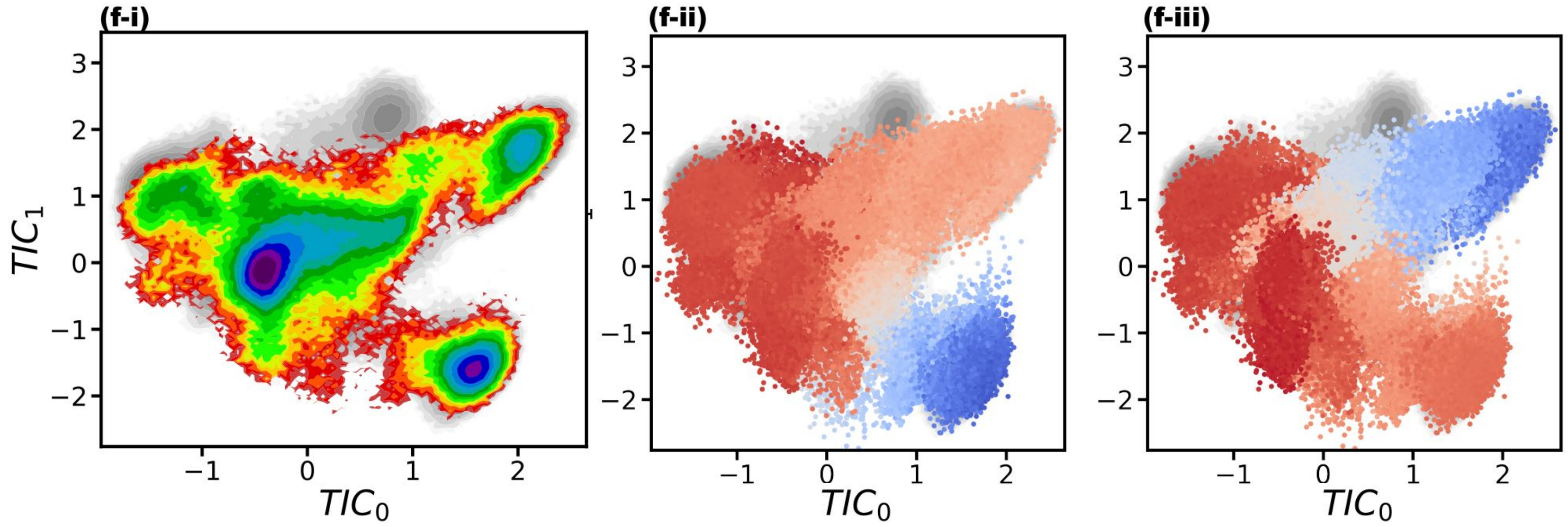


WLALL pentapeptide

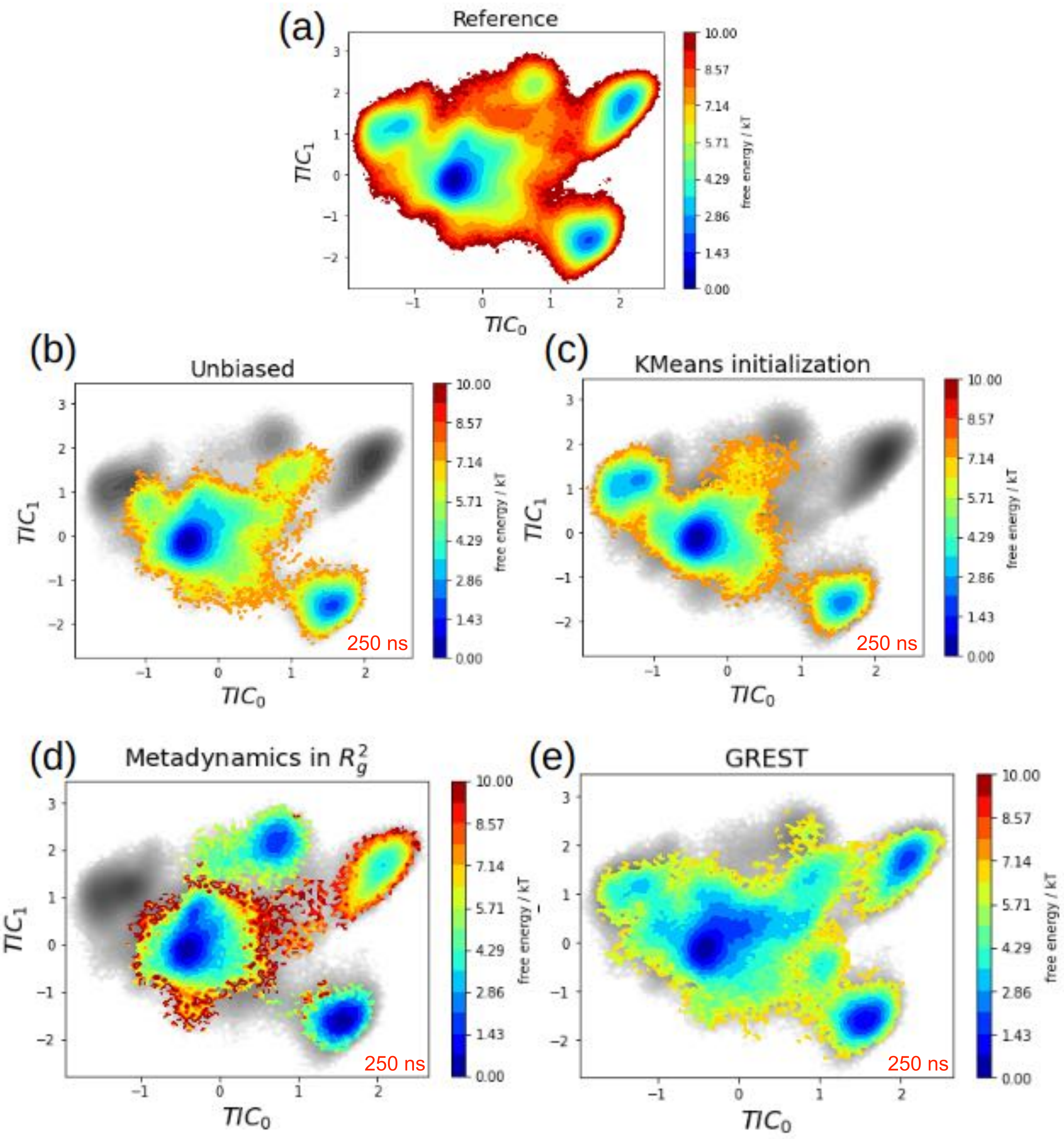


Round 5

25 ns



WLALL pentapeptide



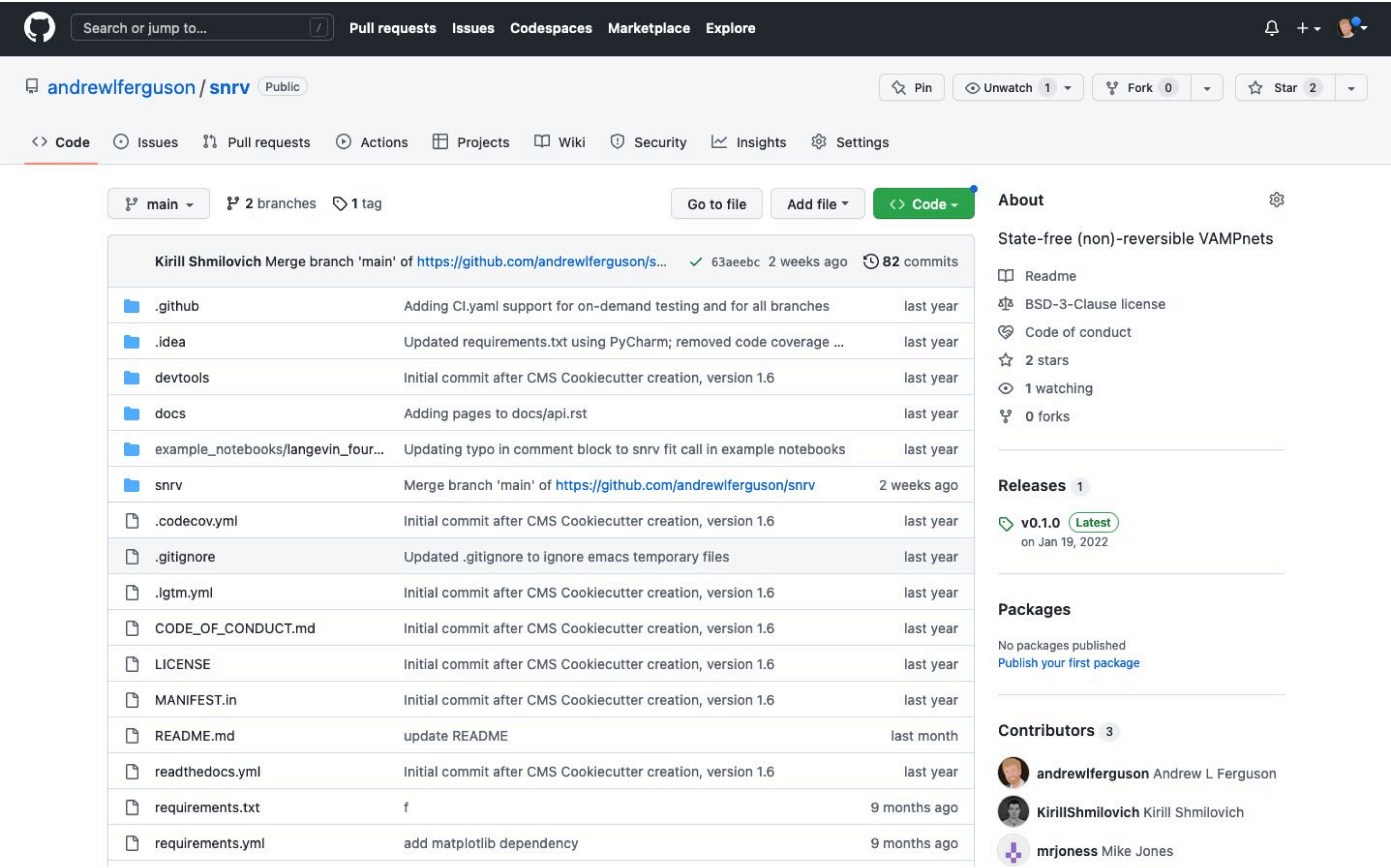
2/5 wells

3/5 wells

4/5 wells

5/5 wells

Python package



Kirill Shmilovich Merge branch 'main' of https://github.com/andrewlferguson/s... ✓ 63aeebc 2 weeks ago 82 commits

.github	Adding CI.yaml support for on-demand testing and for all branches	last year
.idea	Updated requirements.txt using PyCharm; removed code coverage ...	last year
devtools	Initial commit after CMS Cookiecutter creation, version 1.6	last year
docs	Adding pages to docs/api.rst	last year
example_notebooks/langevin_four...	Updating typo in comment block to snrv fit call in example notebooks	last year
snrv	Merge branch 'main' of https://github.com/andrewlferguson/snrv	2 weeks ago
.codecov.yml	Initial commit after CMS Cookiecutter creation, version 1.6	last year
.gitignore	Updated .gitignore to ignore emacs temporary files	last year
.lgtm.yml	Initial commit after CMS Cookiecutter creation, version 1.6	last year
CODE_OF_CONDUCT.md	Initial commit after CMS Cookiecutter creation, version 1.6	last year
LICENSE	Initial commit after CMS Cookiecutter creation, version 1.6	last year
MANIFEST.in	Initial commit after CMS Cookiecutter creation, version 1.6	last year
README.md	update README	last month
readthedocs.yml	Initial commit after CMS Cookiecutter creation, version 1.6	last year
requirements.txt	f	9 months ago
requirements.yml	add matplotlib dependency	9 months ago

About

State-free (non)-reversible VAMPnets

- Readme
- BSD-3-Clause license
- Code of conduct
- 2 stars
- 1 watching
- 0 forks

Releases 1

v0.1.0 Latest
on Jan 19, 2022

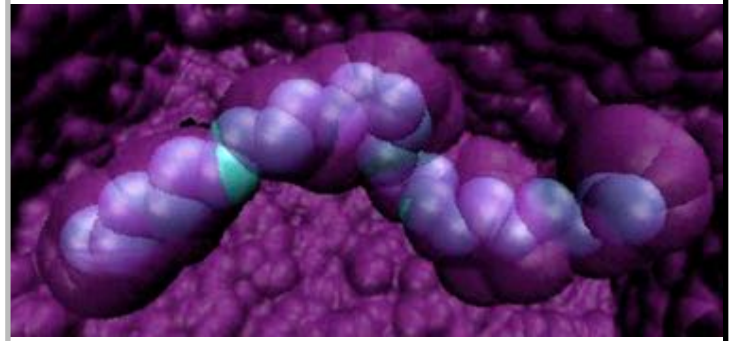
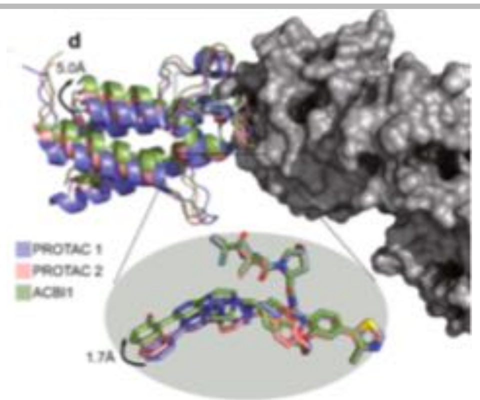
Packages

No packages published
Publish your first package

Contributors 3

- **andrewlferguson** Andrew L Ferguson
- **KirillShmilovich** Kirill Shmilovich
- **mrjoness** Mike Jones

Conclusions & Outlook

Girsanov reweighting enables rigorous iterative learning of and enhanced sampling in **slow molecular CVs**	$W(\omega_{t\to t+\tau}) = g(x_t) \times M(\omega_{t\to t+\tau})$
Slow CVs reveal **important dynamical motions** and **optimal for enhanced sampling & long-time kinetic models**	$\tilde{\lambda}_k = \frac{\langle u_k \vert \mathscr{T}(\tau) \circ u_k \rangle_\pi}{\langle u_k \vert u_k \rangle_\pi}$
Incorporation of **solvent coordinates** respecting permutational invariance	
Application to **larger protein systems**	

https://doi.org/10.1101/2021.09.26.461830