Enhanced sampling with autoencoders

Gabriel STOLTZ

(CERMICS, Ecole des Ponts & MATHERIALS team, Inria Paris)

Joint work with Z. Belkacemi (Sanofi & ENPC), P. Gkeka, M. Bianciotto, H. Minoux (Sanofi), T. Pigeon (Inria & IFPEN), Wei Zhang (FU Berlin) and T. Lelièvre (ENPC & Inria)

CECAM workshop "Quantum2 on machine learning enhanced sampling", November 2023

Outline

- **A (short/biased) review of machine learning approaches for CV**
- **Constructing CVs with autoencoders**[1]
  - Preliminaries: definitions, training, etc.
  - An interpretation in terms of conditional expectations
  - Constructing transition paths
- **Applications** (alanine dipeptide, chignolin, HSP90)
  - Free energy biasing and iterative learning[2]
  - A semi-supervised approach for complex systems[3]

---

[1] Lelièvre/Pigeon/Stoltz/Zhang, *arXiv preprint* **2310.03492**
[2] Belkacemi/Gkeka/Lelièvre/Stoltz, *J. Chem. Theory Comput.* **18** (2022)
[3] Belkacemi/Bianciotto/Minoux/Lelièvre/Stoltz/Gkeka, *J. Chem. Phys.* (2023)

ML approaches for finding CV

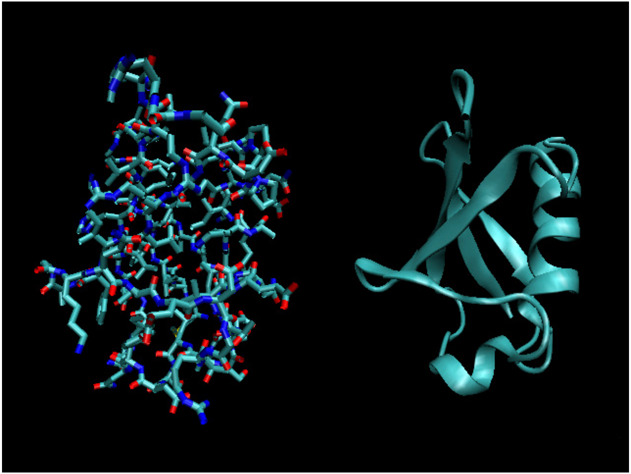
(A biased perspective on some) References

- **ML reviews in MD** (biased towards dimensionality reduction, not force fields)
  - A. Gliemlo, B. Husic, A. Rodriguez, C. Clementi, F. Noé, A. Laio, *Chem. Rev.* **121**(16), 9722-9758 (2021)
  - P. Gkeka *et al.*, *J. Chem. Theory Comput.* **16**(8), 4757-4775 (2020)
  - F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, *Annu. Rev. Phys. Chem.* **71**, 361-390 (2020)
  - A.L. Ferguson, *J. Phys.: Condens. Matter* **30**, 04300 (2018)
  - M. Chen, *Eur. Phys. J. B* **94**, 211 (2021)

- **More general ML references**
  - P. Mehta, M. Bukov, C.-H. Wang, A.G.R.Day, C. Richardson, C.K. Fisher, D.J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists, *Physics Reports* **810**, 1-124 (2019)
  - I. Goodfellow, Y. Bengio, A. Courville *Deep Learning* (MIT Press, 2016) `http://www.deeplearningbook.org`
  - K.P. Murphy, *Probabilistic Machine Learning: An Introduction* (MIT Press, 2022)

Statistical physics (1)

What is the structure of the protein? What are its typical conformations, and what are the transition pathways from one conformation to another?



Statistical physics (2)

- Microstate of a classical system of N particles:

$$(q,p) = (q_1,\dots,q_N,\ p_1,\dots,p_N) \in \mathcal{E} = (a\mathbb{T})^{3N} \times \mathbb{R}^{3N}$$

Positions q (configuration), momenta p (to be thought of as $M\dot{q}$)

- Hamiltonian $H(q,p) = V(q) + \displaystyle\sum_{i=1}^{N} \frac{p_i^2}{2m_i}$ (physics is in V)

Macrostate: Boltzmann–Gibbs probability measure (NVT)

$$\mu(dq\,dp) = Z_{\mathrm{NVT}}^{-1}\,\mathrm{e}^{-\beta H(q,p)}\,dq\,dp, \qquad \beta = \frac{1}{k_{\mathrm{B}}T}$$

- Typical evolution equations: Langevin dynamics (friction $\gamma > 0$)

$$\begin{cases} dq_t = M^{-1}p_t\,dt \\ dp_t = -\nabla V(q_t)\,dt - \gamma M^{-1}p_t\,dt + \sqrt{2\gamma\beta^{-1}}\,dW_t \end{cases}$$

Reaction coodinates (RC) / collective variables (CV)

- Reaction coordinate $\xi : (a\mathbb{T})^D \to \mathbb{R}^d$ with $d \ll D$
- Ideally: $\xi(q_t)$ captures the slow part of the dynamics
- **Free energy** computed on $\Sigma(z) = \{q \in (a\mathbb{T})^D \,|\, \xi(q) = z\}$ (foliation)

$$F(z) = -\frac{1}{\beta} \ln \left(\int_{\Sigma(z)} \mathrm{e}^{-\beta V(q)} \, \delta_{\xi(q)-z}(dq) \right)$$

- Various methods: TI, FEP, ABF, metadynamics, etc[4]

[4] Lelièvre/Rousset/Stoltz, *Free Energy Computations: A Mathematical Perspective* (Imperial College Press, 2010)

Some representative approaches for finding CV (1)

- Chemical/physical intuition (distances, angles, RMSDs, coordination numbers, etc)

- **Short list of data-oriented approaches** (depending on the data at hand...)
  - [supervised learning] separate metastable states
  - [unsupervised/static] distinguish linear models (PCA) and nonlinear ones (e.g. based on autoencoders such as MESA[5])
  - [unsupervised/dynamics] operator based approaches (VAC, EDMD, diffusion maps, MSM; incl. tICA and VAMPNets)

(Huge literature! I am not quoting precise references here because the list would be too long)

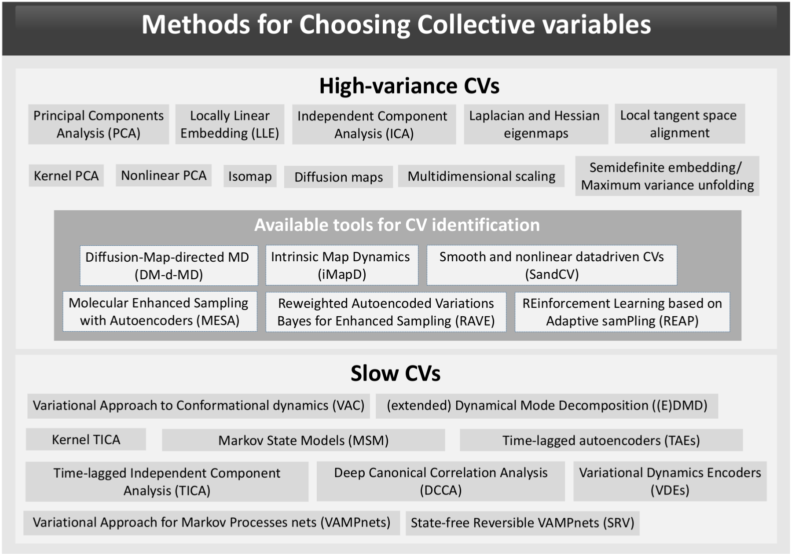
- Other classifications[6,7] possible, e.g. **slow vs. high variance CV**

[5] W. Chen and A.L. Ferguson, *J. Comput. Chem.* 2018; W. Chen, A.R. Tan, and A.L. Ferguson, *J. Chem. Phys.* 2018

[6] P. Gkeka et al., *J. Chem. Theory Comput.* (2020)

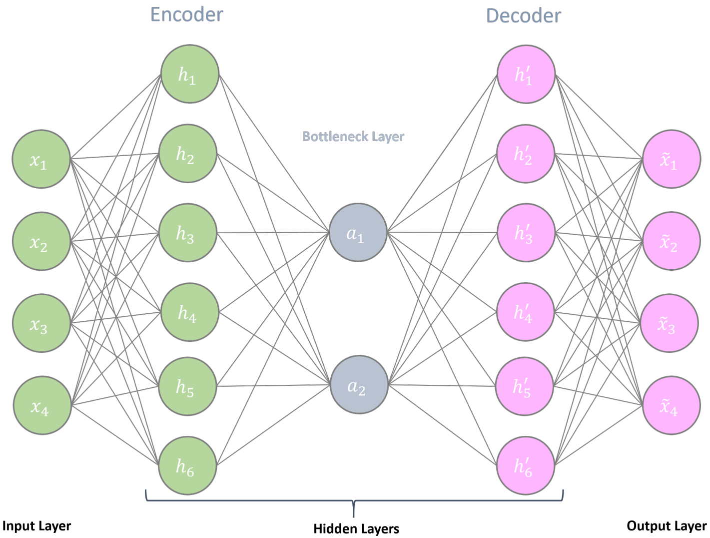
[7] A. Gliemlo *et al.*, *Annu. Rev. Phys. Chem.* (2021)

Some representative approaches for finding CV (2)



Constructing CVs with autoencoders

Bottleneck autoencoders (1)



Bottleneck autoencoders (2)

- Data space $\mathcal{X} \subseteq \mathbb{R}^D$, bottleneck space $\mathcal{A} \subseteq \mathbb{R}^d$ with $d < D$

$$f(x) = f_{\mathsf{dec}}\Big(f_{\mathsf{enc}}(x)\Big)$$

where $f_{\mathsf{enc}} : \mathcal{X} \to \mathcal{A}$ and $f_{\mathsf{dec}} : \mathcal{A} \to \mathcal{X}$

Collective variable = encoder part

$$\xi = f_{\mathrm{enc}}$$

- Fully connected neural network, symmetrical structure, $2L$ layers

- Parameters $\mathbf{p} = \{p_k\}_{k=1,\dots,K}$ (bias vectors b_ℓ and weights matrices W_ℓ)

$$f_{\mathbf{p}}(x) = g_{2L}\left[b_{2L} + W_{2L} \dots g_1(b_1 + W_1 x)\right],$$

with activation functions g_ℓ

(examples: $\tanh(x)$, ReLU $\max(0,x)$, sigmoid $\sigma(x) = 1/(1+\mathrm{e}^{-x})$, etc)

Training autoencoders

• Theoretically: minimization problem in $\mathcal{P} \subset \mathbb{R}^K$

$$\mathbf{p}_\mu \in \underset{\mathbf{p} \in \mathcal{P}}{\operatorname{argmin}} \, \mathcal{L}(\mu, \mathbf{p}),$$

with cost function

$$\mathcal{L}(\mu, \mathbf{p}) = \mathbb{E}_\mu(\|X - f_{\mathbf{p}}(X)\|^2) = \int_{\mathcal{X}} \|x - f_{\mathbf{p}}(x)\|^2 \, \mu(dx)$$

• In practice, access only to a sample: **minimization of empirical cost**

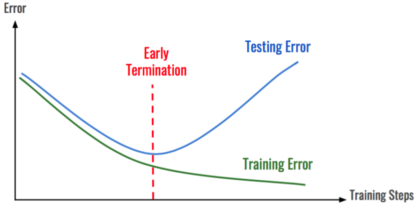
$$\mathcal{L}(\hat{\mu}, \mathbf{p}) = \frac{1}{N} \sum_{i=1}^{N} \|x^i - f_{\mathbf{p}}(x^i)\|^2, \qquad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \delta_{x^i}$$

• **Typical choices:** canonical measure μ, data points x^i postprocessed from positions q (alignement to reference structure, centering, reduction to backbone carbon atoms, etc)

Some elements on training neural networks

- Many local minima...

- **Actual procedure:**
  - "Early stopping": stop when validation loss no longer improves[8]



  - Choice of optimization method[9], here Adam
  - No added regularization here (e.g. ℓ^1/ℓ^2, dropout, etc)

[8] See Section 7.8 in [Goodfellow/Bengio/Courville]
[9] See Chapter 8 in [Goodfellow/Bengio/Courville]

Some properties of autoencoders

Three viewpoints on the loss function (1/2)

Idealized setting: $f_{\mathrm{enc}} : \mathcal{X} \to \mathcal{Z}$ and $f_{\mathrm{dec}} : \mathcal{Z} \to \mathcal{X}$ measurable

$$\mathcal{F} = \{ f = f_{\mathrm{dec}} \circ f_{\mathrm{enc}},\ f_{\mathrm{enc}} \in \mathcal{F}_{\mathrm{enc}},\ f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}} \}$$

Usual loss: $\inf_{f \in \mathcal{F}} \mathbb{E}\left[\| X - f(X) \|^2 \right]$

Principal manifold formulation: fix decoder, minimize over encoders

$$\begin{aligned} \inf_{f \in \mathcal{F}} \mathbb{E}\left[\| X - f(X) \|^2 \right] &= \inf_{f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}}} \left\{ \inf_{f_{\mathrm{enc}} \in \mathcal{F}_{\mathrm{enc}}} \mathbb{E}\left[\| X - f_{\mathrm{dec}} \circ f_{\mathrm{enc}}(X) \|^2 \right] \right\} \\ &= \inf_{f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}}} \mathbb{E}\left[\left\| X - f_{\mathrm{dec}} \circ h^\star_{f_{\mathrm{dec}}}(X) \right\|^2 \right] \end{aligned}$$

with ideal encoder $h^\star_{f_{\mathrm{dec}}}(x) \in \mathrm{argmin}_{z \in \mathcal{Z}} \| x - f_{\mathrm{dec}}(z) \|$

Hastie/Stützle (1986), Tibshirani (1992)
Venturoli/Vanden–Eijnden (2009)
Gerber/Whitaker, *J. Mach. Learn. Res.* (2013); Gerber, *arXiv preprint* **2104.05000**

Three viewpoints on the loss function (2/2)

Formulation with conditional expectation: fix encoder, minimize over decoders

$$\begin{aligned}\inf_{f\in\mathcal{F}} \mathbb{E}\left[\|X - f(X)\|^2\right] &= \inf_{f_{\mathrm{enc}}\in\mathcal{F}_{\mathrm{enc}}} \left\{ \inf_{f_{\mathrm{dec}}\in\mathcal{F}_{\mathrm{dec}}} \mathbb{E}\left[\|X - f_{\mathrm{dec}}\circ f_{\mathrm{enc}}(X)\|^2\right]\right\} \\ &= \inf_{f_{\mathrm{enc}}\in\mathcal{F}_{\mathrm{enc}}} \mathbb{E}\left[\left\|X - g^\star_{f_{\mathrm{enc}}}\circ f_{\mathrm{enc}}(X)\right\|^2\right]\end{aligned}$$

with ideal decoder $g^\star_{f_{\mathrm{enc}}}(z) = \mathbb{E}[\,X \mid f_{\mathrm{enc}}(X) = z]$

Alternative interpretations of the reconstruction error

$$\begin{aligned}\mathbb{E}\left[\left\|X - g^\star_{f_{\mathrm{enc}}}\circ f_{\mathrm{enc}}(X)\right\|^2\right] &= \mathrm{Var}(X) - \mathrm{Var}\left[\mathbb{E}(X|f_{\mathrm{enc}}(X))\right] \\ &= \mathbb{E}\left[\mathrm{Var}(X|f_{\mathrm{enc}}(X))\right]\end{aligned}$$

First equality by developing square, second by conditioning on $f_{\mathrm{enc}}(X)$

Numerical illustration

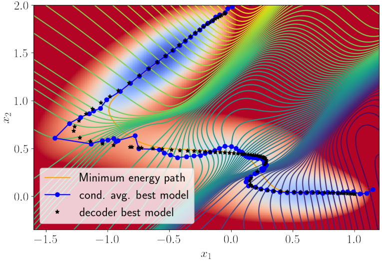
Practical implication: minimizing reconstruction loss amounts to...

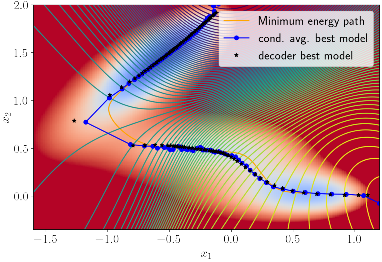
- minimizing intraclass dispersion

  (small spread of data points for $f_{\rm enc}$ given around the mean)

- maximizing interclass dispersion

  (the mean values associated with $f_{\rm enc}$ given should be as spread out as possible)





Left: topology (2, 5, 5, 1, 5, 5, 2). Right: topology (2, 5, 5, 1, 20, 20, 2)

Additional properties

- Necessary condition for critical points of the loss functional on f_{enc}

$$\left[x - g^\star_{f_{\mathrm{enc}}}(f_{\mathrm{enc}}(x))\right]^\top \partial_{z_j} g^\star_{f_{\mathrm{enc}}}(f_{\mathrm{enc}}(x)) = 0, \qquad 1 \leqslant j \leqslant d, \ x \in \mathrm{Supp}(\mu)$$

Low temperature limit: $\{g^\star_{f_{\mathrm{enc}}}(z)\}_{z\in[z_A,z_B]}$ minimum energy path[10]

- **Formulation with conditional expectations for other models:**
  - clustering
  - PCA (= autoencoders with identity activation functions)

- **Various extensions:**
  - change reference measure to better take transition states into account
  - multiple transition paths: single encoder and several decoders
  - possibly add some regularization terms

[10] Venturoli/Vanden–Eijnden (2009)

Free energy biasing
and iterative learning

Extended systems

- Computing $\nabla\xi$ already difficult, higher order derivatives is worse

- **Extended system** strategy : $V_{\mathrm{ext}}(q,\lambda) = V(q) + \dfrac{\kappa}{2}\big(\xi(q)-\lambda\big)^2$

- Free energy for the (simple) collective variable $\xi_{\mathrm{ext}}(q,\lambda) = \lambda$

$$
\begin{aligned}
F_\kappa(z) &= -\frac{1}{\beta}\ln\int_{\mathcal{D}} \mathrm{e}^{-\beta V_{\mathrm{ext}}(q,z)}\,dq + C \\
&= -\frac{1}{\beta}\ln\int\left(\int_{\Sigma(\zeta)} \mathrm{e}^{-\beta V(q)}\delta_{\xi(q)-\zeta}(dq)\right)\mathrm{e}^{-\beta\kappa(\zeta-z)^2/2}\,d\zeta + C \\
&= -\frac{1}{\beta}\ln\int \mathrm{e}^{-\beta F(\zeta)}\chi_\kappa(z-\zeta)\,d\zeta + \widetilde{C}, \qquad \chi_\kappa(s) = \left(\frac{\beta\kappa}{2\pi}\right)^{d/2}\mathrm{e}^{-\beta\kappa s^2/2} \\
&\xrightarrow[\kappa\to+\infty]{} F(z)
\end{aligned}
$$

Calls for taking κ large

Extended ABF

Extended overdamped Langevin dynamics (κ limits Δt...)

$$\begin{cases} dq_t = \Big[- \nabla V(q_t) + \kappa(\xi(q_t) - \lambda_t)\nabla\xi(q_t)\Big] dt + \sqrt{2\beta^{-1}} dW_t^q \\ d\lambda_t = -\kappa[\lambda_t - \xi(q_t)]\, dt + \sqrt{2\beta^{-1}}\, dW_t^\lambda \end{cases}$$

Bias by the free energy: add $F'_\kappa(\lambda)$ = steady state conditional average of $\kappa(\lambda - \xi(q))$

Extended ABF overdamped Langevin dynamics

$$\begin{cases} dq_t = \Big[- \nabla V(q_t) + \kappa(\xi(q_t) - \lambda_t)\nabla\xi(q_t)\Big] dt + \sqrt{2\beta^{-1}} dW_t^q \\ d\lambda_t = \kappa[\xi(q_t) - \mathbb{E}(\xi(q_t) \,|\, \lambda_t)]\, dt + \sqrt{2\beta^{-1}}\, dW_t^\lambda \end{cases}$$

In practice, $\mathbb{E}(\xi(q_t) \,|\, \lambda_t)$ is estimated by $\dfrac{\displaystyle\int_0^t \delta_\varepsilon(\lambda_s - \Lambda)\xi(q_s)\, ds}{\max\left(\eta, \displaystyle\int_0^t \delta_\varepsilon(\lambda_s - \Lambda)\, ds\right)}$

Iterative training on modified target measures

- Interesting systems are metastable (no spontaneous exploration of phase space)
Sample according to a biased distribution $\widetilde{\mu}$ (importance sampling)

- Need for reweighting[11] $w(q) = \mu(q)/\widetilde{\mu}(q)$

$$\mathcal{L}(\widehat{\mu}_{\mathsf{wght}}, \mathbf{p}) = \sum_{i=1}^{N} \widehat{w}_i \|q^i - f_{\mathbf{p}}(q^i)\|^2, \qquad \widehat{w}_i = \frac{\mu(q^i)/\widetilde{\mu}(q^i)}{\sum_{j=1}^{N} \mu(q^j)/\widetilde{\mu}(q^j)}$$

- **Free-energy biasing:** $\mu(q,\lambda) \propto \mathrm{e}^{-\beta V_{\mathrm{ext}}(q,\lambda)}$ and $\widetilde{\mu}(q,\lambda) \propto \mu(q,\lambda)\mathrm{e}^{\beta F_\kappa(\lambda)}$

- Iteration between free energy biasing for ξ fixed and retraining of $\xi = f_{\mathrm{enc}}$

- Convergence: (linear) regression to assess whether $\xi_k \approx \Phi(\xi_{k-1})$

[11] As done in RAVE for instance, see Ribeiro/Bravo/Wang/Tiwary (2018), Wang/Ribeiro/Tiwary (2019)

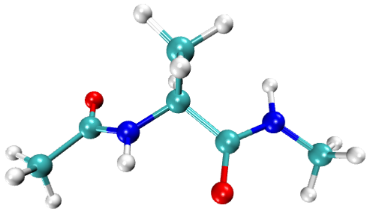
Alanine dipeptide

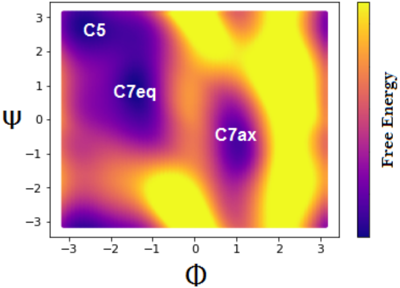
- **Molecular dynamics:**

openmm with openmm-plumed to link it with plumed
colvar module for eABF and computation of free energies[12]
timestep 1 fs, friction $\gamma = 1$ ps^{-1} in Langevin dynamics

- **Machine learning:**

keras for autoencoder training
input = carbon backbone (realignement to reference structure and centering)
neural network: topology 24-40-2-40-24, tanh activation functions



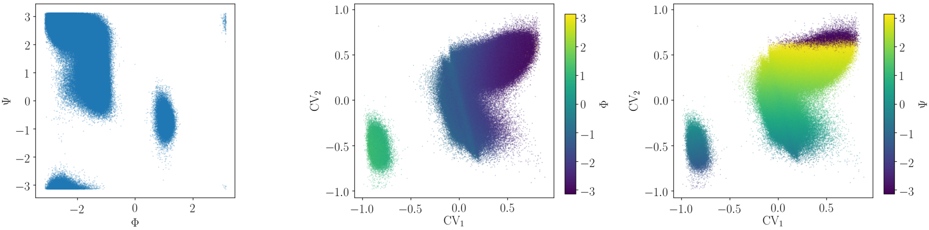


[12] See also Chen/Liu/Feng/Fu/Cai/Shao/Chipot, *J. Chem. Inf. Model.* (2022)

Ground truth computation

Long trajectory (1.5 μs), $N = 10^6$ (frames saved every 1.5 ps)

CV close to dihedral angles Φ, Ψ

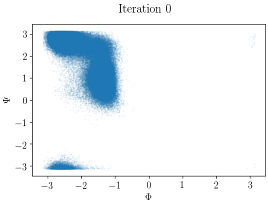


Quantify $s_{\min} = 0.99$ for $N = 10^5$ using a bootstraping procedure

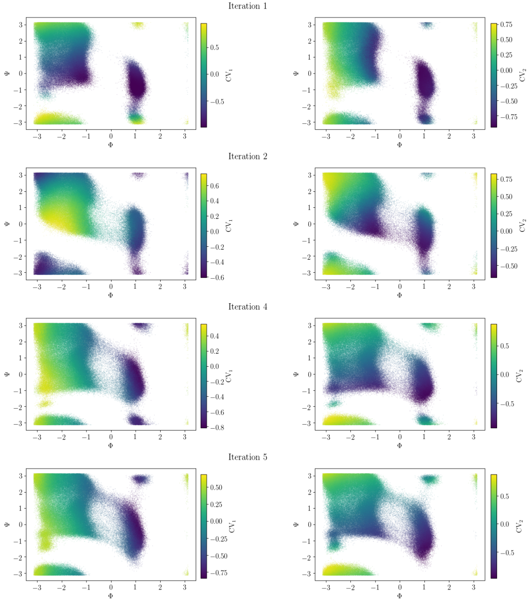
For the iterative algorithm: 10 ns per iteration

(compromise between times not too short to allow for convergence of the free energy, and not too large in order to alleviate the computation cost)

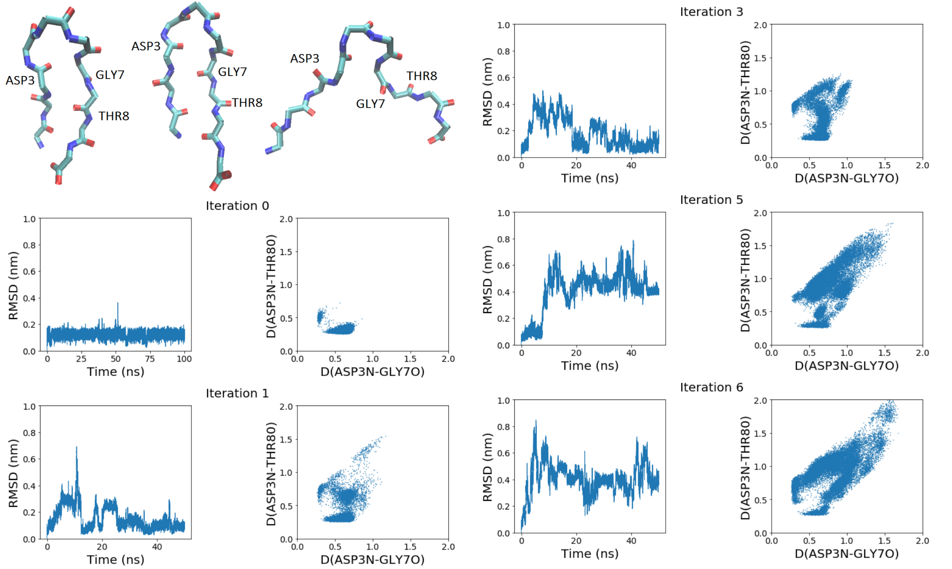
Results for the iterative algorithm



iter.	regscore	(Φ, Ψ)
0	–	0.922
1	0.872	0.892
2	0.868	0.853
3	0.922	0.973
4	0.999	0.972
5	0.999	0.970
6	0.999	0.971
7	0.999	0.967
8	0.998	0.966
9	0.999	0.968

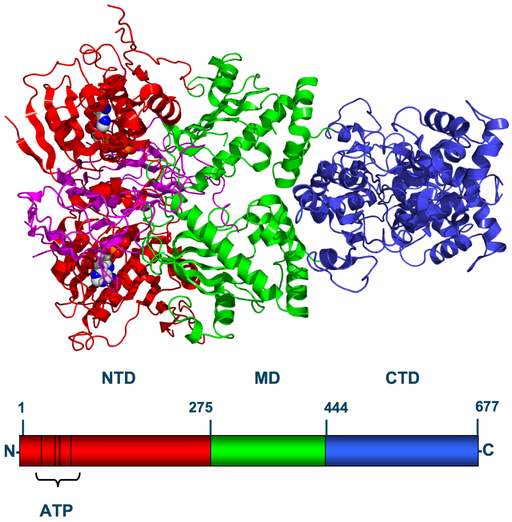


Chignolin (Folded/misfolded/unfolded states)



A semi-supervised approach for complex systems

Case study: HSP90

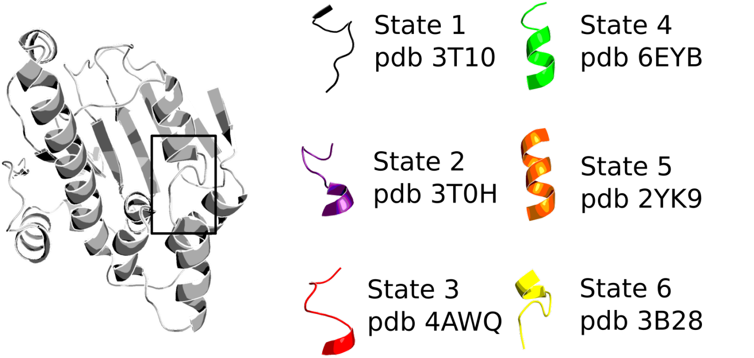


(picture from https://en.wikipedia.org/wiki/File:Hsp90_schematic_2cg9.png)

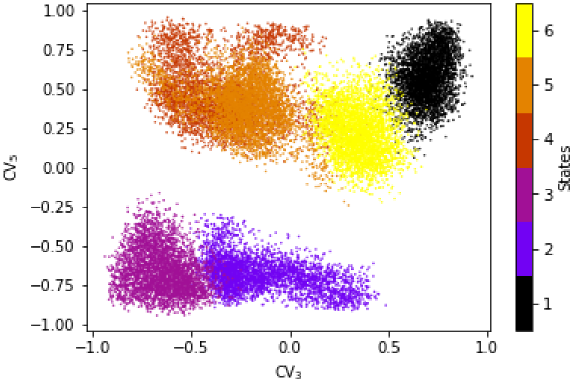
Chaperone protein assisting other proteins to fold properly and stabilizing them against stress, including proteins required for tumor growth

⟶ look for inhibitors (e.g. targeting binding region of ATP; focus only on the N-terminal domain)

Semi-supervised approach



Local sampling from 6 known conformations taken from the protein databank ; 10×20 ns trajectories each (in fact, States 4 and 5 can be merged)



Autoencoder

- input = 207 C carbons
- 621-100-k-100-621
- bottleneck $2 \leqslant k \leqslant 10$

Best heuristic:

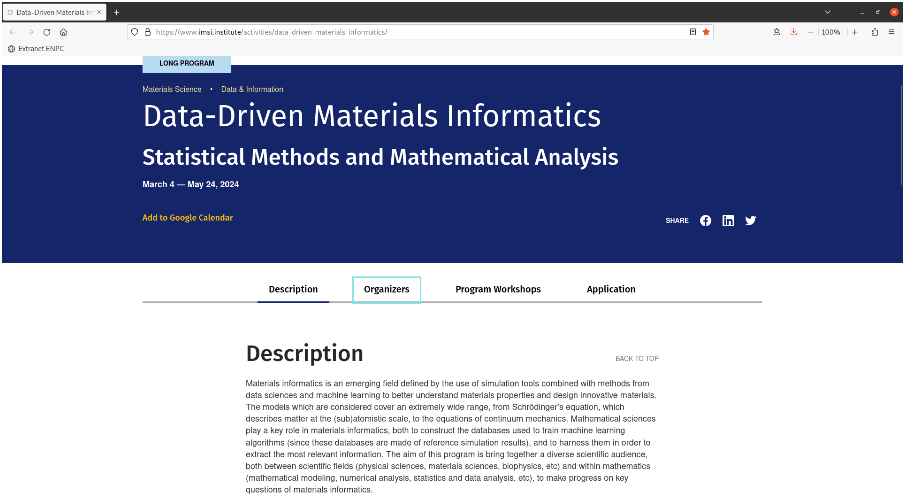
- $k = 5$
- further reduce with coordinate projection

Free energy biasing with the resulting CV allows to observe transitions between states

On-going work on making these choices more automatic

Interested in interfaces between ML and MD?

Program March-May 2024 at University of Chicago


Data-Driven Materials Inf ×

https://www.imsi.institute/activities/data-driven-materials-informatics/

Extranet ENPC

LONG PROGRAM

Materials Science • Data & Information

Data-Driven Materials Informatics

Statistical Methods and Mathematical Analysis

March 4 — May 24, 2024

Add to Google Calendar

SHARE

Description | Organizers | Program Workshops | Application

Description

BACK TO TOP

Materials informatics is an emerging field defined by the use of simulation tools combined with methods from data sciences and machine learning to better understand materials properties and design innovative materials. The models which are considered cover an extremely wide range, from Schrödinger's equation, which describes matter at the (sub)atomistic scale, to the equations of continuum mechanics. Mathematical sciences play a key role in materials informatics, both to construct the databases used to train machine learning algorithms (since these databases are made of reference simulation results), and to harness them in order to extract the most relevant information. The aim of this program is bring together a diverse scientific audience, both between scientific fields (physical sciences, materials sciences, biophysics, etc) and within mathematics (mathematical modeling, numerical analysis, statistics and data analysis, etc), to make progress on key questions of materials informatics.