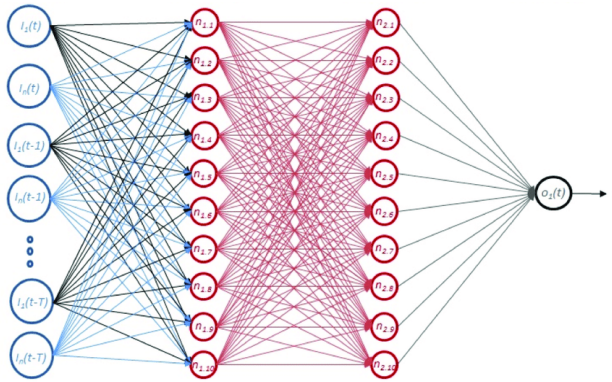
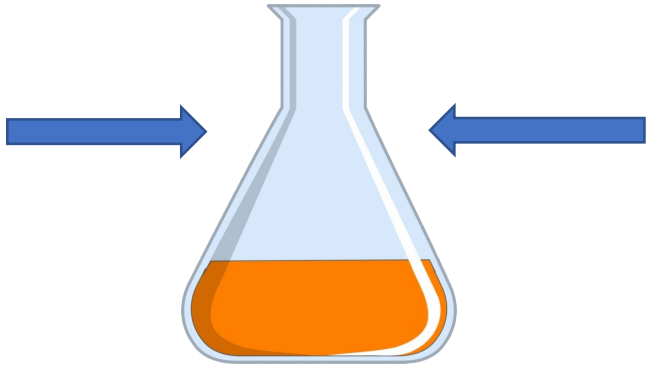
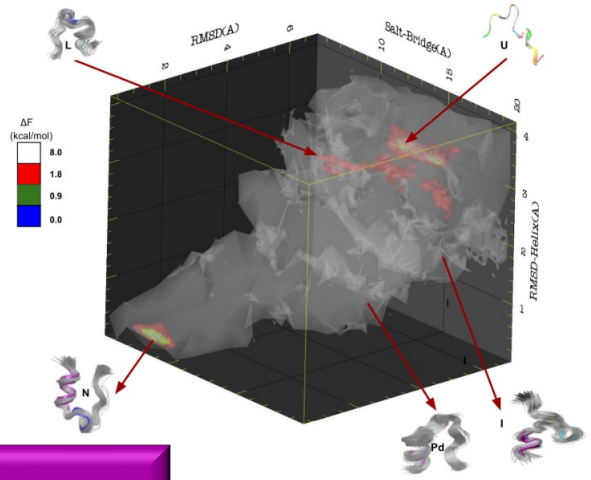
From A to B via a synthesis of rare-event sampling and machine learning







(c)

IPAM Workshop on Learning and Emergence in Molecular Systems, January (2023)

Mark E. Tuckerman
Chair of Chemistry
Dept. of Chemistry and Courant Institute of Mathematical Sciences
Simons Center for Computational Physical Chemistry
New York University, 100 Washington Square East, NY 10003
NYU-ECNU Center for Computational Chemistry at NYU Shanghai 200062, China

□ □ □ □ □ □ -□ □ □ □ □ □ □ □ □ □ □ □ □ □

National Science Foundation
MRSEC
MATERIALS RESEARCH SCIENCE AND ENGINEERING CENTERS

New York University

上海纽约大学
NYU SHANGHAI

NEW YORK UNIVERSITY MRSEC

Acknowledgements

External

- Serdal Kirmizialtin (NYU AD)
- Charlles Abreu (Redesign Science)

Postdocs

- Nawavi Naleem (NYU AD)

Students

- Muchen Tong (NYU)
- Shitanshu Bajpai (IIT Kanpur)
- Luke Dai (UC Berkeley)

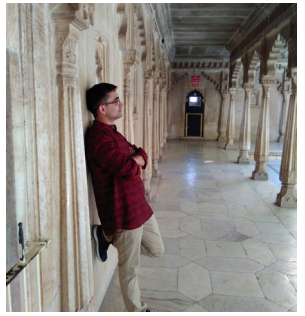
Funding

جامعة نيويورك أبوظبي

NYU ABU DHABI

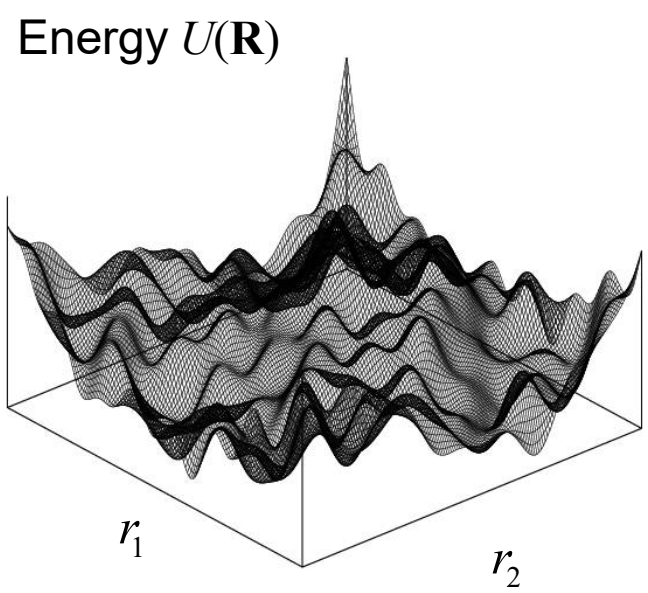


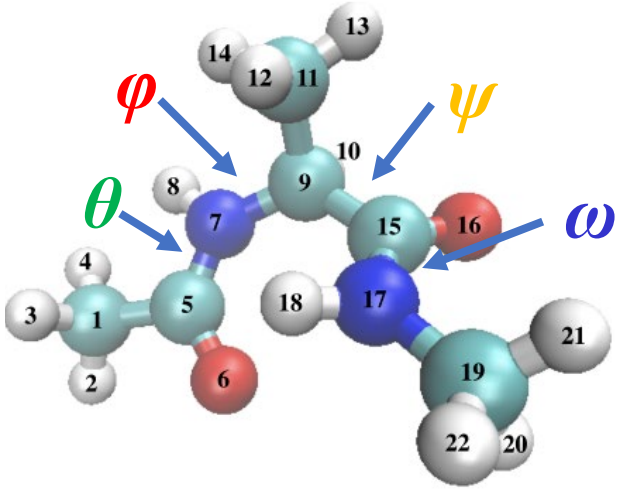


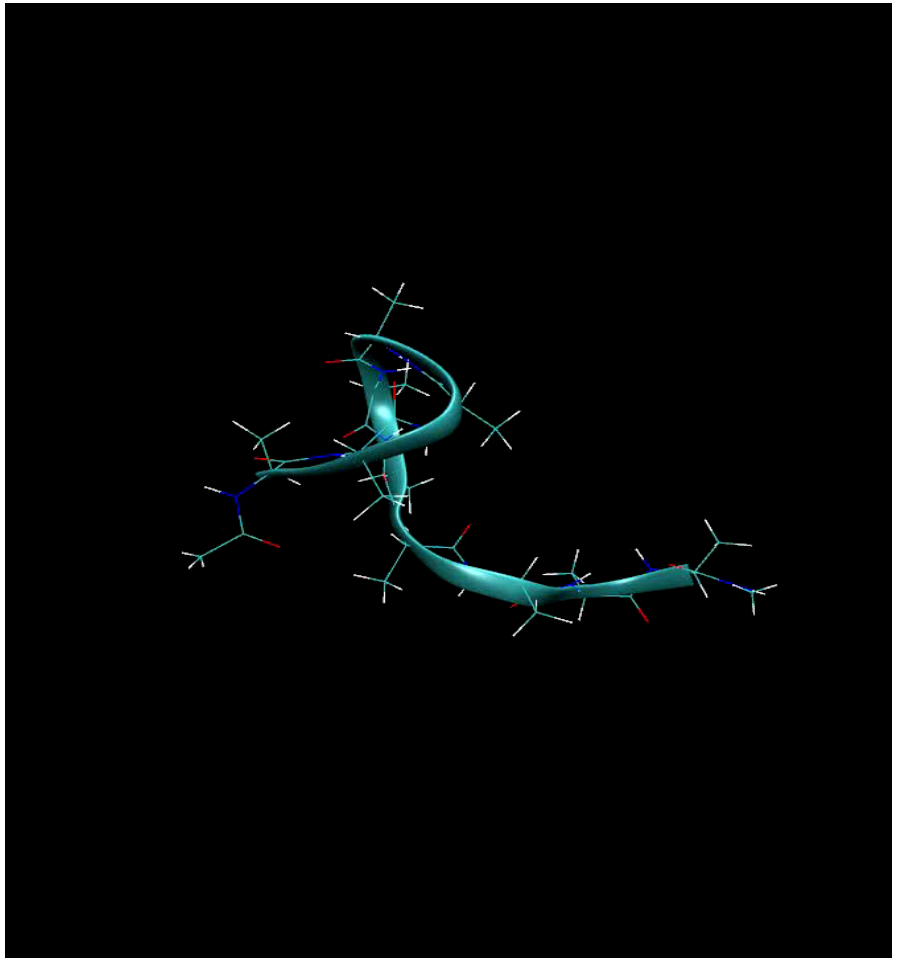



@GroupTuckerman

Exemplifying rare events: Alanine decamer







Rare-event MD simulation targeting collective variables [Cuendet et al. *JCP* (2018)]

$$\Delta A_{\text{hairpin}\to\text{helix}} \approx -40 \text{ kJ/mol}$$

Using deactivated morphing [Park et al. *JCP* (2008)]:

$$\Delta A_{\text{hairpin}\to\text{helix}} = -42 \pm 4 \text{ kJ/mol}$$

Driven Adiabatic free energy (temperature-accelerated molecular) dynamics (d-AFED/TAMD)

L. Rosso, P. Minary, Z. Zhu, MET *J. Chem. Phys*. **116**, 4389 (2002), Margliano and Vanden-Eijnden, *Chem. Phys. Lett.* **426**, 168 (2006); J. B. Abrams, MET, *J. Phys. Chem. B* **112**, 14752 (2008).
For prediction of crystal structures: Yu and MET *Phys. Rev. Lett*. **107**, 015701 (2011); Yu *et al. J. Chem. Phys*. **150**, 214109 (2014)

Suppose n collective variables characterize a free energy landscape of interest

$$q_\alpha(\mathbf{R}): \mathbb{R}^{dN} \to \mathbb{R}, \qquad \alpha = 1,\ldots,n < dN$$

Canonical probability distribution and free energy surface:

$$P(s_1,\ldots,s_n,T) = \int \mathrm{d}\mathbf{R}\; e^{-\beta U(\mathbf{R})} \prod_{\alpha=1}^{n} \delta(q_\alpha(\mathbf{R}) - s_\alpha)$$

$$A(s_1,\ldots,s_n,T) = -k_B T \ln P(s_1,\ldots,s_n,T)$$

Write δ-functions as product of Gaussians:

$$\delta\left(q_\alpha(\mathbf{R}) - s_\alpha\right) = \lim_{\{\kappa_\alpha \to \infty\}} \left(\frac{\beta\kappa_\alpha}{2\pi}\right)^{1/2} \exp\left[-\frac{\beta\kappa_\alpha}{2}\left(q_\alpha(\mathbf{R}) - s_\alpha\right)^2\right]$$

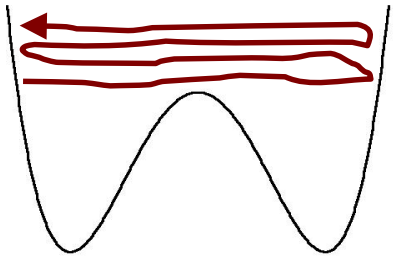
Dynamically sample Gaussian centers $s_1,\ldots,s_n$ by introducing a kinetic energy for them
And extending the phase space, which gives the following Hamiltonian:

$$\mathcal{H}(\mathbf{R},\mathbf{P},s,p_s) = H(\mathbf{R},\mathbf{P}) + \sum_{\alpha=1}^{n} \frac{p_{s_\alpha}^2}{2m_\alpha} + \sum_{\alpha=1}^{n} \frac{\kappa_\alpha}{2}\left(q_\alpha(\mathbf{R}) - s_\alpha\right)^2$$

Driven Adiabatic free energy (temperature-accelerated molecular) dynamics (d-AFED/TAMD)

Ensure "on the fly" sampling of the marginal: $m_\alpha \square m_i$

Accelerate sampling via temperature: $T_s \square T$ for extended variables



Adiabatically decoupled equations of motion:

$$M_i \ddot{\mathbf{R}}_i = -\frac{\partial U}{\partial \mathbf{R}_i} + \sum_\alpha \kappa_\alpha \left(s_\alpha - q_\alpha(\mathbf{R})\right)\frac{\partial q_\alpha}{\partial \mathbf{R}_i} + \text{heat bath}(T)$$

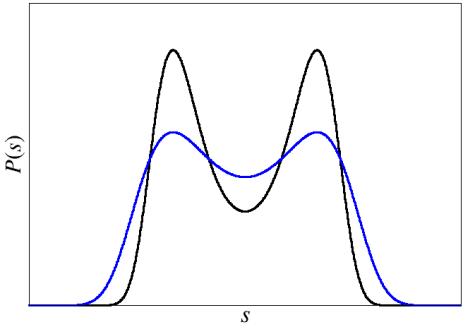
$$m_\alpha \ddot{s}_\alpha = -\kappa_\alpha \left(s_\alpha - q_\alpha(\mathbf{R})\right) + \text{heat bath}(T_s)$$

Under adiabatic conditions, we generate a distribution $P_{\text{adb}}^{(\{\kappa\})}(s_1,\ldots,s_n,T_s)$

$$\lim_{\{\kappa\to\infty\}} P_{\text{adb}}^{(\{\kappa\})}(s_1,\ldots,s_n,T_s,T) = e^{-\beta_s A(s_1,\ldots,s_n)} = \left[e^{-\beta A(s_1,\ldots,s_n)}\right]^{\beta_s/\beta} = \left[P(s_1,\ldots,s_n,T)\right]^{T/T_s}$$

$$\lim_{\{\kappa\to\infty\}} \left[\ln P_{\text{adb}}^{(\{\kappa\})}(s_1,\ldots,s_n,T_s,T)\right] = \frac{T}{T_s}\ln P(s_1,\ldots,s_n,T)$$

$$\lim_{\{\kappa\to\infty\}} \left[-k_B T_s \ln P_{\text{adb}}^{(\{\kappa\})}(s_1,\ldots,s_n,T_s,T)\right] = -k_B \cancel{T_s}\frac{T}{\cancel{T_s}}\ln P(s_1,\ldots,s_n,T) = A(s_1,\ldots,s_n,T)$$



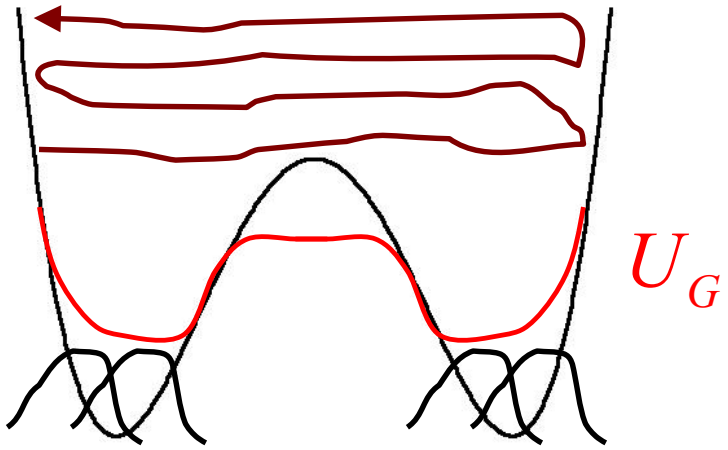
Combining these methods

M. Chen, M. Cuendet, and MET *J. Chem. Phys.* **137**, 024102 (2012); S. Paul, N. N. Nair, H. Vashisth *Mol. Simulat.* **45**, 1273 (2019)

Let $s_1, \tilde{\mathbf{s}} \in \mathbf{s}$

Apply metadynamics-like bias in the $\tilde{\mathbf{s}}$ subspace:

$$U_G(\tilde{\mathbf{s}},t) = \sum_i h e^{-\left\|\tilde{\mathbf{s}}-\tilde{\mathbf{s}}_G(t_i)\right\|/2\sigma^2}$$



Hamiltonian: ("UFED" method)

$$\mathcal{H}(\mathbf{R},\mathbf{P},\mathbf{s},\mathbf{p}_s) = H(\mathbf{R},\mathbf{P}) + \sum_{\alpha=1}^{n}\frac{p_{s_\alpha}^2}{2m_\alpha} + \sum_{\alpha=1}^{n}\frac{\kappa_\alpha}{2}\left(q_\alpha(\mathbf{R})-s_\alpha\right)^2 + U_G(\tilde{\mathbf{s}},t)$$

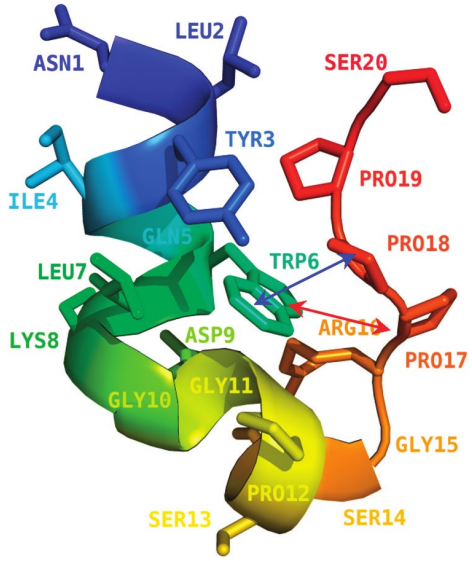
Hamiltonian: ("TASS" method)

$$\mathcal{H}(\mathbf{R},\mathbf{P},\mathbf{s},\mathbf{p}_s) = H(\mathbf{R},\mathbf{P}) + \sum_{\alpha=1}^{n}\frac{p_{s_\alpha}^2}{2m_\alpha} + \sum_{\alpha=1}^{n}\frac{\kappa_\alpha}{2}\left(q_\alpha(\mathbf{R})-s_\alpha\right)^2 + U_G(\tilde{\mathbf{s}},t) + \frac{1}{2}k_{ub}\left(s_1-\sigma\right)^2$$

All now available in OpenMM! https://ufedmm.readthedocs.io/en/latest/ Publication: Bajpai et al. *J. Comp. Chem.* (submitted)

20-residue miniprotein:

Asn-Leu-Tyr-Ile-Gln-Trp-Leu
-Lys-Asp-Gly-Gly-Pro-Ser-Ser
-Gly-Arg-Pro-Pro-Pro-Ser



Simulated using TASS with 8 CVs:

UFED: RMSD of the alpha carbons.

UFED: R_g of alpha carbons.

D-AFED/TAMD coordinates:

1. RMSD of helical residues (2-8).
2. RMSD of hydrophobic core (6, 17-19).
3. Salt-bridge distance between Asp9-Arg16.
4. End-to-end distance.
5. Dihedral angle correlation

$$q(\mathbf{R}) = \sum_i \left[1 + \cos\left(\phi_i - \psi_i\right)\right]$$

6. α-helical similarity:

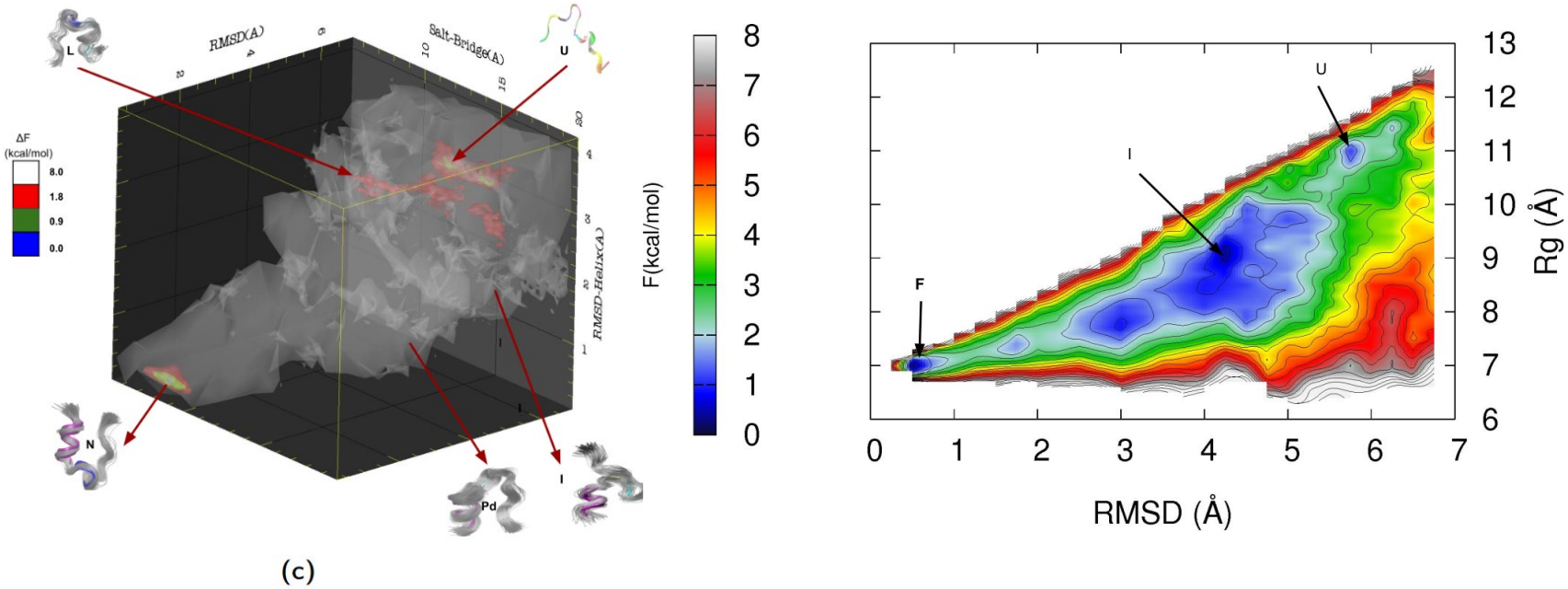
$$q(\mathbf{R}) = \sum_i \left[1 + \cos\left(\phi_i - \phi_i^{ref}\right)\right]$$

CV set from Juraszek and Bolhuis *PNAS* (2006), *Biophys J.* (2008)

Folding free energy of the Trp-cage in water

20-residue miniprotein:

Asn-Leu-Tyr-Ile-Gln-Trp-Leu
-Lys-Asp-Gly-Gly-Pro-Ser-Ser
-Gly-Arg-Pro-Pro-Pro-Ser



(c)

Simulated using TASS with 8 CVs:

UB(25): RMSD of the alpha carbons

UFED: R_g of alpha carbons.

D-AFED/TAMD coordinates:

1. RMSD of helical residues (2-8).
2. RMSD of hydrophobic core (6, 17-19).
3. Salt-bridge distance between Asp9-Arg16.
4. End-to-end distance.
5. Dihedral angle correlation

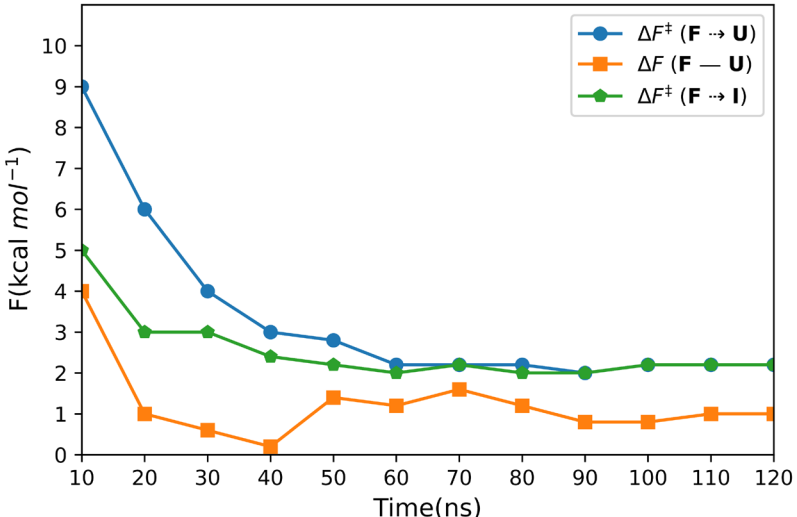
$$q(\mathbf{R}) = \sum_i \left[1 + \cos\left(\phi_i - \psi_i\right)\right]$$

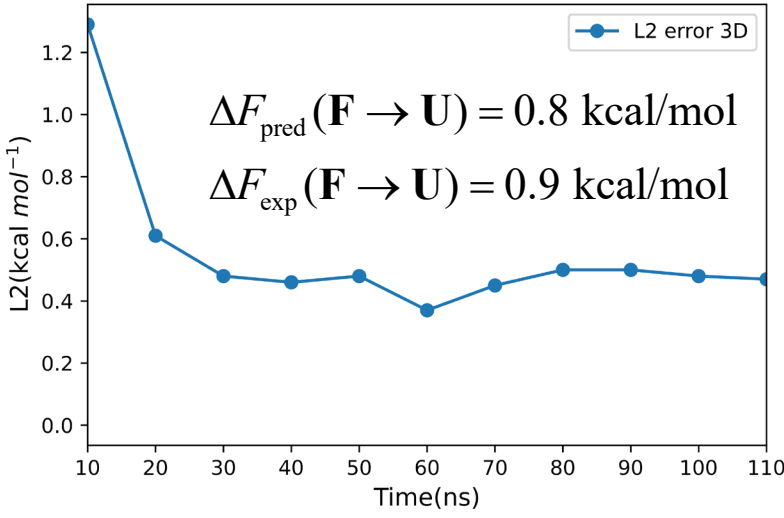
6. α-helical similarity:

$$q(\mathbf{R}) = \sum_i \left[1 + \cos\left(\phi_i - \phi_i^{ref}\right)\right]$$

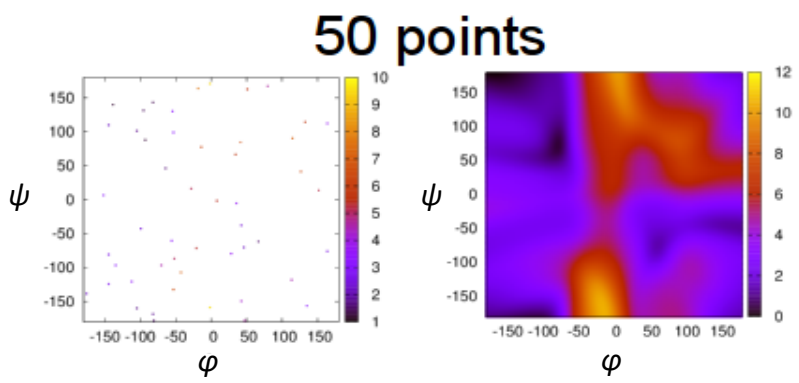
S. Bajpai, C. R. A. Abreu, B. Petkov, N. Nair, MET (in prep)

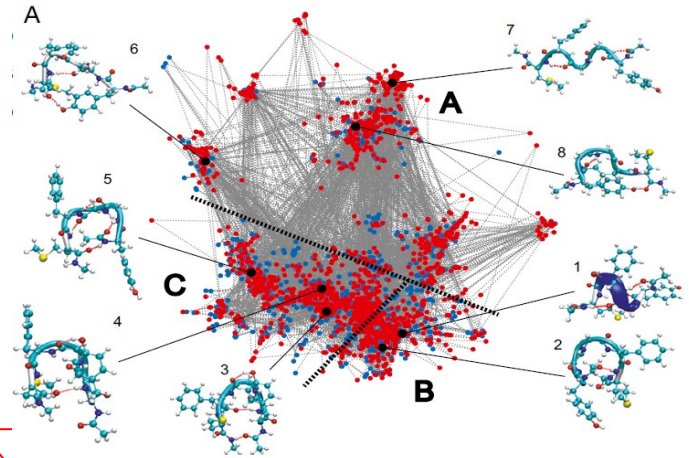
Exp: Streicher & Makhatadze *Biochem* (2007)

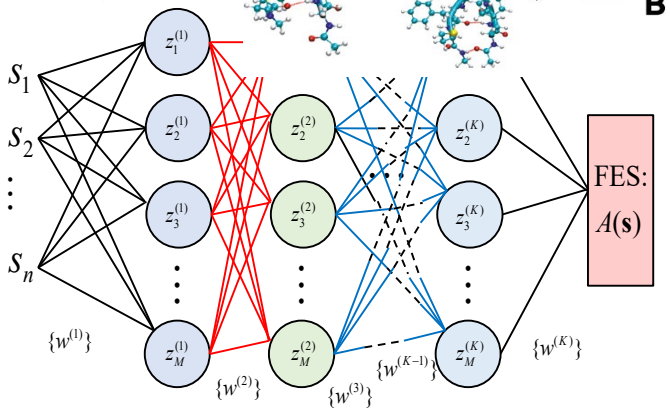


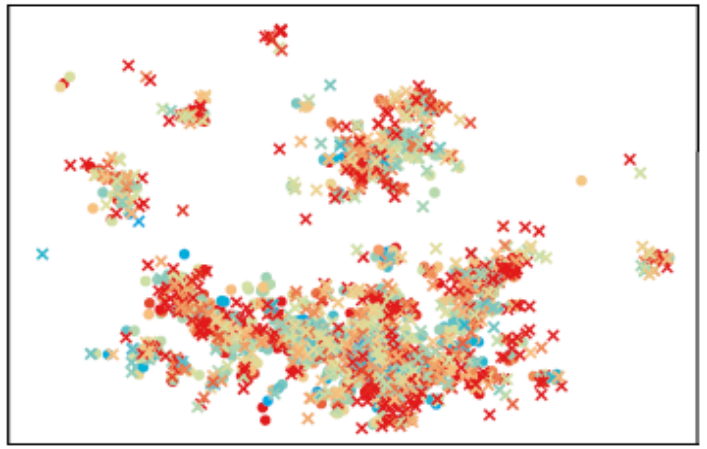


“Emergence”: Training neural networks to represents FESs



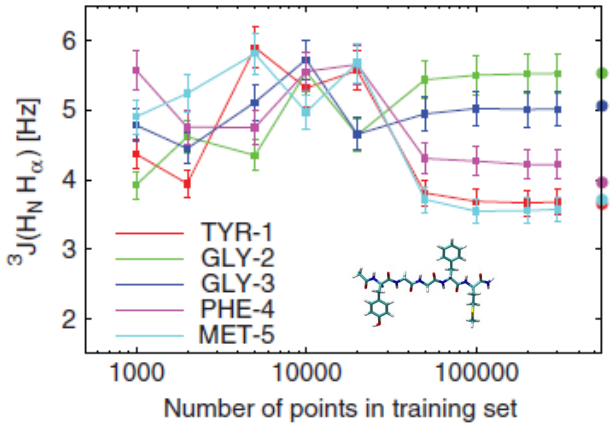






$$\langle O(\mathbf{r}) \rangle = \frac{\int d^n s \, \langle O(\mathbf{r}) \rangle_{\mathbf{r}} (s) \, e^{-\beta A_{\mathrm{ML}}(s)}}{\int d^n s \, e^{-\beta A_{\mathrm{ML}}(s)}}$$

From 10-dimensional FES

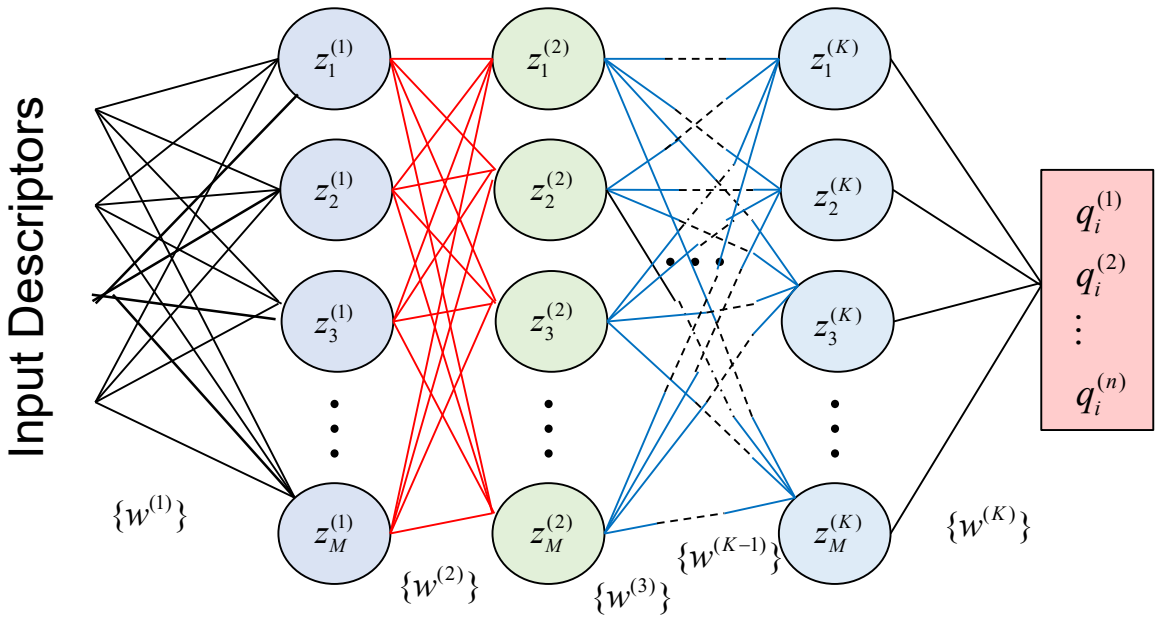


The trained networks can be used to compute observables

Finding optimal reaction coordinates aided machine learning

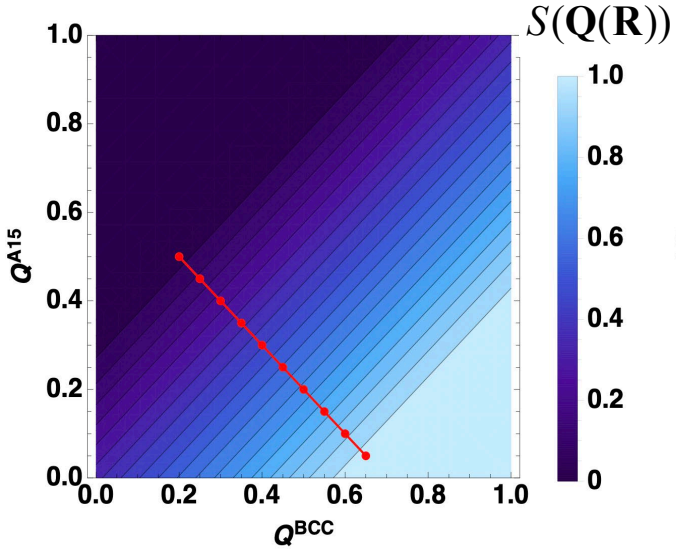
Paths in classification space:

J. Rogal *et al*. *Phys. Rev. Lett*. **123** 245701 (2019)



$$\mathbf{Q}(\mathbf{R}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{q}_i(\mathbf{R})$$

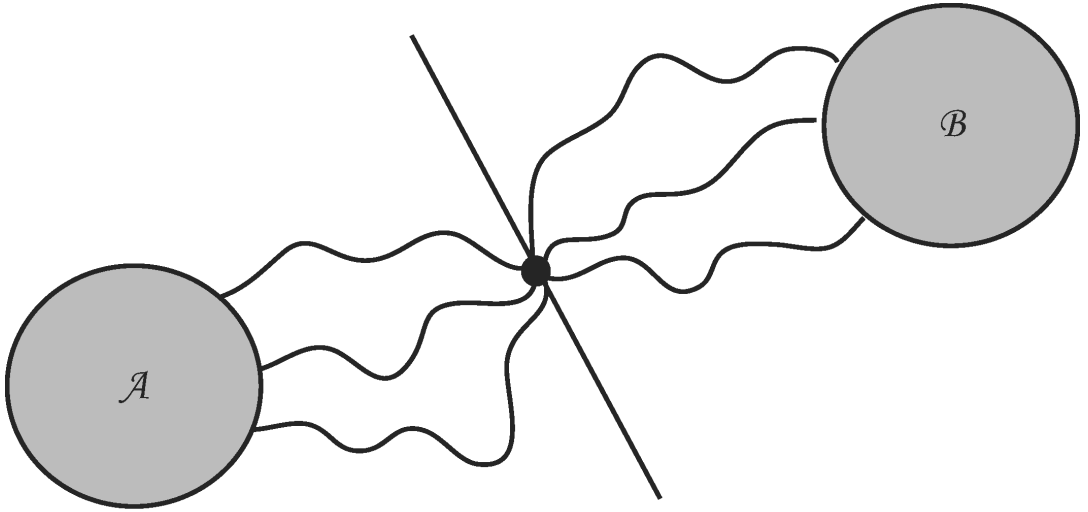
Path CVs in classification space.



Committor learning:

$p_{\mathcal{B}}(\mathbf{R})$ is the probability of a trajectory initiated at $\mathbf{R}$ reaching $\mathcal{B}$ before $\mathcal{A}$.

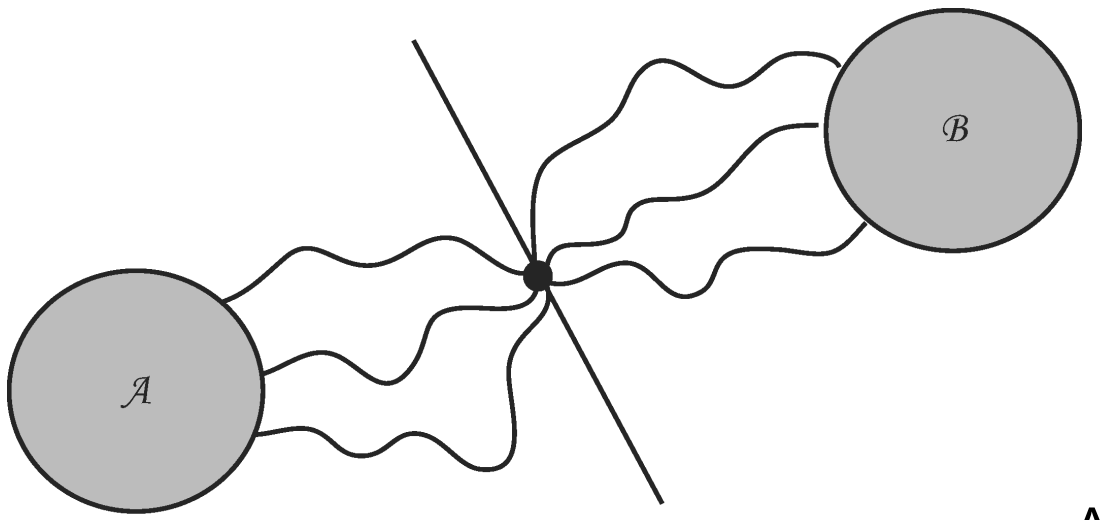
$p_{\mathcal{B}}(\mathbf{R})$ is an ideal reaction coordinate for $\mathcal{A}$ to $\mathcal{B}$ transition.



$p_{\mathcal{B}} = 1/2$

Isocommittor surface

Ma and Dinner (2005), Peters and Trout (2006), Peters et al. (2007), Lechner et al. (2010), Jung et al. (2019), Mori et al. (2020), Jung et al. (2021), Kikutsuji et al. (2022),



Suppose we have a large set of putative CVs $q_\alpha(\mathbf{R})$. Represent $p_{\mathcal{B}}(\mathbf{R})$ as a linear combination of these:

$$p_{\mathcal{B}}(\mathbf{R};\mathbf{w}) = \sum_{\alpha=1}^{n} w_\alpha q_\alpha(\mathbf{R}) + w_0$$

(Method 1)

Alternatively, $p_{\mathcal{B}}(\mathbf{R})$ can be approximated via a function $\pi_{\mathcal{B}}(r(\mathbf{R}))$, where $r(\mathbf{R})$ is a putative reaction coordinate that leads to sigmoidal behavior of $\pi_{\mathcal{B}}(r(\mathbf{R}))$. A model form for $\pi_{\mathcal{B}}(r(\mathbf{R}))$ is

$$\pi_{\mathcal{B}}(r(\mathbf{R})) = \frac{1}{1+e^{-r(\mathbf{R})}}$$

Represent $r(\mathbf{R})$ as a linear combination of the CVs:

$$r(\mathbf{R};\mathbf{w}) = \sum_{\alpha=1}^{n} w_\alpha q_\alpha(\mathbf{R}) + w_0$$

(Method 2)

Learning models for finding an optimal reaction coordinate: Assume N training values of $p_B^*(\mathbf{R}_k)$

- Binary classification (cross-entropy) model (CREM):

$$E(\mathbf{w}) = -\sum_{k=1}^{N} p_{\mathcal{B}}^*(\mathbf{R}_k) \ln p_{\mathcal{B}}(\mathbf{R}_k, \mathbf{w}) - \sum_{k=1}^{N} \left(1 - p_{\mathcal{B}}^*(\mathbf{R}_k)\right) \ln\left[1 - p_{\mathcal{B}}(\mathbf{R}_k, \mathbf{w})\right] + \mathrm{Re}\, g(\mathbf{w}) \qquad \textbf{(Method 1)}$$

$$E(\mathbf{w}) = -\sum_{k=1}^{N} p_B^*(\mathbf{R}_k) \ln \pi_B(r(\mathbf{R}_k, \mathbf{w})) - \sum_{k=1}^{N} \left(1 - p_B^*(\mathbf{R}_k)\right) \ln\left[1 - \pi_B(r(\mathbf{R}_k, \mathbf{w}))\right] + \mathrm{Re}\, g(\mathbf{w}) \qquad \textbf{(Method 2)}$$

- Standard regression model:

$$E(\mathbf{w}) = \frac{1}{2N} \sum_{k=1}^{N} \left| p_{\mathcal{B}}(\mathbf{R}, \mathbf{w}) - p_{\mathcal{B}}^*(\mathbf{R}_k) \right|^2 + \mathrm{Re}\, g(\mathbf{w}) \qquad \textbf{(Method 1)}$$

$$E(\mathbf{w}) = \frac{1}{2N} \sum_{k=1}^{N} \left| \pi_{\mathcal{B}}(r_k(\mathbf{R}, \mathbf{w})) - p_{\mathcal{B}}^*(\mathbf{R}_k) \right|^2 + \mathrm{Re}\, g(\mathbf{w}) \qquad \textbf{(Method 2)}$$

$$\mathrm{Reg}(\mathbf{w}) = \begin{cases} \lambda \|\mathbf{w}\| & \text{Lasso (LASR)} \\ \lambda\rho \|\mathbf{w}\| + \beta(1-\rho) \|\mathbf{w}\|^2 & \text{Elastic Net (ELAN)} \end{cases}$$

Learning models

List of standard regressors:

- Lasso (LASR)
- Elastic Net (ELAN)
- Lasso Least Angle (LLAR)
- Huber (HUBR)
- Linear (LINR)
- Ridge (RIDR)
- Bayeian (BAYR)
- Orthogonal Mathing Pursuit (ORMP)
- Passive-Aggressive (PAGR)

Neighbors/trees:

- Decision Tree (DECT)
- Random Forest (RAFR)
- Extra Trees (EXTR)

Splitting rule based on variance reduction:

$$\sigma^2 = \frac{\sum_{k=1}^{n}\left(p^*_{\mathcal{B}}(\mathbf{R}_k) - \bar{p}_{\mathcal{B}}\right)^2}{n}$$

Boosters:

- AdaBoost (ADAB)
- Gradient Boost (GRBR)
- Light Gradient Boost (LGBM)

Boosting:

$$\pi_B^{(0)}(\mathbf{R}) = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

for $m = 1$ to M

$$\bar{p}_{B,k} = p_{B,k} - \pi_B^{(m-1)}(\mathbf{R}_k)$$

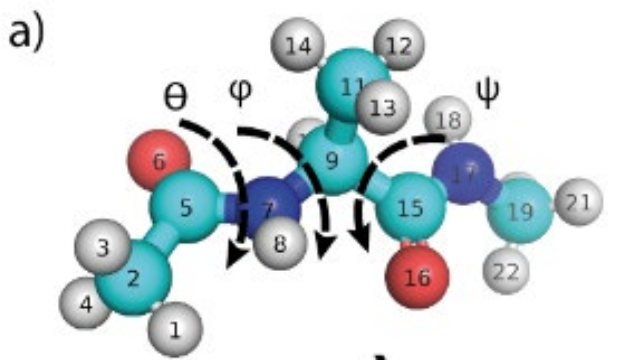
$$(\mathbf{w}_m, \beta_m) = \arg\min_{\beta,\mathbf{w}} \sum_{k=1}^{N} (\bar{p}_{\mathcal{B},k} - \pi_{\mathcal{B},k}^{(m-1)} - \beta h(\mathbf{R}_k; \mathbf{w}))^2$$

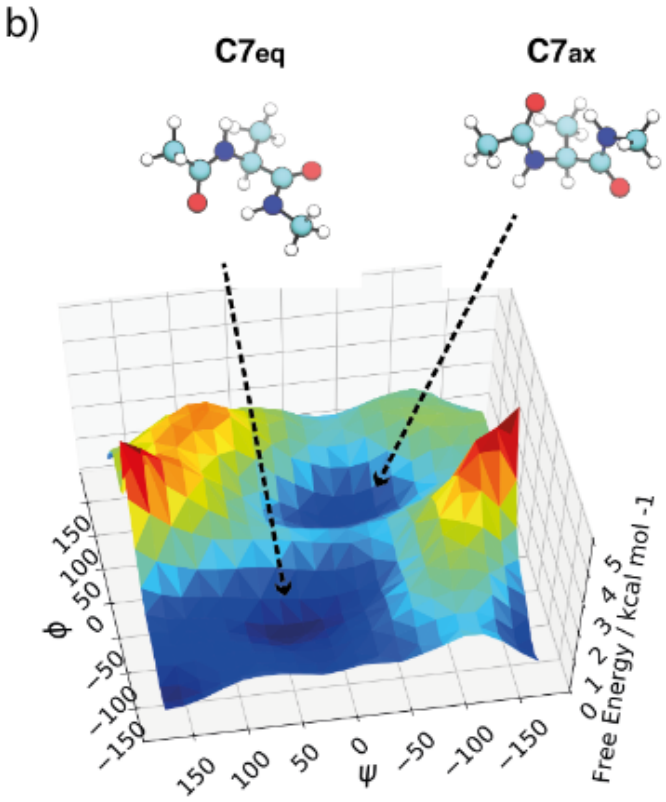
$$\pi_B^{(m)}(\mathbf{R}) = \pi_B^{(m-1)}(\mathbf{R}) + l_r \beta_m h(\mathbf{R}; \mathbf{w}_m)$$

endfor

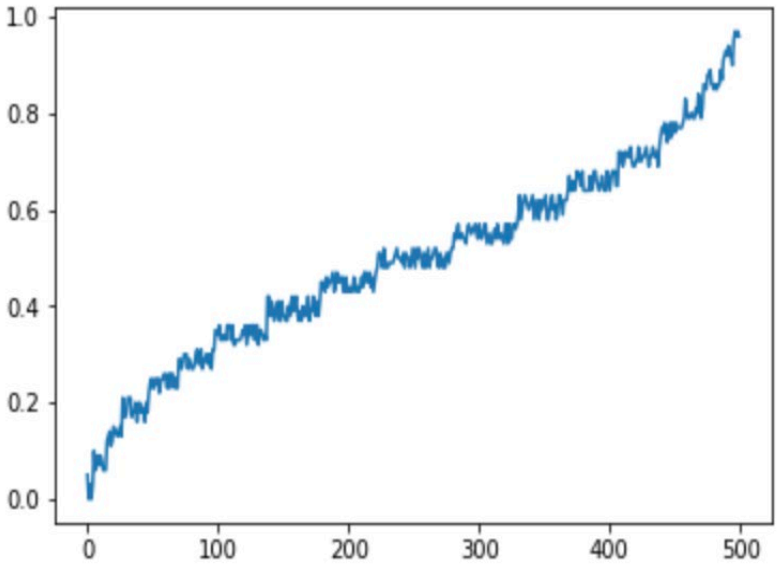
Reaction coordinate of C7eq -> C7ax conformational transition in Ala dipeptide in vacuum

Alanine dipeptide



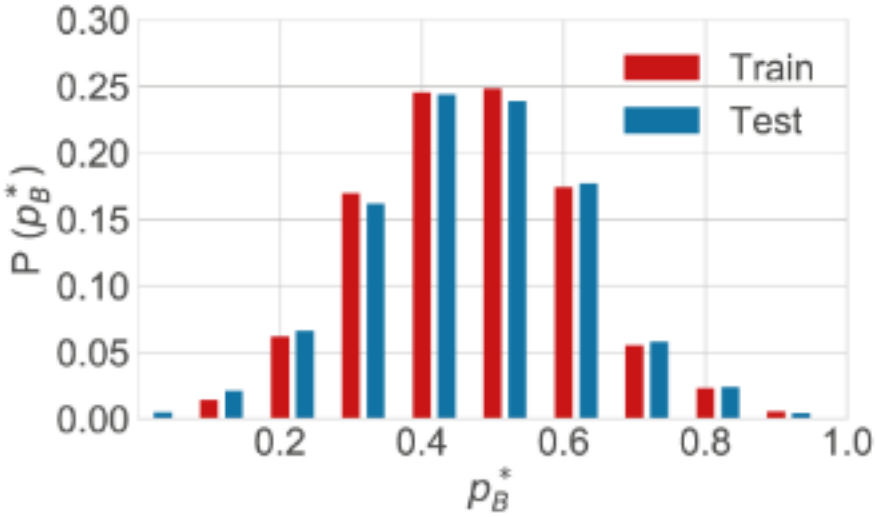


$p_B^*(\mathbf{r}_k)$



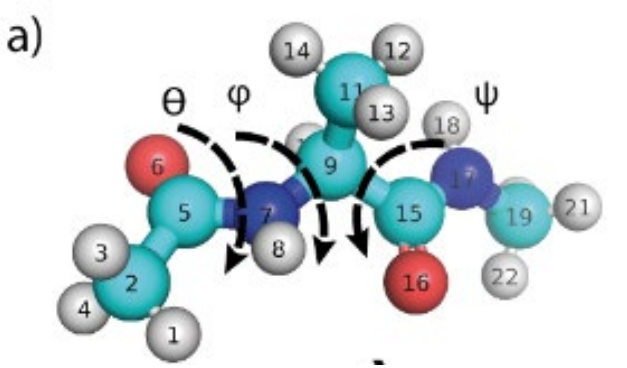
Training data point index

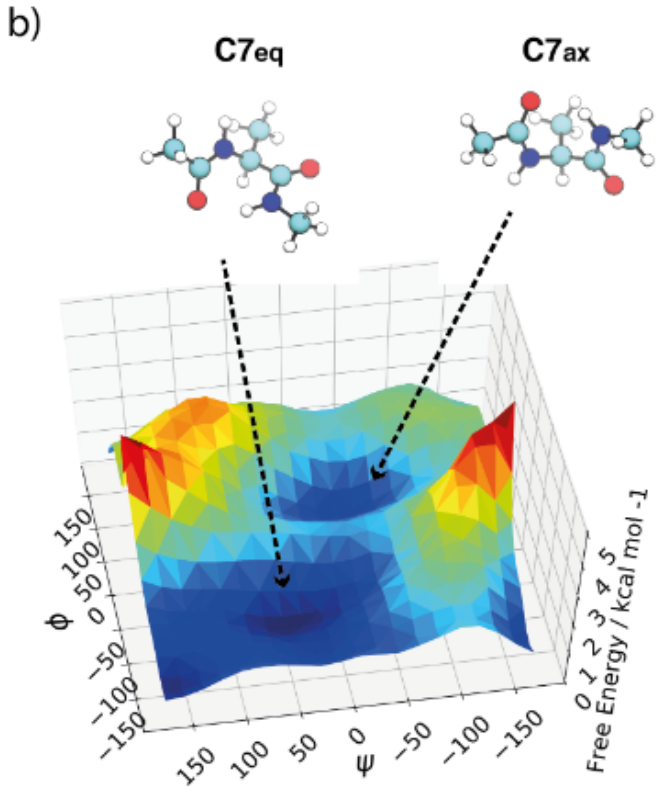
Committor values sampled using aimless shooting algorithm.



Reaction coordinate of C7eq -> C7ax conformational transition in Ala dipeptide

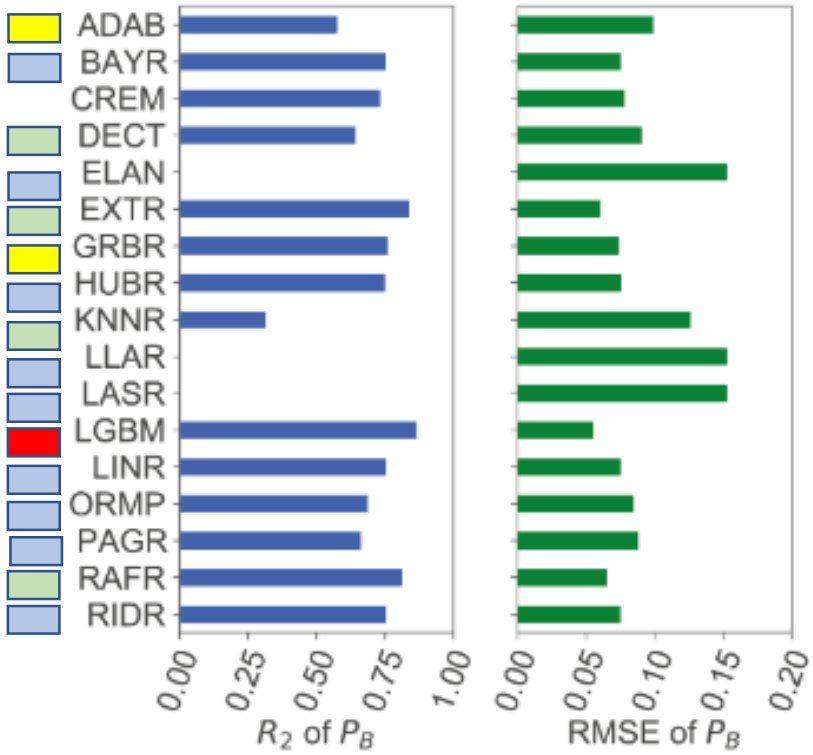
Alanine dipeptide

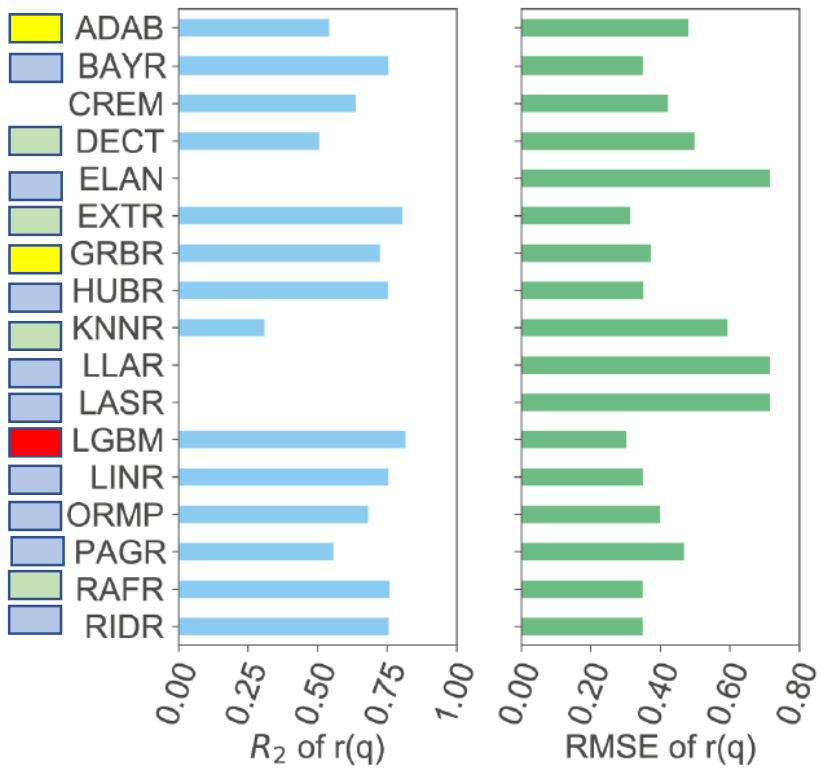




$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{45} w_{i-1} \cos q_i + \sum_{i=46}^{90} w_{i-1} \sin q_i$$

Enhanced sampling of (φ, ψ)





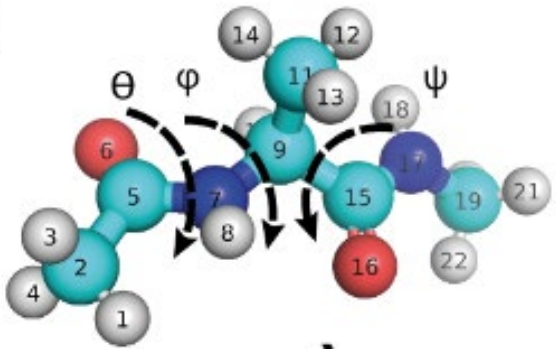
Method 1

Method 2

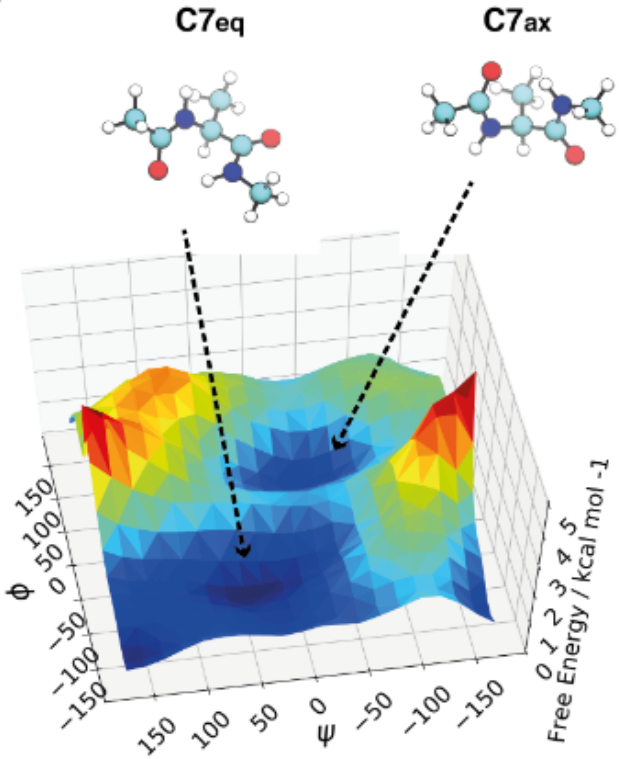
Reaction coordinate of C7eq -> C7ax conformational transition in Ala dipeptide

Alanine dipeptide

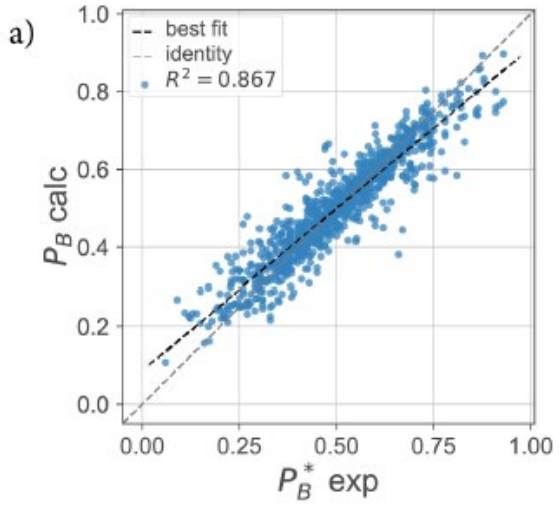
a)

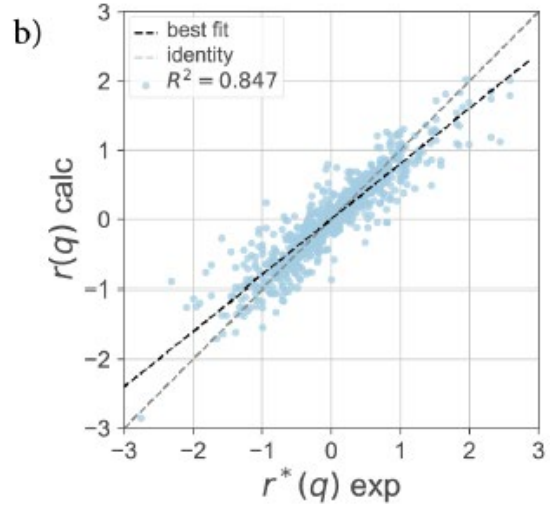


b)

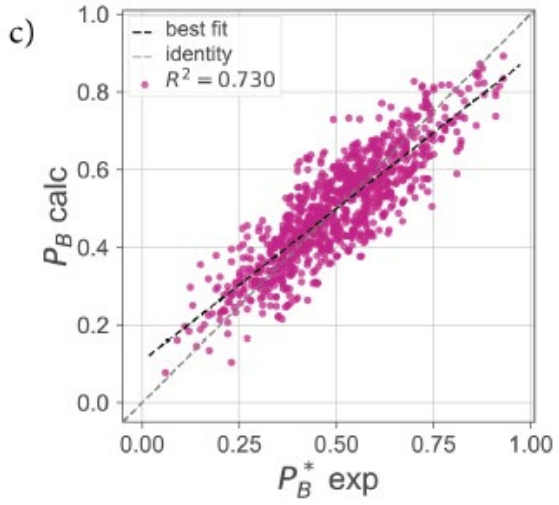


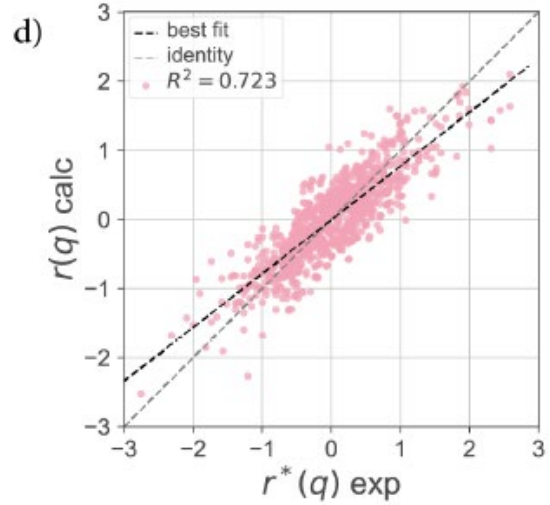
$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{45} w_{i-1} \cos q_i + \sum_{i=46}^{90} w_{i-1} \sin q_i$$





LGBM



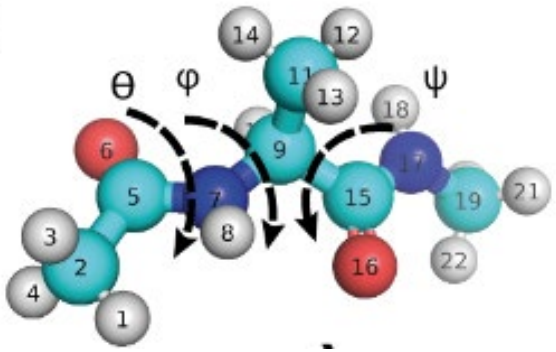


CREM

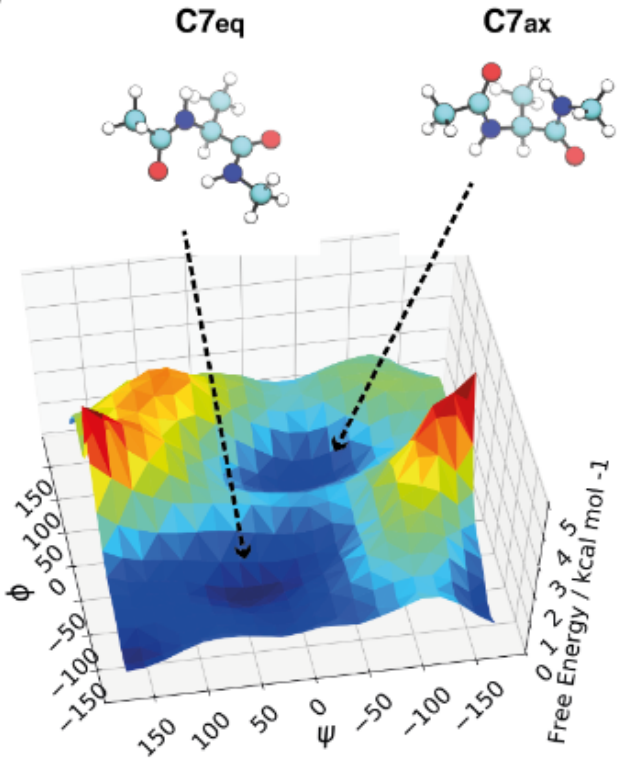
Feature extraction forC7eq -> C7ax conformational transition in Ala dipeptide

Alanine dipeptide

a)



b)

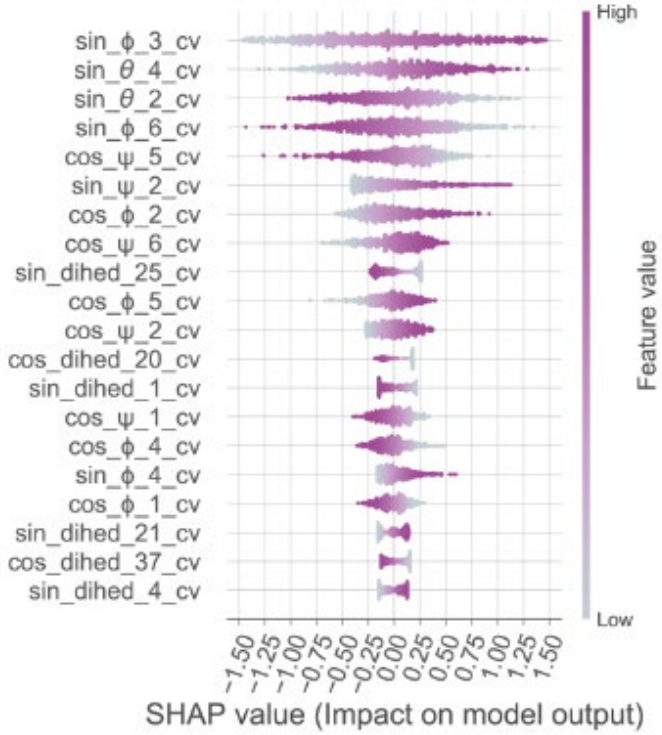


$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{45} w_{i-1} \cos q_i + \sum_{i=46}^{90} w_{i-1} \sin q_i$$

SHAP (SHapley Additive ePlanation – L. S. Shapley in *Contributions to the Theory of Games* (2016)).

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \left(\begin{matrix} |F| \\ |S| \end{matrix} \right)^{-1} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

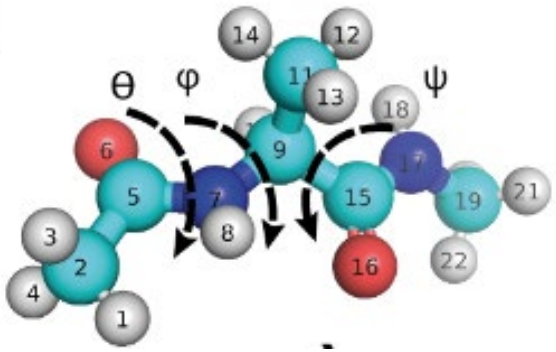
Each dot in the SHAP plot is an observation in our data set. "High" and "low" are high and low values of the features. Positive and negative values describe whether high or low values of the feature push the predictions toward one basin or the other (0 or 1 commitor probability values).



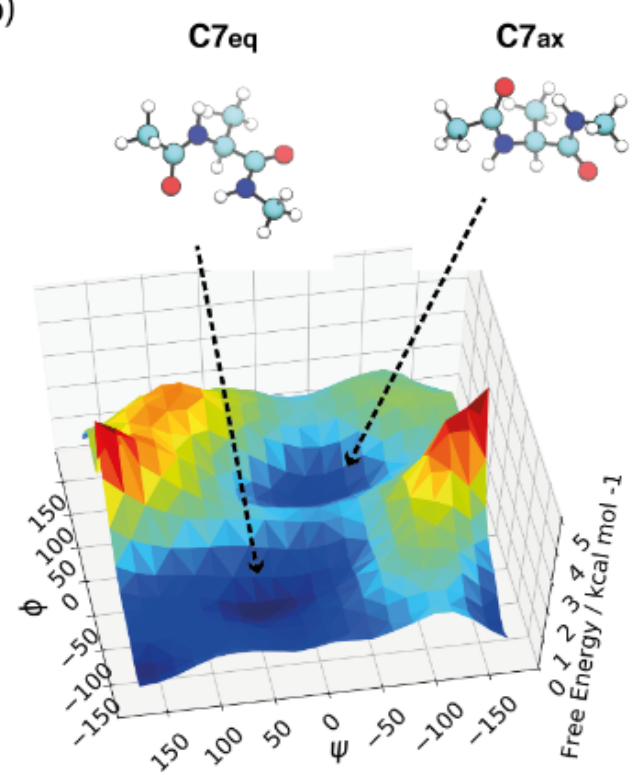
Feature extraction forC7eq -> C7ax conformational transition in Ala dipeptide

Alanine dipeptide

a)



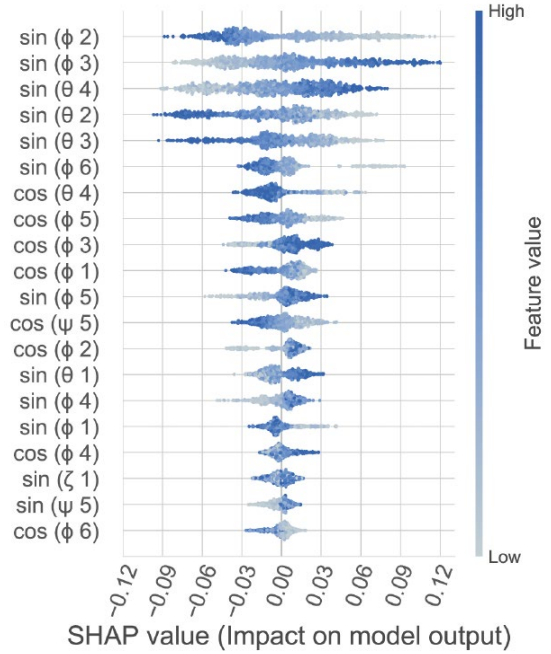
b)



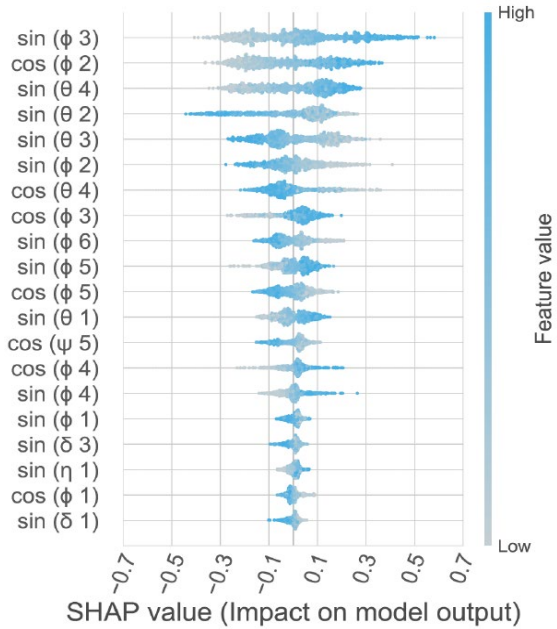
$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{45} w_{i-1} \cos q_i + \sum_{i=46}^{90} w_{i-1} \sin q_i$$

Dihedral Index	Abbriv.	1	2	3	4	Dihedral Index	Abbriv.	1	2	3	4
1	β_1	1	2	5	6	24	γ_8	15	9	11	13
2	β_2	1	2	5	7	25	γ_9	15	9	11	14
3	β_3	3	2	5	6	26	ψ_1	7	9	15	16
4	β_4	3	2	5	7	27	ψ_2	7	9	15	17
5	β_5	4	2	5	6	28	ψ_3	10	9	15	16
6	β_6	4	2	5	7	29	ψ_4	10	9	15	17
7	θ_1	2	5	7	8	30	ψ_5	11	9	15	16
8	θ_2	2	5	7	9	31	ψ_6	11	9	15	17
9	θ_3	6	5	7	8	32	δ_1	9	15	17	18
10	θ_4	6	5	7	9	33	δ_2	9	15	17	19
11	ϕ_1	5	7	9	10	34	δ_3	16	15	17	18
12	ϕ_2	5	7	9	11	35	δ_4	16	15	17	19
13	ϕ_3	5	7	9	15	36	ϵ_1	15	17	19	20
14	ϕ_4	8	7	9	10	37	ϵ_2	15	17	19	21
15	ϕ_5	8	7	9	11	38	ϵ_3	15	17	19	22
16	ϕ_6	8	7	9	15	39	ϵ_4	18	17	19	20
17	γ_1	7	9	11	12	40	ϵ_5	18	17	19	21
18	γ_2	7	9	11	13	41	ϵ_6	18	17	19	22
19	γ_3	7	9	11	14	42	ζ_1	2	7	5	6
20	γ_4	10	9	11	12	43	η_1	5	9	7	8
21	γ_5	10	9	11	13	44	κ_1	9	17	15	16
22	γ_6	10	9	11	14	45	λ_1	15	19	17	18
23	γ_7	15	9	11	12						

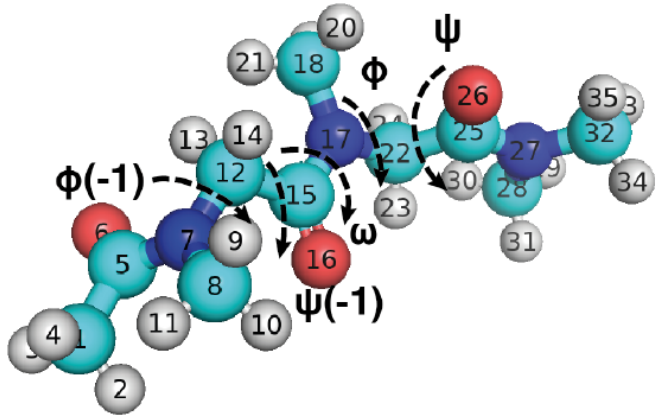
Meth 1



Meth 2



Reaction coordinate of conformational transition of disarcosine in GB solvent



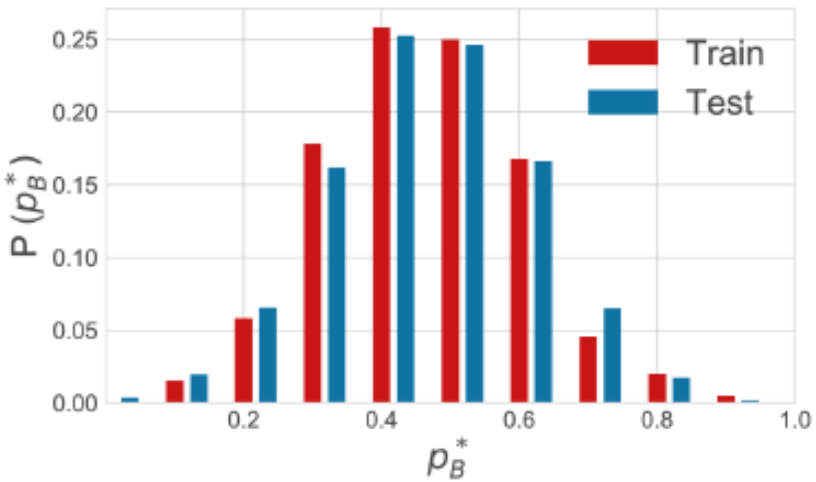
X-15-17-Y → ω

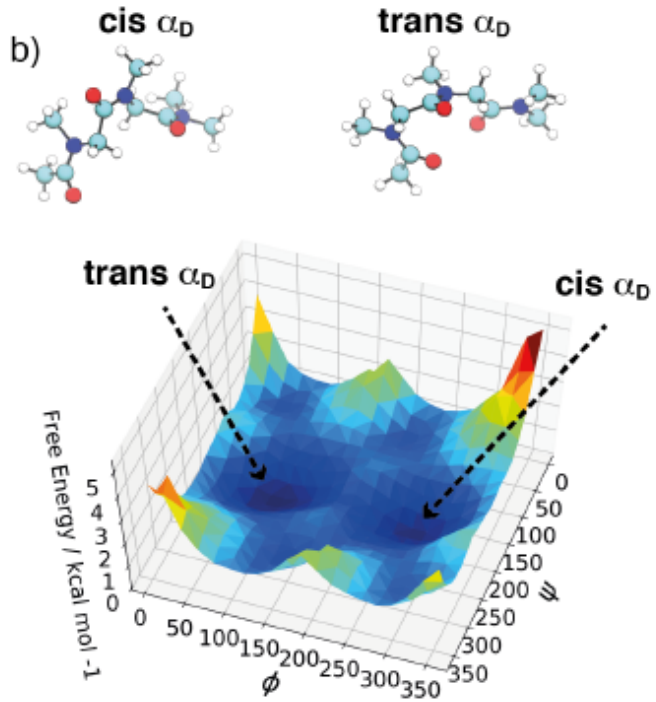
X-17-22-Y → φ

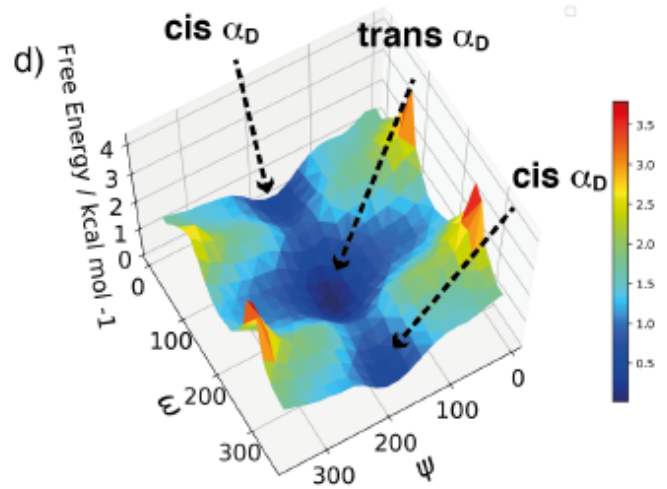
X-22-25-Y → ψ

X-7-12-Y → φ(-1)

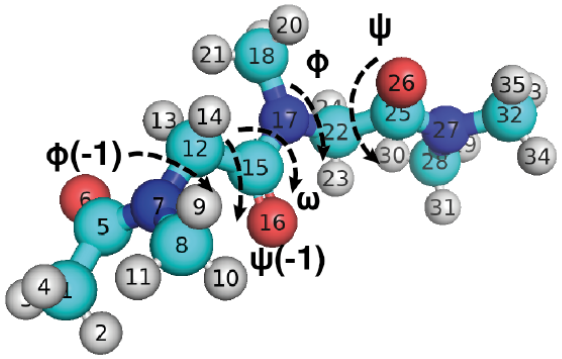
X-12-15-Y → ψ(-1)







Reaction coordinate of trans->cis conformational transition in disarcosine in GB solvent



X-15-17-Y → ω

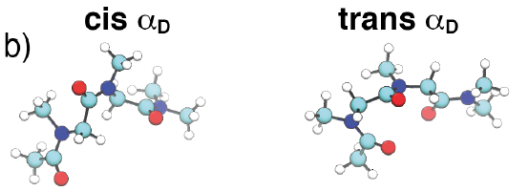
X-17-22-Y → φ

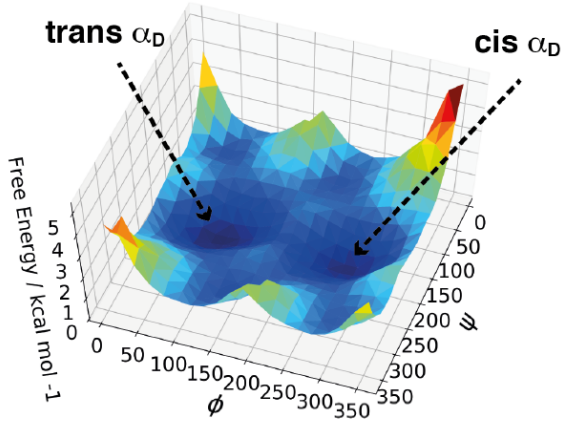
X-22-25-Y → ψ

X-7-12-Y → φ(-1)

X-12-15-Y → ψ(-1)

b)

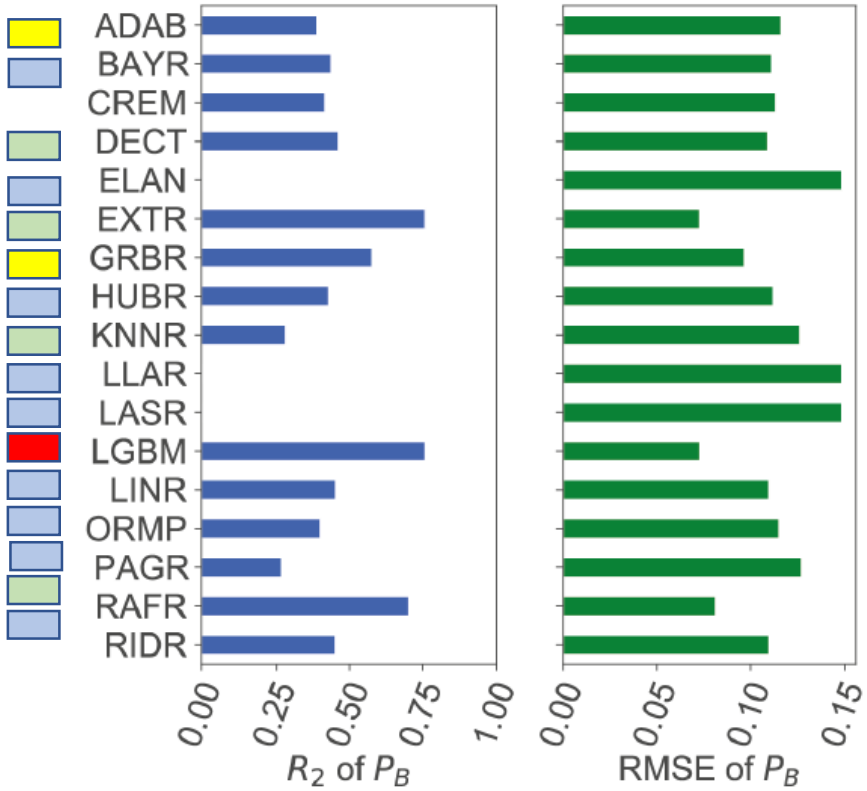




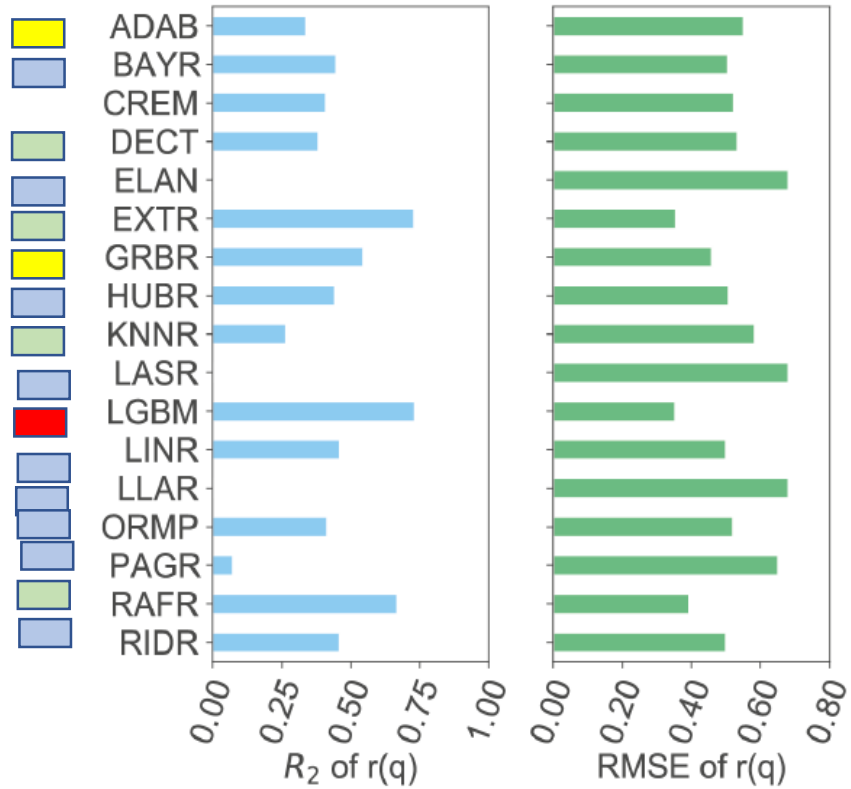
$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{66} w_{i-1}\cos q_i + \sum_{i=67}^{132} w_{i-1}\sin q_i$$

Enhanced sampling of

$$(\varphi_2, \varphi(-1)_2, \psi_1, \psi(-1)_1, \omega_1)$$

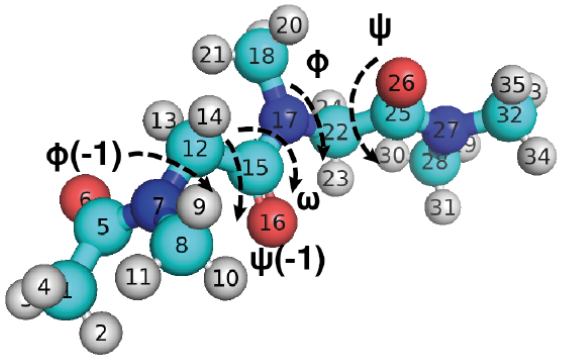


Method 1



Method 2

Reaction coordinate of trans->cis conformational transition in disarcosine



$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{66} w_{i-1}\cos q_i + \sum_{i=67}^{132} w_{i-1}\sin q_i$$

Enhanced sampling of

$$(\varphi_2, \varphi(-1)_2, \psi_1, \psi(-1)_1, \omega_1)$$

LGBM

X-15-17-Y → ω

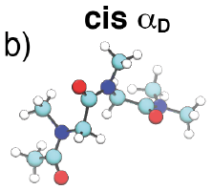
X-17-22-Y → φ

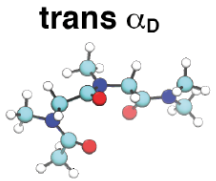
X-22-25-Y → ψ

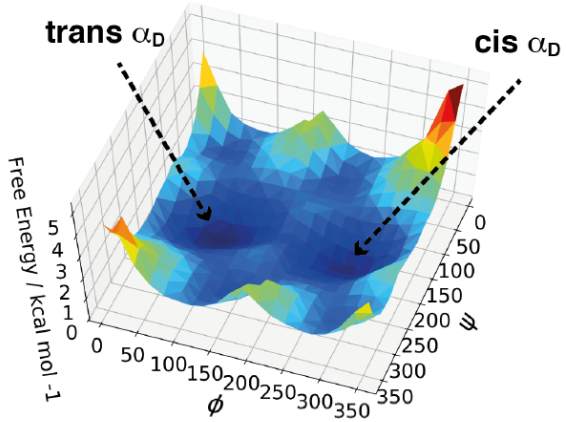
X-7-12-Y → φ(-1)

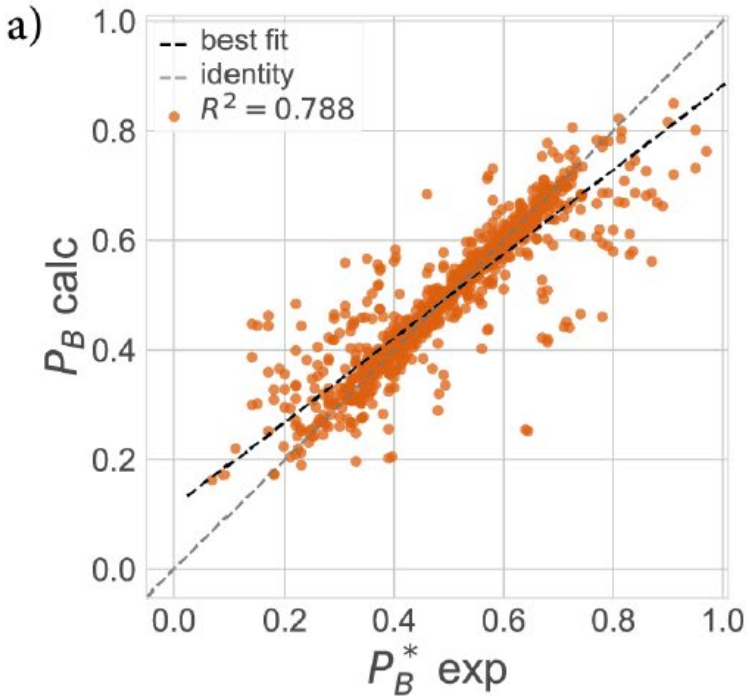
X-12-15-Y → ψ(-1)

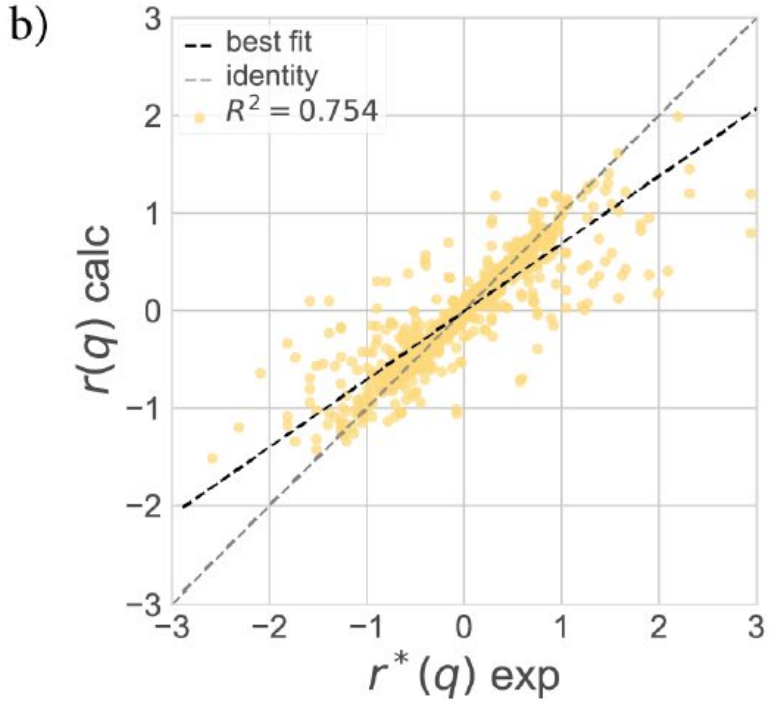
b)



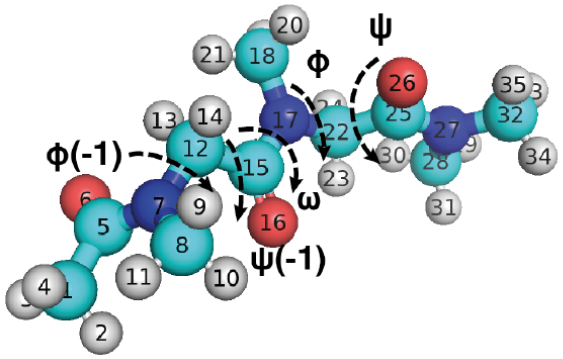








Reaction coordinate of trans->cis conformational transition in disarcosine



X-15-17-Y → ω

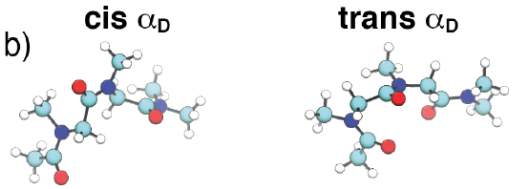
X-17-22-Y → φ

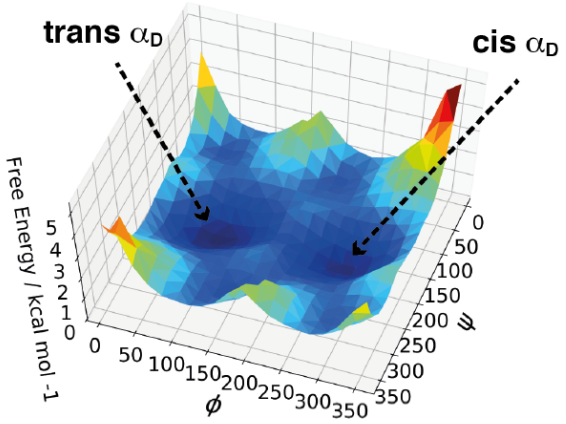
X-22-25-Y → ψ

X-7-12-Y → φ(-1)

X-12-15-Y → ψ(-1)

b)

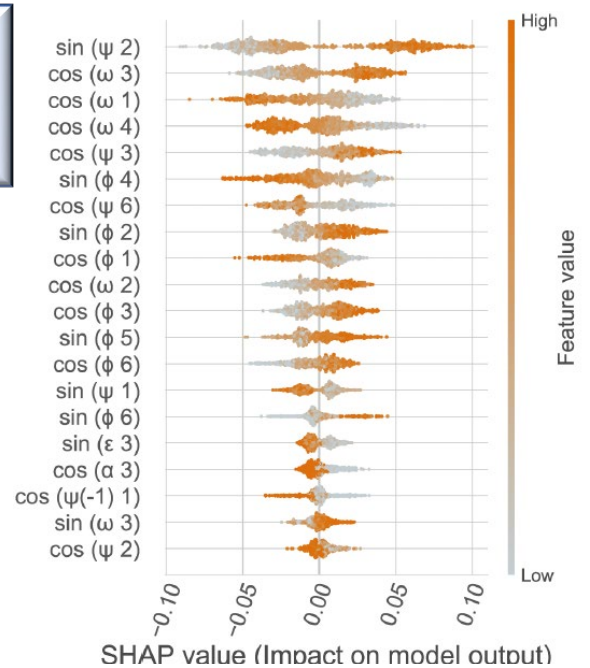




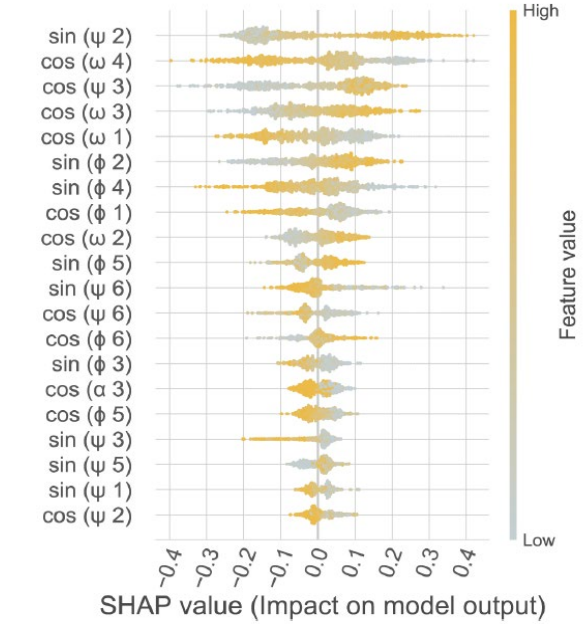
$$\left.\begin{matrix} r(\mathbf{q},\mathbf{w}) \\ p_{\mathcal{B}}(\mathbf{q},\mathbf{w}) \end{matrix}\right\} = b + \sum_{i=1}^{66} w_{i-1} \cos q_i + \sum_{i=67}^{132} w_{i-1} \sin q_i$$

Dihedral Index	Abbriv.	1	2	3	4	Dihedral Index	Abbriv.	1	2	3	4
1	α_1	1	5	7	12	34	ϕ_2	15	17	22	25
2	α_2	1	5	7	8	35	ϵ_3	15	17	18	21
3	β_1	2	1	5	7	36	ϕ_3	15	17	22	24
4	β_2	2	1	5	6	37	ω_3	16	15	17	22
5	β_3	3	1	5	7	38	ω_4	16	15	17	18
6	β_4	3	1	5	6	39	ψ_1	17	22	25	27
7	β_5	4	1	5	7	40	ψ_2	17	22	25	26
8	β_6	4	1	5	6	41	ϕ_4	18	17	22	25
9	γ_1	5	7	8	10	42	ϕ_5	18	17	22	24
10	$\phi(-1)_1$	5	7	12	13	43	ϕ_6	18	17	22	23
11	γ_2	5	7	8	9	44	ζ_1	19	18	17	22
12	$\phi(-1)_2$	5	7	12	15	45	ζ_2	20	18	17	22
13	γ_3	5	7	8	11	46	ζ_3	21	18	17	22
14	$\phi(-1)_3$	5	7	12	14	47	η_1	22	25	27	28
15	α_3	6	5	7	12	48	η_2	22	25	27	32
16	α_4	6	5	7	8	49	ψ_3	23	22	25	26
17	$\psi(-1)_1$	7	12	15	17	50	ψ_4	23	22	25	27
18	$\psi(-1)_2$	7	12	15	16	51	ψ_5	24	22	25	27
19	$\phi(-1)_4$	8	7	12	15	52	ψ_6	24	22	25	26
20	$\phi(-1)_5$	8	7	12	14	53	κ_1	25	27	32	33
21	$\phi(-1)_6$	8	7	12	13	54	λ_1	25	27	28	29
22	δ_1	9	8	7	12	55	κ_2	25	27	32	35
23	δ_2	10	8	7	12	56	λ_2	25	27	28	31
24	δ_3	11	8	7	12	57	κ_3	25	27	32	34
25	ω_1	12	15	17	22	58	λ_3	25	27	28	30
26	ω_2	12	15	17	18	59	η_3	26	25	27	32
27	$\psi(-1)_3$	13	12	15	17	60	η_4	26	25	27	28
28	$\psi(-1)_4$	13	12	15	16	61	κ_4	28	27	32	35
29	$\psi(-1)_5$	14	12	15	17	62	κ_5	28	27	32	34
30	$\psi(-1)_6$	14	12	15	16	63	κ_6	28	27	32	33
31	ϵ_1	15	17	18	20	64	μ_1	29	28	27	32
32	ϕ_1	15	17	22	23	65	μ_2	30	28	27	32
33	ϵ_2	15	17	18	19	66	μ_3	31	28	27	32

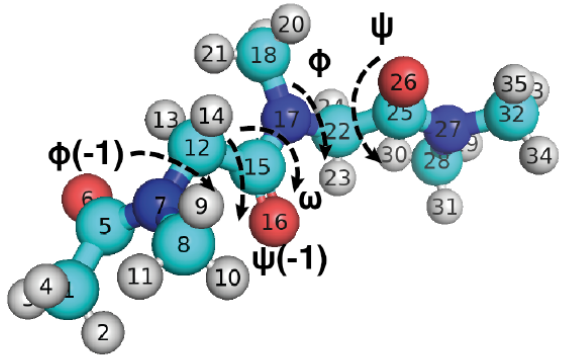
Meth 1



Meth 2



Mechanism of trans->cis conformational transition in disarcosine



X-15-17-Y → ω

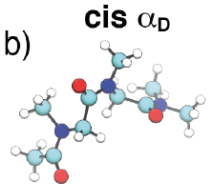
X-17-22-Y → φ

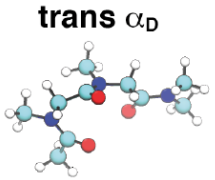
X-22-25-Y → ψ

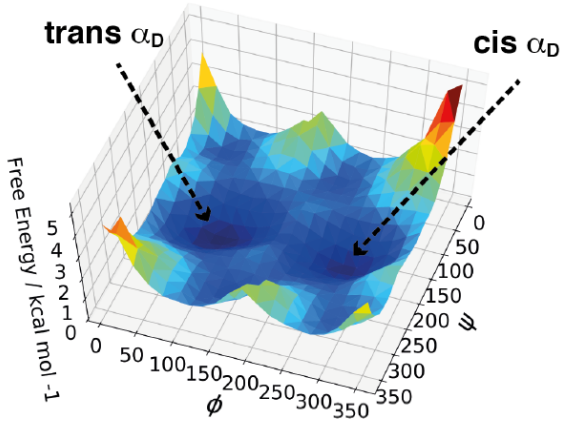
X-7-12-Y → φ(-1)

X-12-15-Y → ψ(-1)

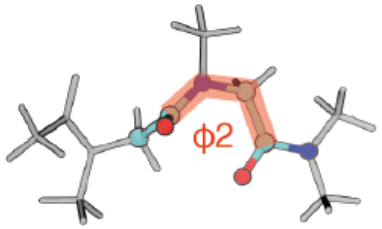
b)

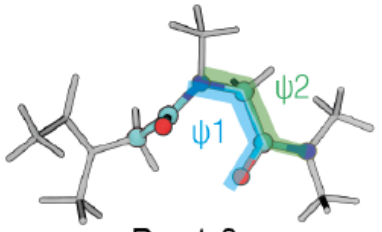




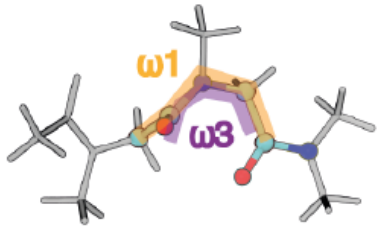


b)

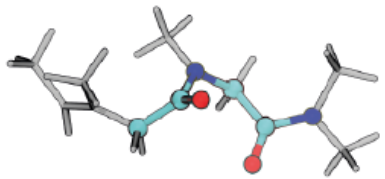




P_B=1.0

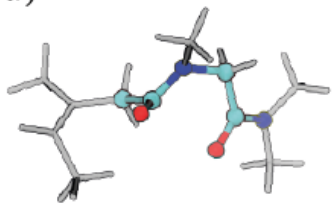


c)



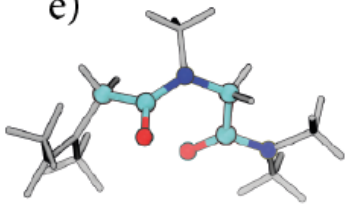
P_B=0.75

d)



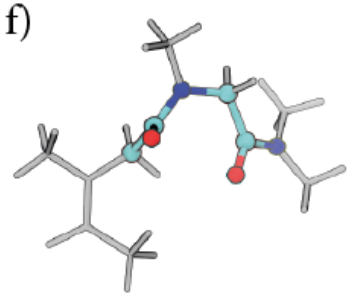
P_B=0.50

e)

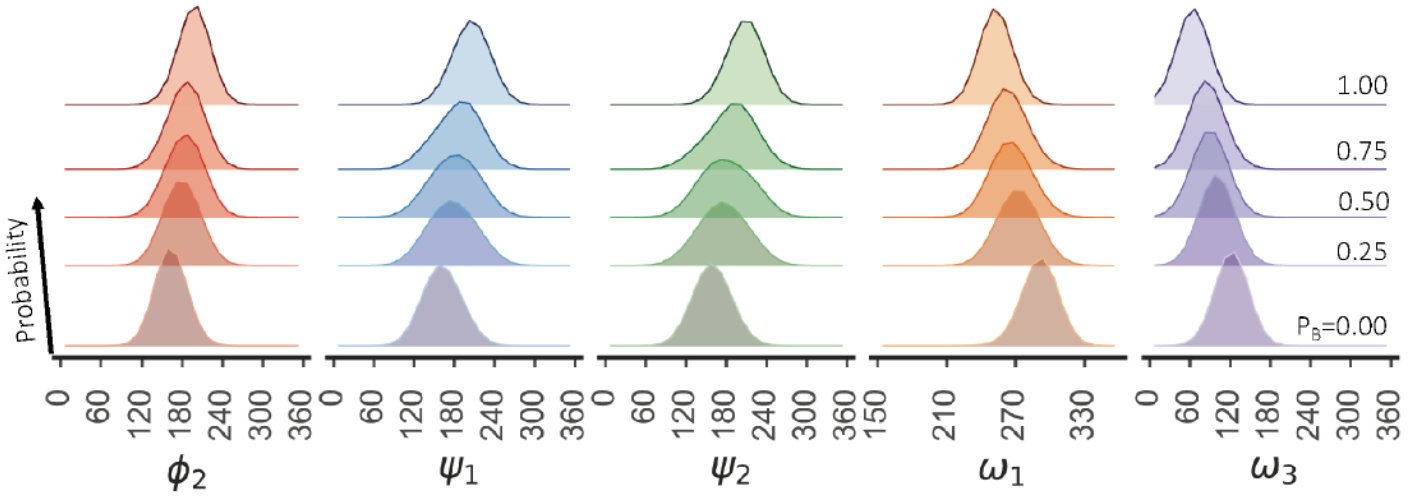


P_B=0.25

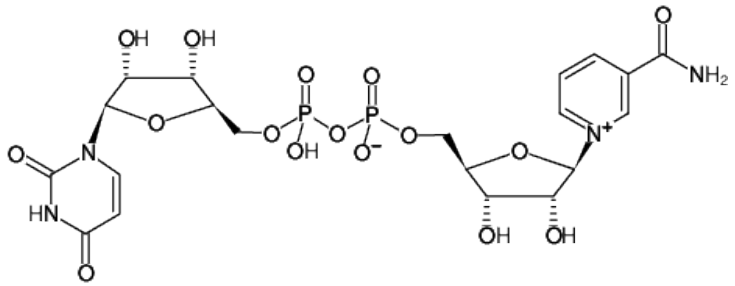
f)

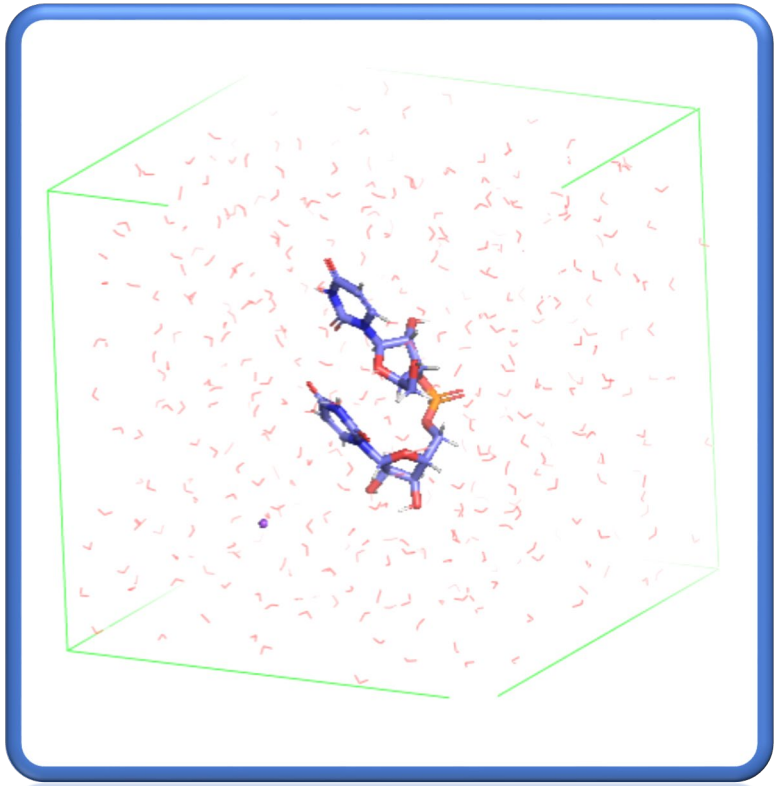


P_B=0.0



Conformational transitions of uracil dinucleotide in explicit water





CVs include sines and cosines of 17 dihedral angles plus 152 symmetry functions of the form:

$$G_2^i = \sum_j e^{-\eta(R_{ij}-R_s)^2} \cdot f_c(R_{ij}),$$

Try out schemes for automatically selecting descriptors:
G. Imbalzano *et al*. *JCP* (2018).

Conclusions

- Interoperability of d-AFED/TAMD, metadynamics, and umbrella sampling leads to an algorithm ideally suited for use with machine learning method for regressing high-dimensional FESs, as it sweeps over the full landscape, producing an initially sparse sampling (the density of which grows with simulation length) that regression methods can easily interpolate.

- From a large set of redundant CVs, committor learning is possible via either a direct mapping of the CVs to the committor or via a model of the committor in terms of a reaction coordinate expressed in terms of the CVs.

- For the systems studied, light gradient boosting proved to be the best performing of the machine-learning models investigated.

- SHAP analysis allows the most important features to be extracted, which reduces the number of degrees of freedom needed to express the committor or reaction coordinate.