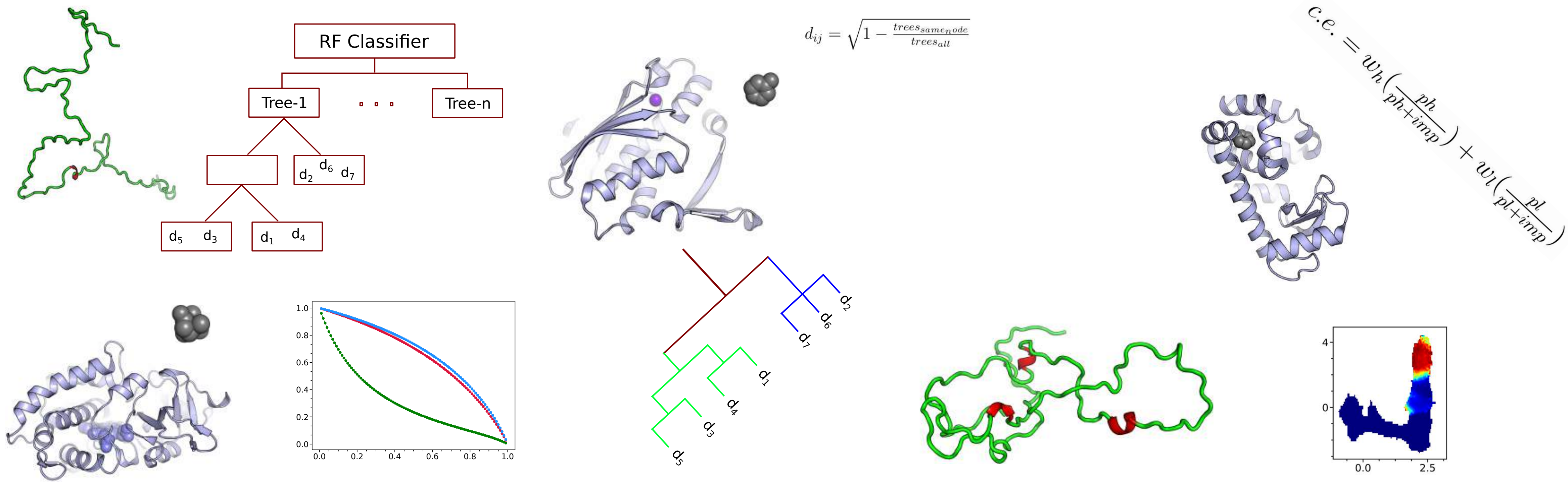
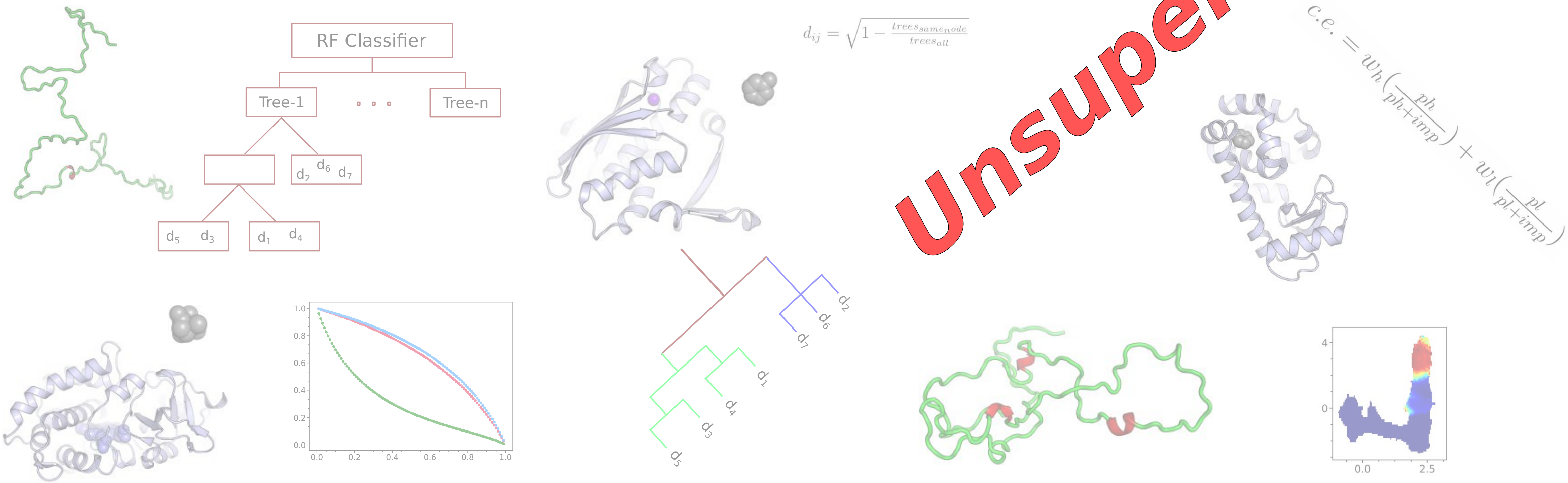
Resolving high dimensional conformational space of proteins

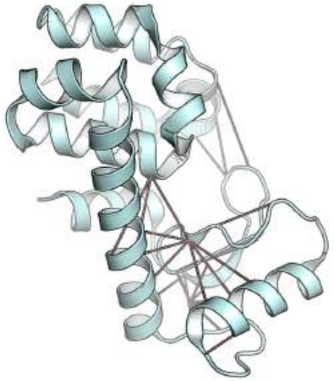


Resolving high dimensional conformational space of proteins

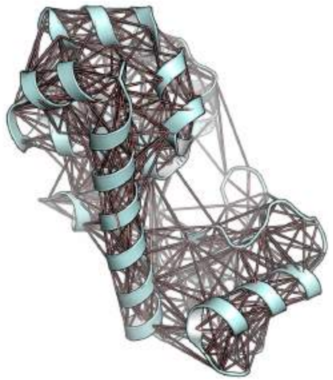
Unsupervised



Resolving Conformational Dynamics in Proteins



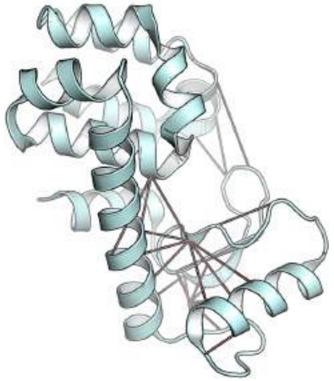
vs



Low dimension

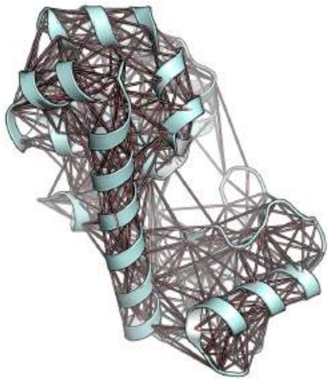
High dimension

Resolving Conformational Dynamics in Proteins

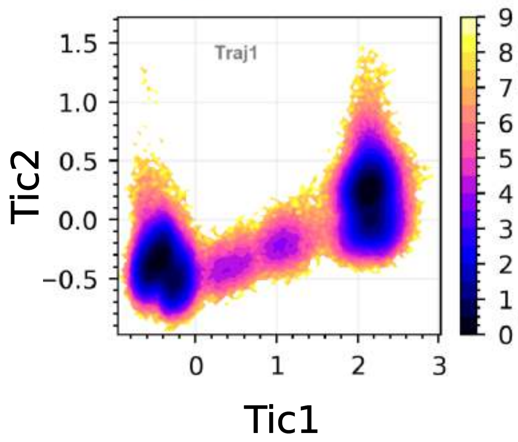


Low dimension

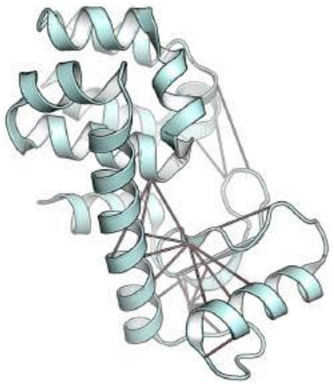
vs



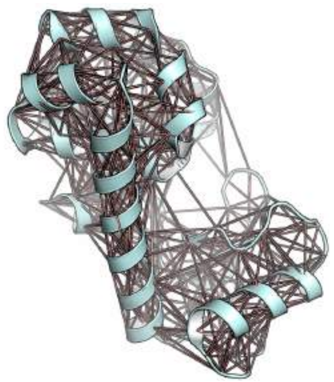
High dimension

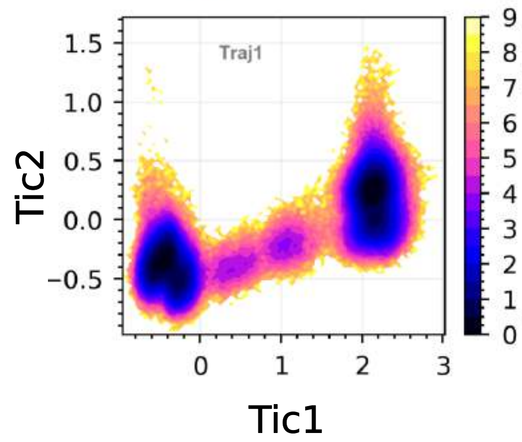


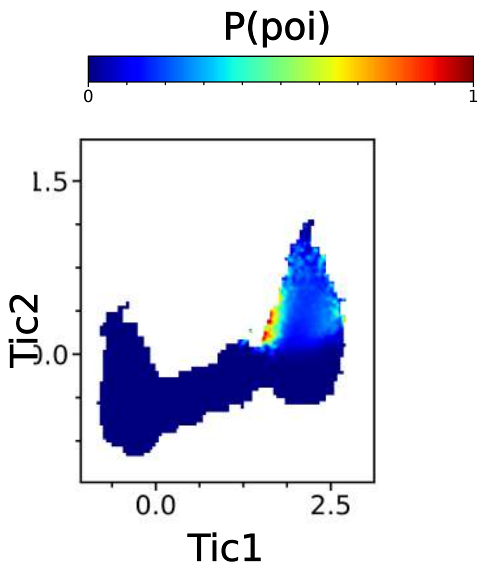
Resolving Conformational Dynamics in Proteins



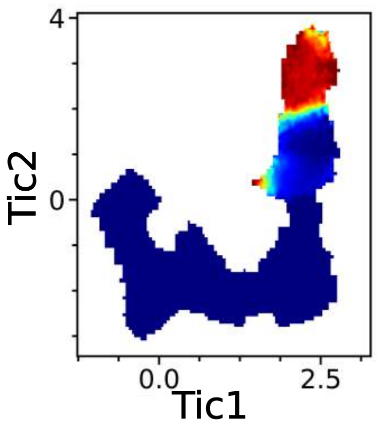
vs







RF



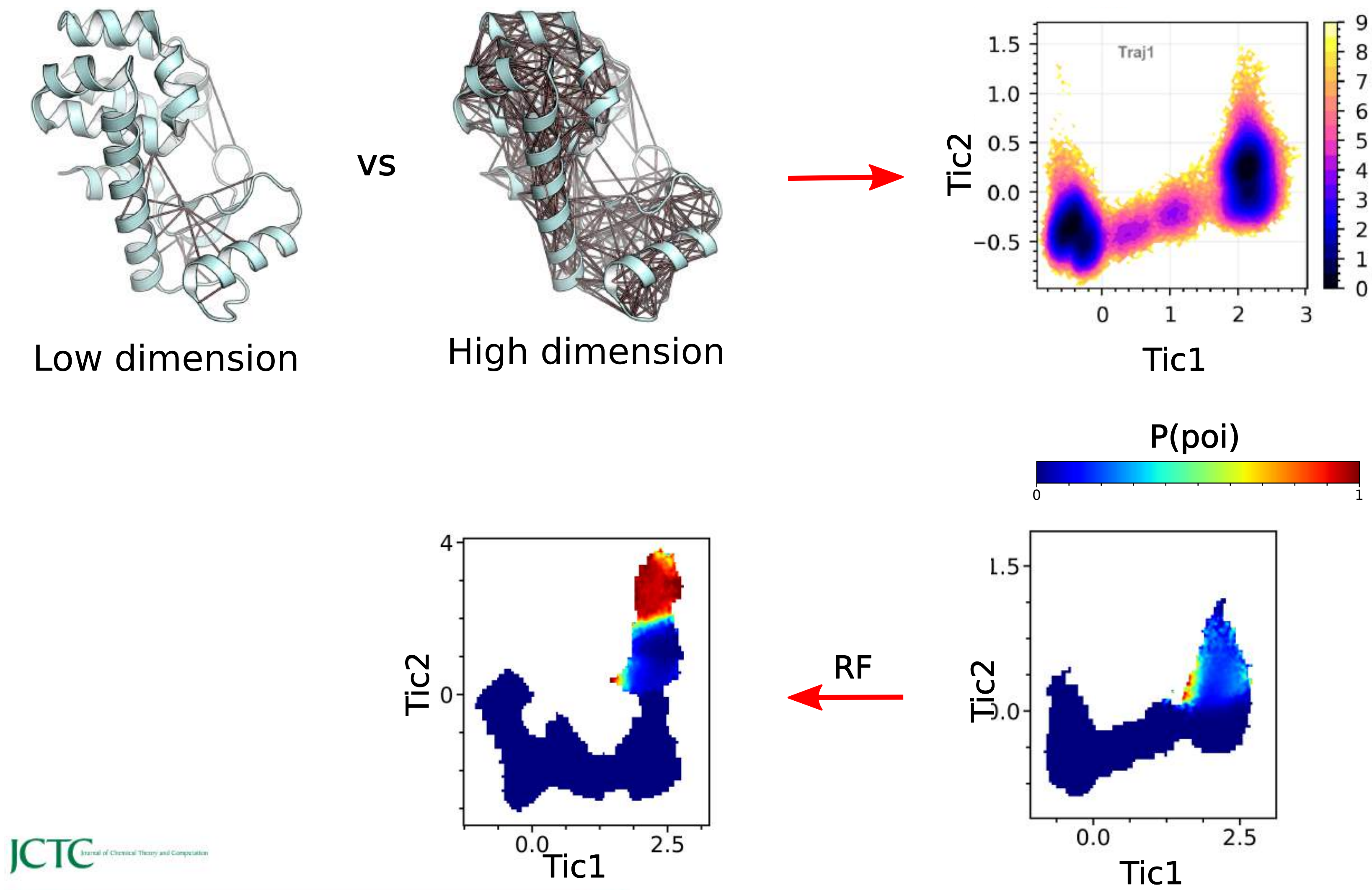
JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Resolving Protein Conformational Plasticity and Substrate Binding via Machine Learning

Navjeet Ahalawat,*,# Mohammad Sahil,# and Jagannath Mondal*

Resolving Conformational Dynamics in Proteins



JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Resolving Protein Conformational Plasticity and Substrate Binding via Machine Learning

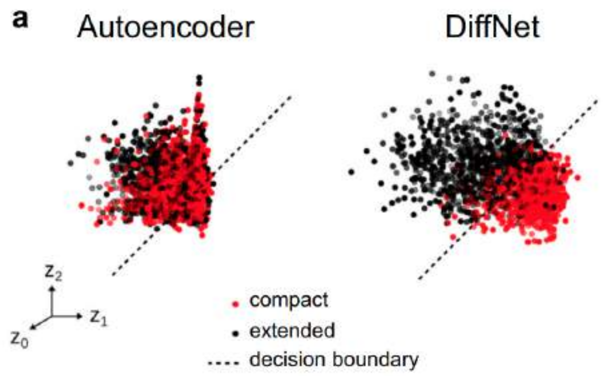
Navjeet Ahalawat,*,# Mohammad Sahil,# and Jagannath Mondal*

nature COMMUNICATIONS

ARTICLE OPEN

Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets

Michael D. Ward[1,2], Maxwell I. Zimmerman[1,2], Artur Meller[1,2], Moses Chung[1,2], S. J. Swamidass[3] & Gregory R. Bowman[1,2]



RESEARCH ARTICLE | SEPTEMBER 16 2021

A deep autoencoder framework for discovery of metastable ensembles in biomacromolecules

Satyabrata Bandyopadhyay ; Jagannath Mondal

Check for updates

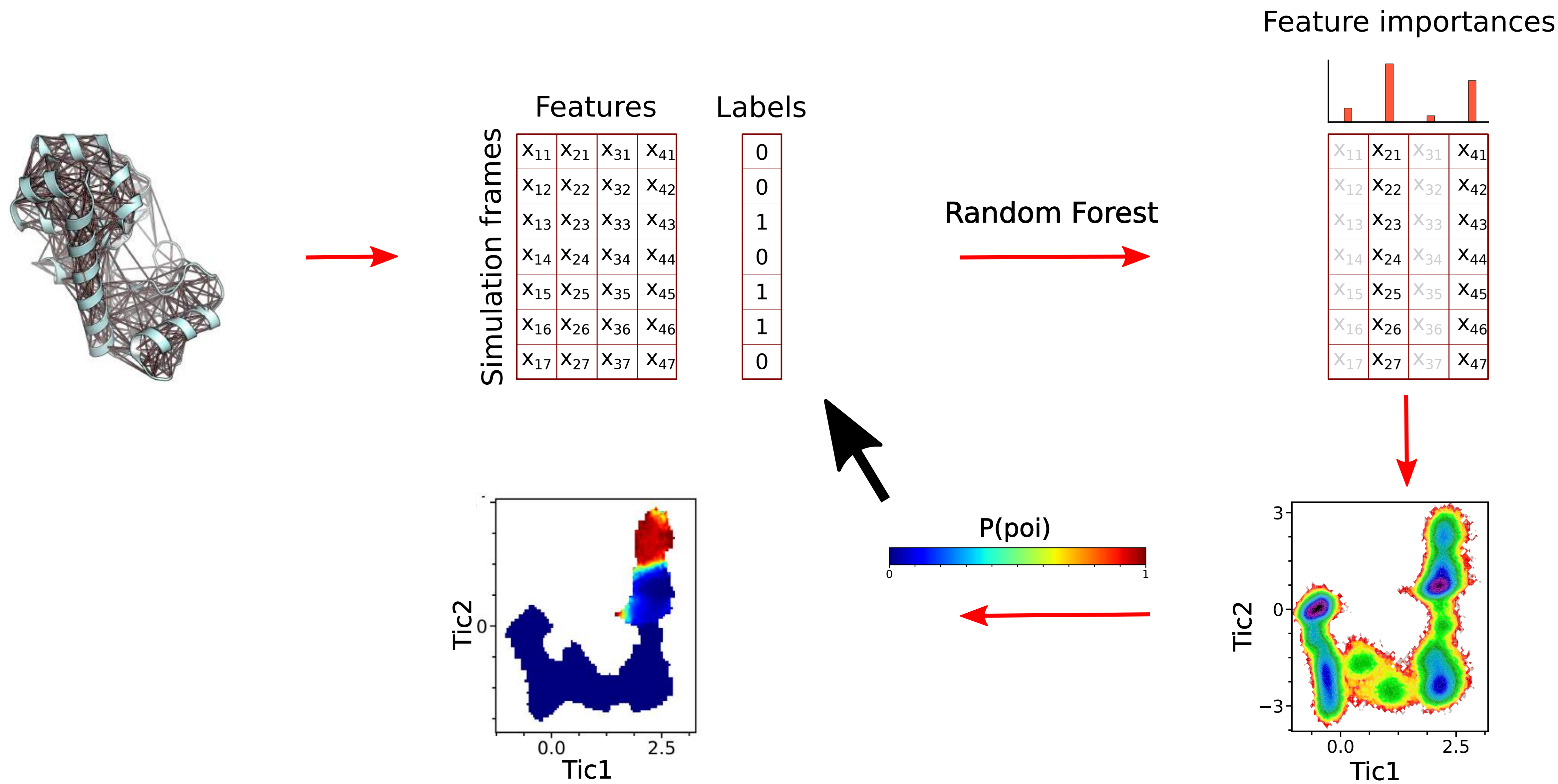
J. Chem. Phys. 155, 114106 (2021)

https://doi.org/10.1063/5.0059965

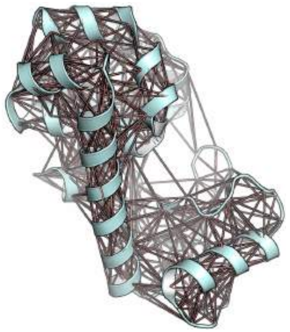
Machine Learning Subtle Conformational Change due to Phosphorylation in Intrinsically Disordered Proteins

Subinoy Adhikari and Jagannath Mondal*

Random Forest - Supervised problem



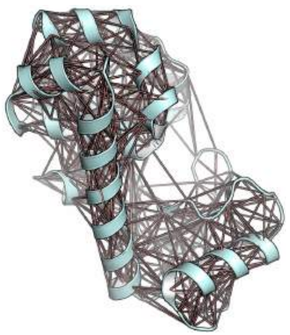
Unsupervised Random Forest



Simulations

	features			
d_1	x_{11}	x_{21}	x_{31}	x_{41}
d_2	x_{12}	x_{22}	x_{32}	x_{42}
d_3	x_{13}	x_{23}	x_{33}	x_{43}
d_4	x_{14}	x_{24}	x_{34}	x_{44}
d_5	x_{15}	x_{25}	x_{35}	x_{45}
d_6	x_{16}	x_{26}	x_{36}	x_{46}
d_7	x_{17}	x_{27}	x_{37}	x_{47}

Unsupervised Random Forest



Simulations

features

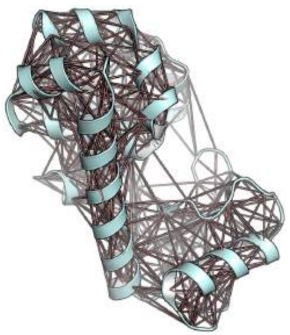
d_1	x_{11}	x_{21}	x_{31}	x_{41}
d_2	x_{12}	x_{22}	x_{32}	x_{42}
d_3	x_{13}	x_{23}	x_{33}	x_{43}
d_4	x_{14}	x_{24}	x_{34}	x_{44}
d_5	x_{15}	x_{25}	x_{35}	x_{45}
d_6	x_{16}	x_{26}	x_{36}	x_{46}
d_7	x_{17}	x_{27}	x_{37}	x_{47}

synthetic data

x_{13}	x_{25}	x_{37}	x_{44}
x_{16}	x_{27}	x_{35}	x_{47}
x_{14}	x_{21}	x_{32}	x_{46}
x_{11}	x_{23}	x_{36}	x_{42}
x_{17}	x_{22}	x_{31}	x_{43}
x_{15}	x_{26}	x_{34}	x_{41}
x_{12}	x_{24}	x_{33}	x_{45}

$$Y_{(m,f)} = X_{(m,f)} \odot (P_1, P_2, ..P_f)_{(1,f)}$$

Unsupervised Random Forest



Simulations

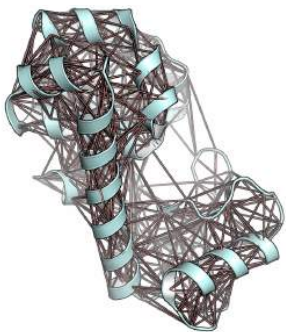
	features				labels
d_1	x_{11}	x_{21}	x_{31}	x_{41}	0
d_2	x_{12}	x_{22}	x_{32}	x_{42}	0
d_3	x_{13}	x_{23}	x_{33}	x_{43}	0
d_4	x_{14}	x_{24}	x_{34}	x_{44}	0
d_5	x_{15}	x_{25}	x_{35}	x_{45}	0
d_6	x_{16}	x_{26}	x_{36}	x_{46}	0
d_7	x_{17}	x_{27}	x_{37}	x_{47}	0

synthetic data

x_{13}	x_{25}	x_{37}	x_{44}	1
x_{16}	x_{27}	x_{35}	x_{47}	1
x_{14}	x_{21}	x_{32}	x_{46}	1
x_{11}	x_{23}	x_{36}	x_{42}	1
x_{17}	x_{22}	x_{31}	x_{43}	1
x_{15}	x_{26}	x_{34}	x_{41}	1
x_{12}	x_{24}	x_{33}	x_{45}	1

$$Y_{(m,f)} = X_{(m,f)} \odot (P_1, P_2, ..P_f)_{(1,f)}$$

Unsupervised Random Forest



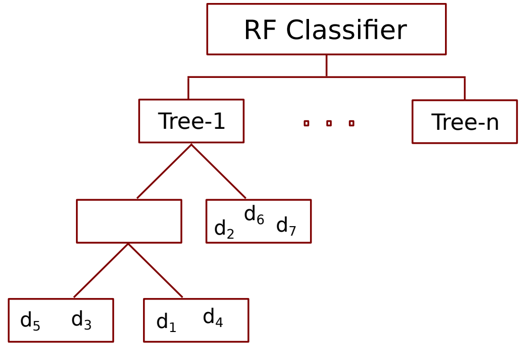
Simulations

	features				labels
d_1	x_{11}	x_{21}	x_{31}	x_{41}	0
d_2	x_{12}	x_{22}	x_{32}	x_{42}	0
d_3	x_{13}	x_{23}	x_{33}	x_{43}	0
d_4	x_{14}	x_{24}	x_{34}	x_{44}	0
d_5	x_{15}	x_{25}	x_{35}	x_{45}	0
d_6	x_{16}	x_{26}	x_{36}	x_{46}	0
d_7	x_{17}	x_{27}	x_{37}	x_{47}	0

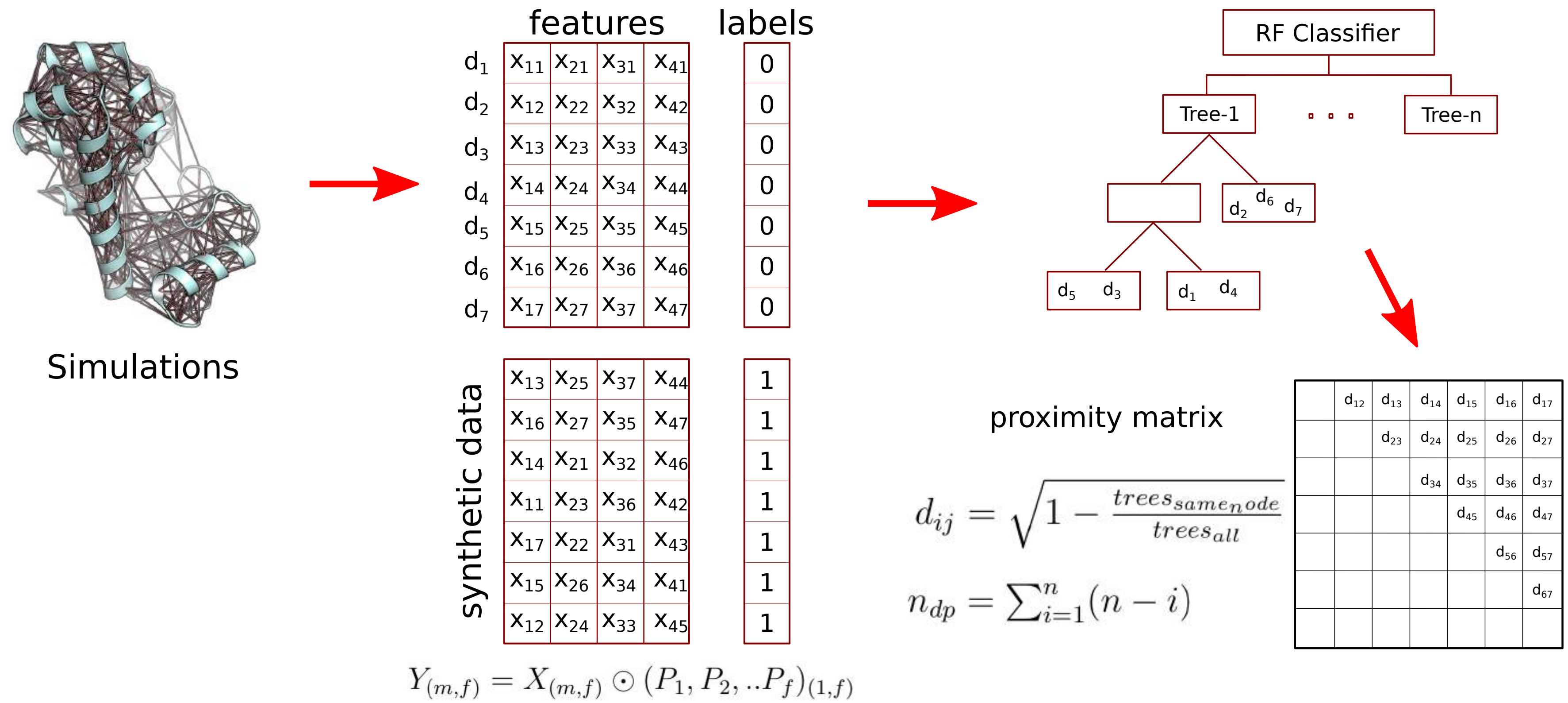
synthetic data

x_{13}	x_{25}	x_{37}	x_{44}	1
x_{16}	x_{27}	x_{35}	x_{47}	1
x_{14}	x_{21}	x_{32}	x_{46}	1
x_{11}	x_{23}	x_{36}	x_{42}	1
x_{17}	x_{22}	x_{31}	x_{43}	1
x_{15}	x_{26}	x_{34}	x_{41}	1
x_{12}	x_{24}	x_{33}	x_{45}	1

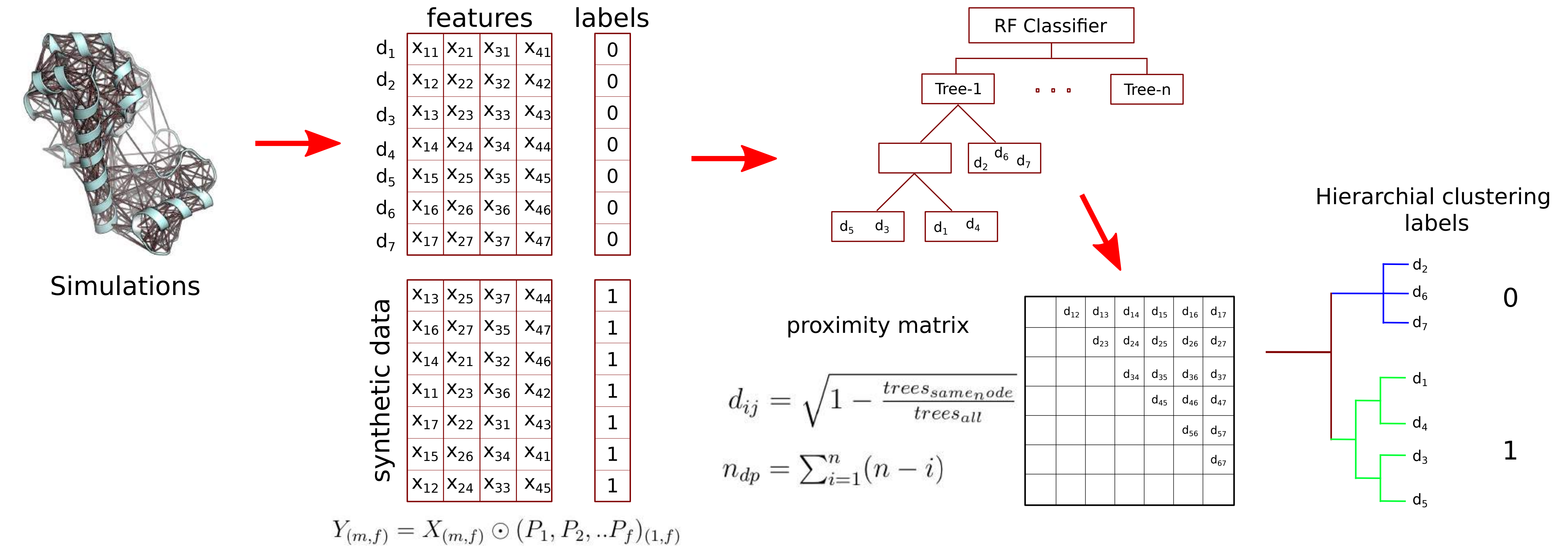
$$Y_{(m,f)} = X_{(m,f)} \odot (P_1, P_2, ..P_f)_{(1,f)}$$



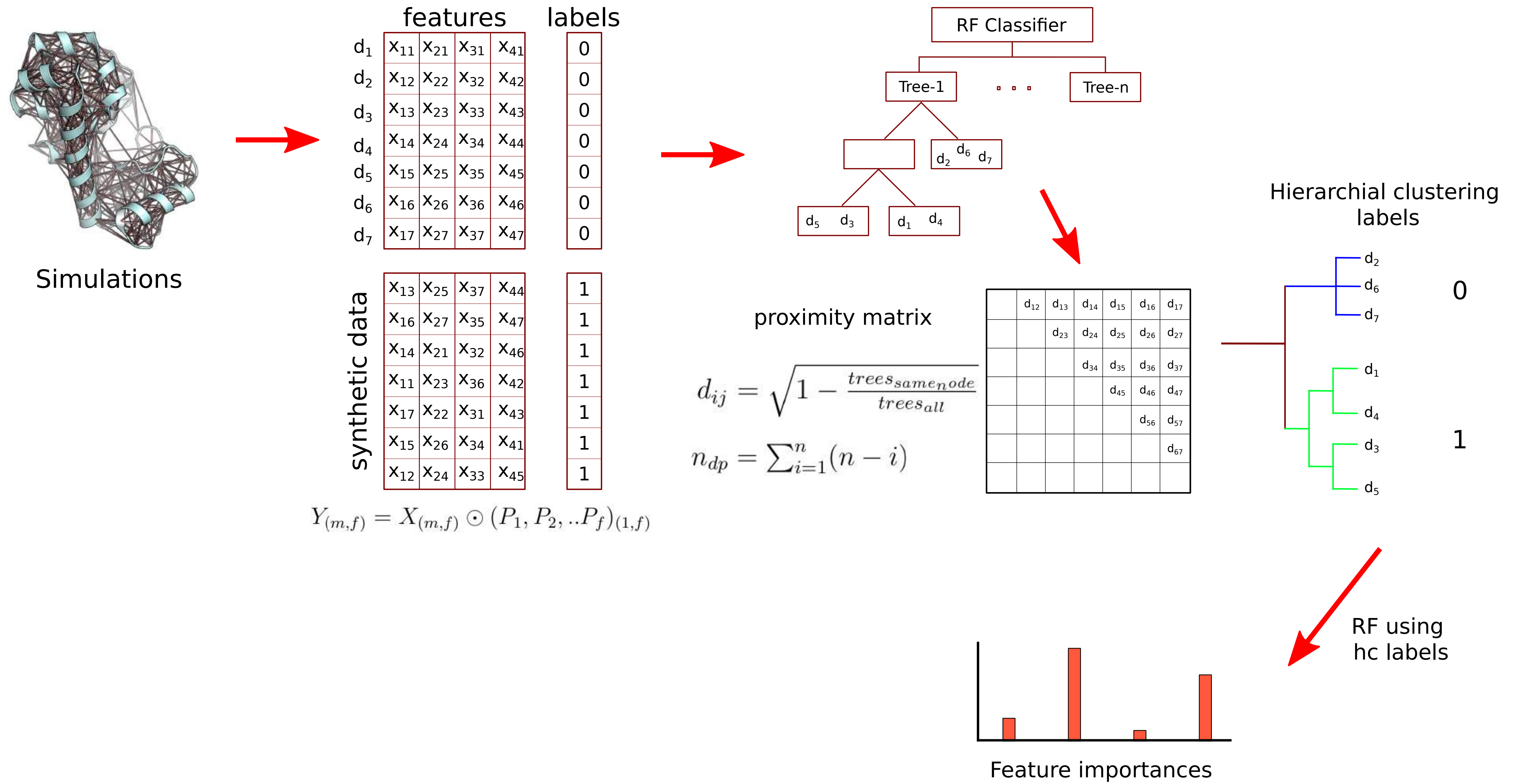
Unsupervised Random Forest



Unsupervised Random Forest

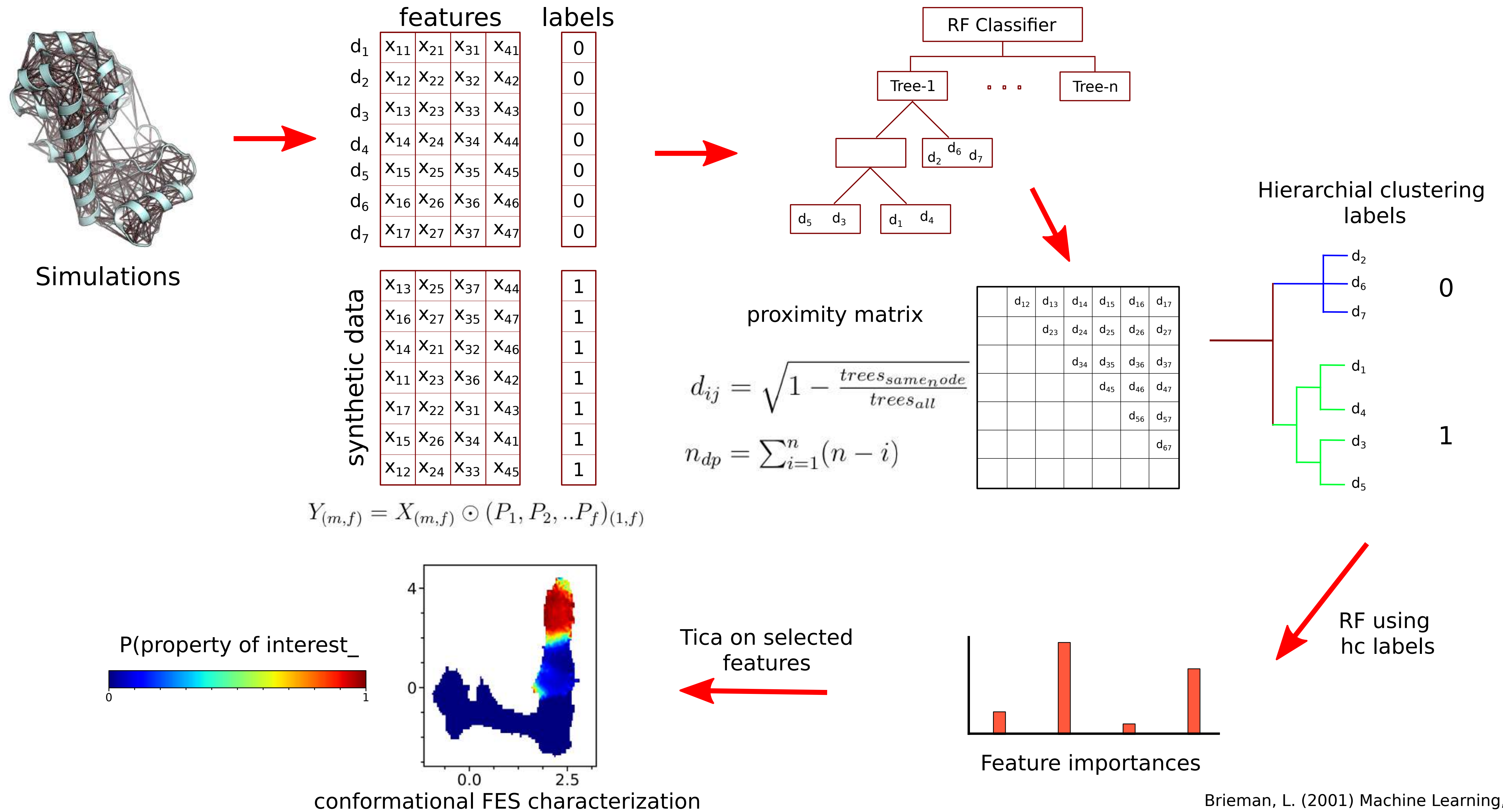


Unsupervised Random Forest



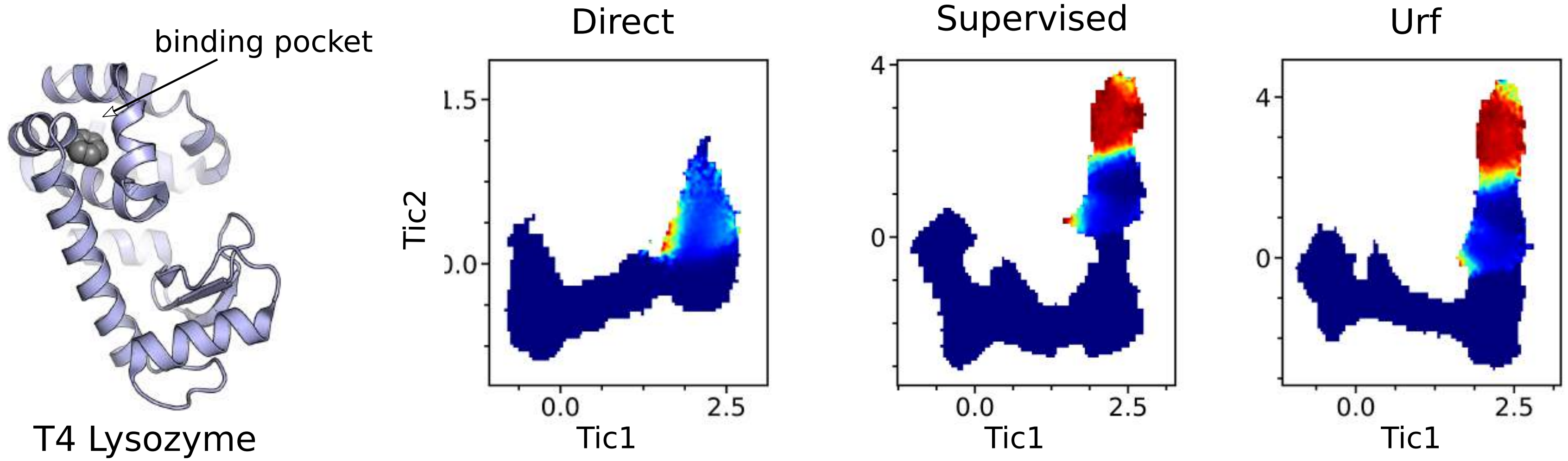
Brieman, L. (2001) Machine Learning, 45, 5-32

Unsupervised Random Forest

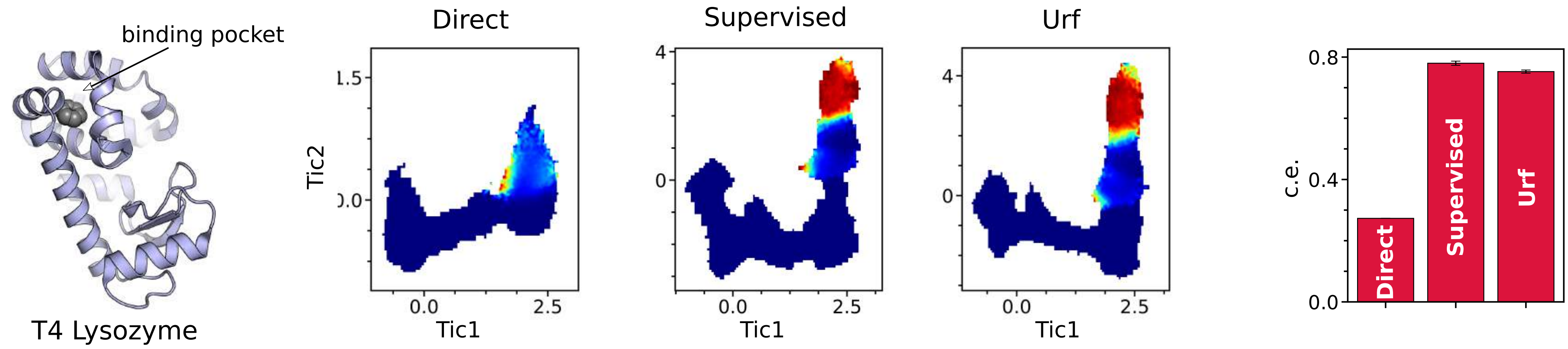


Brieman, L. (2001) Machine Learning, 45, 5-32

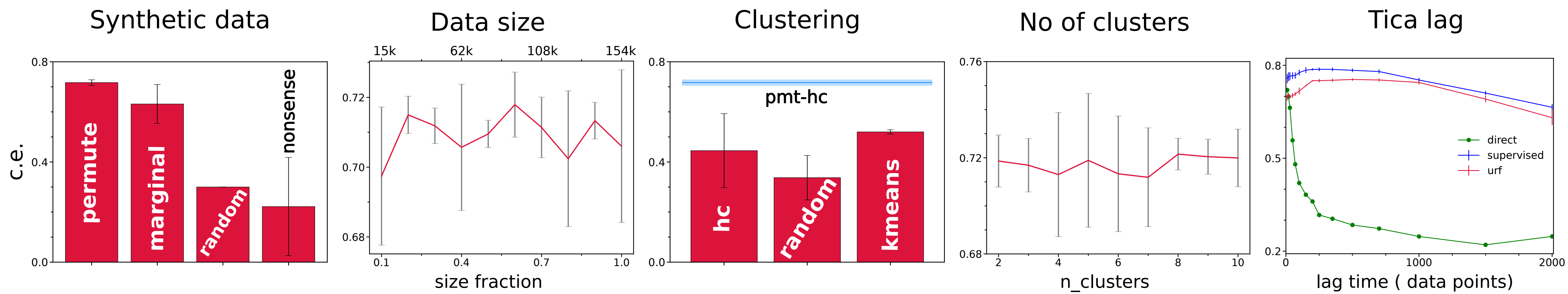
URF can recapitulate supervised results on T4-Lysozyme



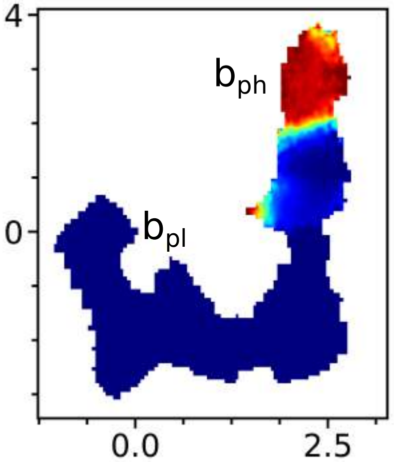
URF can recapitulate supervised results on T4-Lysozyme



All Data (~154k) → permute → pmt-hc → three clusters → tica-lag 700 steps



The Classification extent:: separability of functional states



a fes is a n×m bins

100×100 in this case

divided into zero and non-zero bins (tb)

each tb bin represent probability (0-1) of a particular functional states

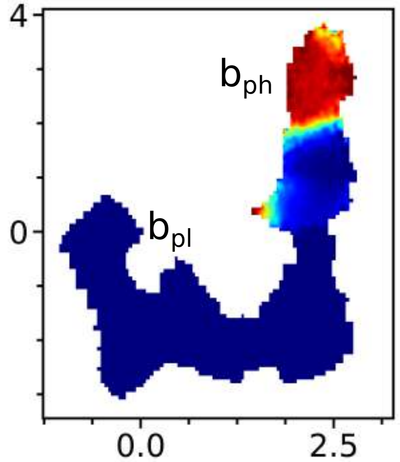
$$b_{ph} = tb >= (1-c) \qquad b_{pl} = tb <= c$$

$$ph = w_p[card(b_{ph}) + \frac{\Sigma(b_{ph}-(1-c))}{card(b_{ph})}] \qquad pl = w_p[card(b_{pl}) + \frac{\Sigma(b_{pl}-c)}{card(b_{pl})}]$$

$$imp = w_{imp}[card((tb < (1-c))\&(tb > c))]$$

$$c.e. = w_h(\frac{ph}{ph+imp}) + w_l(\frac{pl}{pl+imp})$$

The Classification extent:: separability of functional states



a fes is a n×m bins

100×100 in this case

divided into zero and non-zero bins (tb)

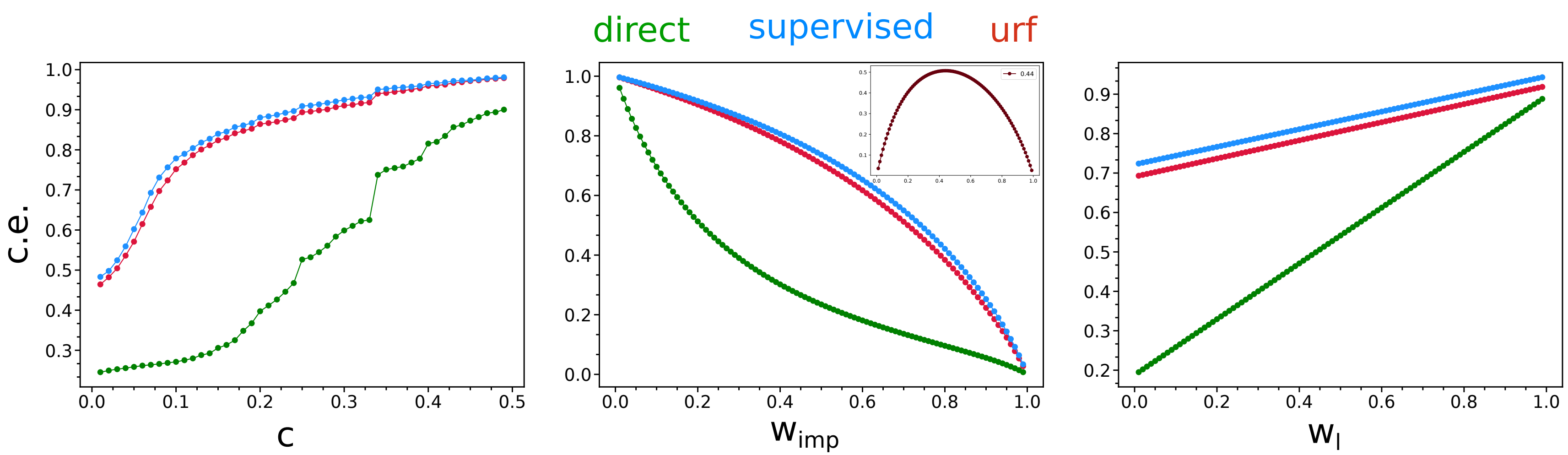
each tb bin represent probability (0-1) of a particular functional states

$$b_{ph} = tb >= (1 - c) \qquad b_{pl} = tb <= c$$

$$ph = w_p[card(b_{ph}) + \frac{\Sigma(b_{ph}-(1-c))}{card(b_{ph})}] \qquad pl = w_p[card(b_{pl}) + \frac{\Sigma(b_{pl}-c)}{card(b_{pl})}]$$

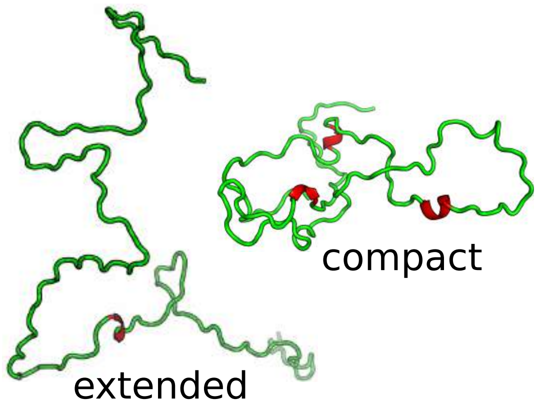
$$imp = w_{imp}[card((tb < (1 - c))\&(tb > c))]$$

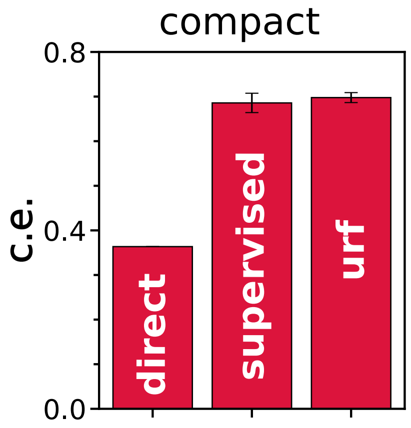
$$c.e. = w_h(\frac{ph}{ph+imp}) + w_l(\frac{pl}{pl+imp})$$



Reproducibility on other systems

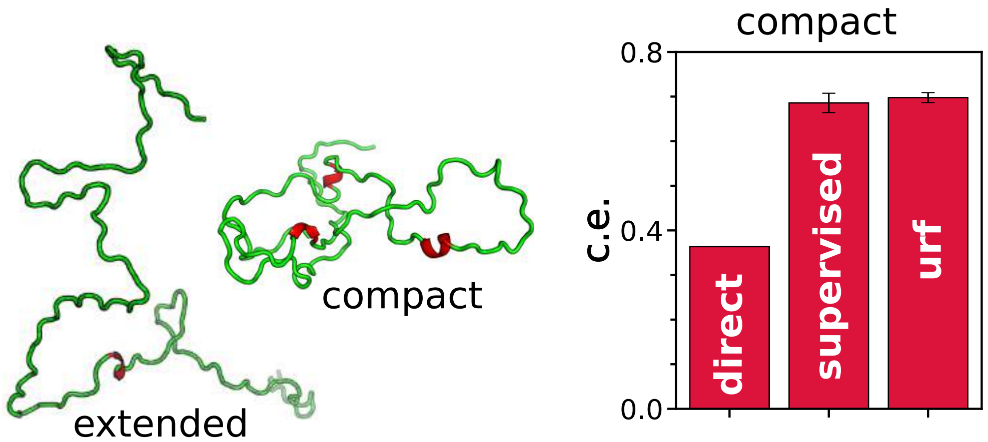
α-synuclien



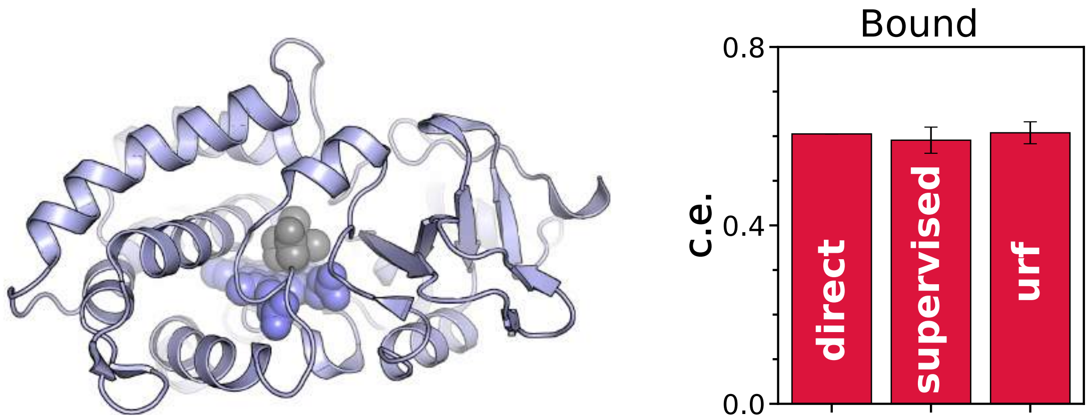


Reproducibility on other systems

α-synuclien

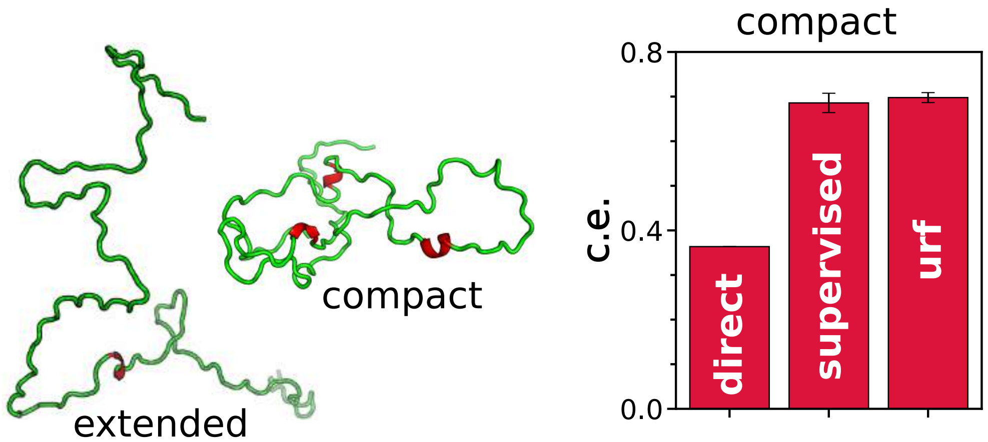


Cytochrome P450

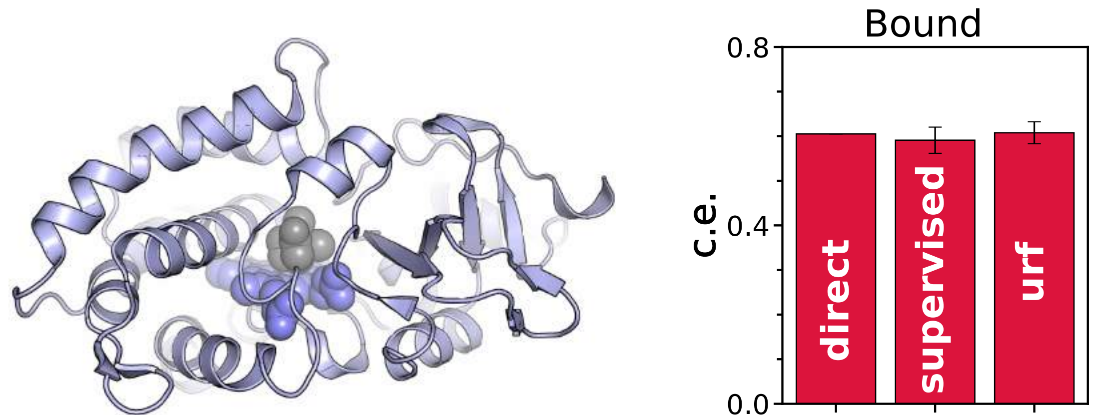


Reproducibility on other systems

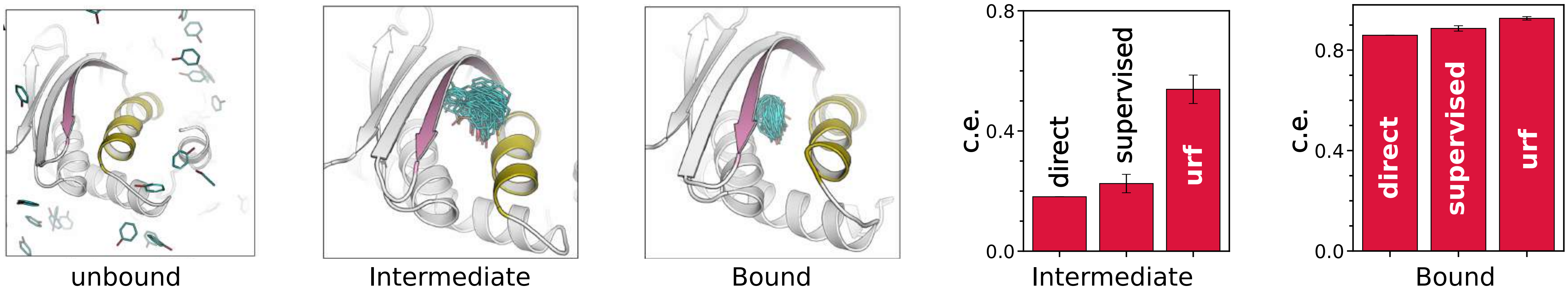
α-synuclien



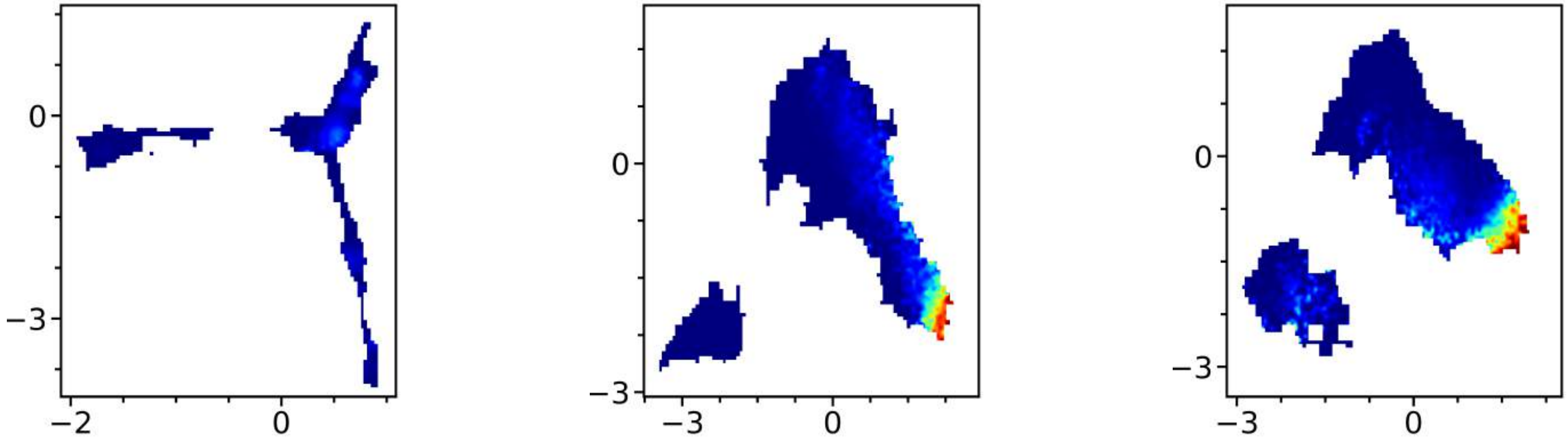
Cytochrome P450



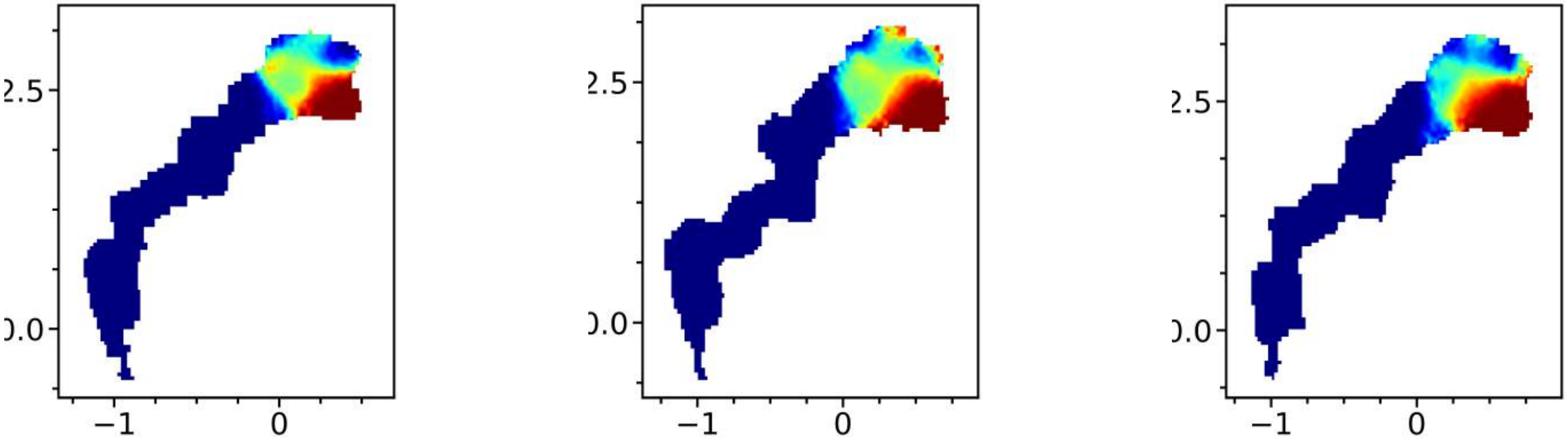
Multi-state Phenol biosensor MopR



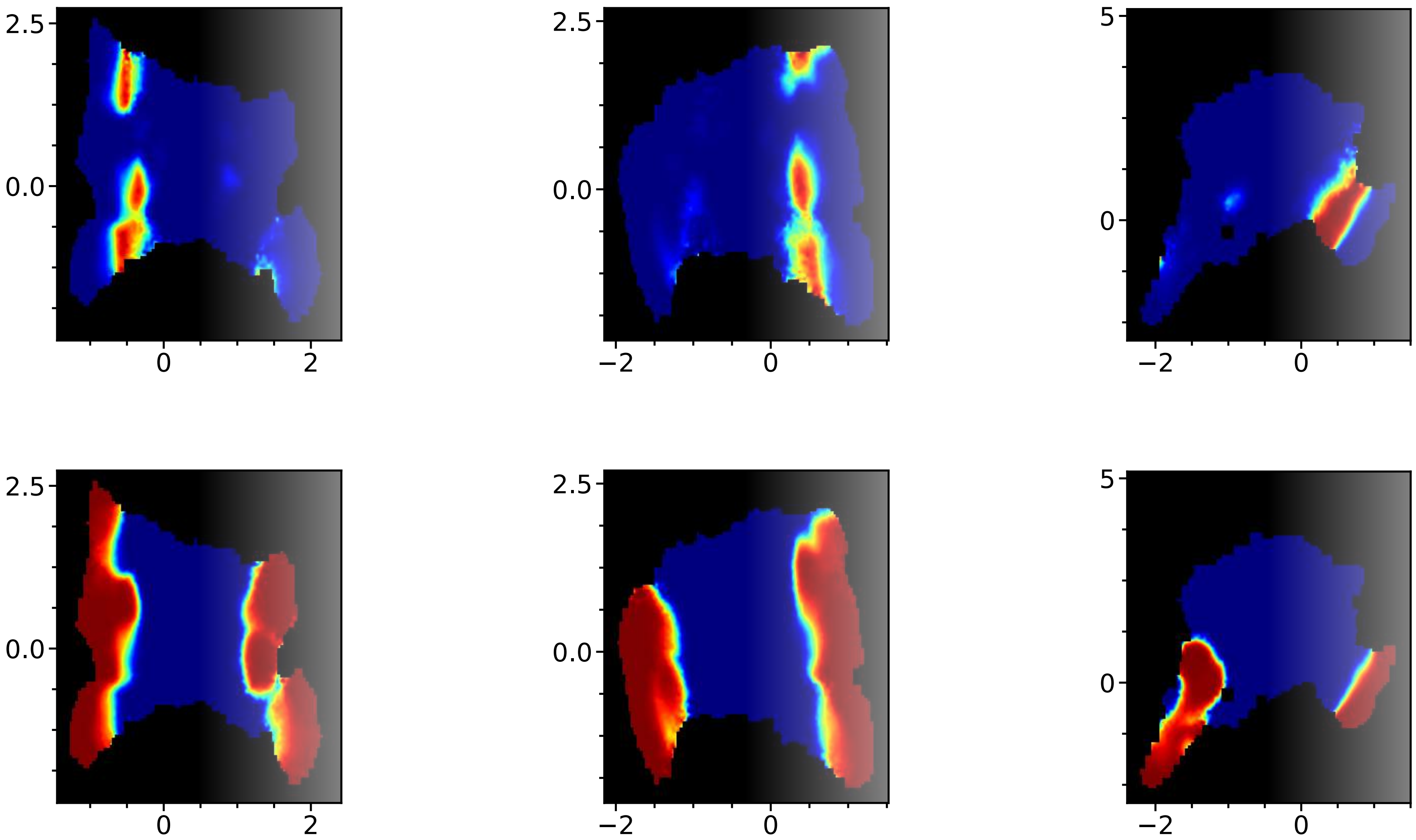
α-synuclien



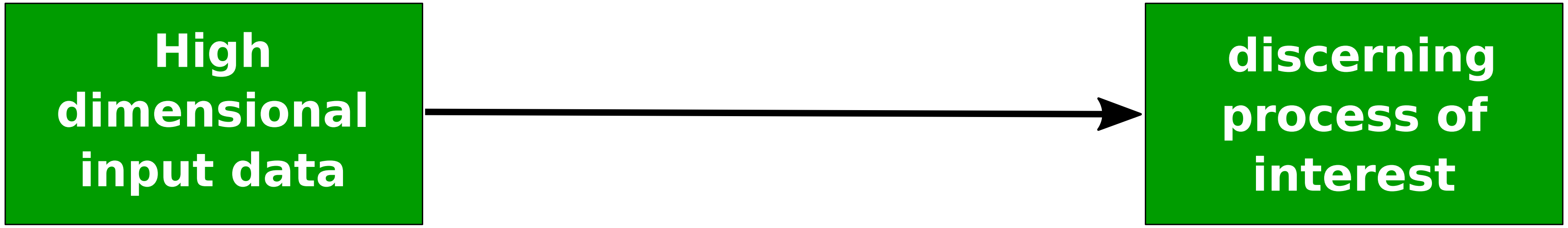
cytochrome P450



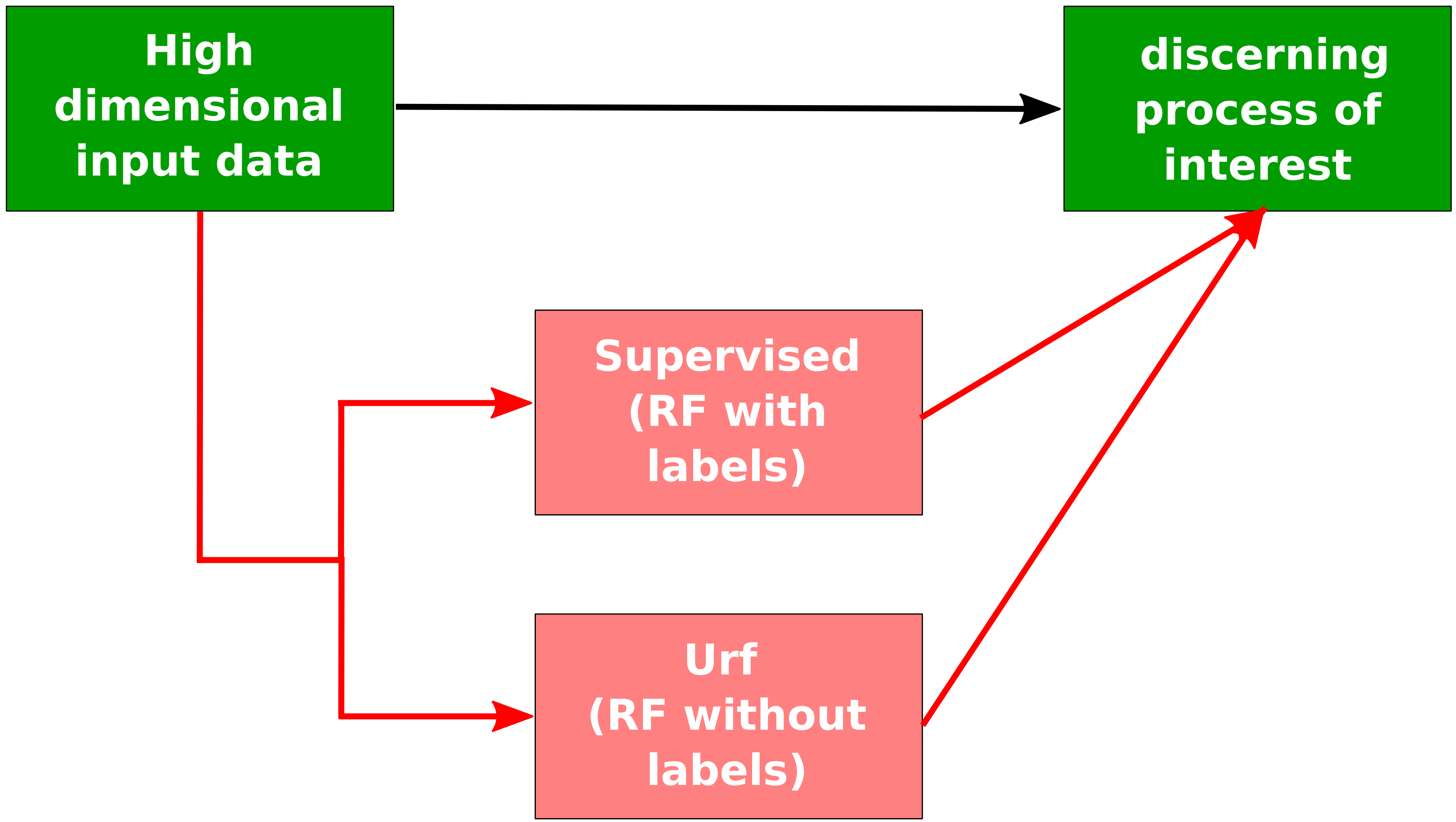
MopR



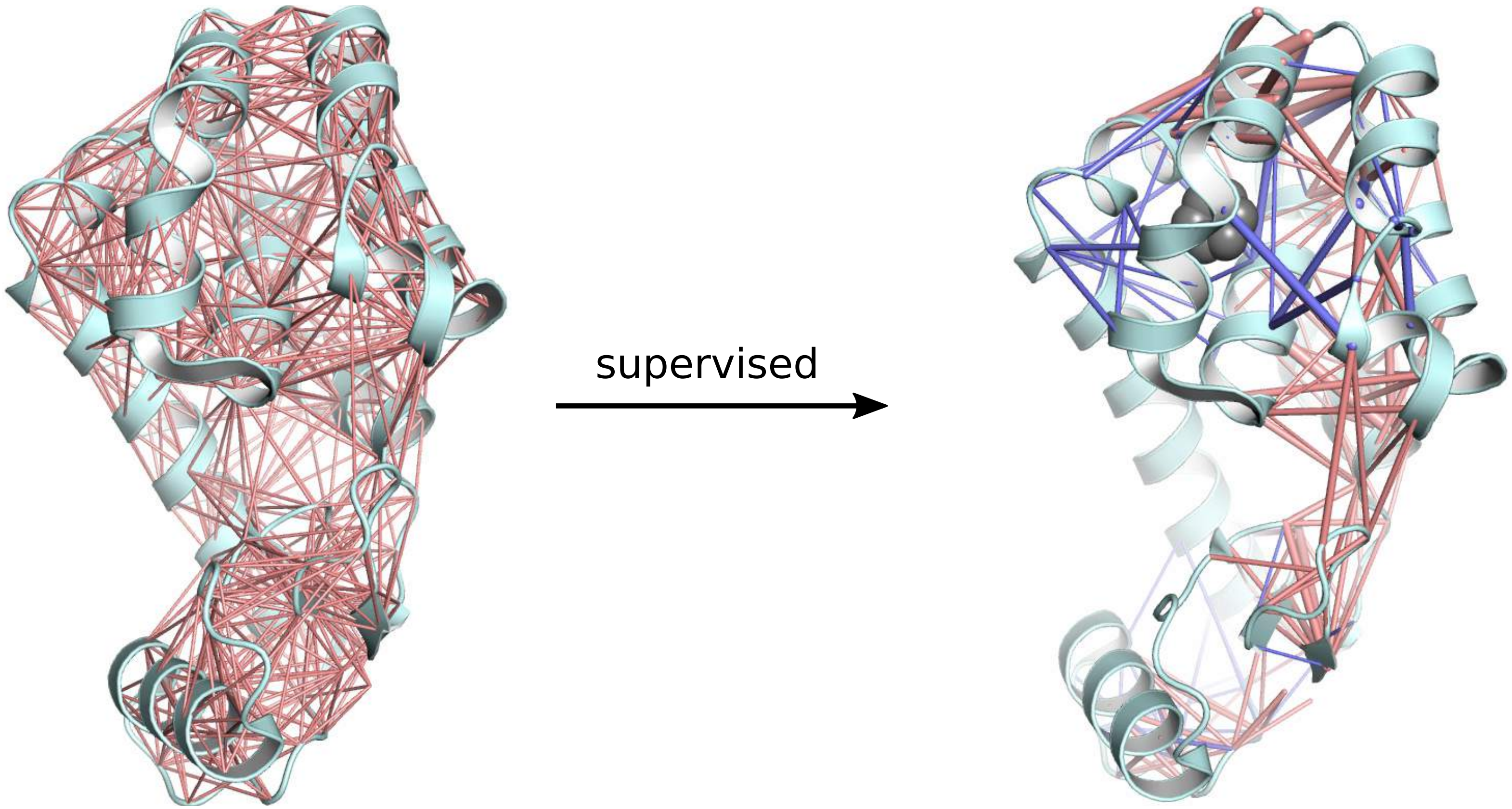
Is all this useful ?????



Is all this useful ?????

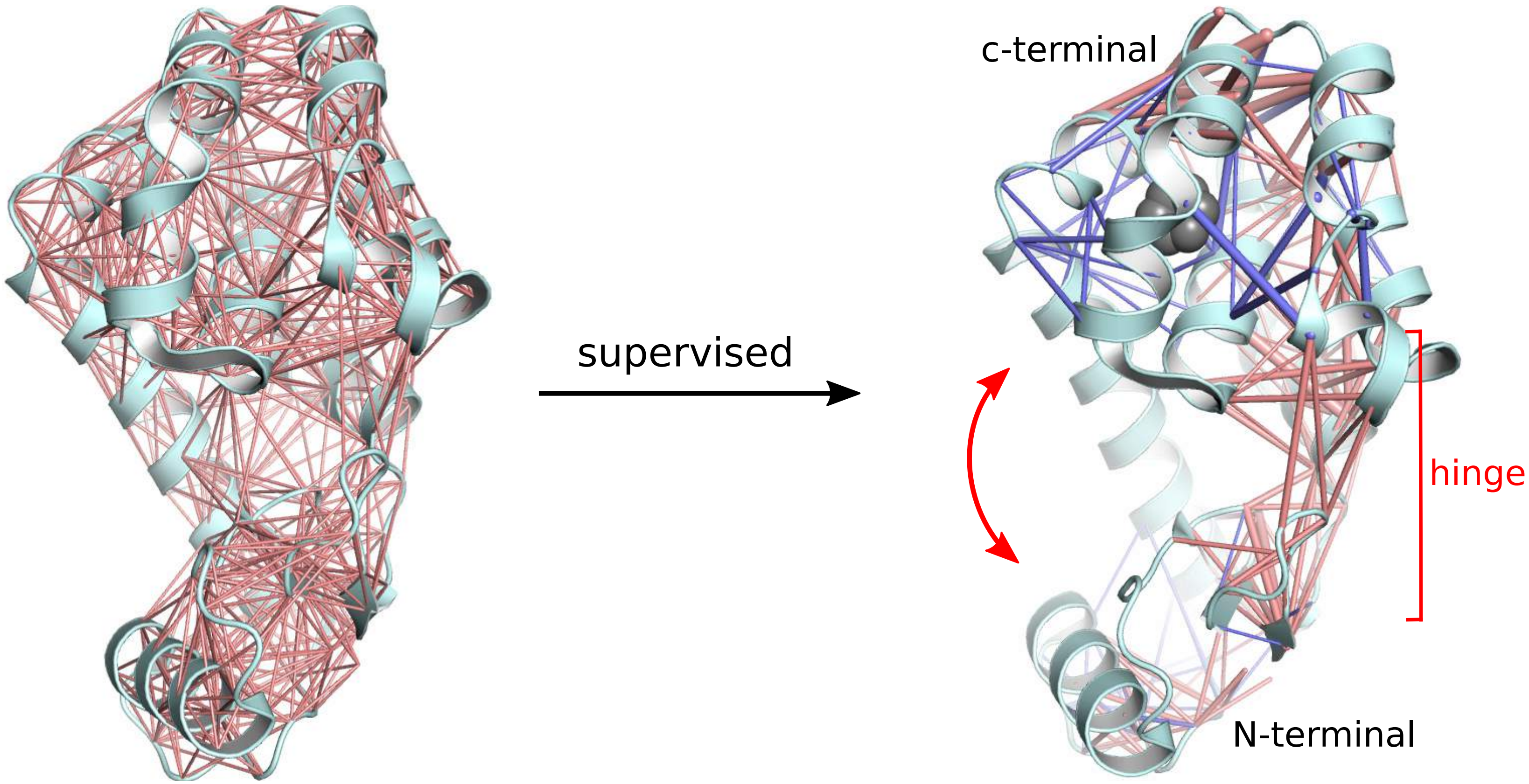


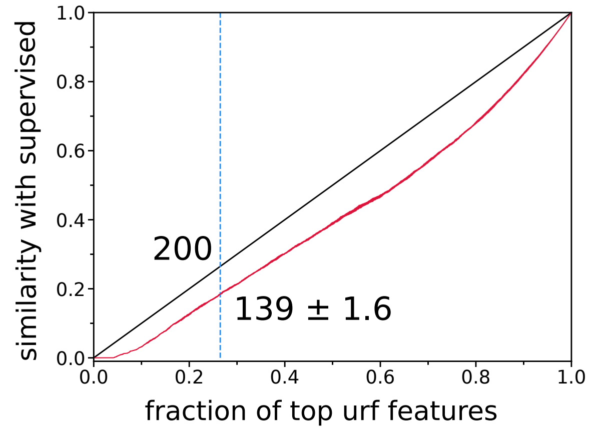
Detecting allosteric network in T4 Lysozyme

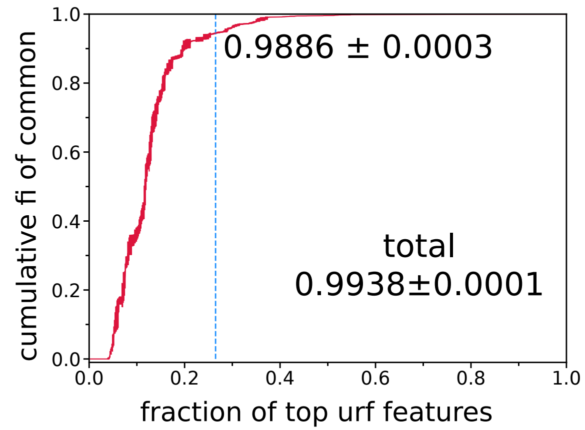


JcTc-2017, 13, 5076-5088
Jmb-2022, 434, 167679
JcTc-2023, 19, 2644-2657

Detecting allosteric network in T4 Lysozyme







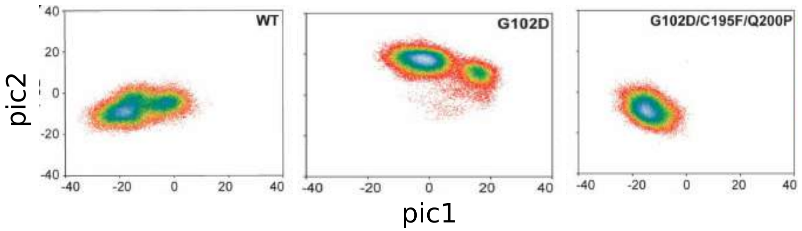
Top Urf features common to supervised are adequate to define T4L allostery based on:

1. feature importances

2. known hinge motions in T4L

JcTc-2017, 13, 5076-5088
Jmb-2022, 434, 167679
JcTc-2023, 19, 2644-2657

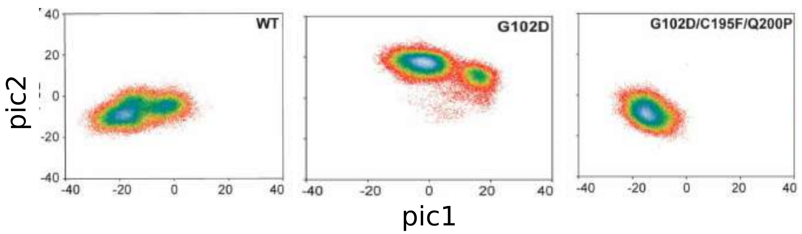
Resolving dihedral pca for dynamic allostery in mopR



using pca shift as a measure of
conformational change in proteins

PNAS-2020, 117(41), 25445-25454

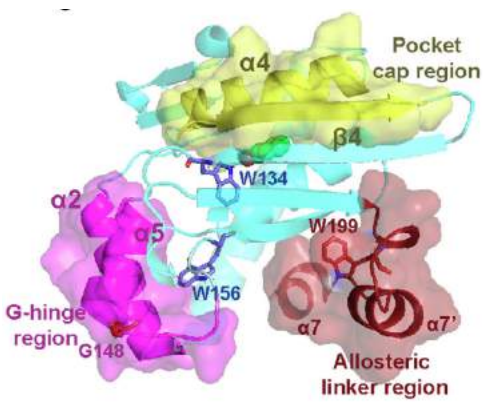
Resolving dihedral pca for dynamic allostery in mopR



using pca shift as a measure of
conformational change in proteins

PNAS-2020, 117(41), 25445-25454

Dynamic allostery in biosensor MopR



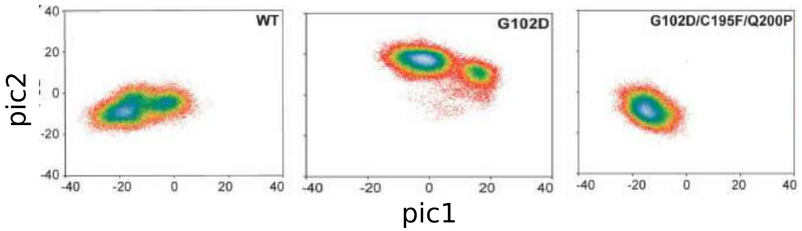
apo-wt → bound-mt

apo-wt → apo-mt

apo-mt ⇄ bound-wt

JBC-2022, 298(10), 102399

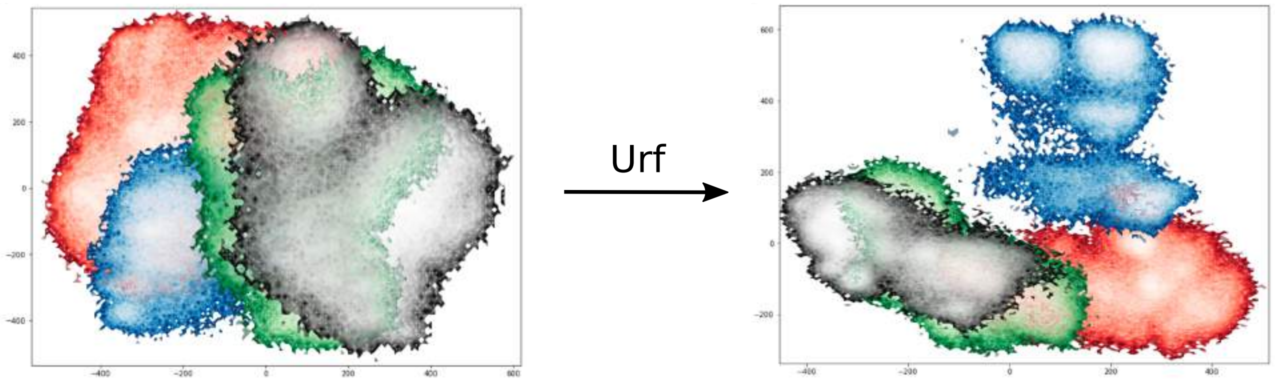
Resolving dihedral pca for dynamic allostery in mopR



using pca shift as a measure of
conformational change in proteins

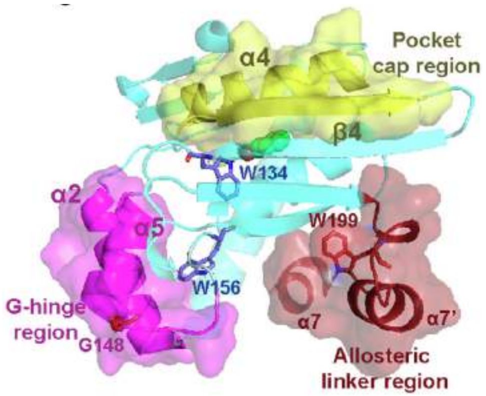
PNAS-2020, 117(41), 25445-25454

Dihedral PCA of MopR simulation ensembles



apo-wt bound-wt apo-mt bound-mt

Dynamic allostery in biosensor MopR



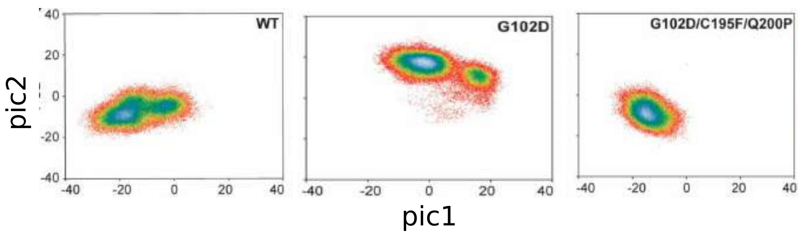
apo-wt → bound-mt

apo-mt

bound-wt

JBC-2022, 298(10), 102399

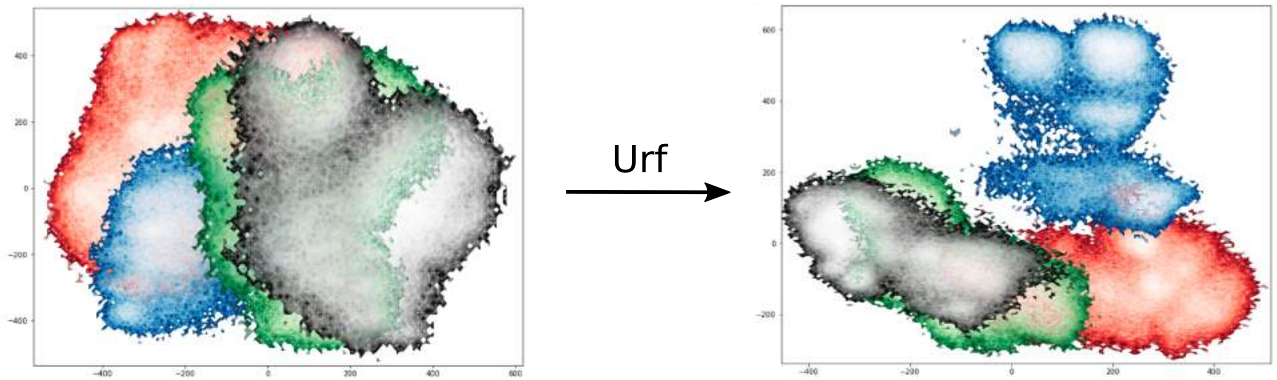
Resolving dihedral pca for dynamic allostery in mopR



using pca shift as a measure of
conformational change in proteins

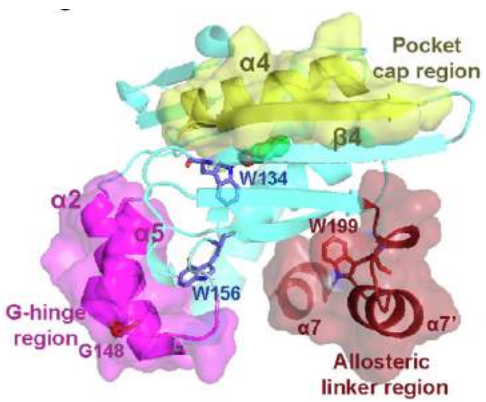
PNAS-2020, 117(41), 25445-25454

Dihedral PCA of MopR simulation ensembles



apo-wt bound-wt apo-mt bound-mt

Dynamic allostery in biosensor MopR

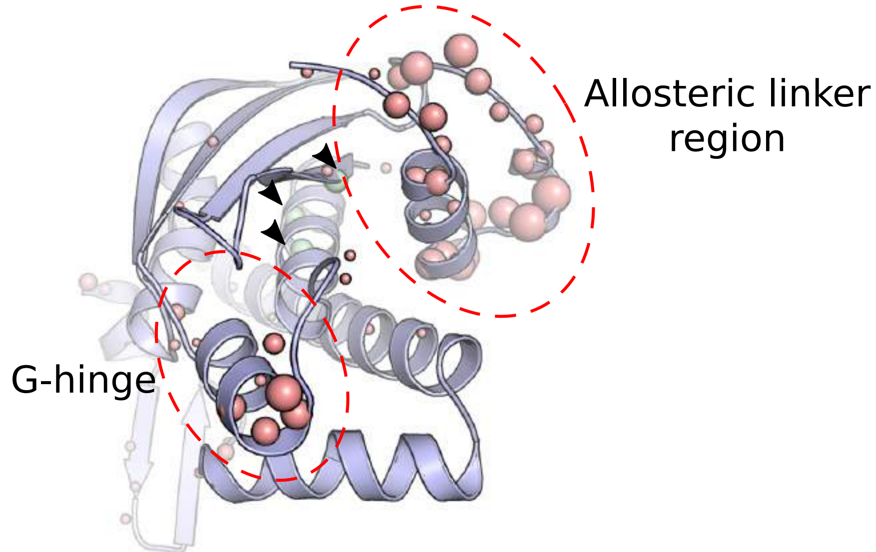


apo-wt → bound-mt

apo-wt → apo-mt

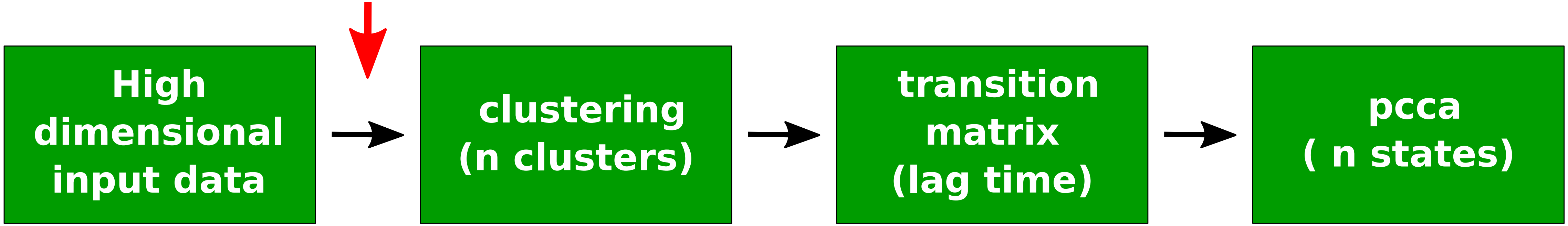
apo-mt ⇄ bound-wt

JBC-2022, 298(10), 102399



residues based on top FI

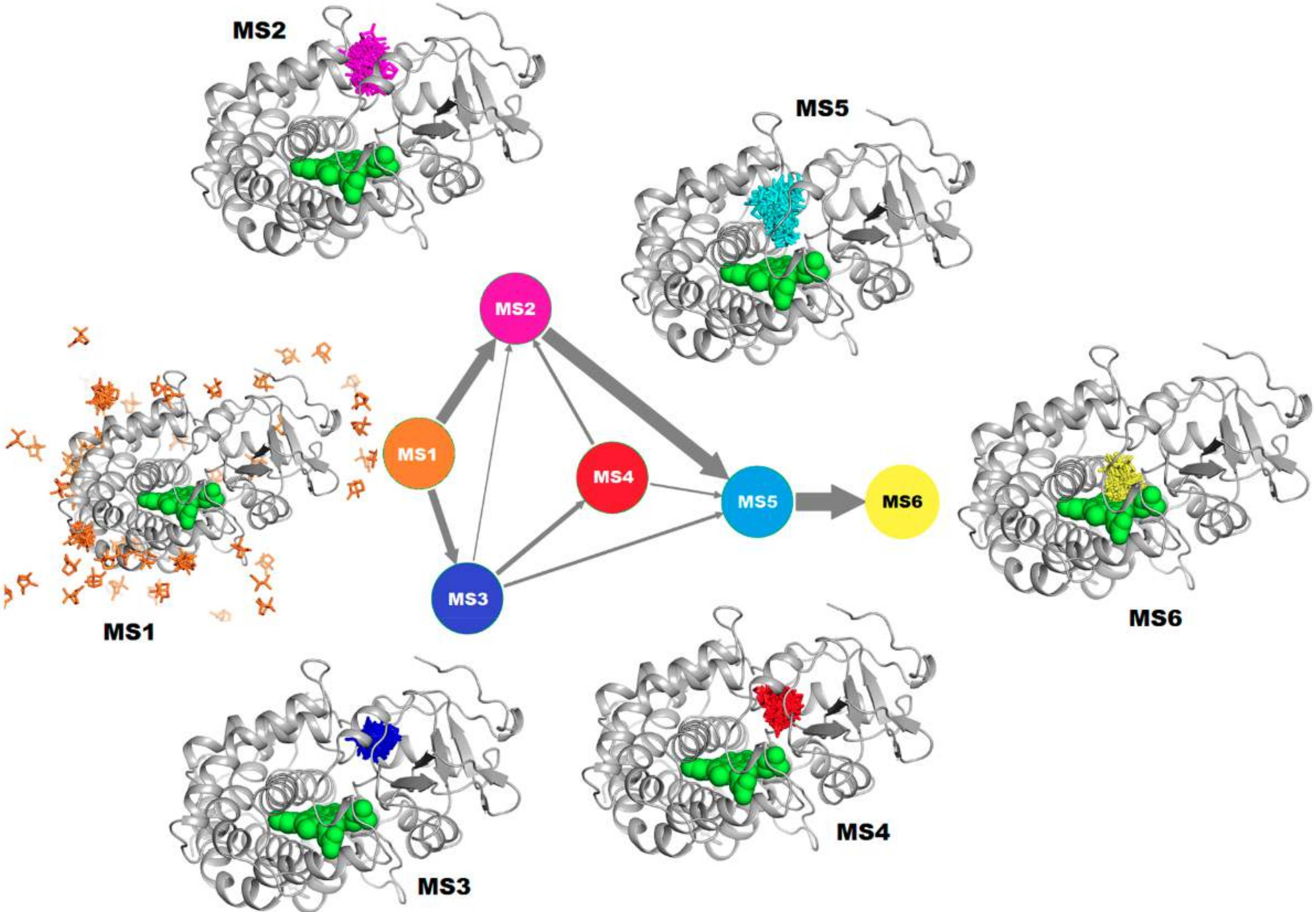
Can we build better MSM with urf ???



Can we build better MSM with urf ???

High dimensional input data	clustering (n clusters)	transition matrix (lag time)	pcca (n states)
	20, 30, 50, 70, 100, 150, 250, 350, 500, 700, 1000, 1500	5, 10 .. 50	3, 4, 5, 6

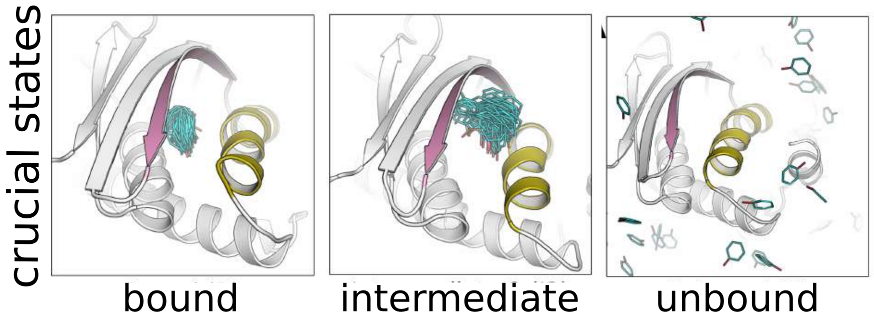
an example MSM

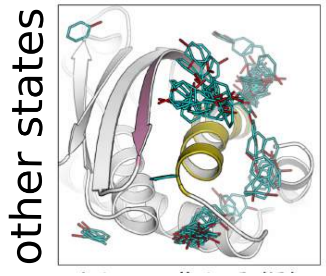


JACS-2018, 140(50), 17743-17752

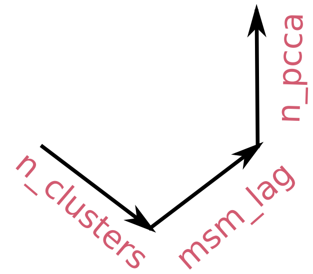
Prelimnary analysis indicates relatively better MSM with Urf

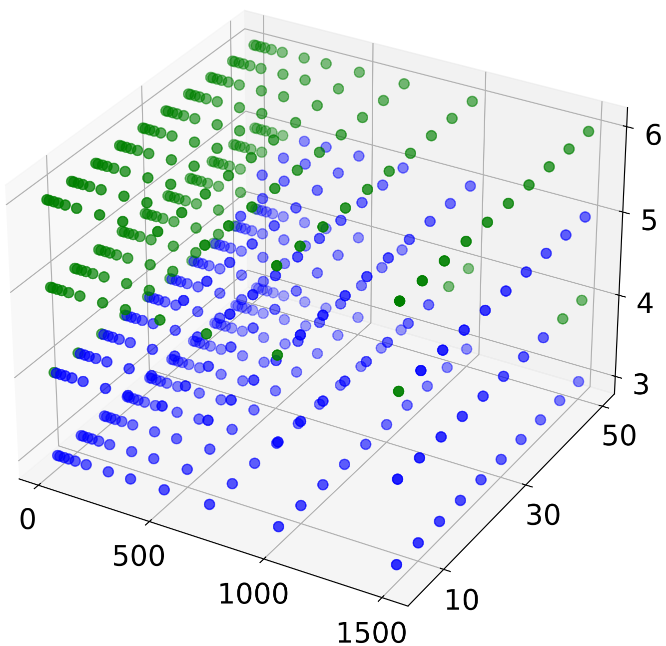
Approach-1: Detecting existence of crucial states





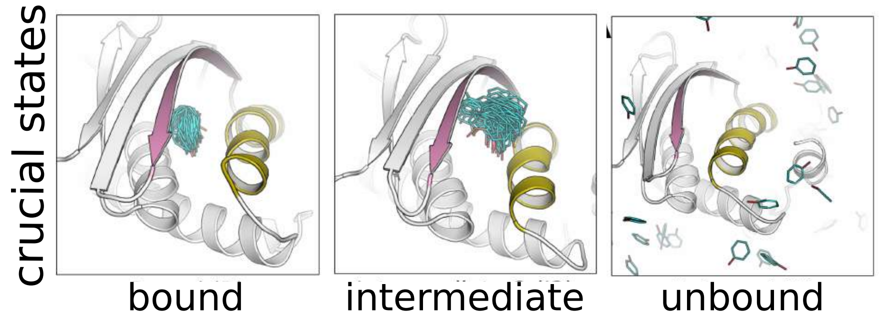
- only 1 state resolved
- any 2 states are resolved
- all three states are resolved

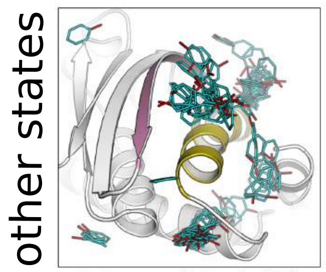


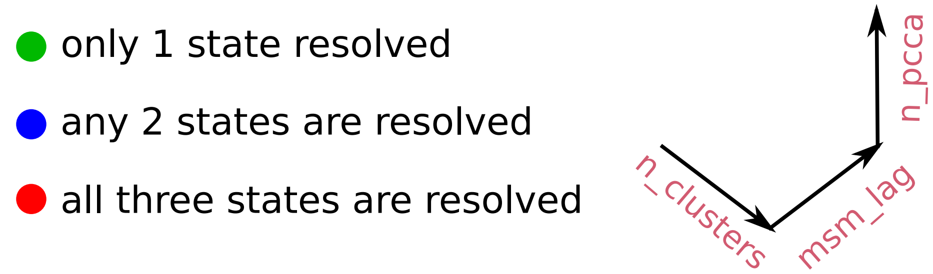


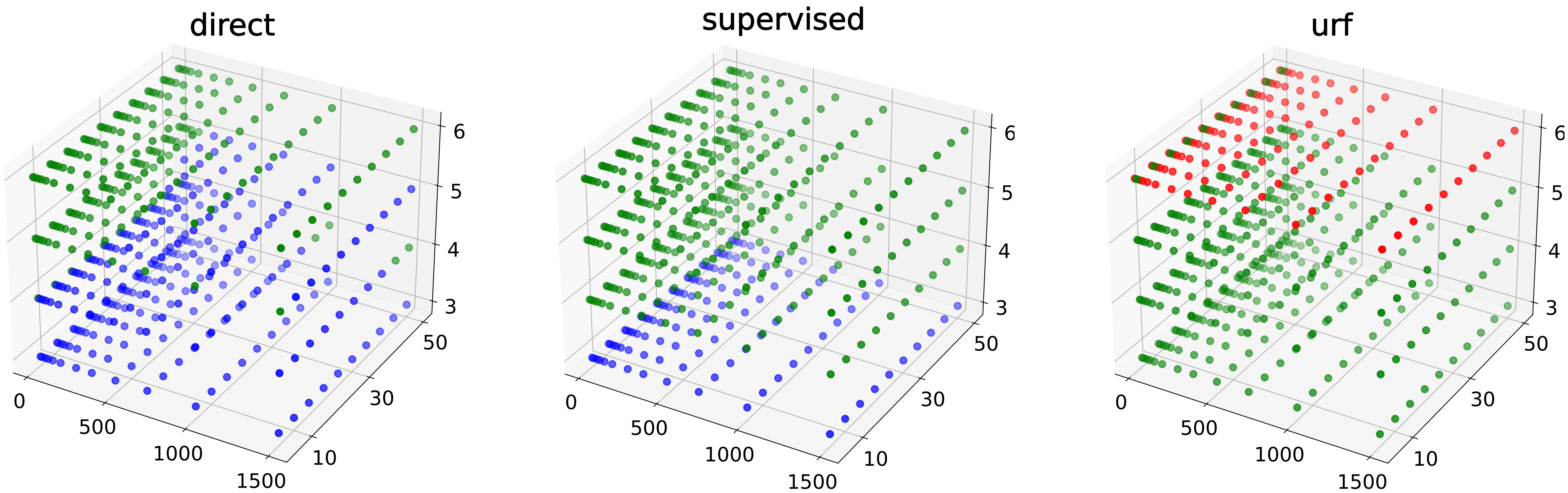
Prelimnary analysis indicates relatively better MSM with Urf

Approach-1: Detecting existence of crucial states



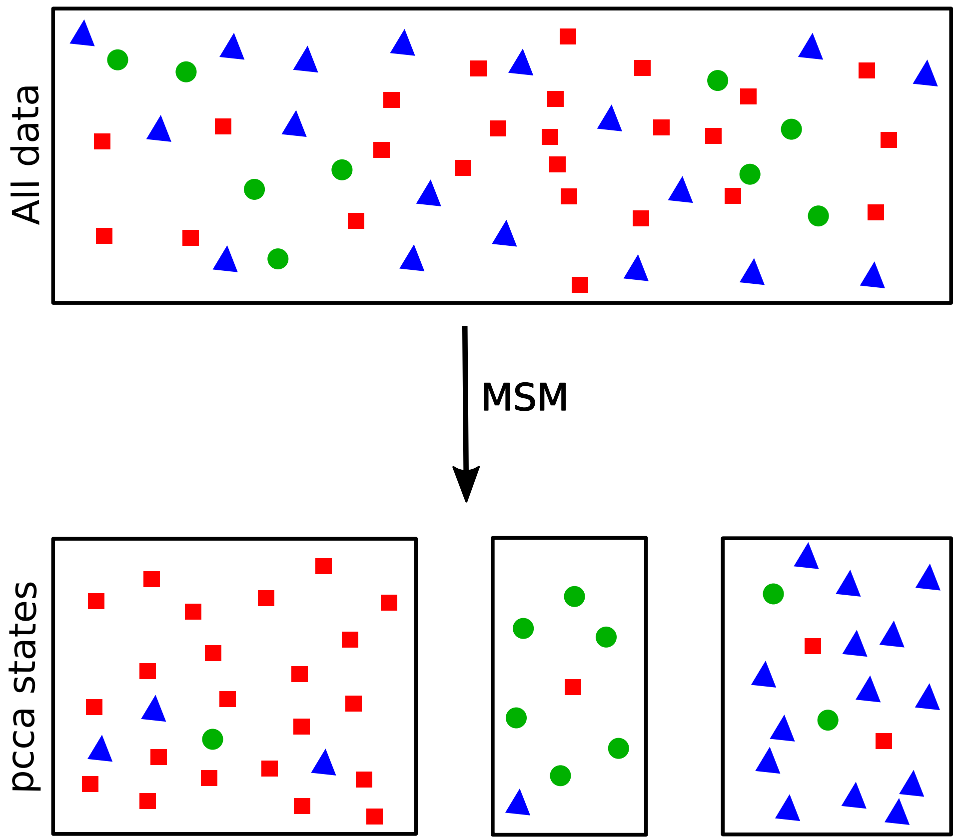






Its on-going ....

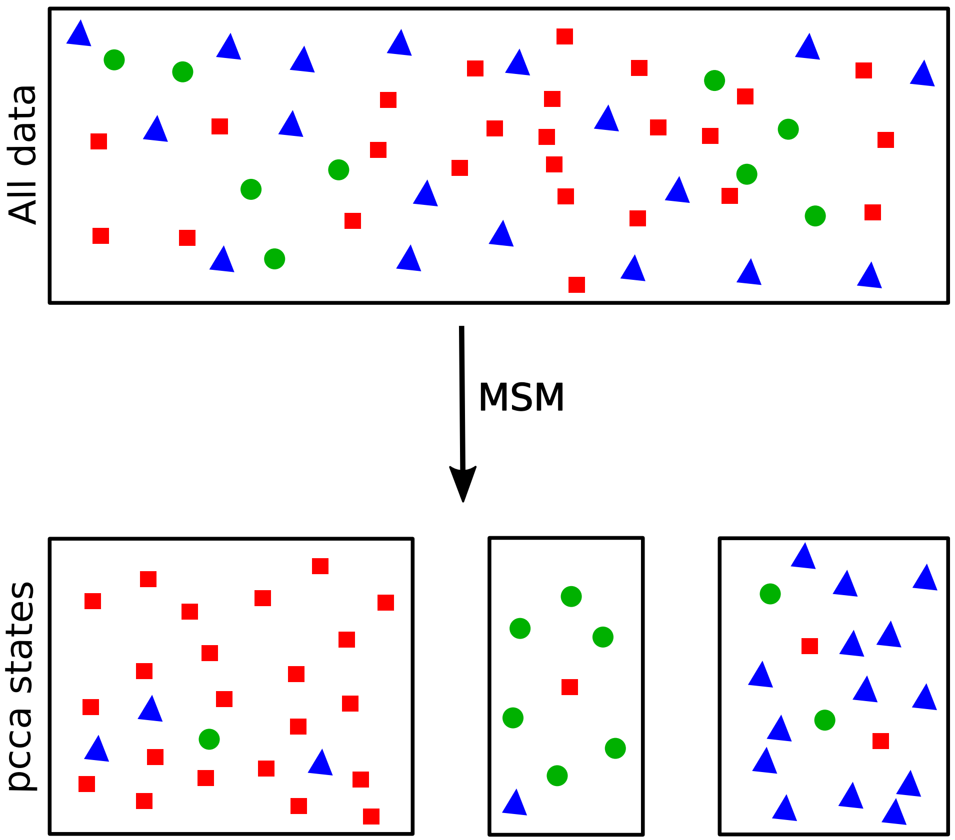
Approach-2: Measuring the impurity in MSM generated metastable states



$$gi = 1 - \Sigma(p_i^2)$$

Its on-going ....

Approach-2: Measuring the impurity in MSM generated metastable states



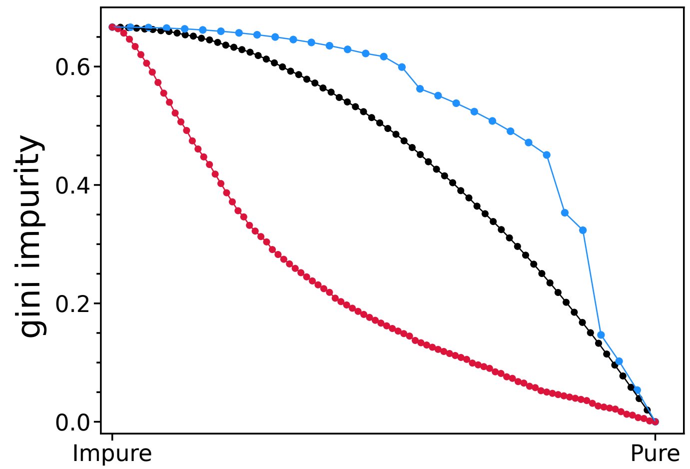
$$gi = 1 - \Sigma(p_i^2)$$

$$wp = \frac{e*p/b}{\Sigma e*p/b}$$

$$gi = 1 - \Sigma(wp_i^2)$$

e = [0.33, 0.33, 0.33]

b = [w1, w2, w3]



[0.33, 0.34, 0.33] → [1.0, 0, 0]

[0.7, 0.05, 0.25] → [1.0, 0, 0]

[0.7, 0.05, 0.25] → [0, 1.0, 0]

Thanks