THE UNIVERSITY of EDINBURGH

# Time Series Analysis with Markov State Models

**Antonia Mey**
School of Chemistry
University of Edinburgh
antonia.mey@ed.ac.uk

ppxasjsm
@ppxasjsm

PyData Meetup

06/09/2018

# About me

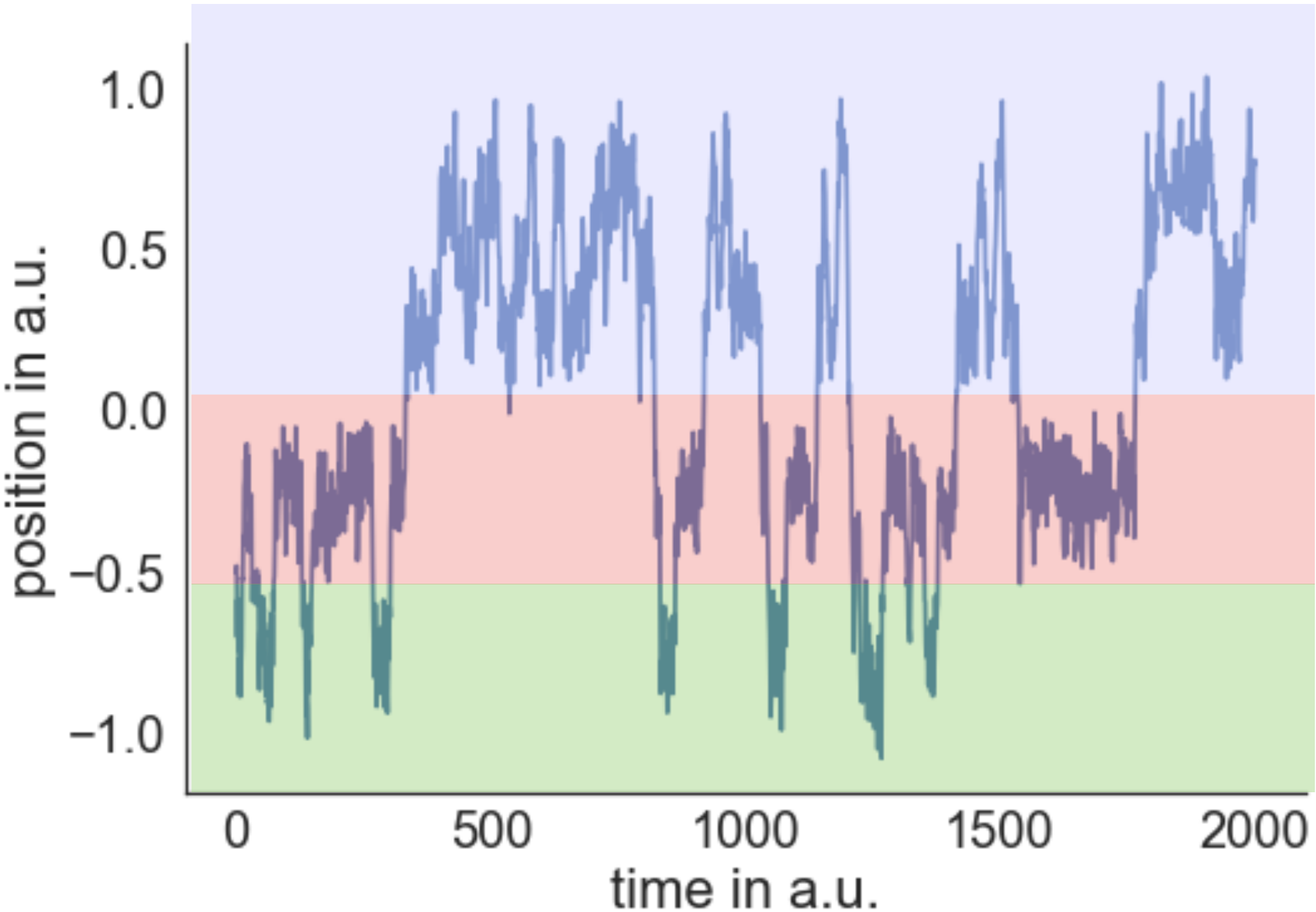
2013 Ph.D. in Physics at the University of Nottingham
2013-2015 Researcher at Freie Universität Berlin (Department of Mathematics)
2015-now Researcher at University of Edinburgh (School of Chemistry)


What am I?

A computational —

insert science related word

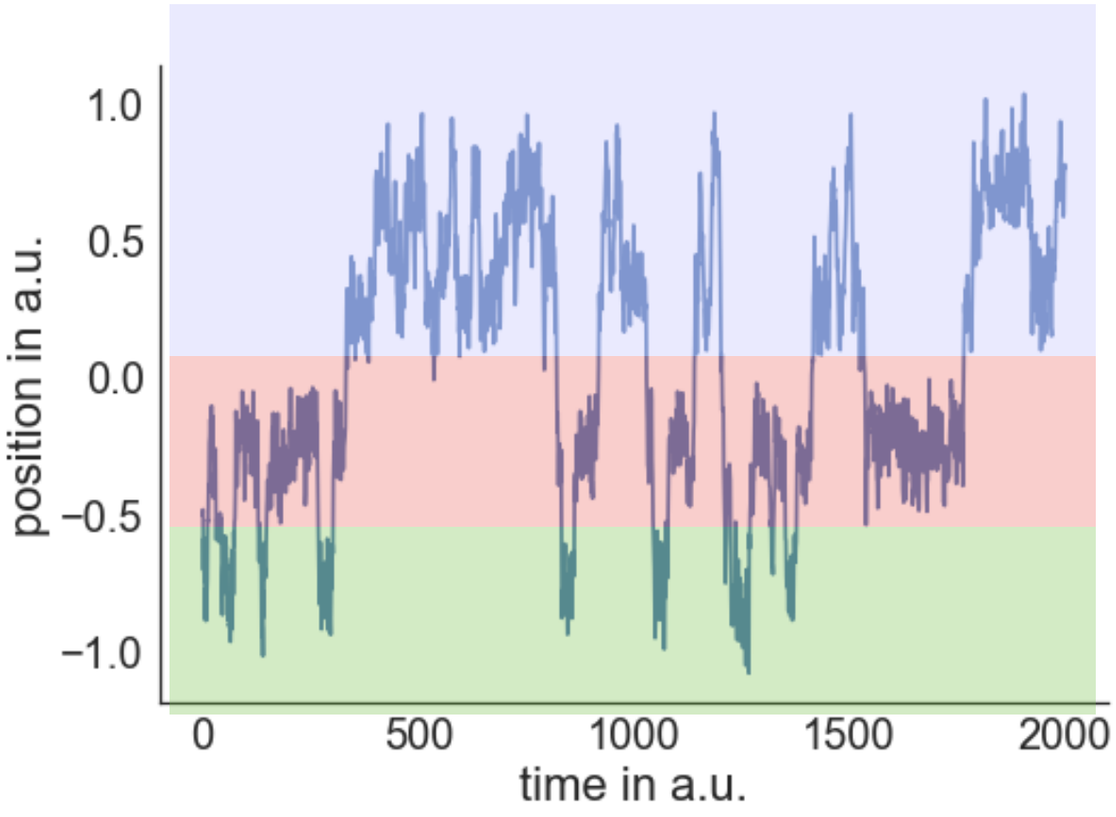
# Timeseries, what now?

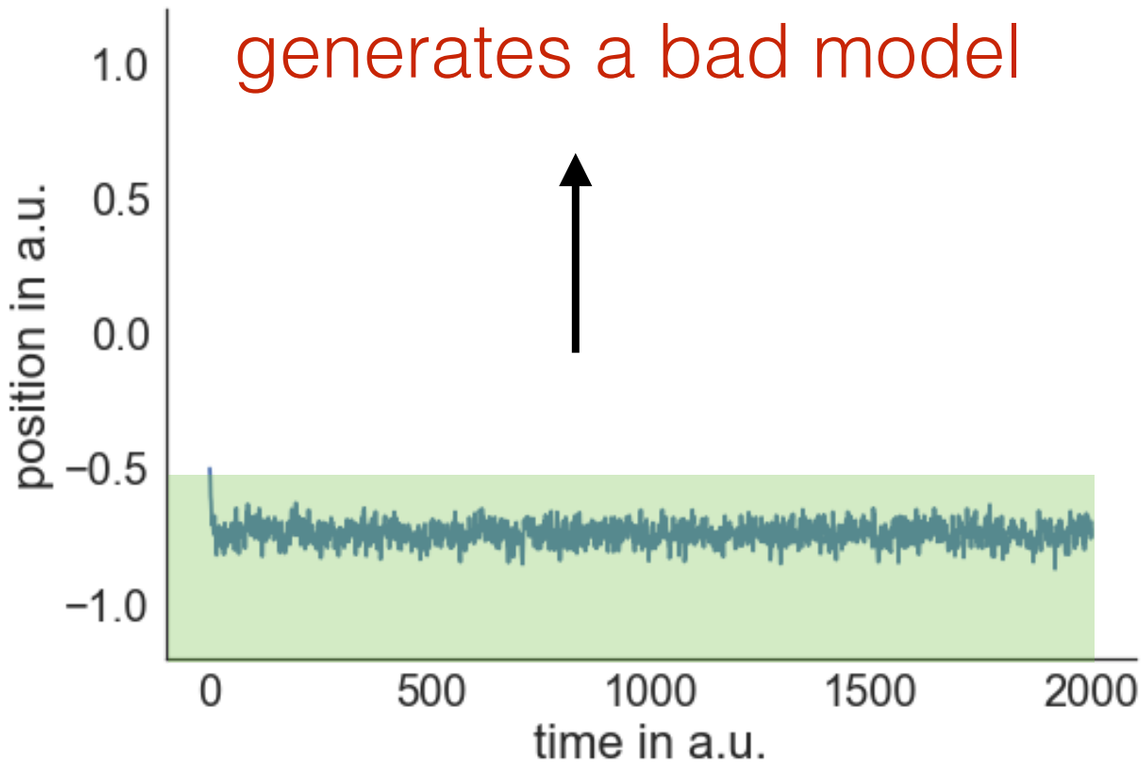



Suppose we have a timeseries and want to identify patterns and then use a model to make future predictions about this timeseries.

One option: model based analysis using a **Markov Model, or Hidden Markov Model**

# Why a Markov Model approach?




- Estimate the residence time in metastable states
- Mean first passage time between state
- Transition path analysis allows to evaluate dominant paths in the system
- Many short timeseries can be used in an aggregate way avoiding bad models in systems with rare events

# How to Markov Model

The basic assumption is that the timeseries can be represented by a Markov Jump process, meaning that the Markov property holds:

$$P(S_i \rightarrow S_j) = P[X_{t+1} \in S_i | X_t \in S_j]$$

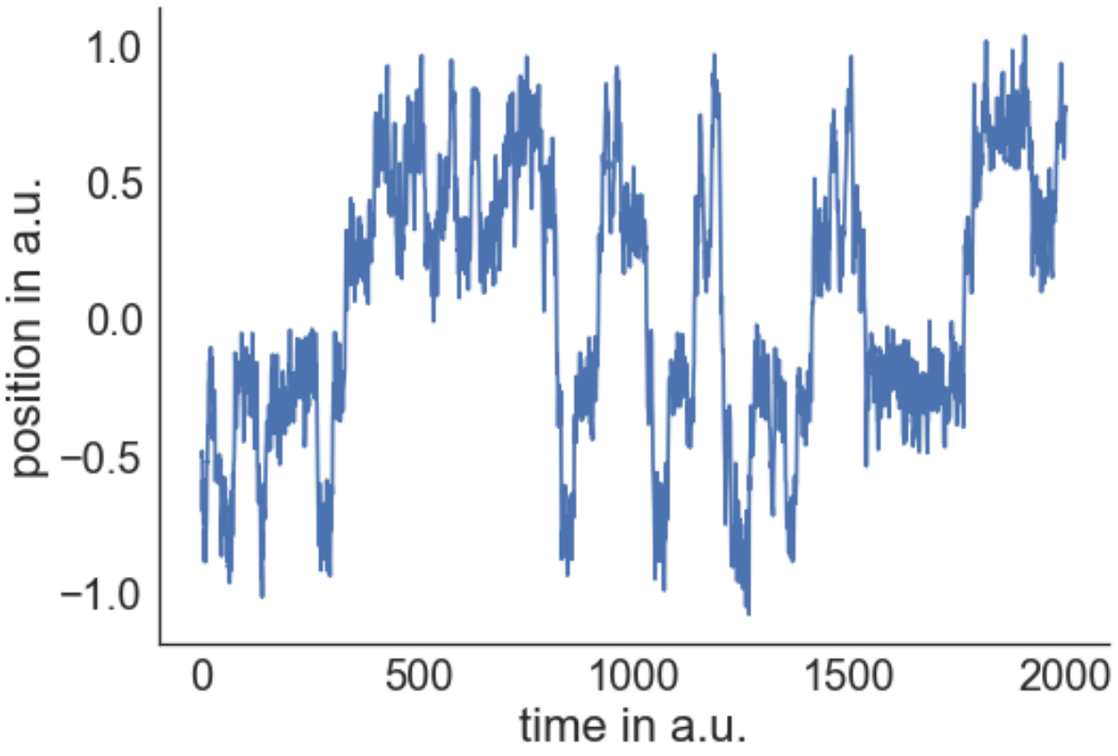
This leads to the following steps of building a Markov model from a timeseries.

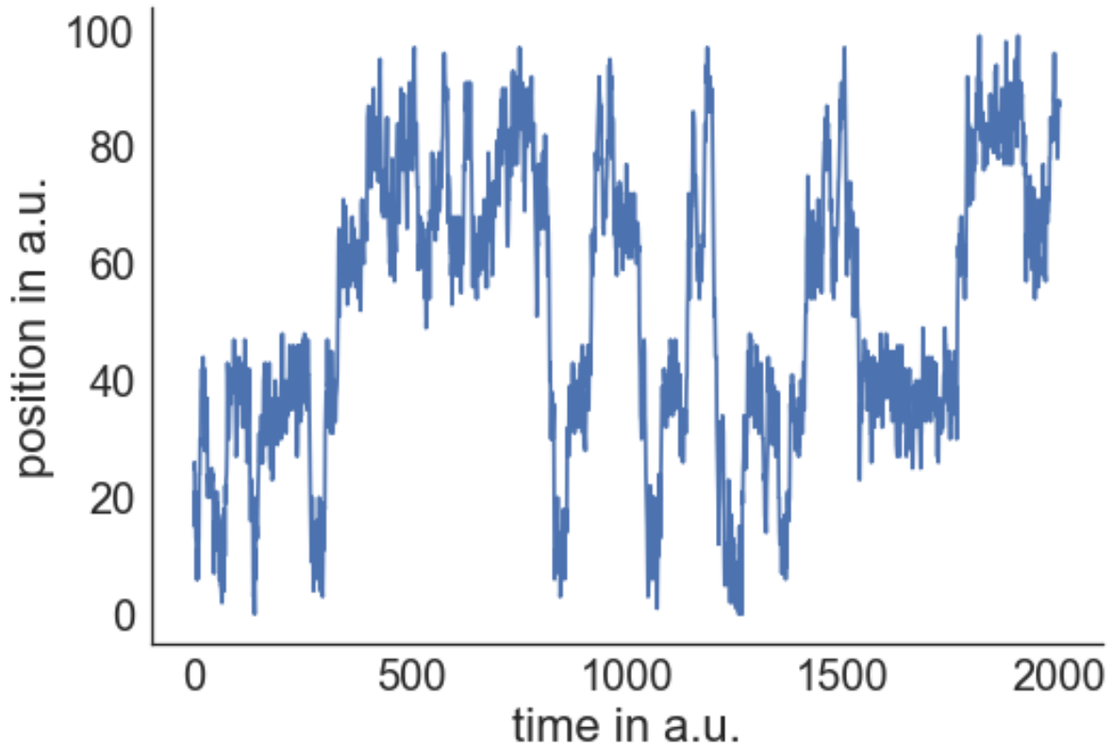
1. Generate timesries/ acquire timeseries
2. Complexity reduction (optional if the system has many dimensions)
3. Discretisation
4. Transition matrix estimation
5. Transition matrix analysis

# 3. Discretisation

## Discretisation

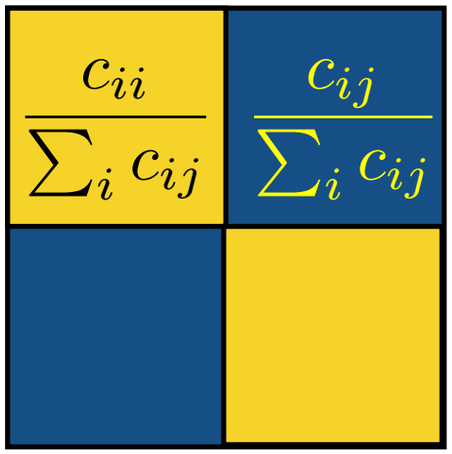
For illustrative purposes: Assume we discretise the position into 100 discrete states



# 4. Transition matrix estimation

## Transition matrix estimation




$= \mathbf{T}$

The transition matrix contains conditional probabilities, of going from state i to state j. Usually, a reversible estimation is used and detailed balance set as a constraint.

$$\pi_i T_{ij} = \pi_j T_{ji}$$

# Reversible estimation

Objective: find the most likely **reversible transition** matrix, based on the **observed counts** using Bayes $\mathbb{P}(\mathbf{T}|\mathbf{C}) = \prod_{i,j} t_{ij}^{c_{ij}}$

We use log-likelihoods instead: $Q = \log \mathbb{P}(\mathbf{T}|\mathbf{C}) = \sum_{i,j} c_{ij} \log t_{ij}$

Maximise the log-likelihood, by taking its derivative and using the constraint, that detailed balance must hold, i.e. $\frac{\partial Q}{\partial x_{ij}} = 0$ and a variable transform $t_{ij} = \frac{x_{ij}}{\sum_k^n x_{ik}}$

$$\frac{\partial Q}{\partial x_{ji}} = \frac{c_{ij} + c_{ji}}{x_{ji}} - \frac{c_i}{x_i} - \frac{c_j}{x_j}$$

set to 0 and the entries can be iterated to convergence.

$$x_{ji} = \frac{c_{ij} + c_{ji}}{\frac{c_i}{x_i} + \frac{c_j}{x_j}} \longrightarrow$$ self consistent iterative update of $x_{ij}$

$x_i = \sum_k x_{ki}$ which is incidentally also the stationary probability of state i.

# Markov Model Analysis I

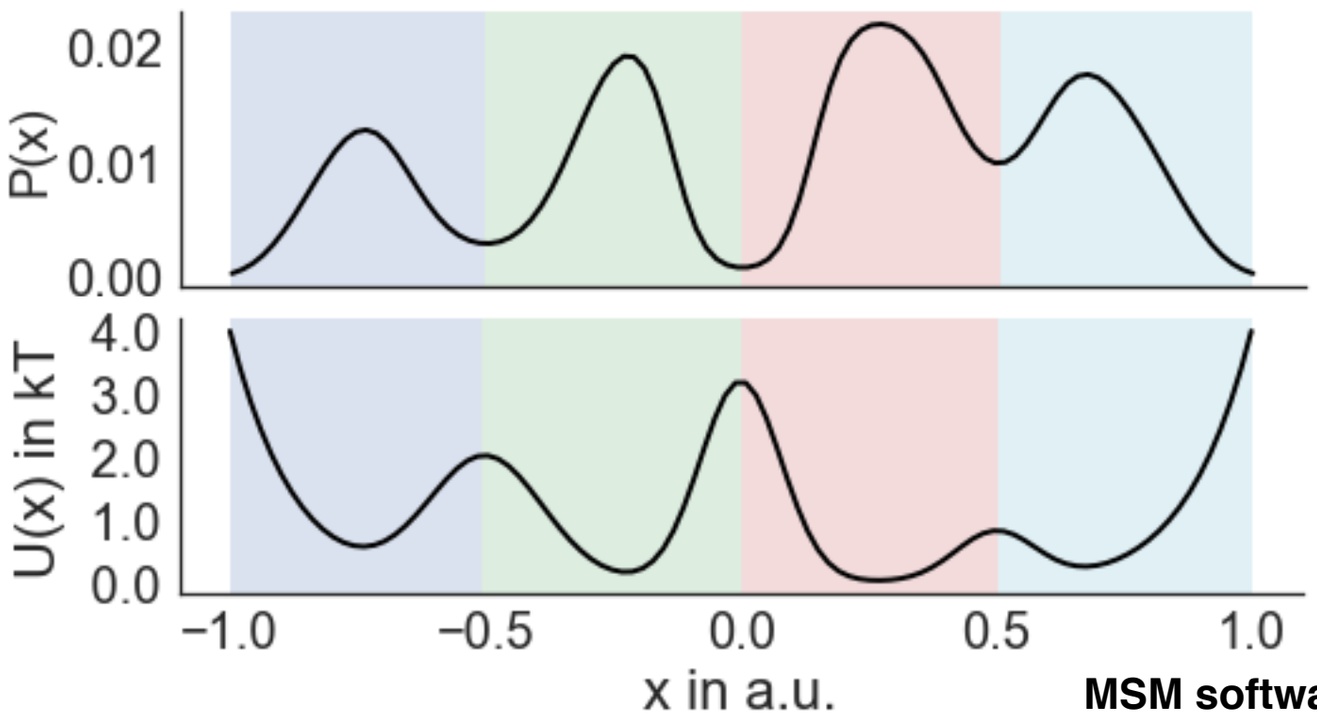
## Transition matrix analysis - stationary properties

For all transition matrices holds:

$\lambda_1 = 1 > \lambda_2 > \lambda_3, \ldots, > \lambda_n$ ⟶ dynamics

↘ stationary behaviour $\mathbf{T}^{\mathrm{T}}\mathbf{l}_1 = 1\mathbf{l}_1$

Also: *MFPT*, can be computed from the transition matrix and the stationary distribution.




stationary distribution

potential

**MSM software:**
Senne, Trendelkamp-Schroer, **Mey**, Schütte, Noé, JCTC, 8, 2223 (2012)
**http:simtk.org/home/emma**

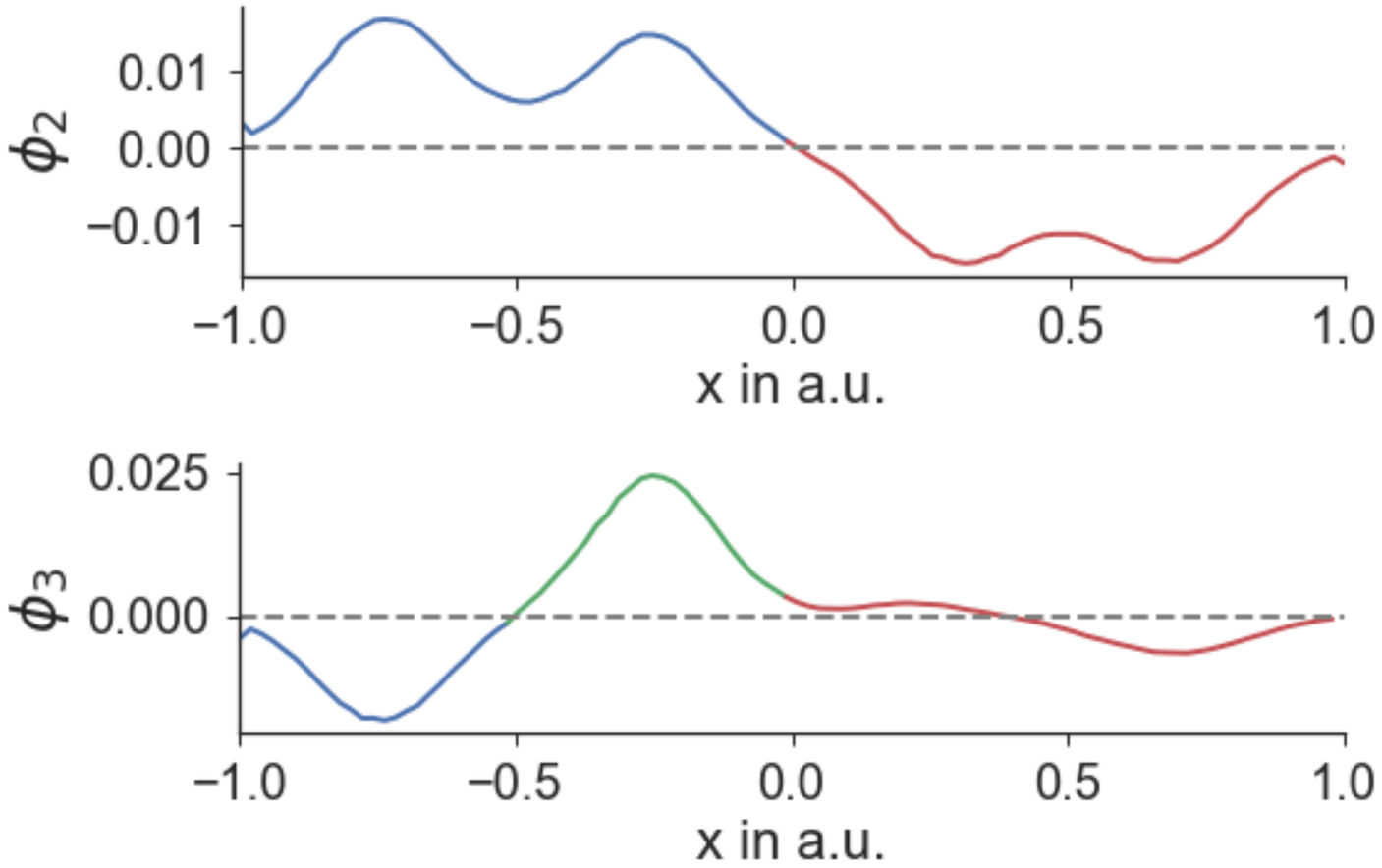
Prinz, J.-H. et al (2011) Markov models of molecular kinetics: Generation and Validation. J. Chem. Phys., 134 . p. 174105.

# Markov Model Analysis II

## Transition matrix analysis - dynamic properties

For all transition matrices holds:

$\lambda_1 = 1 > \boxed{\lambda_2 > \lambda_3, \ldots, > \lambda_n} \longrightarrow$ dynamics

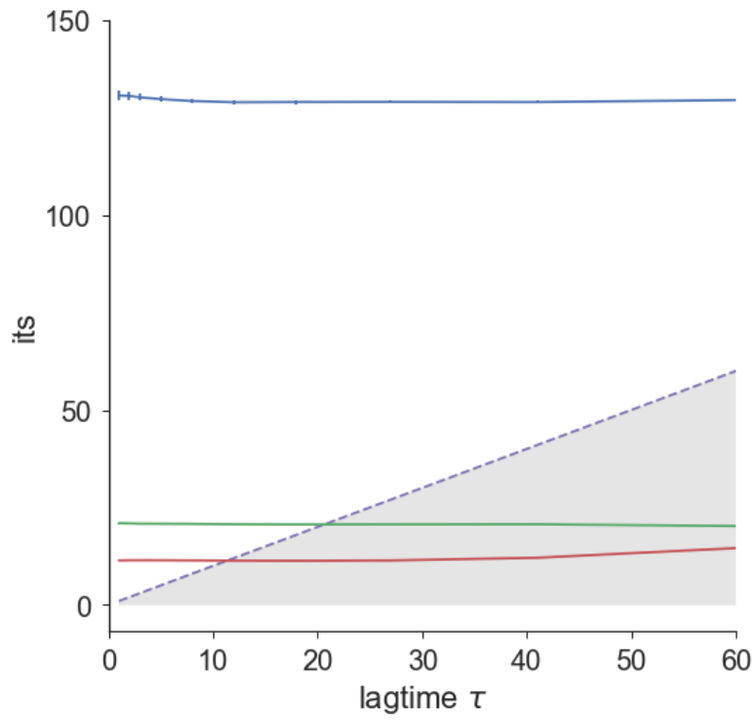



$$\mathbf{T}\mathbf{r}_i = \lambda_i \mathbf{r}_i$$

$$\mathbf{T}^\top \mathbf{l}_i = \lambda_i \mathbf{l}_i$$

$$\boxed{t_i = -\frac{\tau}{\ln \lambda_i}}$$

$$k = t_i^{-1}$$



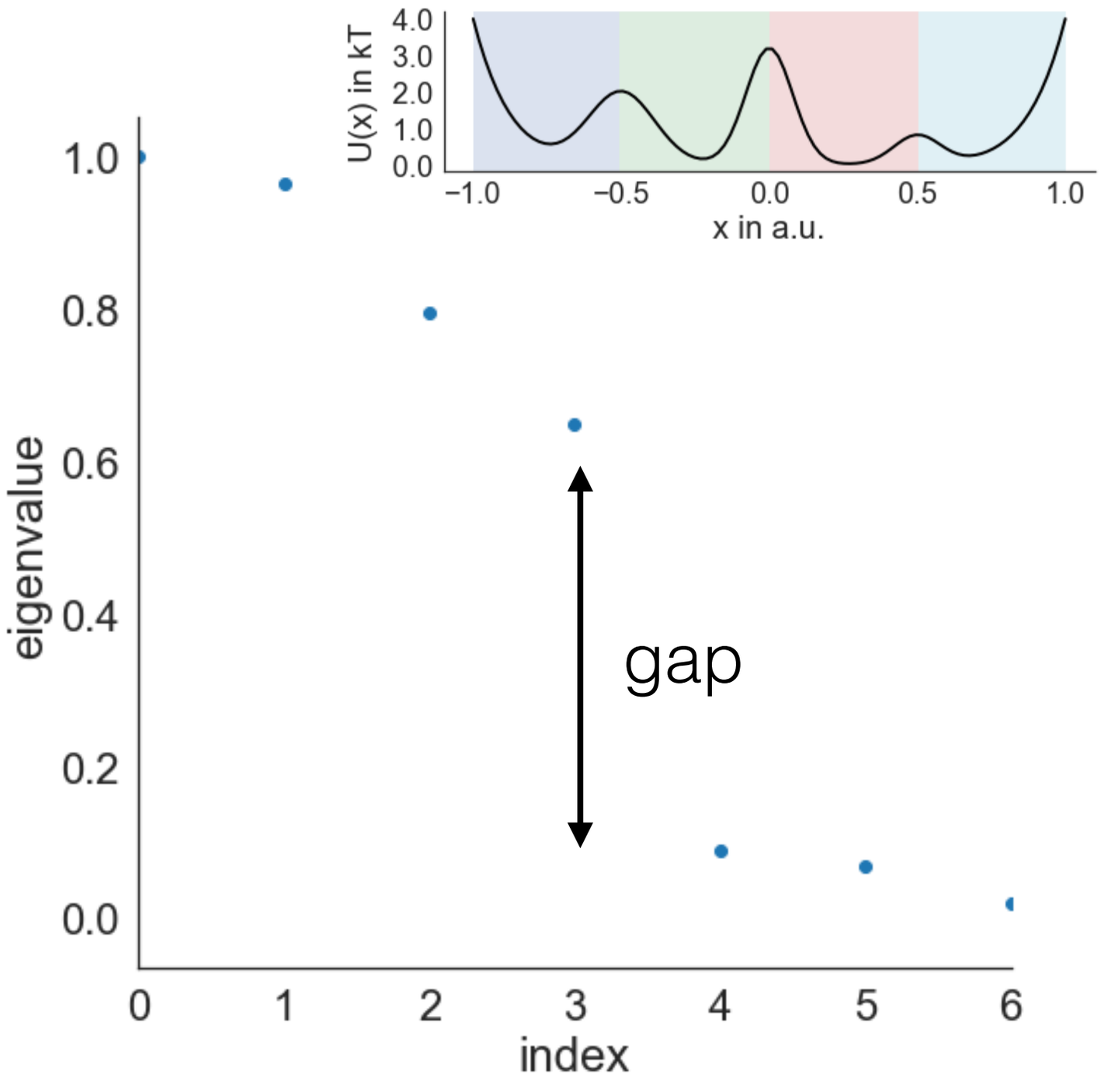

**MSM software:**
Senne, Trendelkamp-Schroer, **Mey**, Schütte, Noé, JCTC, 8, 2223 (2012)
**http:simtk.org/home/emma**
**https://github/markovmodel/pyemma**

Prinz, J.-H. et al (2011) Markov models of molecular kinetics: Generation and Validation. J. Chem. Phys., 134 . p. 174105.

# PyEMMA demo

Go to Jupyter notebook

# Coarse graining




System can be described in terms of slow dynamics between 4 wells.

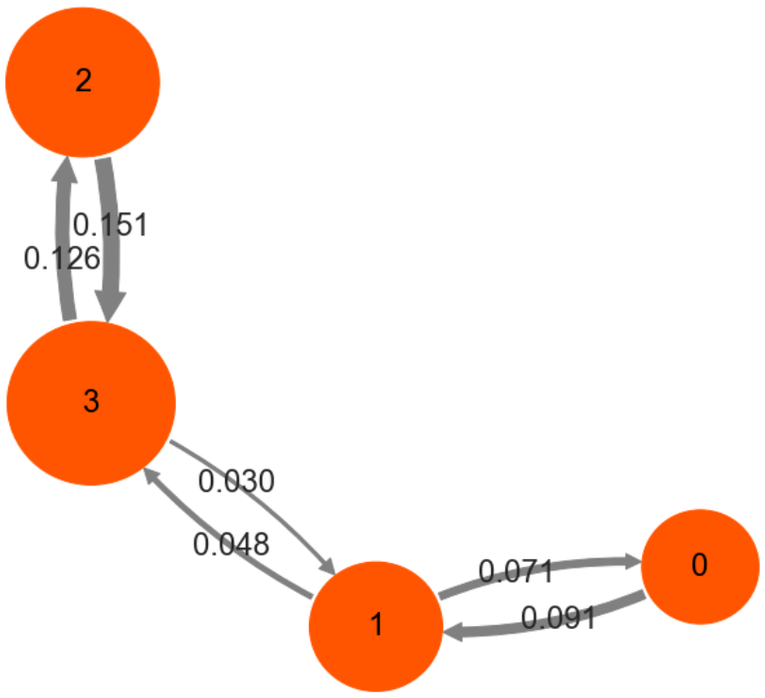
**HMM, PCCA or spectral clustering** can be used to define coarse states.

# Mean first passage time

$$\mathbf{Z} = (\mathbf{I} - \mathbf{P} + \mathbf{\Pi})^{-1}$$

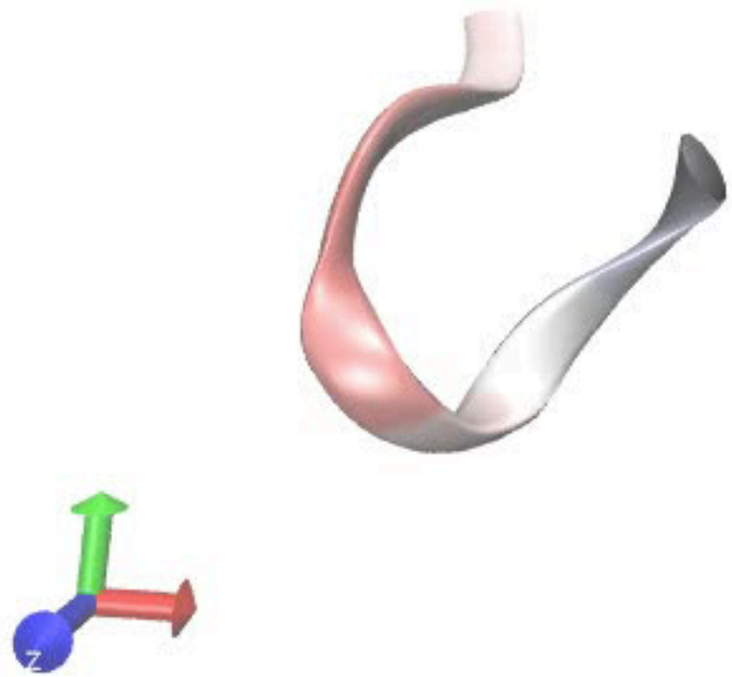
$$m_{ij} = \frac{z_{jj} - z_{ij}}{\pi_j}$$

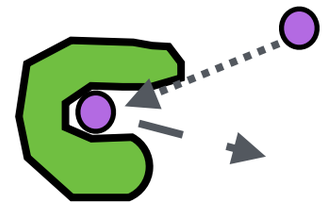
$m_{ij}$ is the mean first passage time of state i to state j.
The inverse of the mean first passage time that is often experimentally measured in biological processes.




Dominant pathways between coarse grained metastable states can be found using transition path theory, allowing to compute net flux between states.

# Molecular Dynamics





**1. Timescale problem:**

Molecular dynamics simulations are relying on integration time steps of 1-4 fs, for accuracy.
$< 10^8$ integration time steps are needed to reach relevant timescales.

**2. Complexity problem:**

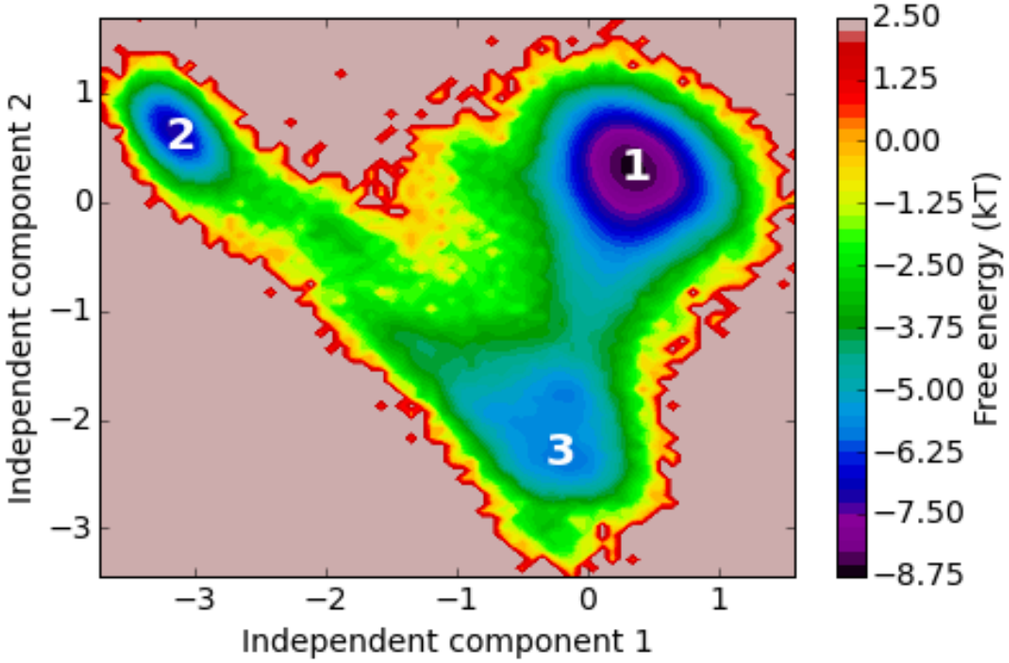
Reduction of 3N coordinates in order to make data more manageable/quantitative:

- clustering of relevant states
- Minimize information loss with the clustering

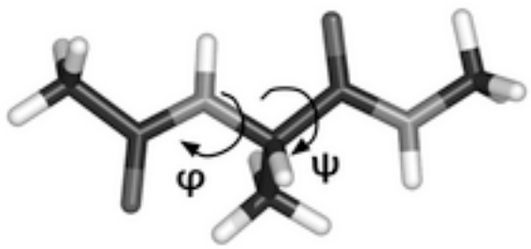
# Dimensionality reduction with TICA

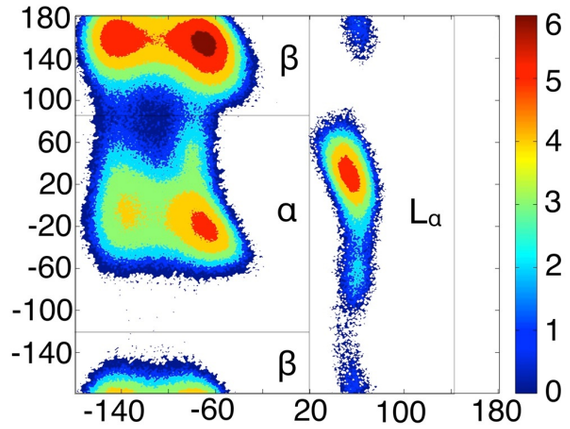
## Complexity reduction and feature selection

No point in clustering on 3N dimensions, there just aren't enough data points

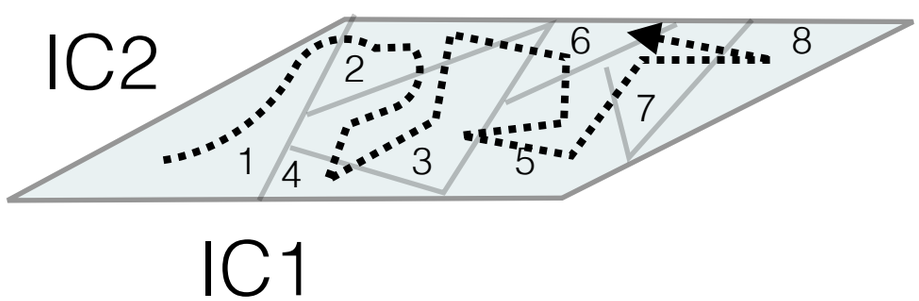



**Dimensionality** reduction using TICA (time lagged independent component analysis) or direct feature selection




## Cluster with your favourite clustering algorithm! (e.g. k-means)

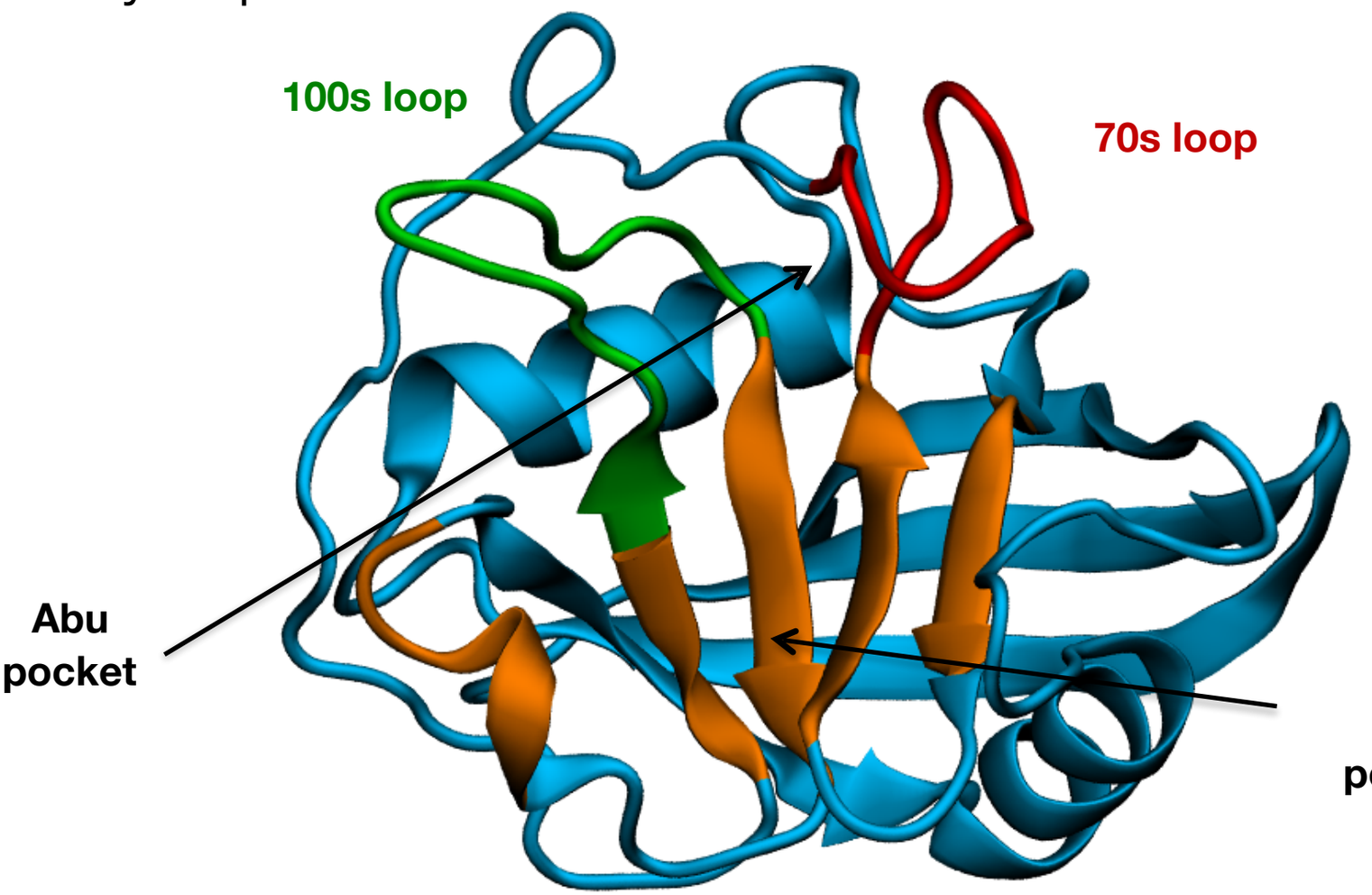



M Schor, **ASJS Mey**, CE MacPhee
Biophysical reviews 8 (4), 429-439

https://arxiv.org/pdf/1506.06259.pdf

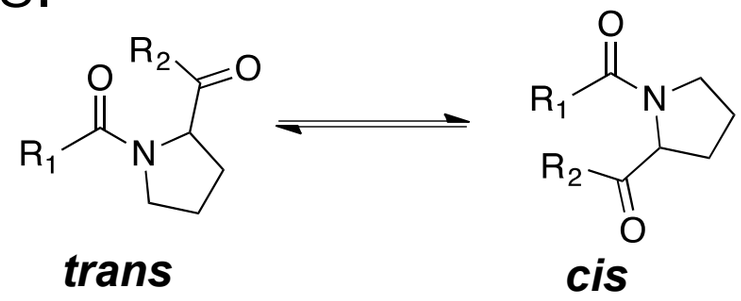
# Cyclophilin A — a heavily studied enzyme

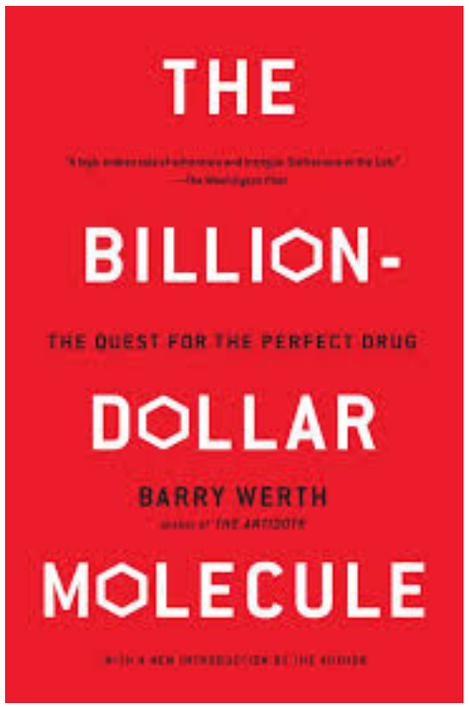
Cyclophilin A:




**Cyclophilin A:**

Hepatitis B, C and HI virus will ‘hijack’ CypA to use as a Chaperone in new virus capsids.

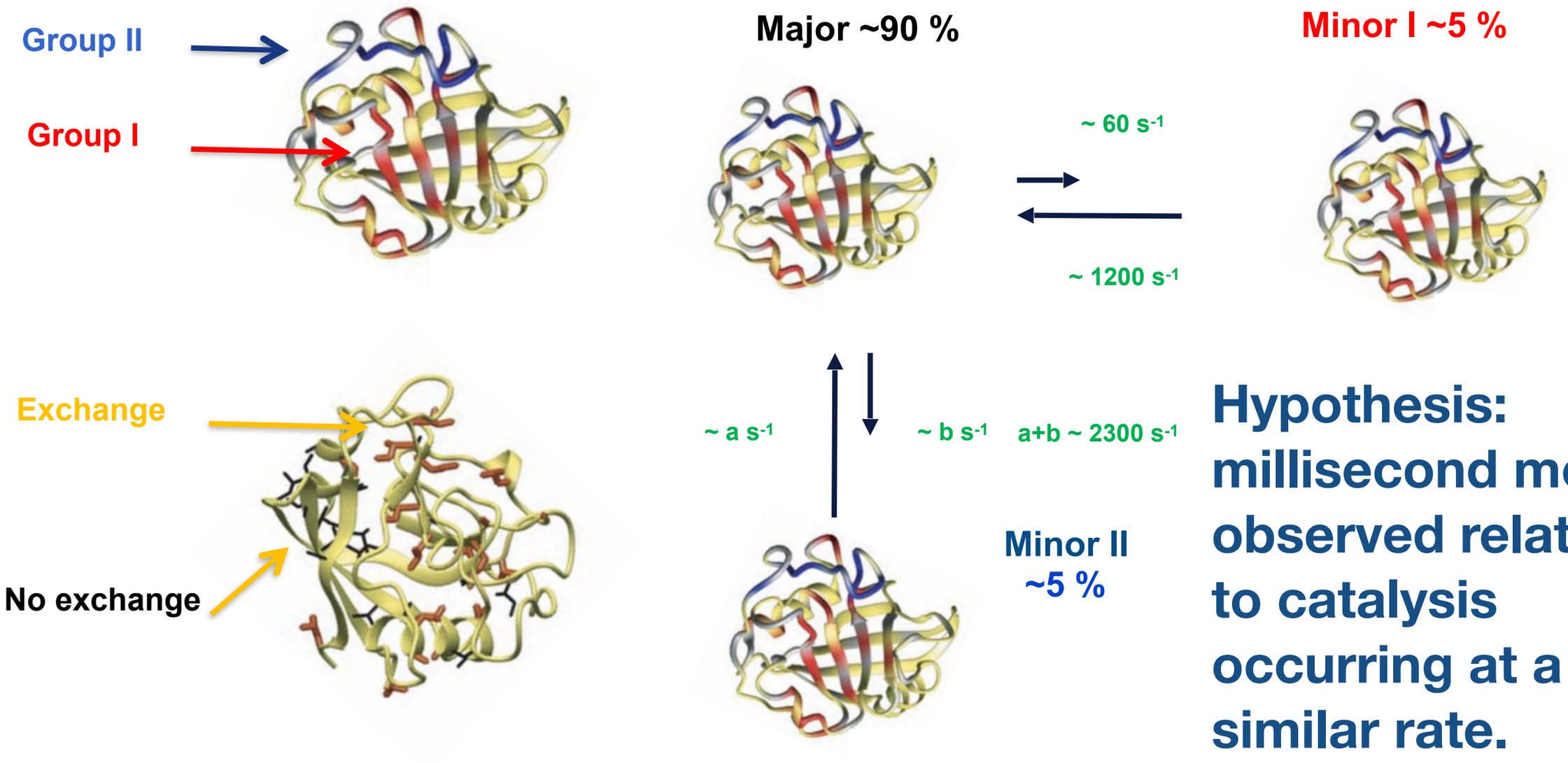



- 17 different isoforms with different functionality
- no isoform specific drug

# Cyclophilin A — a heavily studied enzyme

à NMR relaxation measurements on apo CypA show existence of two slow processes:

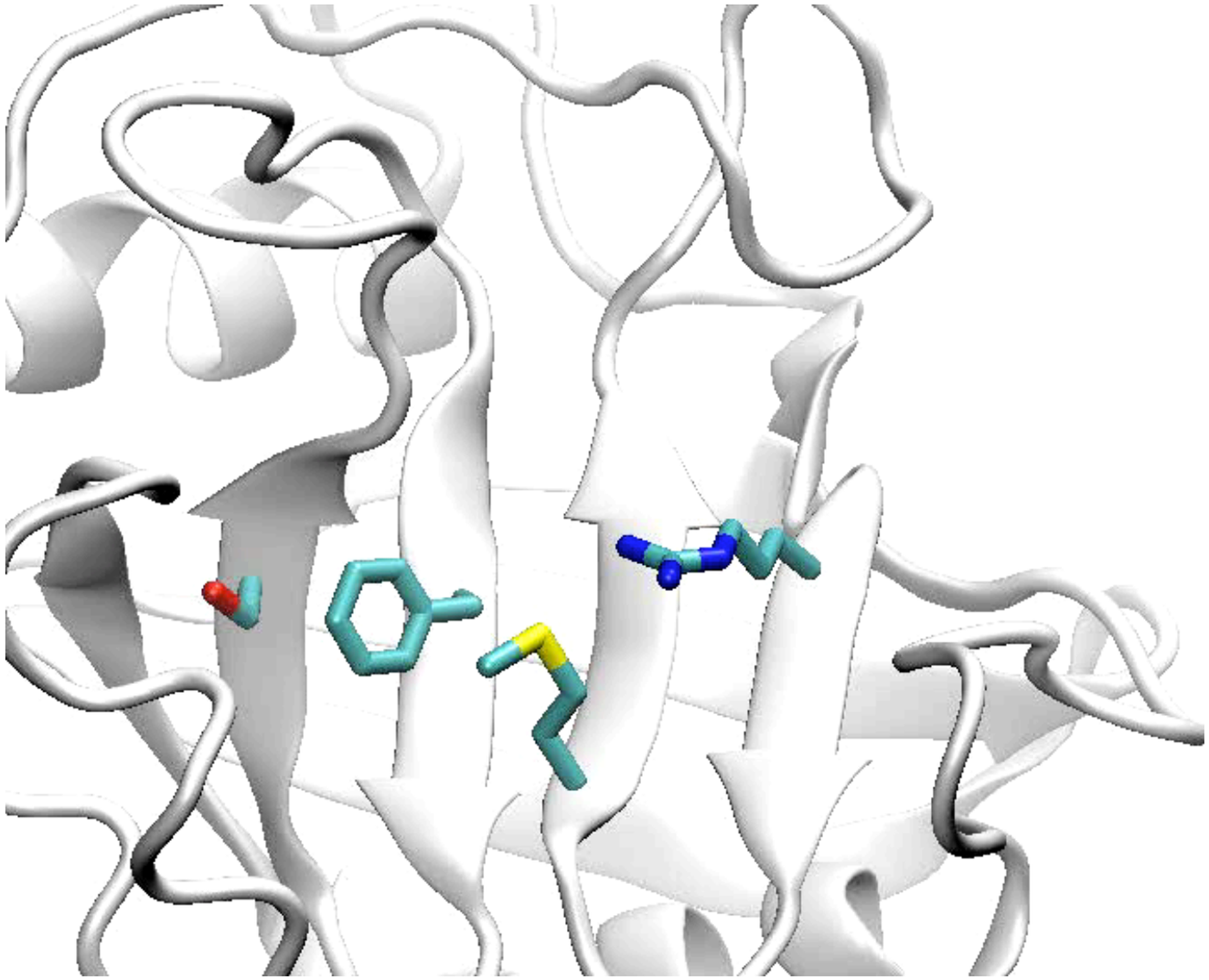


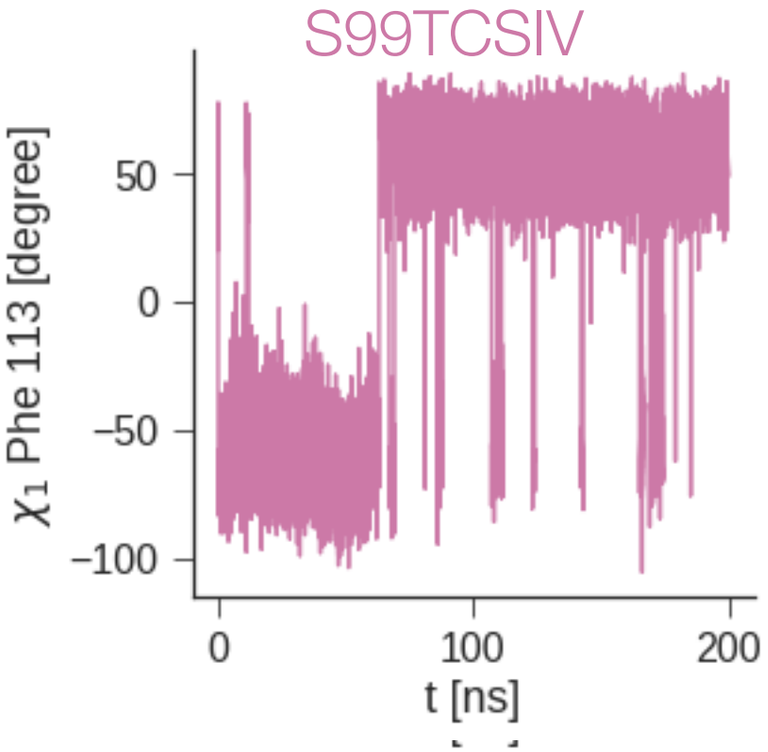

**Hypothesis: millisecond motion observed relates to catalysis occurring at a similar rate.**

Eisenmesser et al. Science (2002)
Eisenmesser et al. Nature (2005)

-> Similar processes observed in substrate-bound CypA

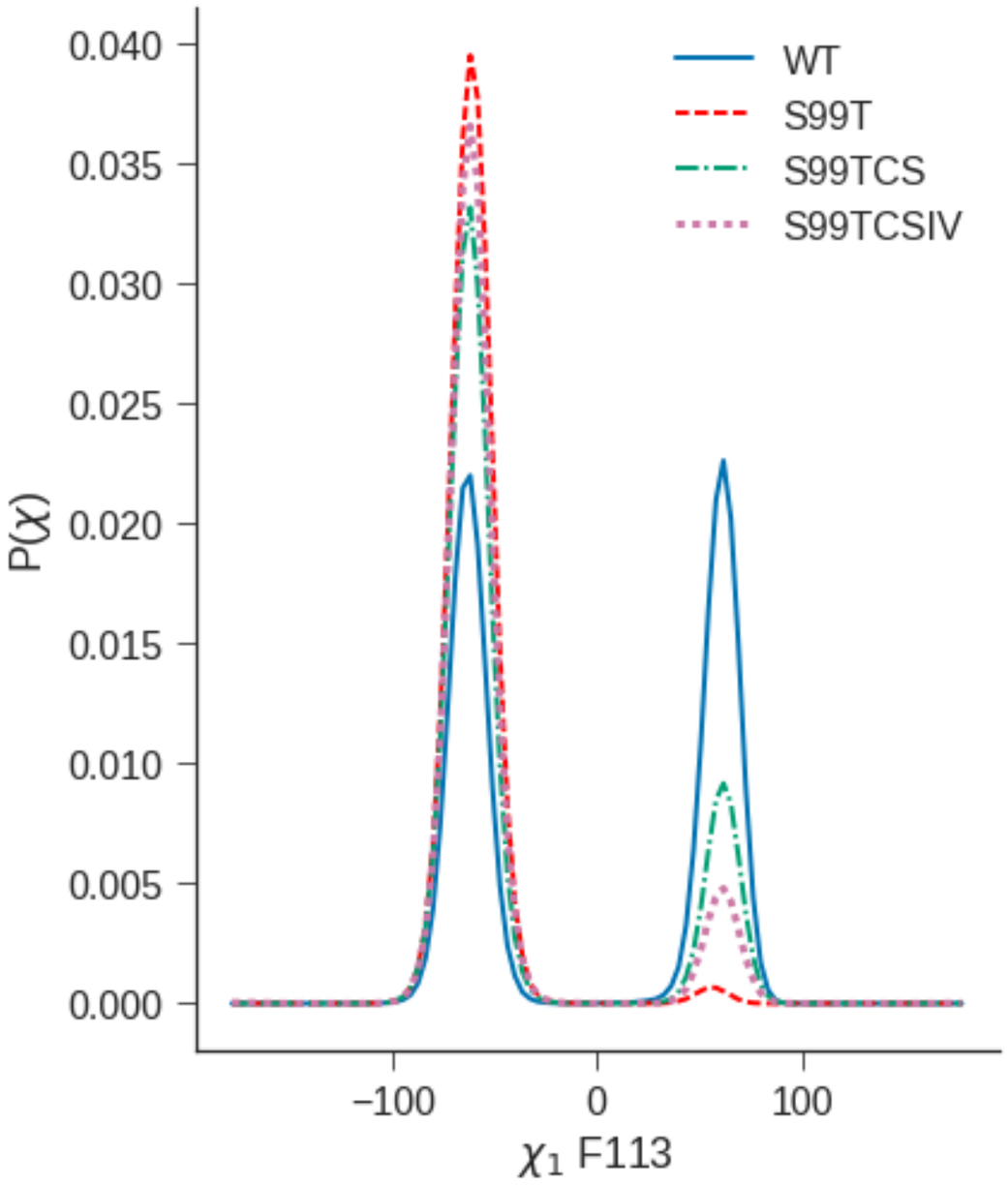
# Dimensionality reduction with TICA





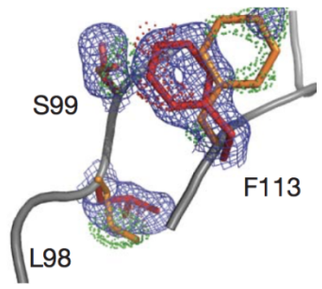
# Dimensionality reduction with TICA

Stationary properties from MSM:




MSM on side chain dihedrals of residues that form part of group 1 residues and include Phe 113, $\chi_1$ $\chi_3$ Arg55, $\chi_1$ $\chi_2$ Met61, Ser/Thr99, Cys115, Ile97, Leu98, Gln 63

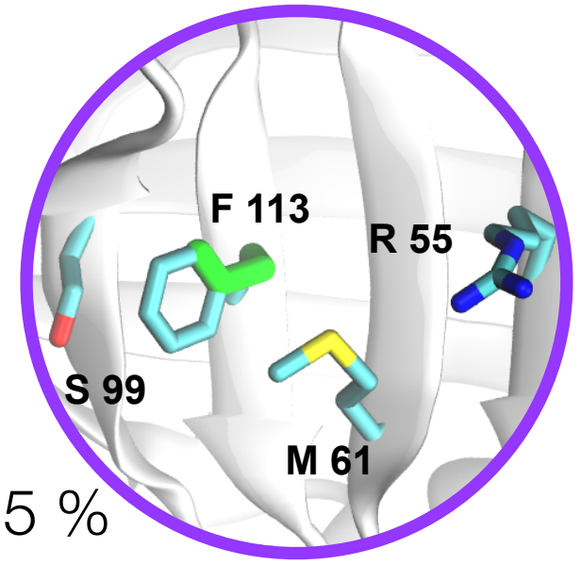
Pop $\chi_1$:




| | MSM | | X-ray occupancy | |
|---|---|---|---|---|
| Pop $\chi_1$: | -60° | +60° | -60° | +60° |
| WT: | 0.51 | 0.49 | 0.37 | 0.63 |
| S99T: | 0.99 | 0.01 | 1 | |
| S99TCS: | 0.8 | 0.2 | 0.79 | 0.21 |
| S99TCSIV: | 0.9 | 0.1 | 1 | |

# Dimensionality reduction with TICA

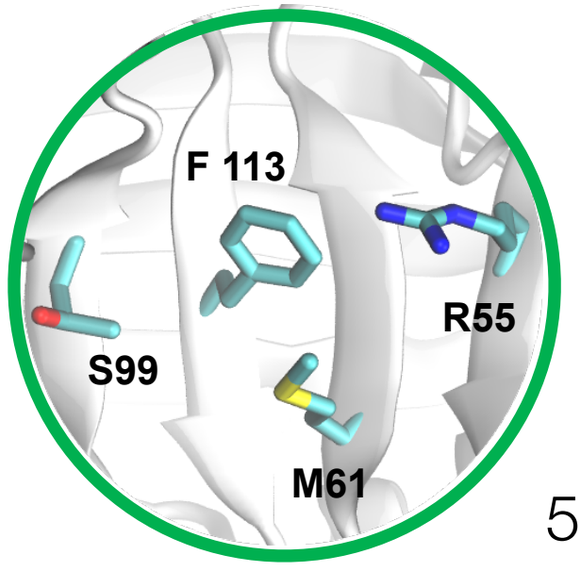
**WT**




49 ± 5 %

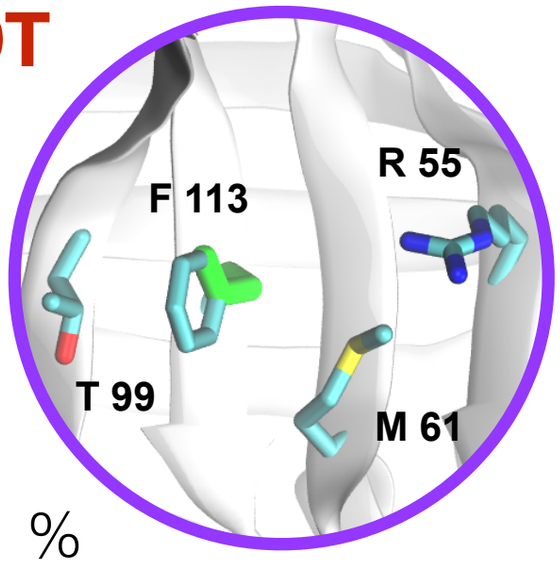
$77 \pm 2\ \mu s^{-1}$

$66 \pm 1\ \mu s^{-1}$




51 ± 5 %

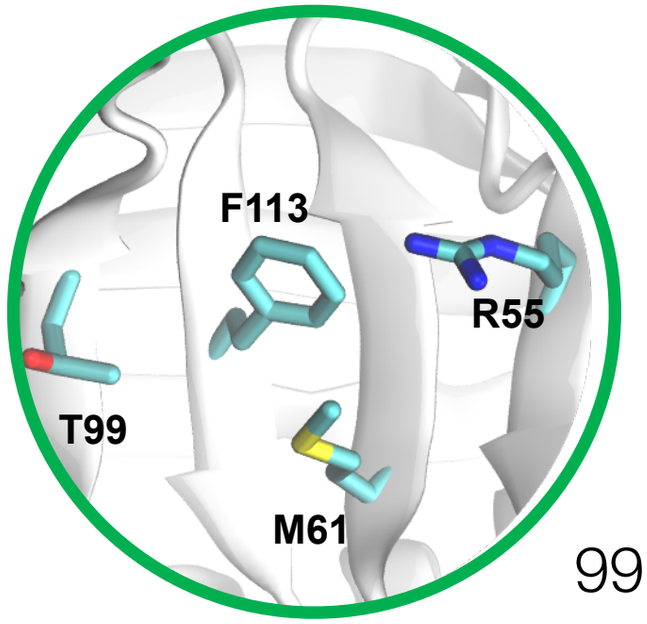
**S99T**




1 ± 1 %

$3.2 \pm 0.2\ \mu s^{-1}$

$310 \pm 10\ \mu s^{-1}$




99 ± 1 %

# Outlook

- Markov models are away to extract equilibrium and dynamic properties from a timeseries.
- In biomolecular simulations this information can be used to guide experiments

Areas of ongoing reasearch

- Use ML to build multiple models and find heuristic for predicting the best model
- Difficult to find a unique metric to measure ‘best’ model
- Feature selection and dimensionality reduction
- Large spread of timescales, makes it difficult to focus on the process you are trying to understand (biomolecules)

# Questions?


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTISTS
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOOKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS SEX SO IMPORTANT
WHY ARE THERE SHARKS IN LAKES
WHY IS THERE PHLEGM
WHY IS YKK ON ALL ZIPPERS
WHY DO IGUANAS DIE
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING
WHY IS GPS FREE
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE LAVA
WHY ARE THERE DINOSAUR GHOSTS
WHY ARE THERE SLAVES IN THE BIBLE
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG