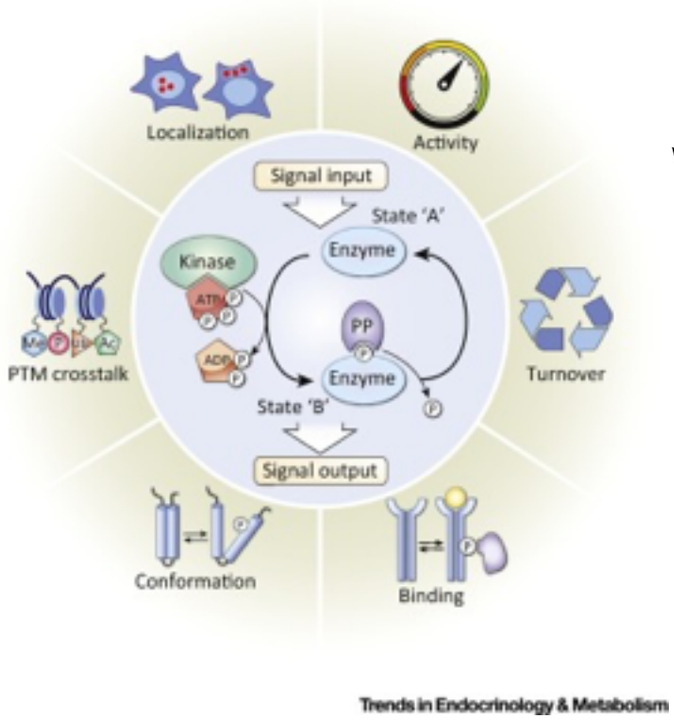
# Discerning protein activity regulation mechanism using "inverse" machine learning

**Sameer Varma**
University of South Florida, Tampa

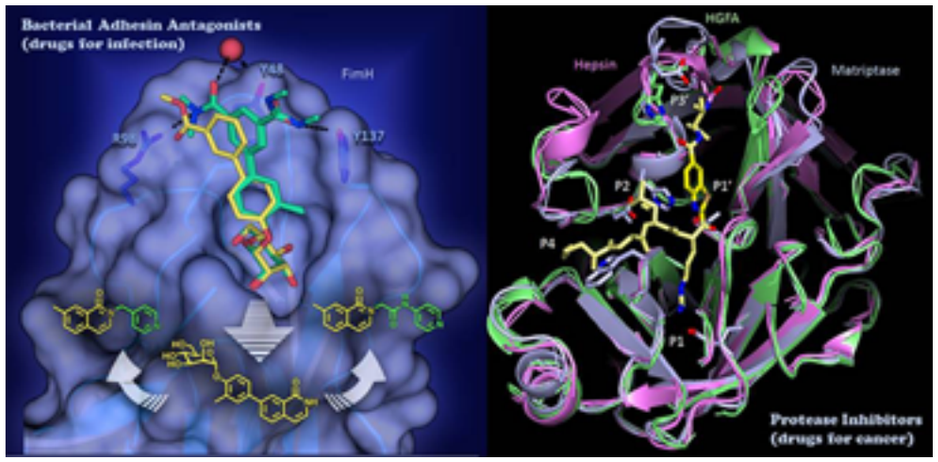
# Protein function

Molecular mechanism

## Cellular role

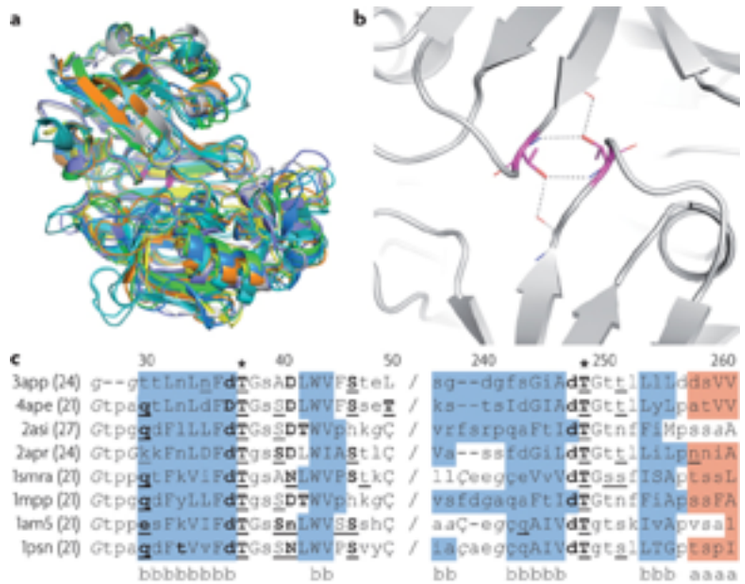



## Therapeutics (diagnostics, drug targets, efficacy)

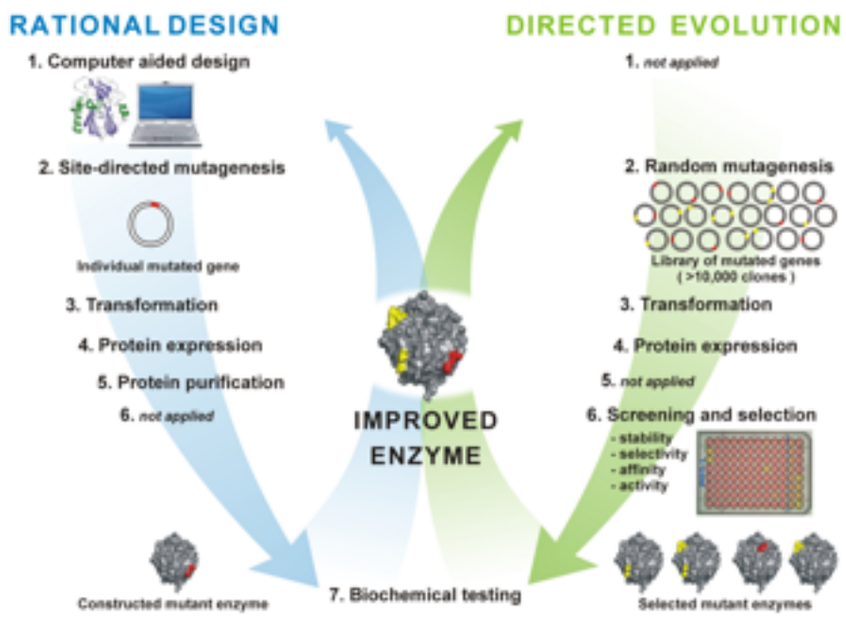



## Nanotechnology


## Evolution




## Bioengineering



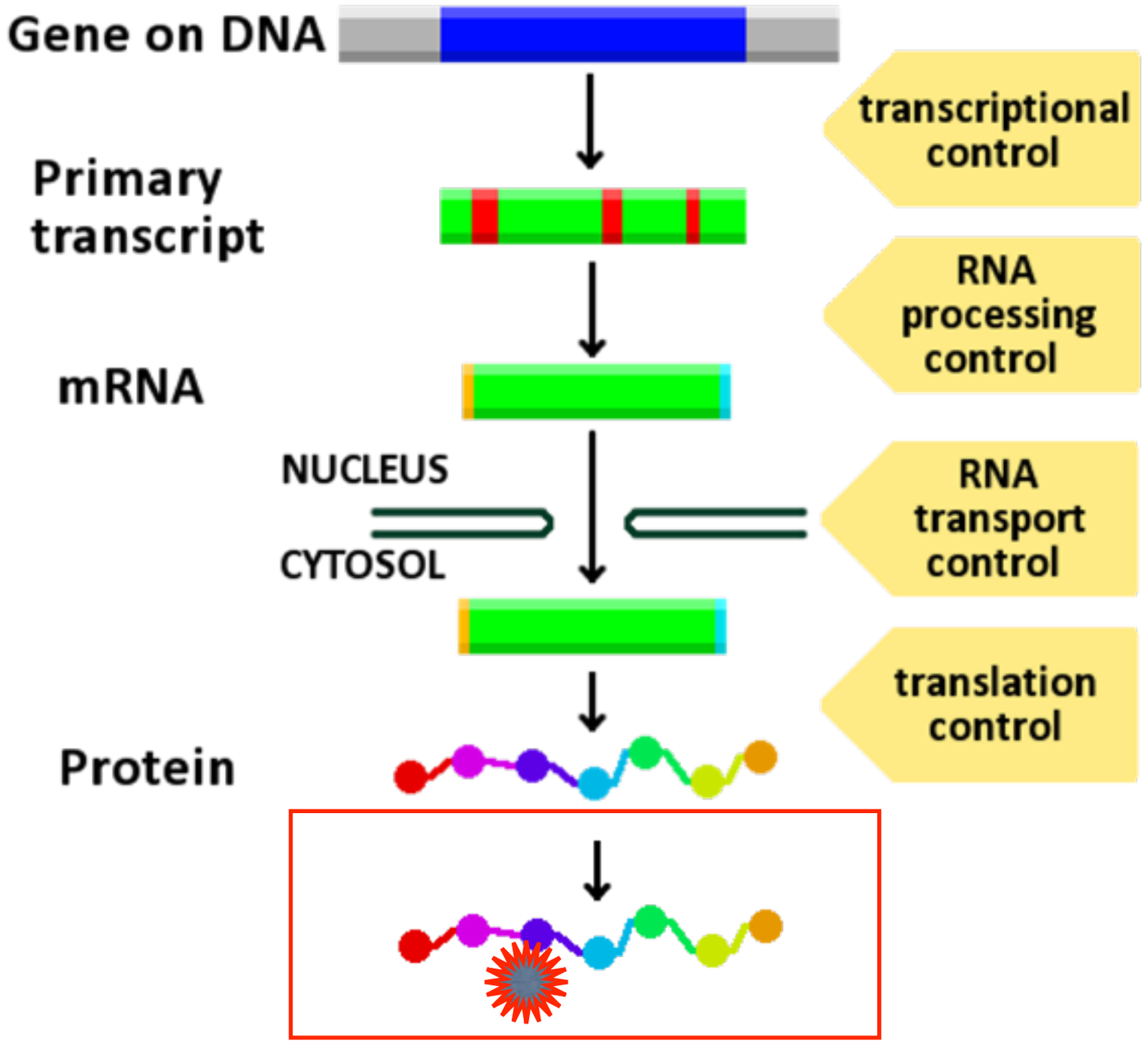
# Protein function regulation

**Protein function is regulated at**

1) transcriptional level

2) translational level

3) post-translational level



# Post-translational regulation of protein function

**Regulators**

(a) non-covalent interactions

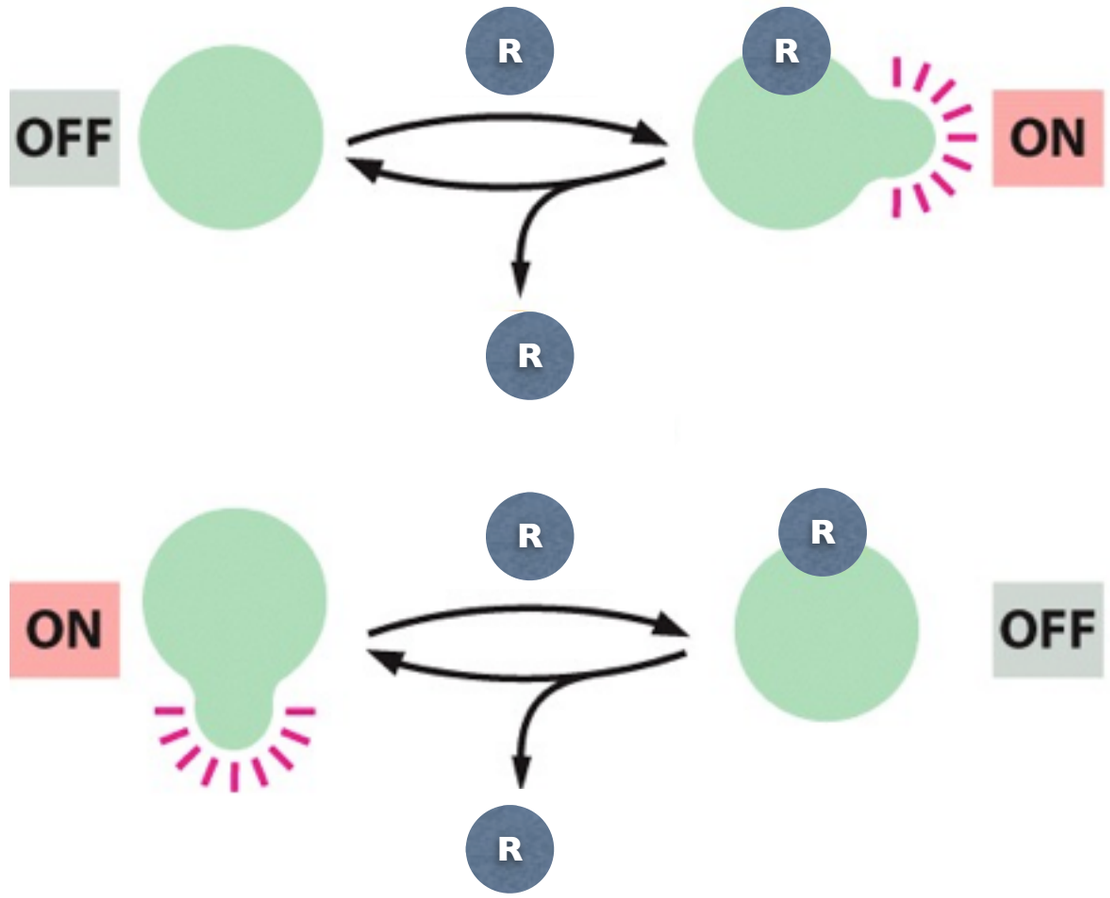
(b) covalent modifications

(c) pH

(d) salt concentrations

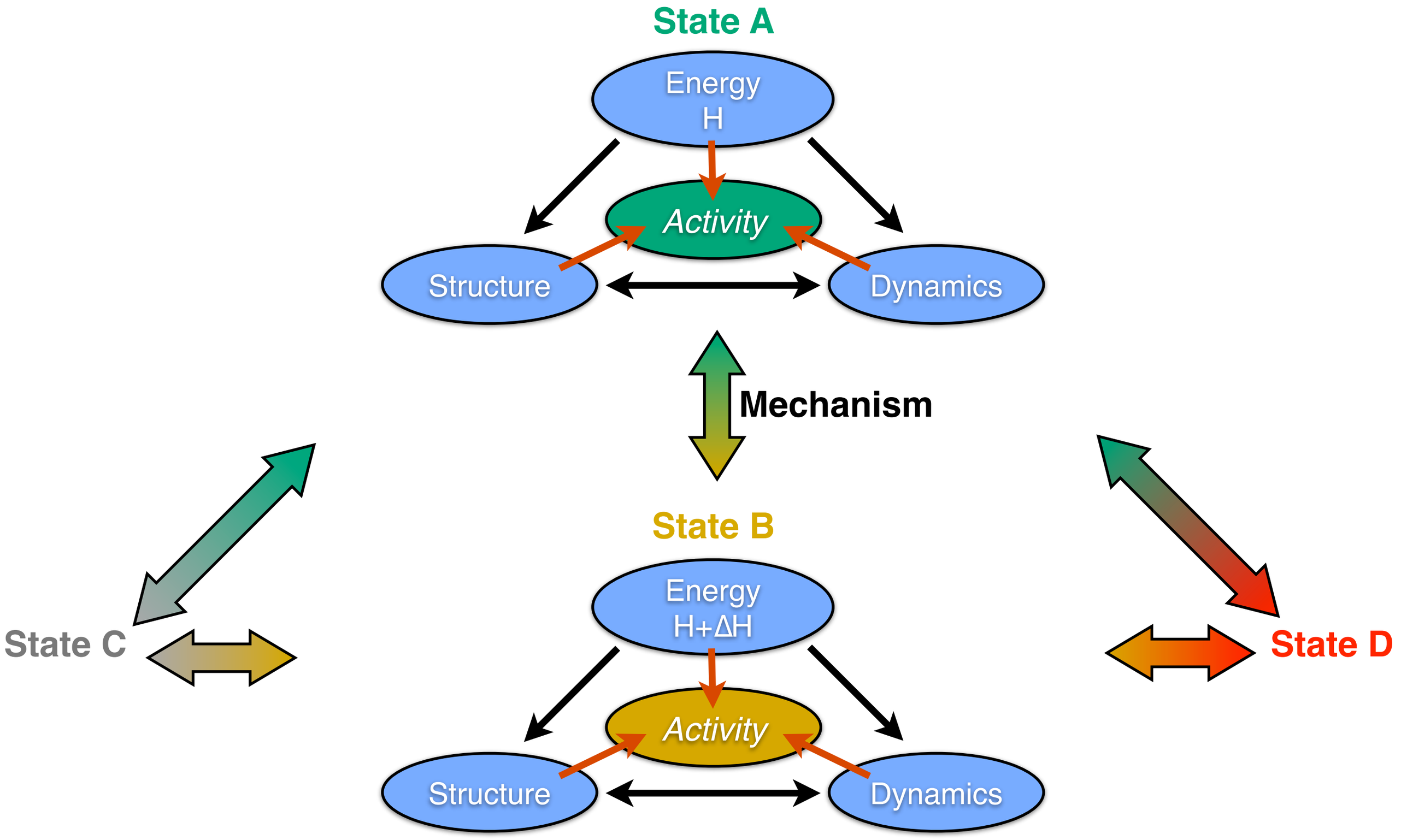
(e) ....

(  ) combination of any of the above

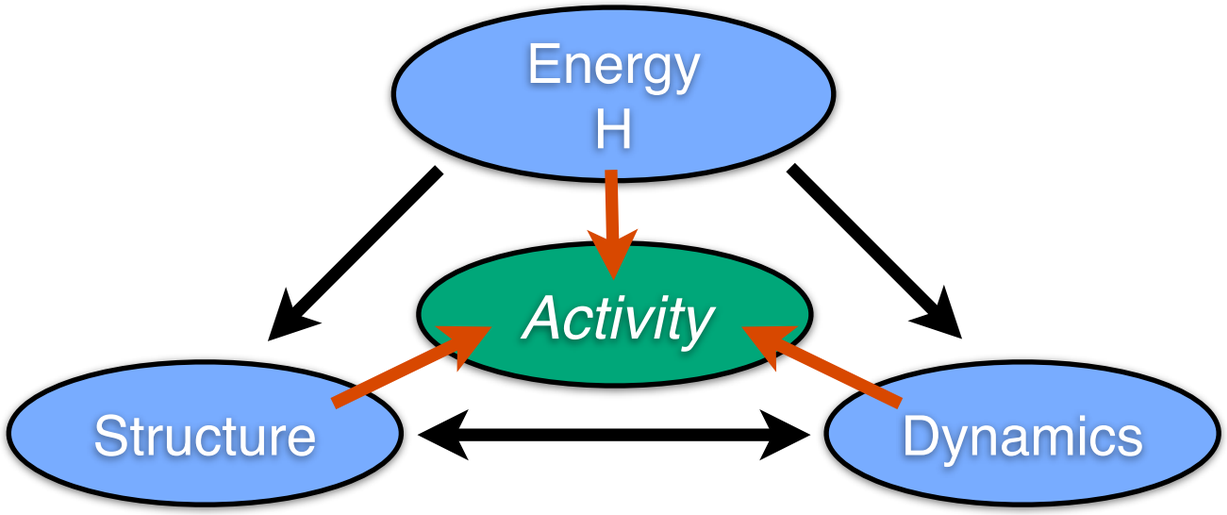


# Mechanism of protein function

Understanding how a protein works requires a systematic assessment of relationships between the structure, dynamics and energetics of its various states



# Pioneers: Molecular simulation methods for biomolecules




## The 1970's revolutionary vision...

Since energy drives structure, dynamics and activity, shouldn't we be able to predict them directly, provided we truly understand the energetics of underlying interactions?

Martin Karplus

Michael Levitt

Arieh Warshel

# Direct estimation of activity from energy

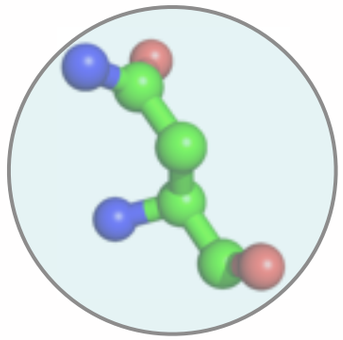
$$H = \sum \frac{p_i^2}{2m_i} + V$$

**Define V**

Example: $V = \sum V_{ij}^{bond} + \sum V_{ijk}^{angle} + \sum V_{ijkl}^{dihedral} + \sum V_{ij}^{electrostatics} + \sum V_{ij}^{vdw}$

Solve Hamilton's equation of motion to get $r_i(t)$

$$\frac{d\vec{r}_i}{dt} = +\frac{\partial H}{\partial \vec{p}_i} \quad , \quad \frac{d\vec{p}_i}{dt} = -\frac{\partial H}{\partial \vec{r}_i}$$

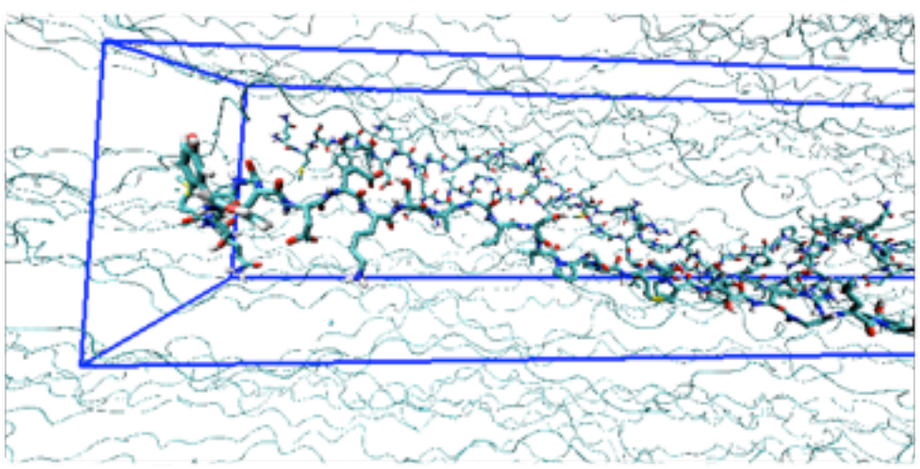


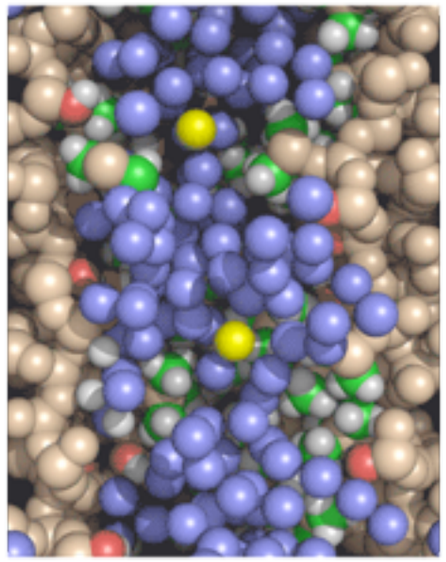
$r_i(t=0)$

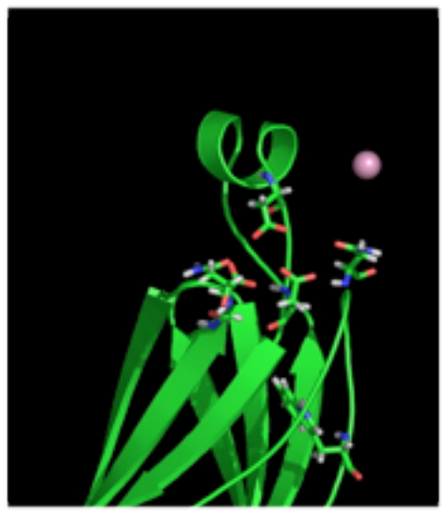
+ Model H

time evolution

$r_i(t)$ → Statistics + Statistical mechanics → Activity/ Mechanism



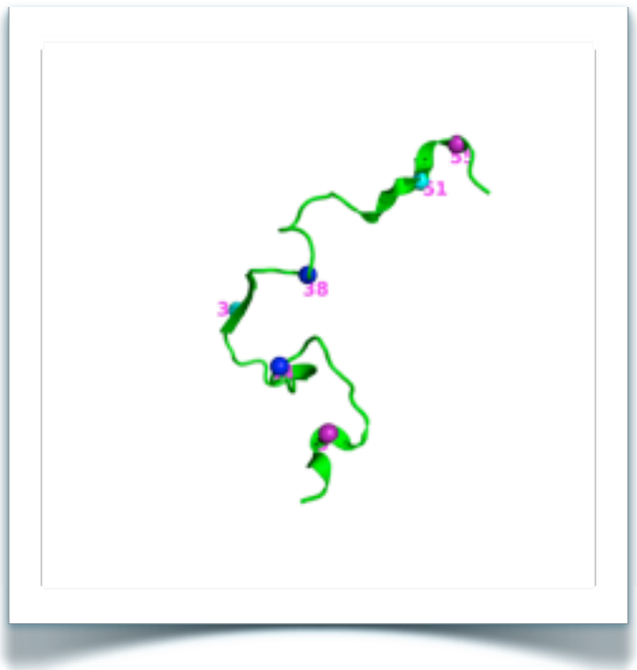


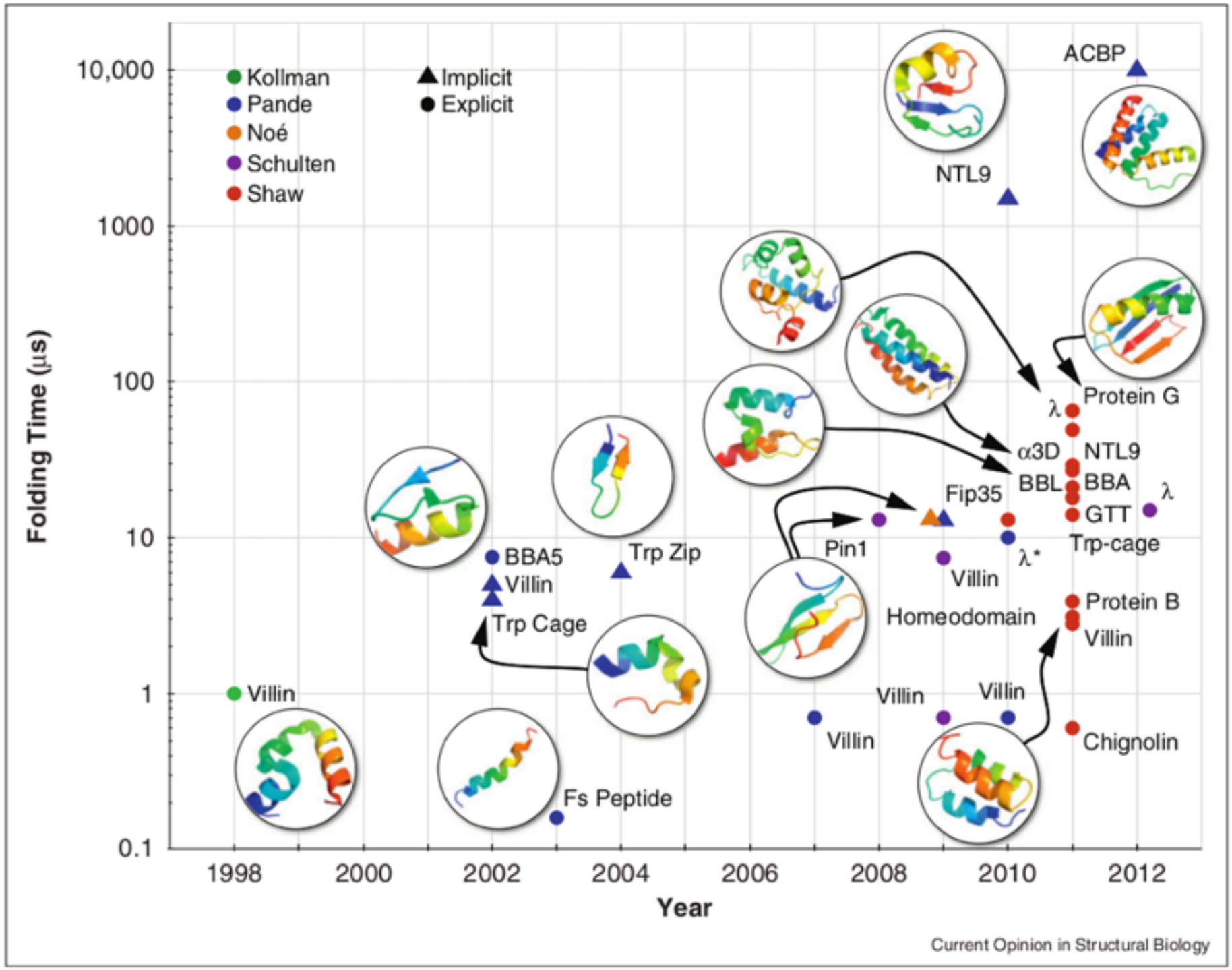


# How well do we understand inter-atomic interactions?

**Reliability test:** *ab initio* protein folding

Qin et al PNAS 2015





# Pioneers

## Nobel Prizes and Laureates

Chemistry Prizes ‹ 2013 ›



**The Nobel Prize in Chemistry 2013**
Martin Karplus, Michael Levitt, Arieh Warshel

# The Nobel Prize in Chemistry 2013

© Nobel Media AB
**Martin Karplus**

Photo: Keilana via Wikimedia Commons
**Michael Levitt**

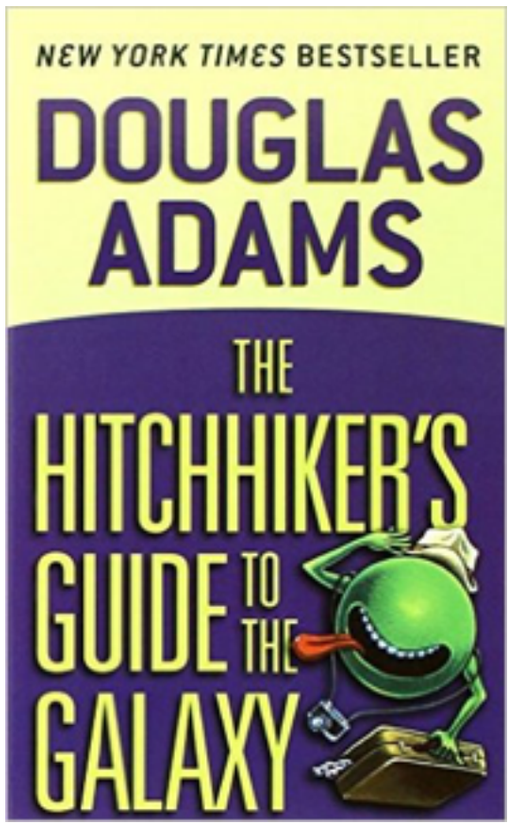
Photo: Wikimedia Commons
**Arieh Warshel**

**The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel** ***"for the development of multiscale models for complex chemical systems"*****.**

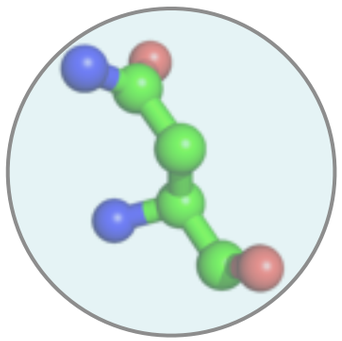
**Does this imply that we can now blindly apply molecular simulations to any protein?**

*NO!*

While there is a lot that we can model, which we couldn't just a decade ago, there remains a lot more that we still cannot...




# Our interests



$r_i(t=0)$

\+ Model H

time evolution

$r_i(t)$

Statistics + Statistical mechanics

Activity/ Mechanism

## Development

**1)** Enhance reliability of atomistic models to describe interactions of proteins with ions

**2)** New methods to study allosteric activation of proteins

**Applications:** Understand the fundamental atomically-detailed principles underlying

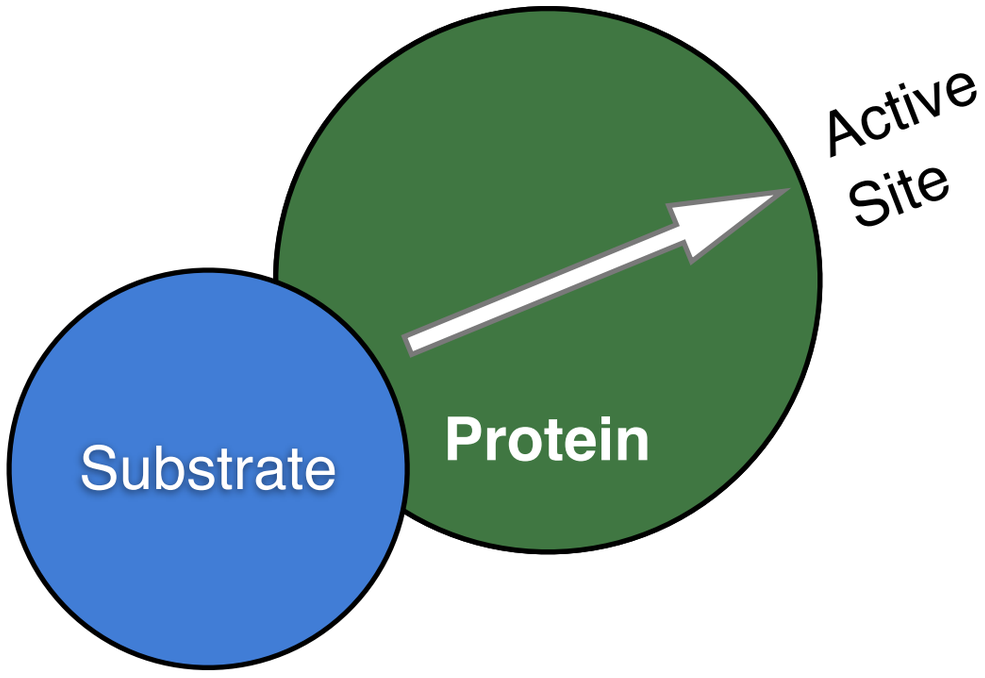
**1)** Selective ion transport by channels

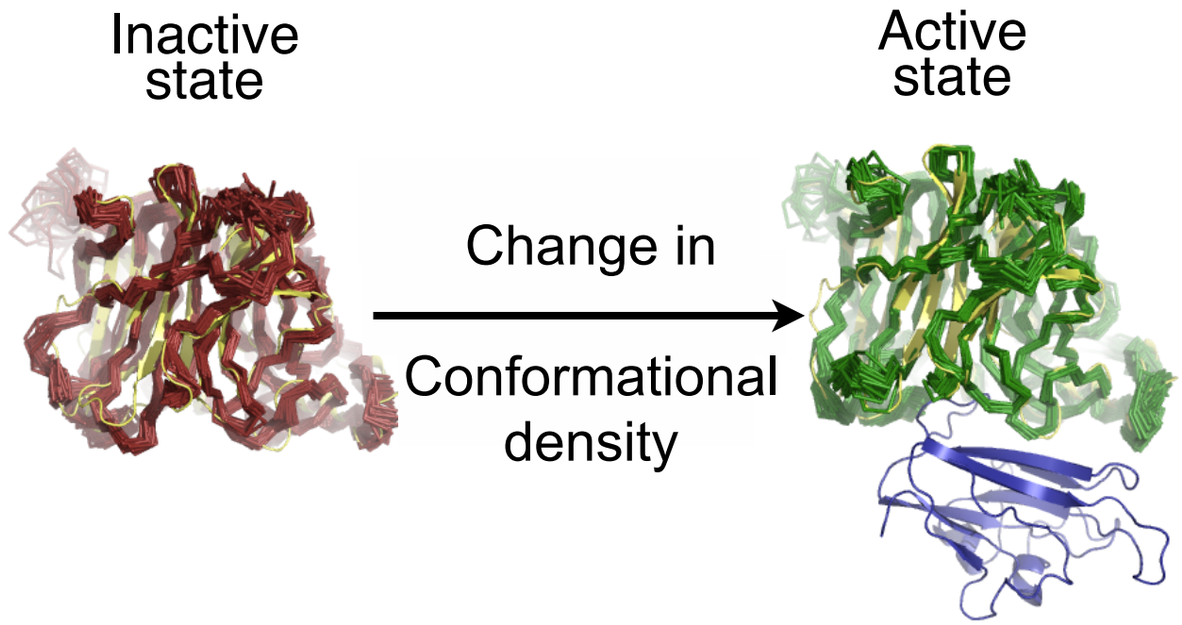
**2)** Protein assembly

**3)** Allosteric regulation of protein activity

# Allostery

## Basic problem



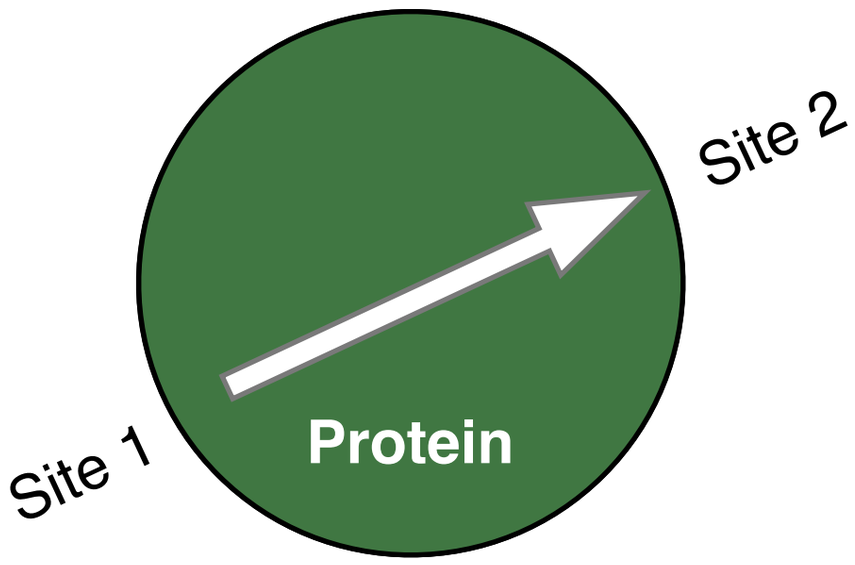

$$\mathbb{R} \xrightarrow{\Delta\mathbb{R}} \mathbb{R}'$$

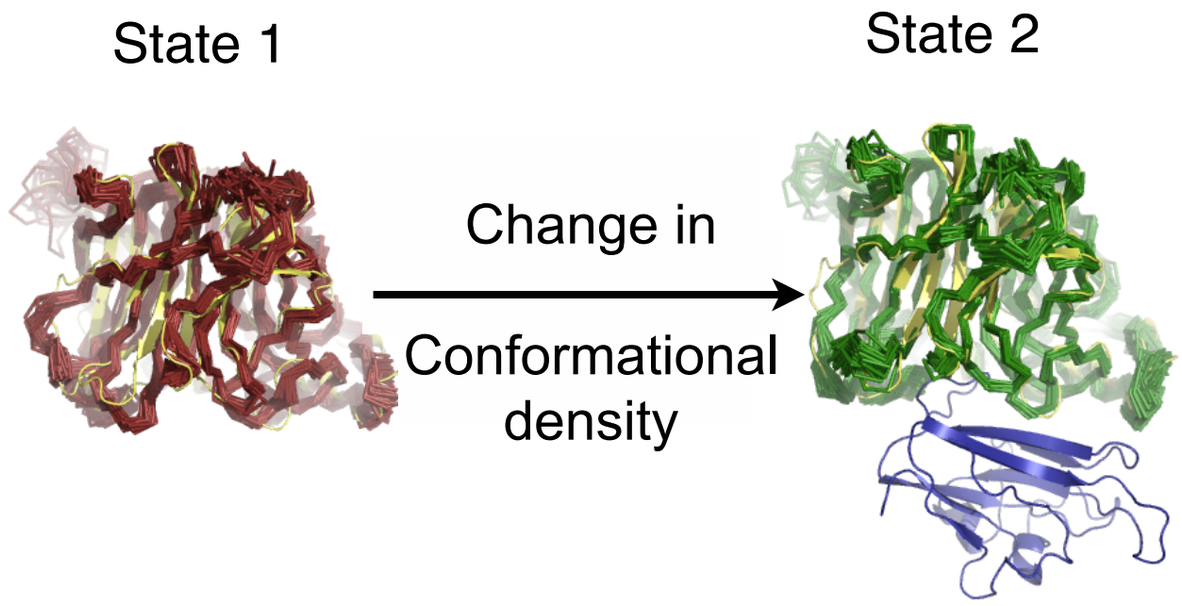
The regulatory signal

$$\Delta\mathbb{R}_{signal} \subset \Delta\mathbb{R}$$

# Allostery

## General problem




$$\mathbb{R} \xrightarrow{\Delta\mathbb{R}} \mathbb{R}'$$

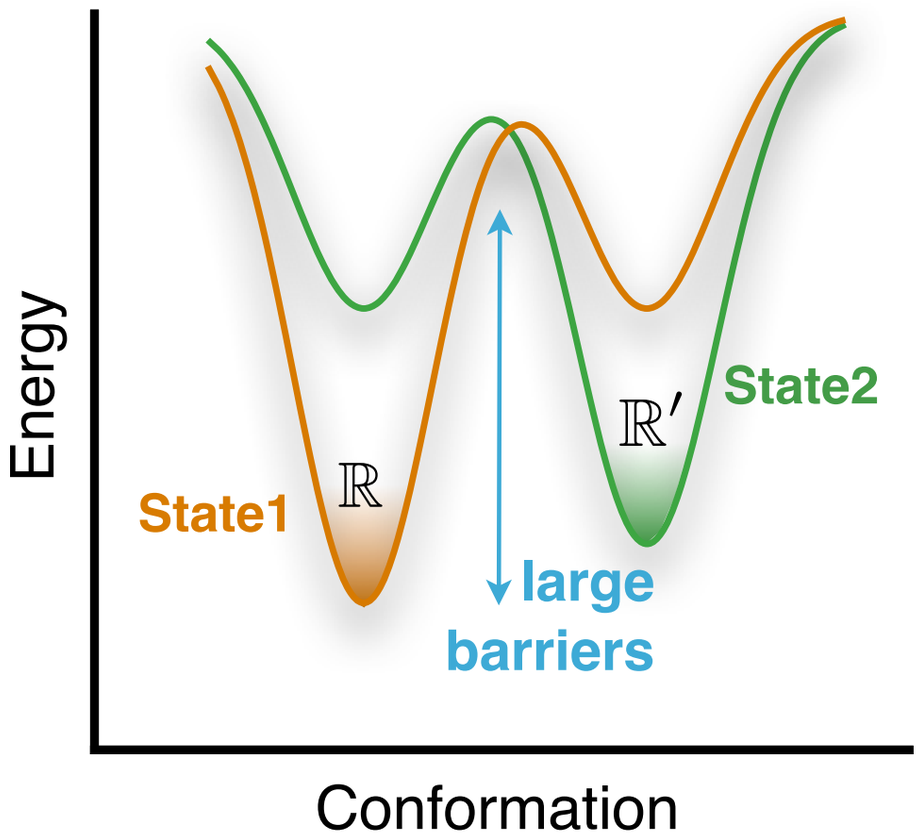
The regulatory signal

$$\Delta\mathbb{R}_{signal} \subset \Delta\mathbb{R}$$

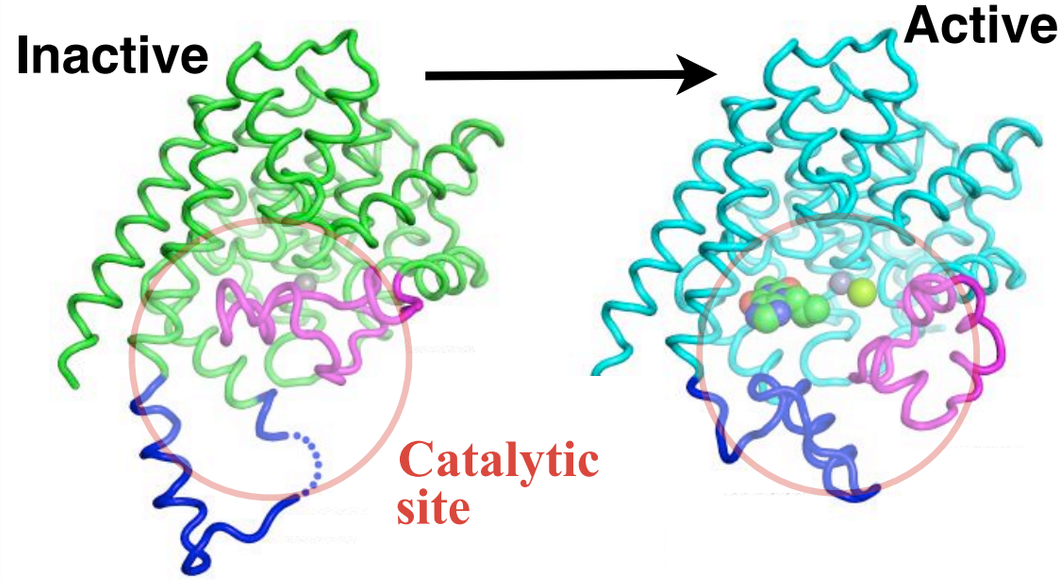
# Allosteric regulation involving large structural changes

For many proteins, their regulatory models can be understood in terms of how their 3D structures, or conformations, change during transition

**(a) Negligible overlap between conformational ensembles of 2 states**




**Example: Phosphodiesterase**



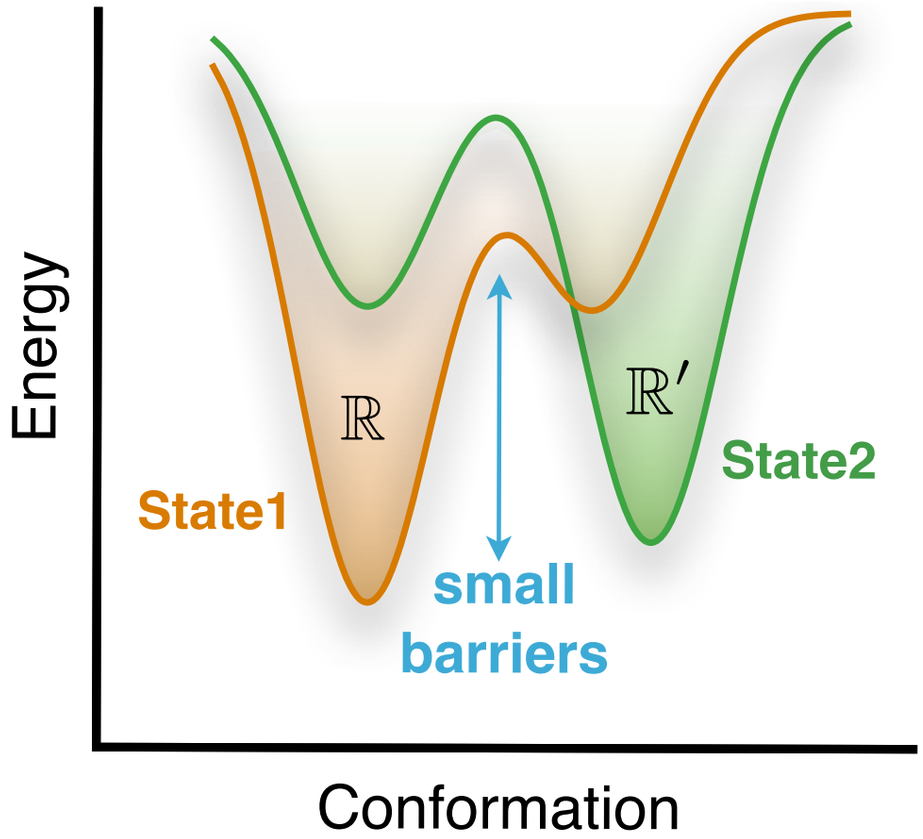

Difference between states discernible by x-ray crystallography

**Many methods are available to study allosteric signaling in such proteins**

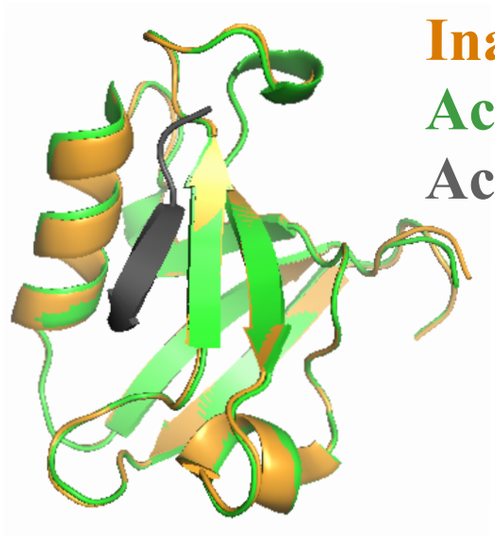
# Allosteric regulation involving small structural changes

For several major pharmaceutical proteins, distinguishing their states **requires** consideration of their finite temperature conformational ensembles

**(b) Discernible overlap between conformational ensembles of 2 states**




**Example: PDZ domain**

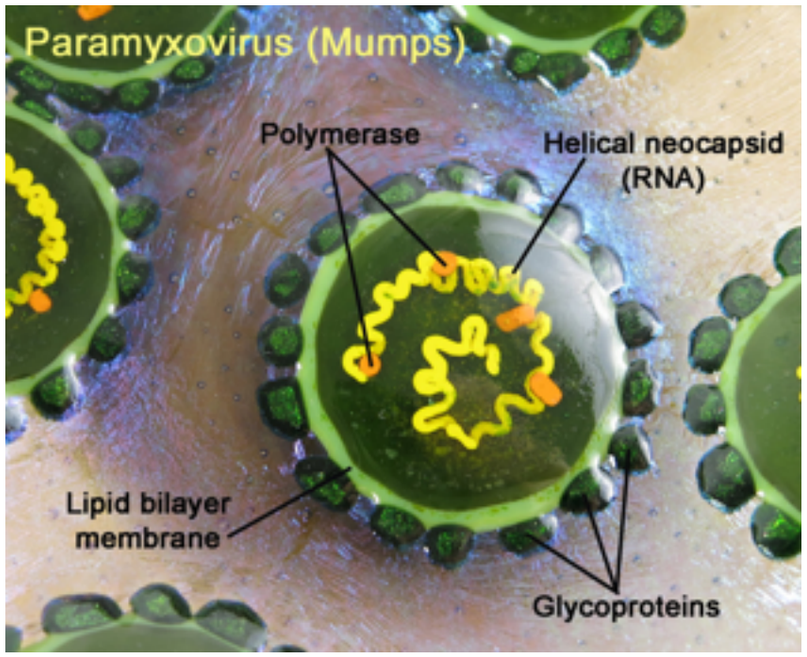



Difference between states (0.4 Å) is beyond the resolution of x-ray crystallography

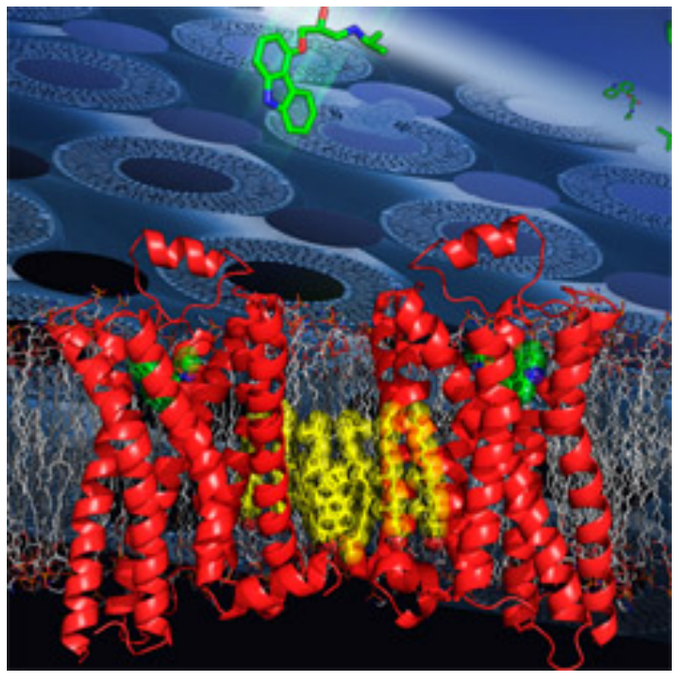
Understanding mechanisms requires accounting for information on both structure and dynamics from multiple states.

# Role of dynamic allostery - examples

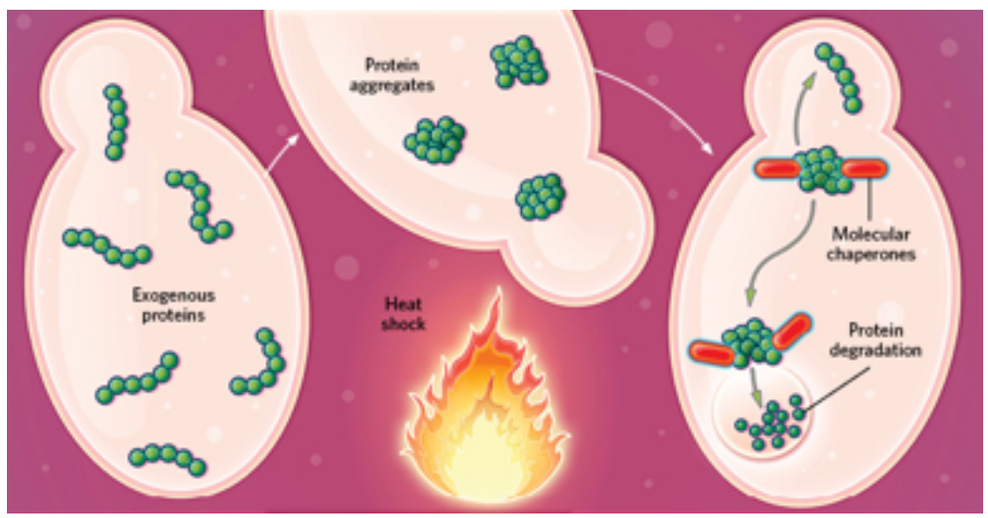
## Fusion regulation in paramyxoviruses



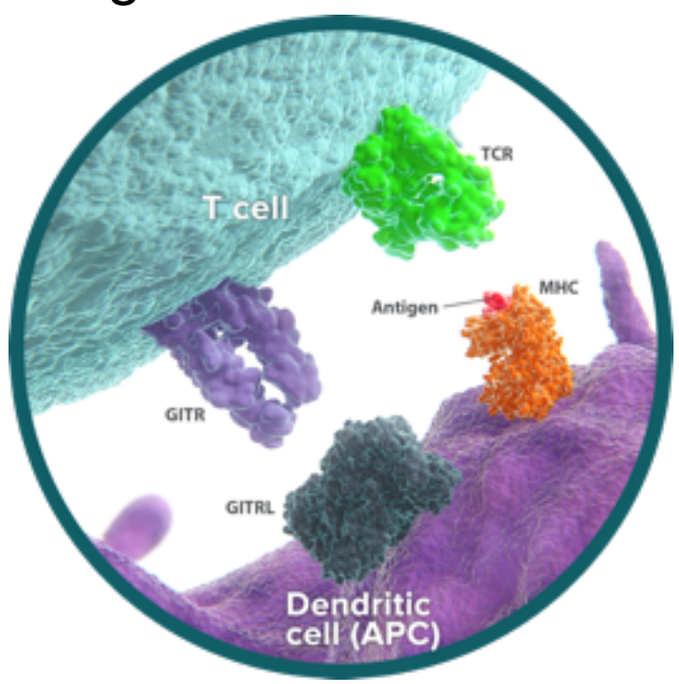

## GPCR regulation by natural compounds and drug molecules



## Heat shock response




## Regulation of Immune response

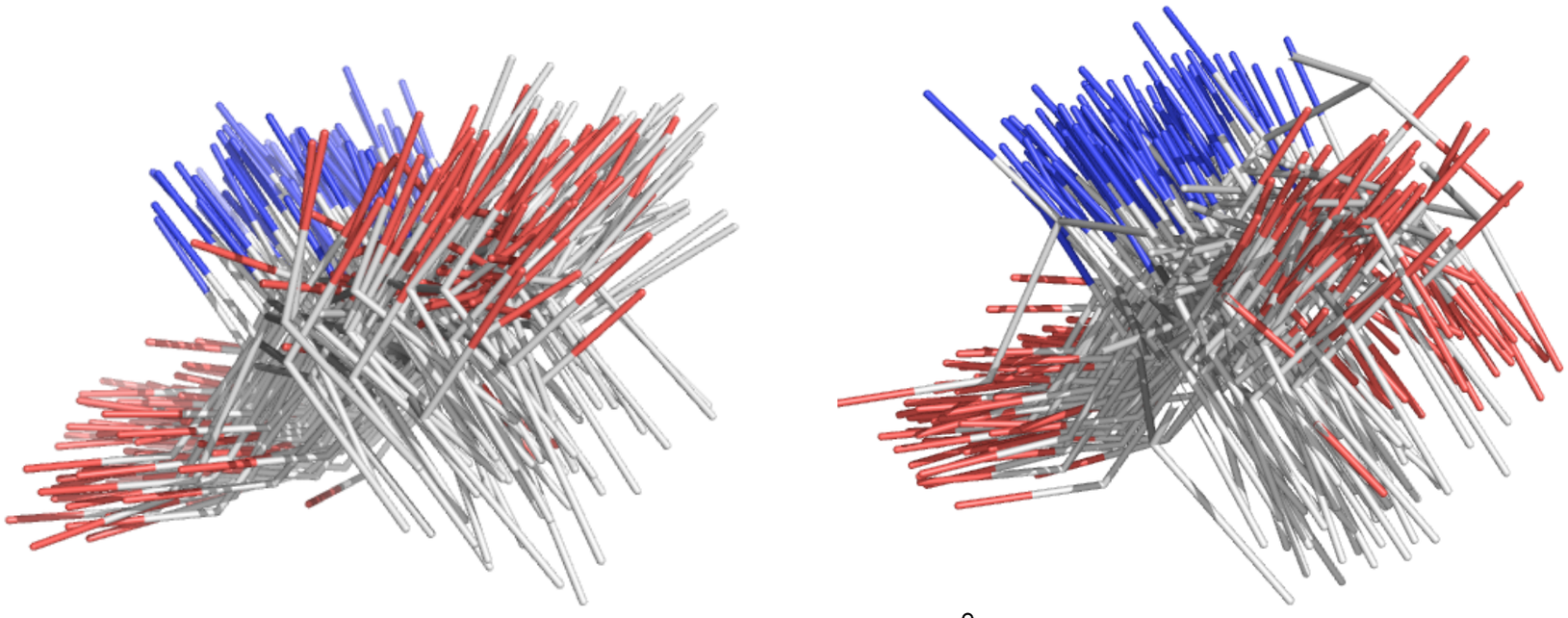


# Methods required?

## 1) A method to quantify differences between conformational ensembles of two states

### Traditional approach

Compare summary statistics of two ensembles



$$\mathbb{R} \xleftrightarrow[\Delta RMSF = 0.03\ \text{Å}]{\Delta CoM = 0.001\ \text{Å}} \mathbb{R}'$$

**Problem with comparing summary statistics**: Need to know the “right set” of summary statistics that differentiate ensembles. But how does one identify such summary statistics beforehand?

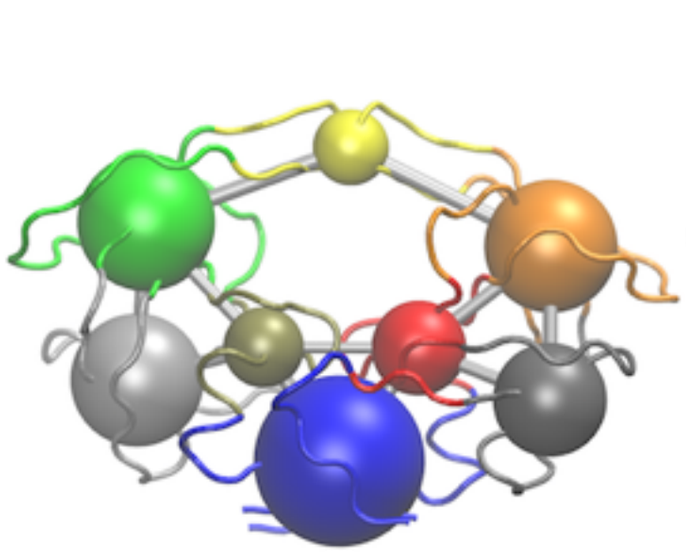
# Methods required?

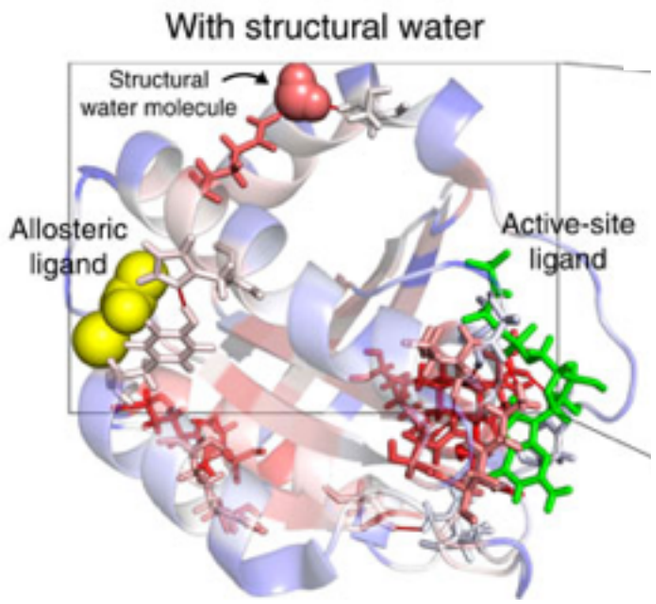
## 2) A method to relate $\Delta\mathbb{R}$ to regulatory signals

### Current strategy

- Construct correlation matrices

- Determine efficient signaling paths from correlation matrices.






Amor et al. Nature Comm. 2015

# Methods required?

## 2) A method to related $\Delta\mathbb{R}$ to regulatory signals

***Assumptions***

(i) structural changes are large, and so conformational fluctuations can be ignored.

OR (ii) ignore $\Delta\mathbb{R}$ and focus only on $\mathbb{R}'$ of only the active state.

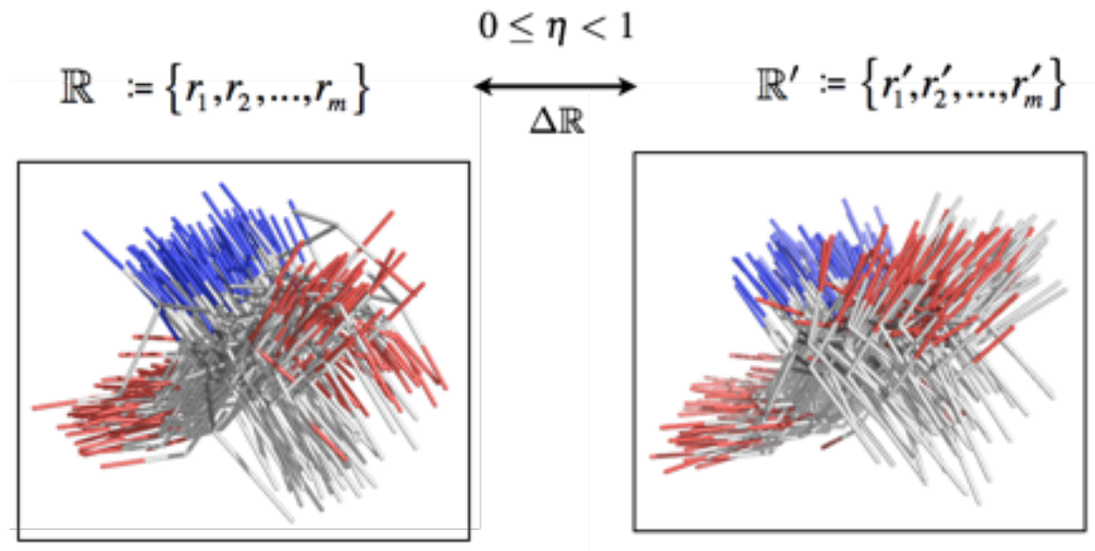
Problem: These assumptions break down for proteins regulated by dynamic allostery

Cilia et al. PLoS Comp Bio. 2012

| Method | Description | Predictions | tot. num. | TP | TN | FP | FN | AUC (Fig. 3A) | AUC (Fig. 3B) | TPR | FPR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fuentes et al 2004 | Experimental measurements | L18, I20, V22, V26, V30, A39, V40, V61, V64, L66, A69, L78, T81, V85 | 14 | n. a. | n. a. | n. a. | n. a. | n. a. | n. a. | n. a. | n. a. |
| Kong and Karplus 2009 | Molecular Dynamics (MD) approach | A12, L18, V26, I41, A45, A46, V58, L59, L66, A69, A74, L78, T81, L89 | 14 | 6 | 6 | 5 | 8 | n. a. | n. a. | 0.43 | 0.45 |
| Gerek and Ozkan 2011 | Perturbation Response Scanning method | L11, L18, I20, V22, T23, I35, V37, A39, V40, I41, A45, A46, V58, L59, A60, V61, L66, A69, A74, V75, T77, L78, T81, V85, L87 | 26 | 11 | 3 | 8 | 3 | n. a. | n. a. | 0.79 | 0.73 |

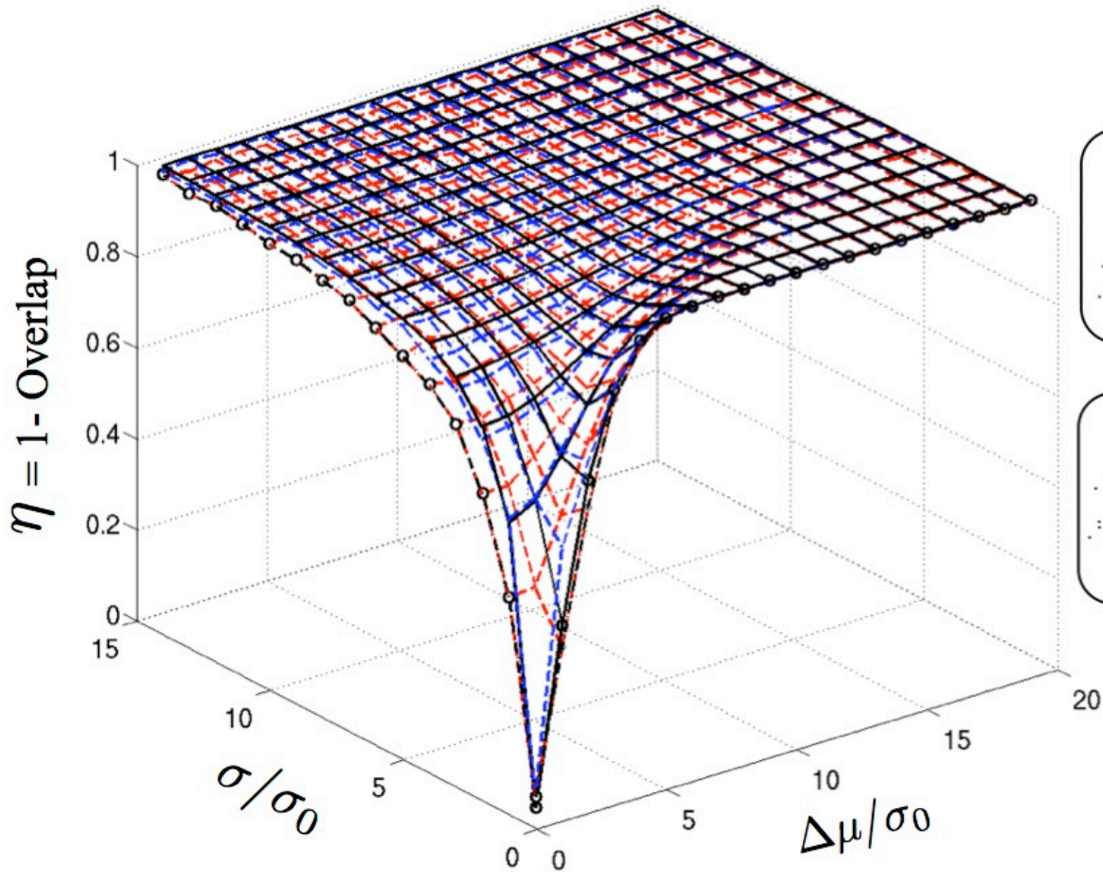
# New method to quantify differences in ensembles

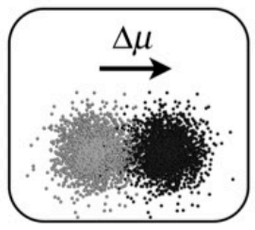
**Instead of comparing summary statistics, compare ensembles directly against each other**



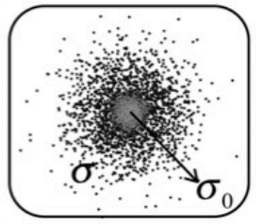

$$\eta = 1 - ||Overlap|| = 1 - ||\mathbb{R} \cap \mathbb{R}'||$$

MAE = 3.2%




Gaussian Distributions with different **means**




Gaussian Distributions with different **widths**

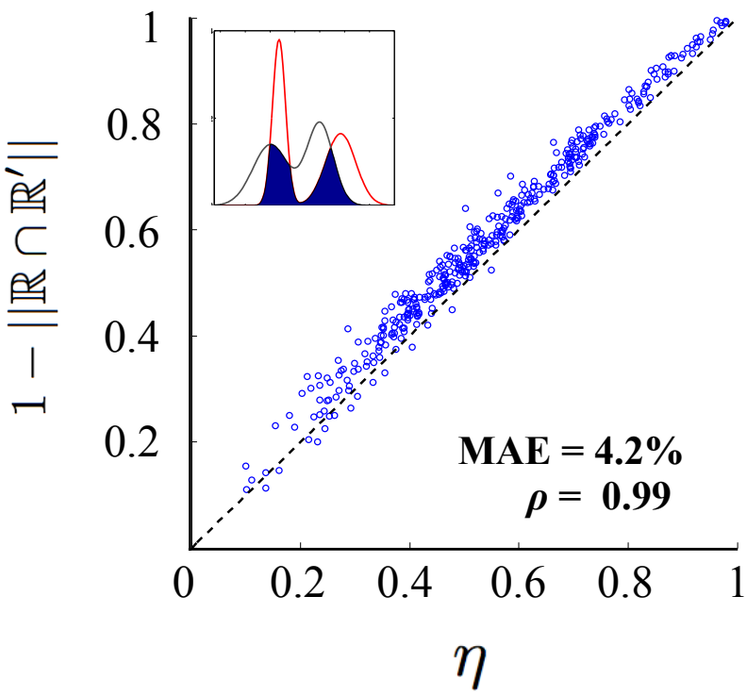
Leighty and Varma, JCTC 2013
Varma, Botlani, Leighty, Proteins 2014

# New method to quantify differences in ensembles

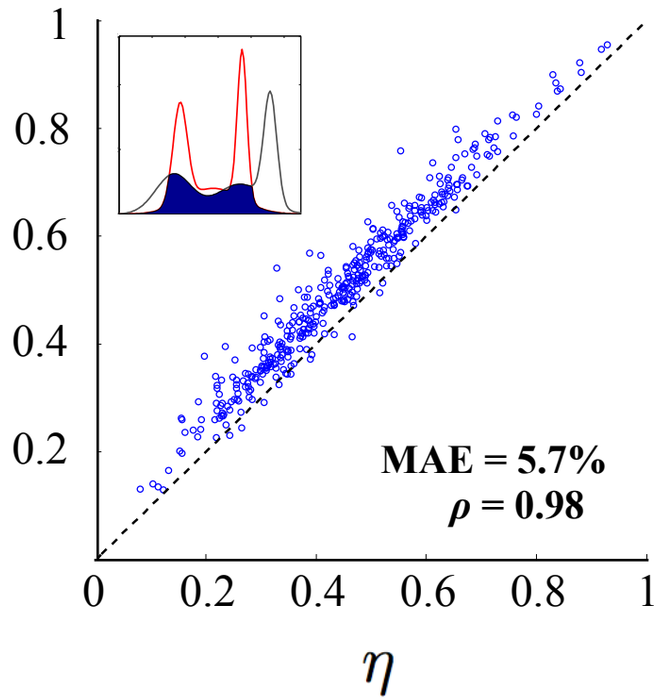
## Works even for multi-modal distributions

$$\eta = 1 - ||\sum_{i=1}^{n} c_i f_i \cap \sum_{j=1}^{n} c'_j f'_j|| = 1 - ||\sum_{i,j=1}^{n} c_i f_i \cap c'_j f'_j||$$

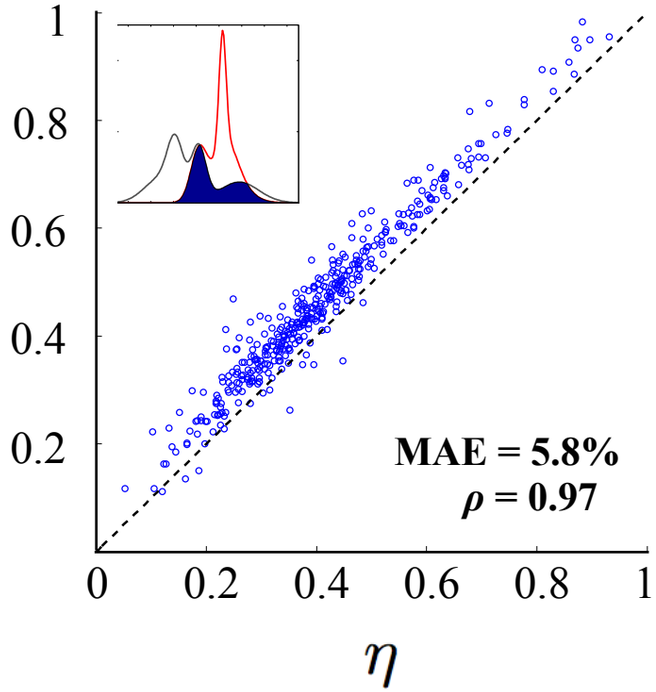
**(a) Bimodal**




**(b) Trimodal**



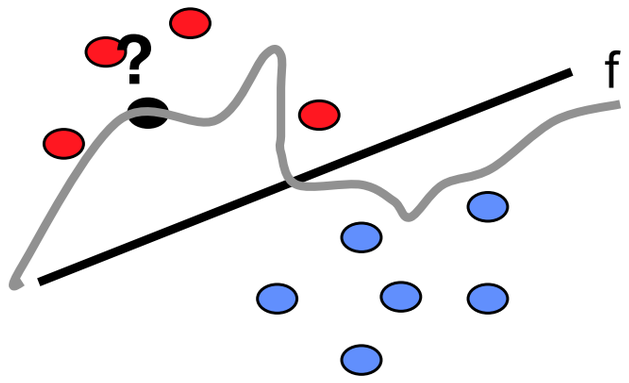

**(c) Quadrimodal**




Dutta, Siddiqui, Botlani, Varma. BJ 2016

# Machine Learning

Typical scenario: learning from data




- Given data set $\mathrm{x}$ and labels $\mathrm{y}$

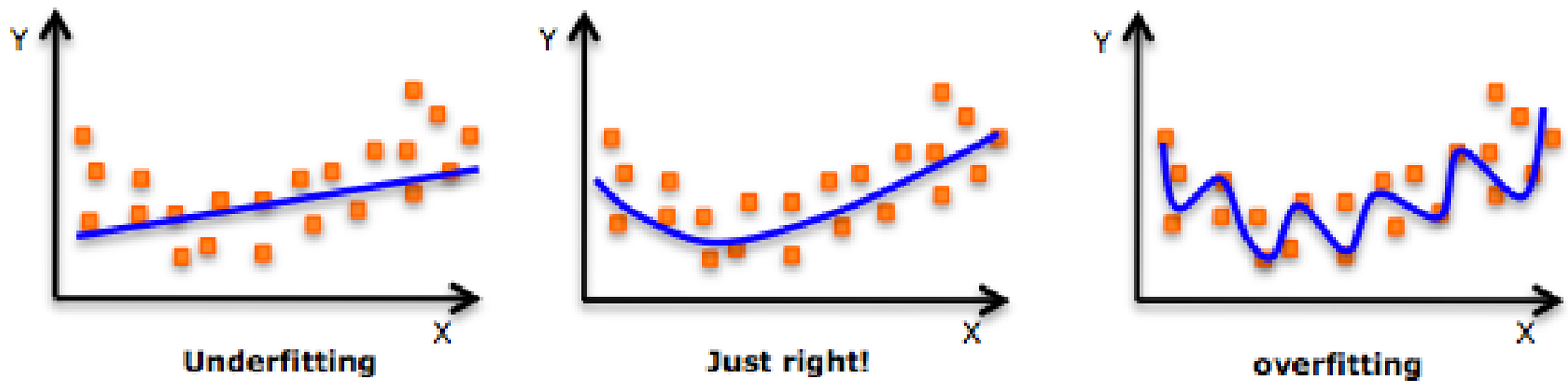
  (generated by some joint probability distribution p(x,y))

- LEARN/INFER underlying unknown mapping

$$\mathrm{y} = f(\mathrm{x})$$

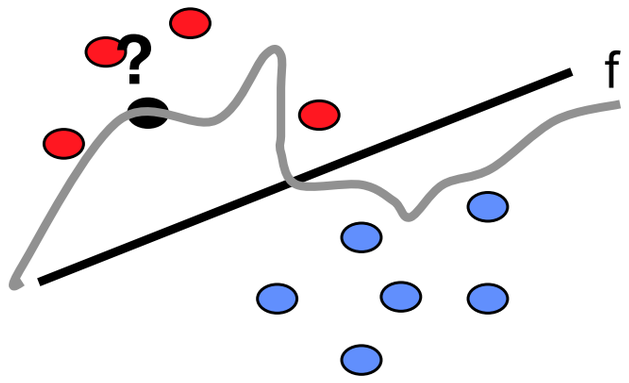
Learn $f$ from examples such that **R**isk of prediction, is minimized

$$R[f] = \int |f(x) - y|^2 \, dP(x,y)$$



# “Inverse” machine learning

Typical scenario: learning from data




- Given data set $\mathrm{x}$ and labels $\mathrm{y}$

  (generated by some joint probability distribution p(x,y))

- LEARN/INFER underlying unknown mapping

$$\mathrm{y} = f(\mathrm{x})$$

Learn $f$ from examples such that **R**isk of prediction, is minimized

**Instead of training $f$ for prediction**

Construct and train $f$ in an appropriate Hilbert space such that

it can be used to derive causalities in physical space

# Support vector machine

A support vector machine is a binary classifier that is trained on a set of instances $\{\mathrm{x}_i\}$ for which their corresponding group identities $y_i = \pm 1$ are known.

It is constructed by defining two hyperplanes

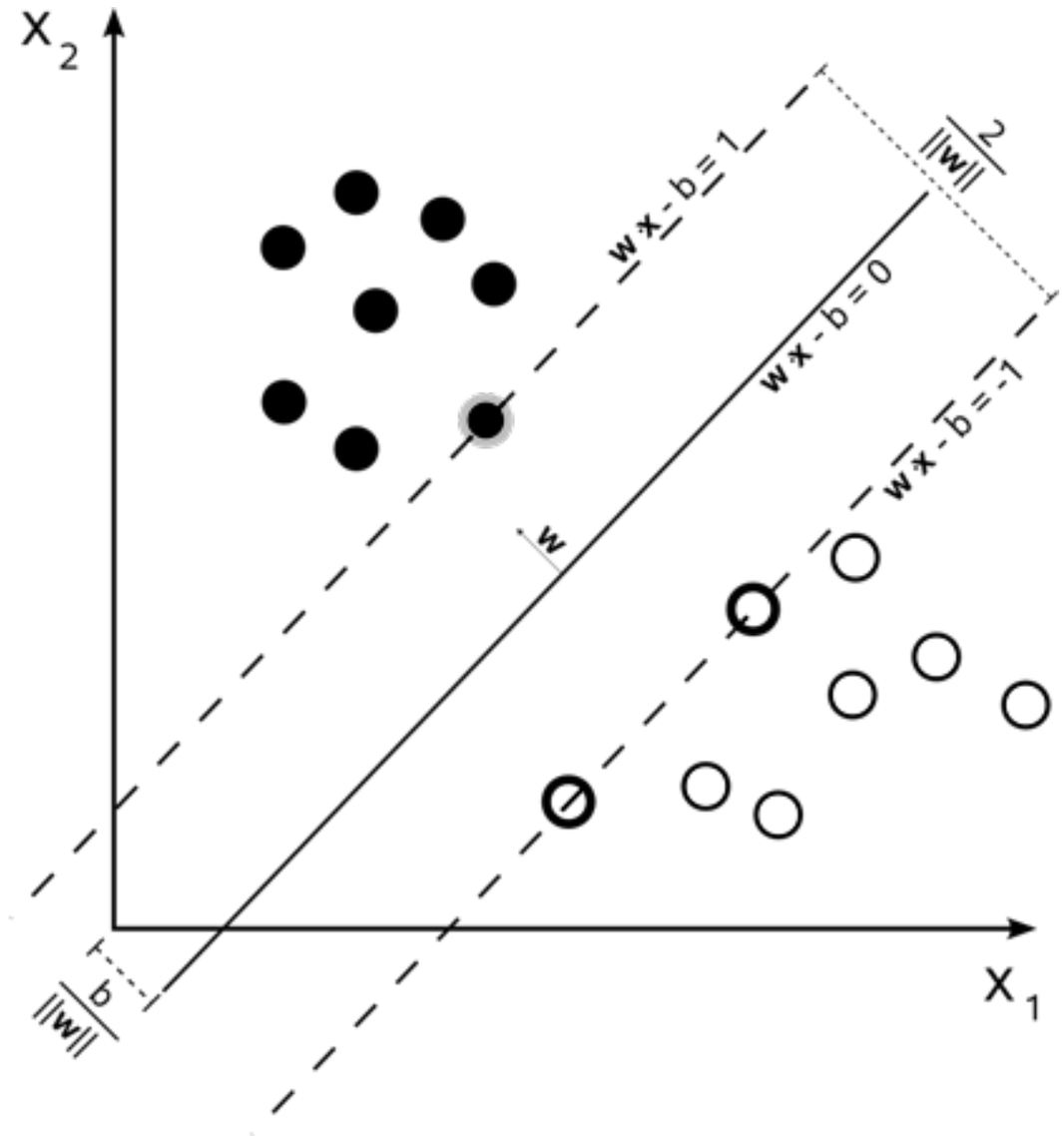
$$y_i(\mathbf{w} \cdot \mathbf{x} - b) = 1$$

that divide the instances into two groups,

$$y_i(\mathbf{w} \cdot \mathbf{x} - b) \geq 1$$

which generates a classification function

$$F(\mathbf{x}) = \sum_{i=1}^{2m} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x})$$



# Support vector machine

**The optimization task**: maximize the distance ($2/\|\mathbf{w}\|$) between the two hyperplanes.

This constrained optimization problem can be cast in terms of Lagrangian multipliers as

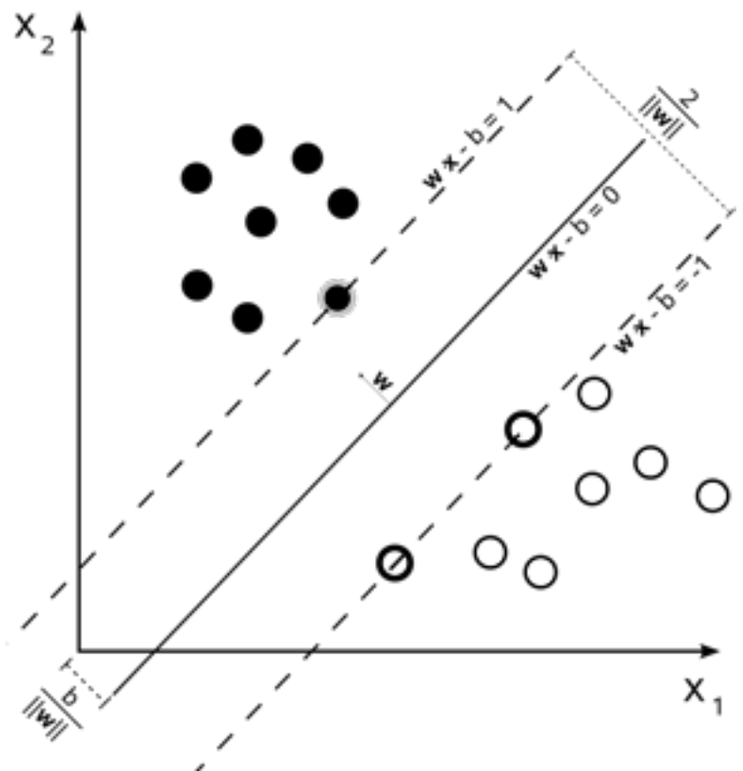
$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{2m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1]$$

where $0 \leq \alpha_i \leq C$ are the Lagrange multipliers, and the limit $C$ is a regularization parameter

Set $\nabla L = 0$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{2m} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{2m} \alpha_i y_i = 0$$



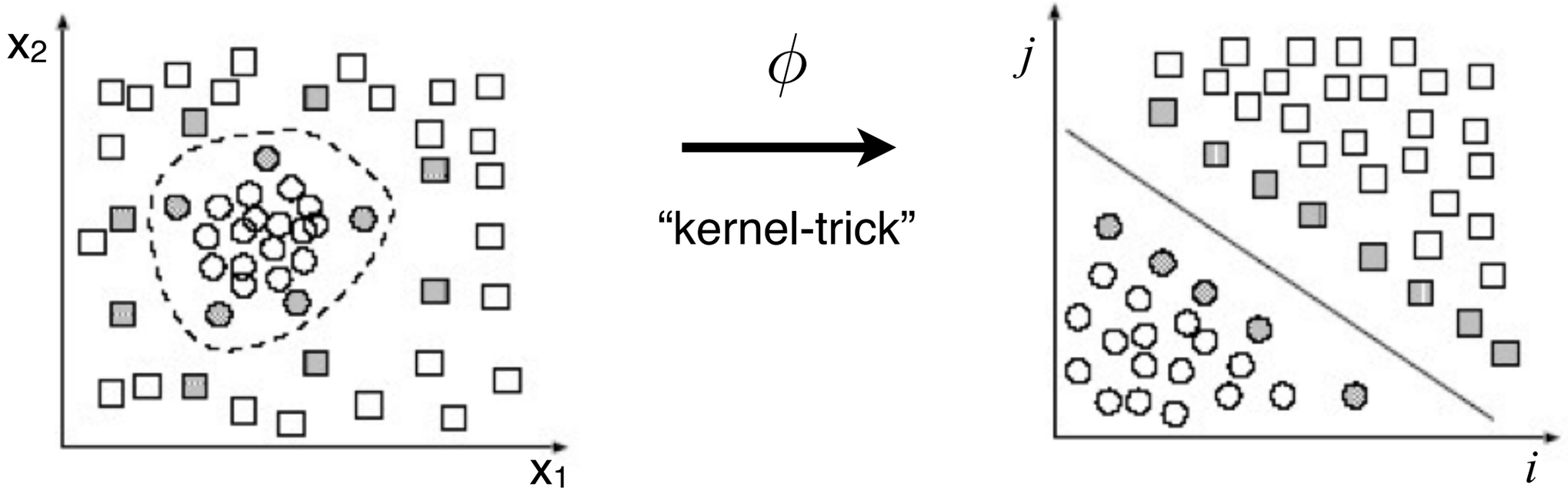
# Support vector machine

$$L = \sum_{i=1}^{2m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

where $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$

Non-linearly separably
in Euclidean space

Linearly separably
in alternative Hilbert space *R*




$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$

$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i).\phi(\mathbf{x}_j) \rangle_R = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

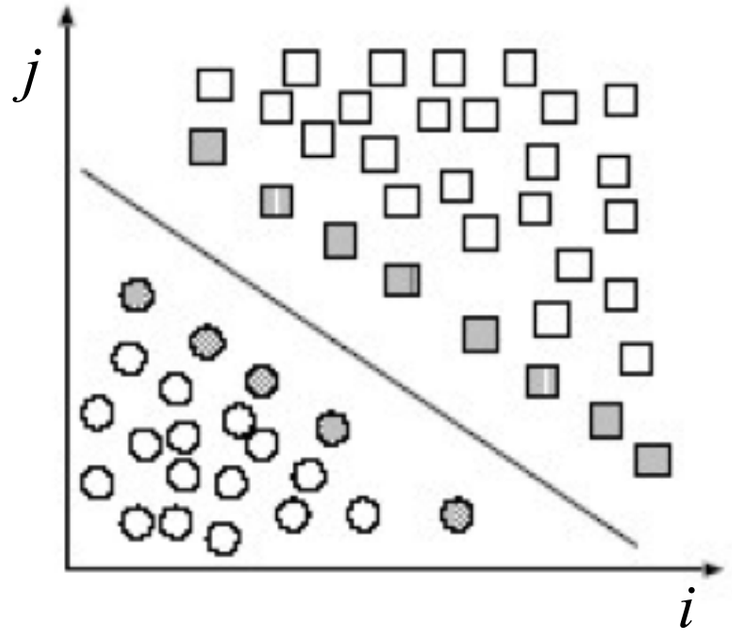
# Support vector machine

Optimization of

$$L = \sum_{i=1}^{2m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

yields two sets of $\alpha_i$:

$\{\alpha_i\} = 0$ and $\{\alpha_i\} > 0$




→ corresponding $\mathbf{x}_i$ are *support vectors* (*s*) that constitute the hyperplane

The number of support vectors are bounded, that is, for type of instance $2 \leq s \leq 2m$.

Use $\{\alpha_i\} > 0$ to construct the classification function:

$$F(\mathbf{x}) = \sum_{i=1}^{2m} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) \longrightarrow \textit{use for prediction}$$

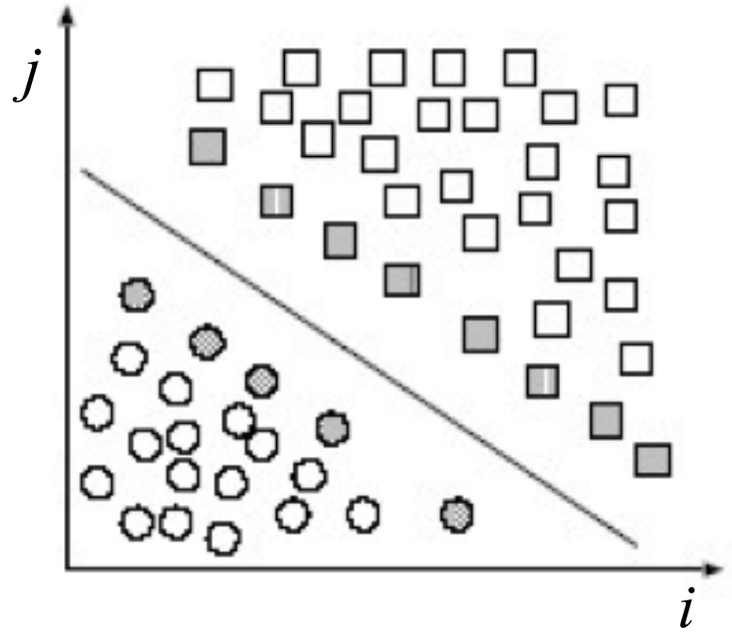
# Support vector machine

Optimization of

$$L = \sum_{i=1}^{2m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

yields two sets of $\alpha_i$:

$\{\alpha_i\} = 0$ and $\{\alpha_i\} > 0$




→ corresponding $\mathbf{x}_i$ are *support vectors* (*s*) that constitute the hyperplane

The number of support vectors are bounded, that is, for type of instance $2 \leq s \leq 2m$.

**It can be expected that**

*Higher the number of support vectors (s), the less discriminable are the two groups in a given Hilbert space.*

*Is it possible to define discriminability between two groups as* $\eta = 1 - \frac{s}{2m}$

# Selection of $C$ and $\gamma$

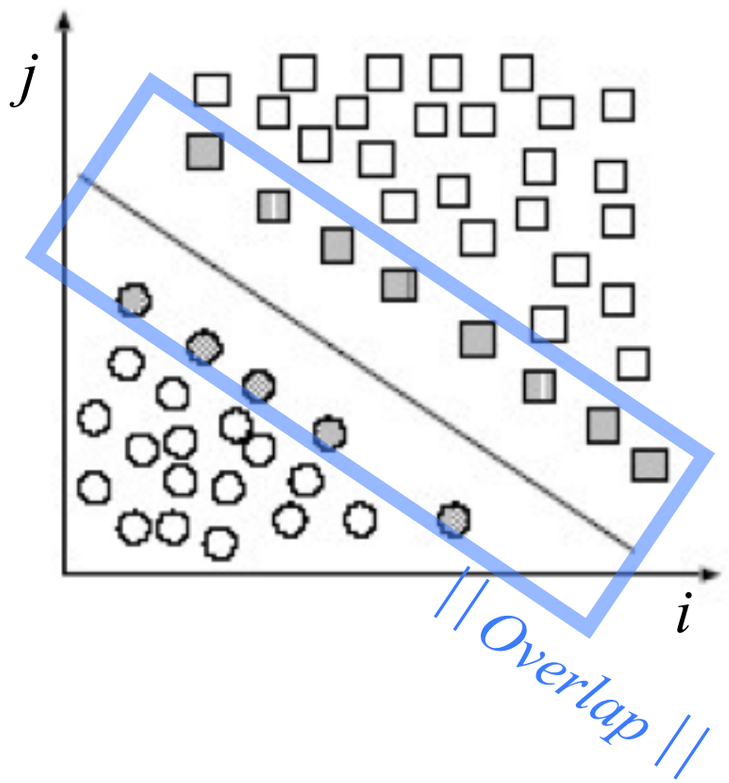
**Possibility?**

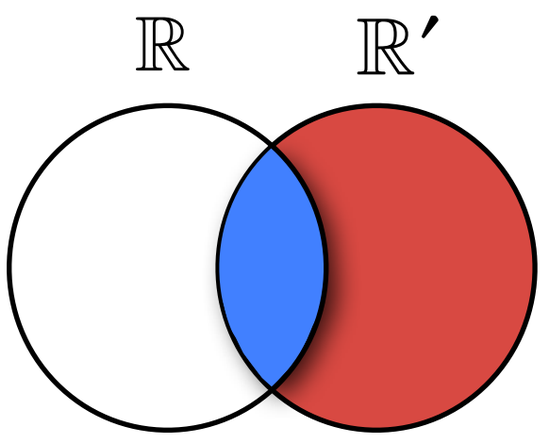
Can we select $C$ and $\gamma$ such that

$$\eta = 1 - s/2m$$

$$= 1 - ||Overlap||$$

$$= 1 - ||\mathbb{R} \cap \mathbb{R}'||$$




Such a definition of discriminability satisfies two conditions

**(i)** $\left|\eta(\mathbb{R} \to \mathbb{R}')\right| = \left|\eta(\mathbb{R}' \to \mathbb{R})\right|$

**(ii)** If $\left|\eta(\mathbb{R} \to \mathbb{R}')\right| = \left|\eta(\mathbb{R}' \to \mathbb{R}'')\right|$, then it does not necessarily imply that $\left|\eta(\mathbb{R} \to \mathbb{R}')\right| = \left|\eta(\mathbb{R} \to \mathbb{R}'')\right|$

Leighty and Varma, JCTC 2013

# Selection of $C$ and $\gamma$

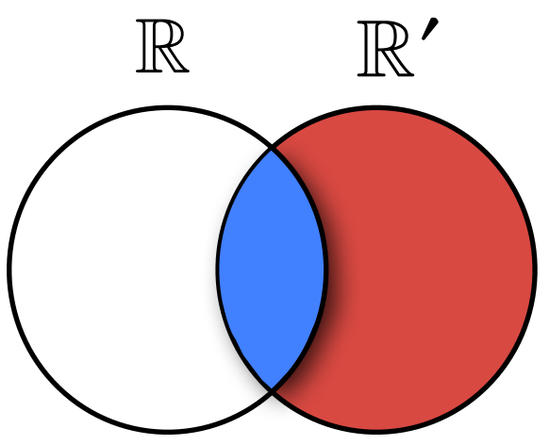
**Possibility?**

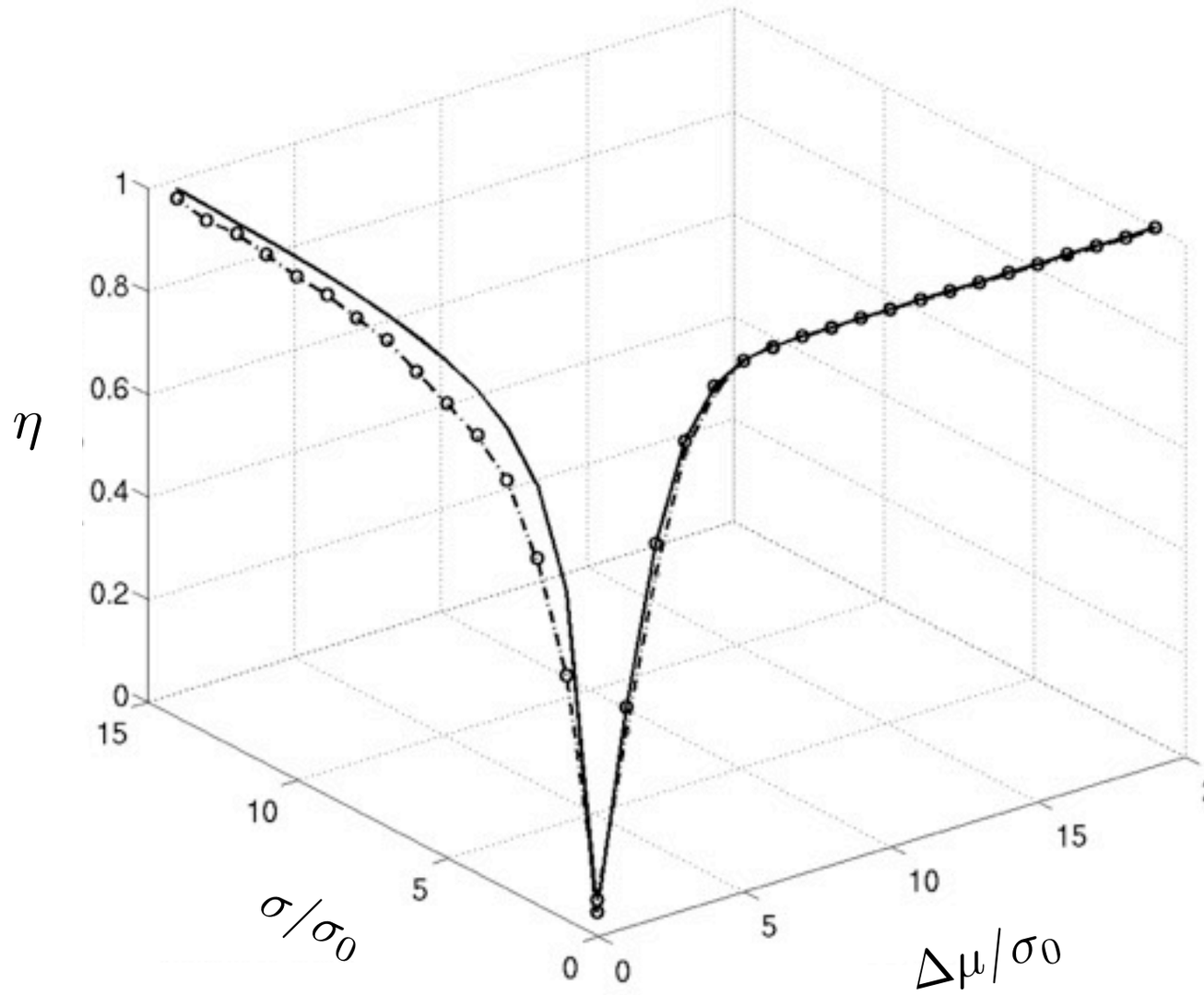
Can we select $C$ and $\gamma$ such that

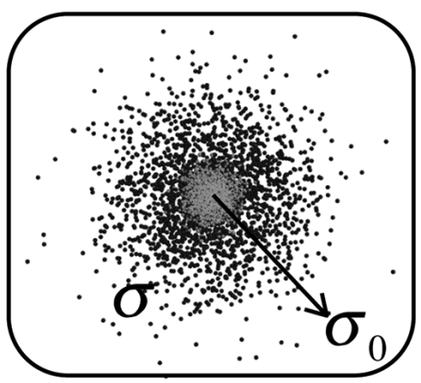
$$\eta = 1 - s/2m$$

$$= 1 - ||Overlap||$$

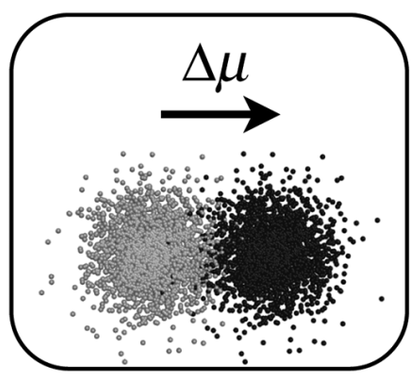
$$= 1 - ||\mathbb{R} \cap \mathbb{R}'||$$



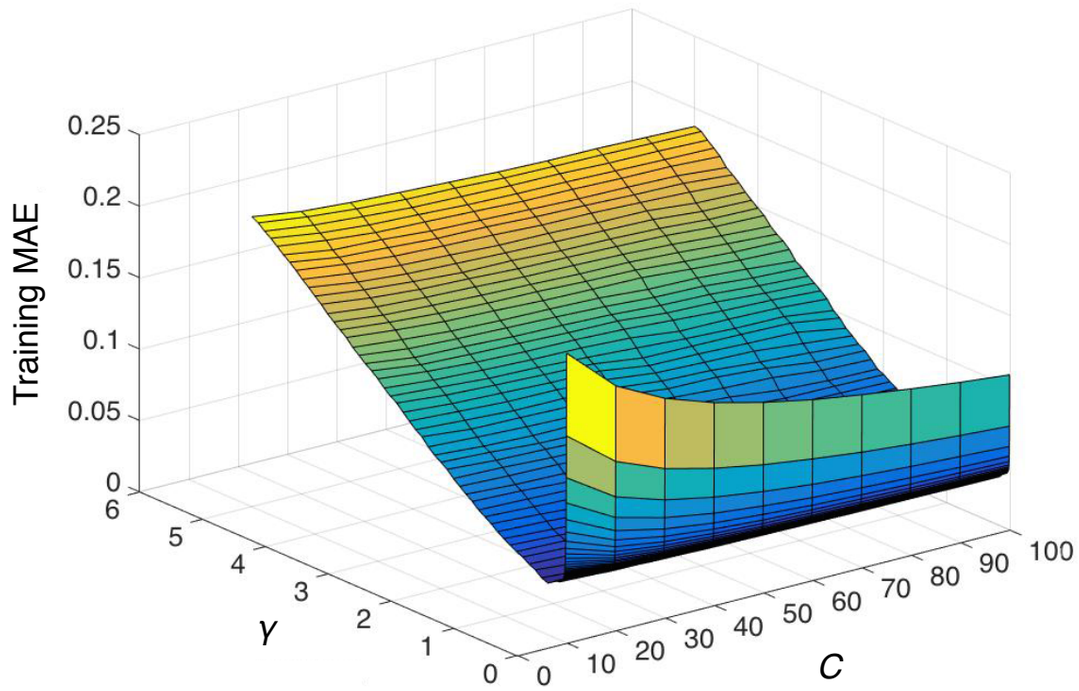

Gaussian Distributions with different **widths**



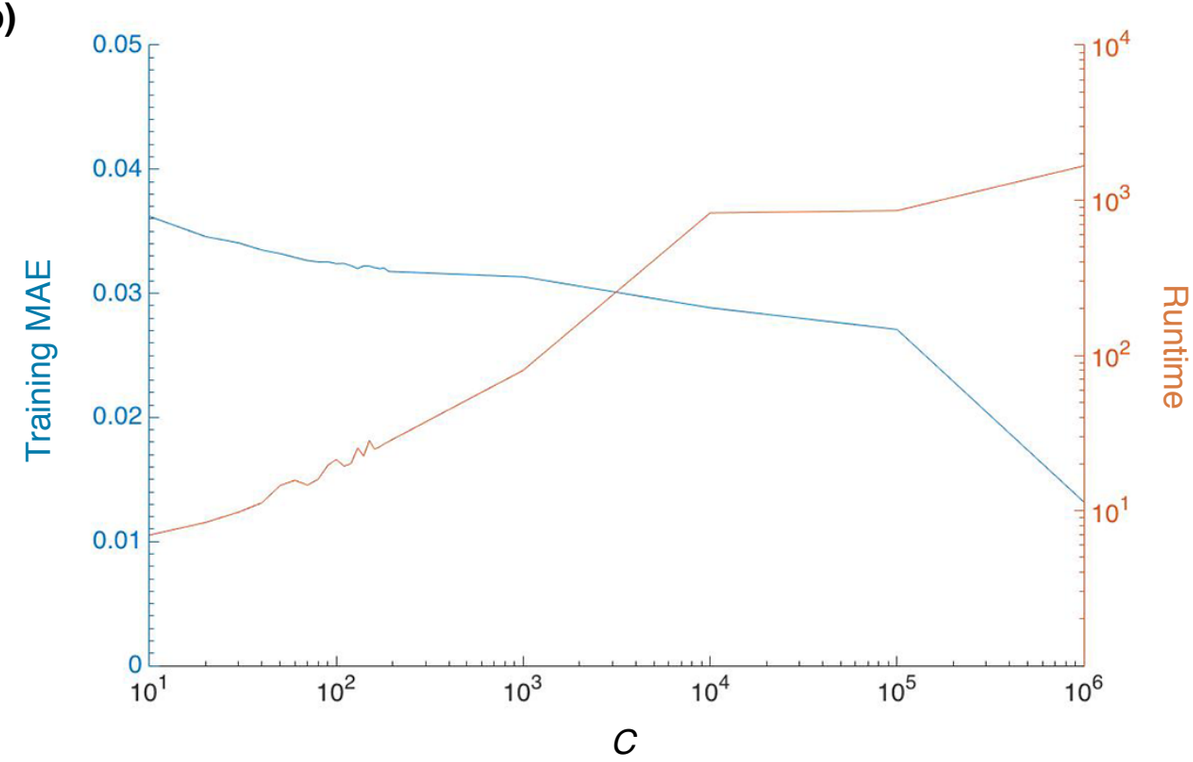

Gaussian distributions with different **means**

# Selection of $C$ and $\gamma$

(a)



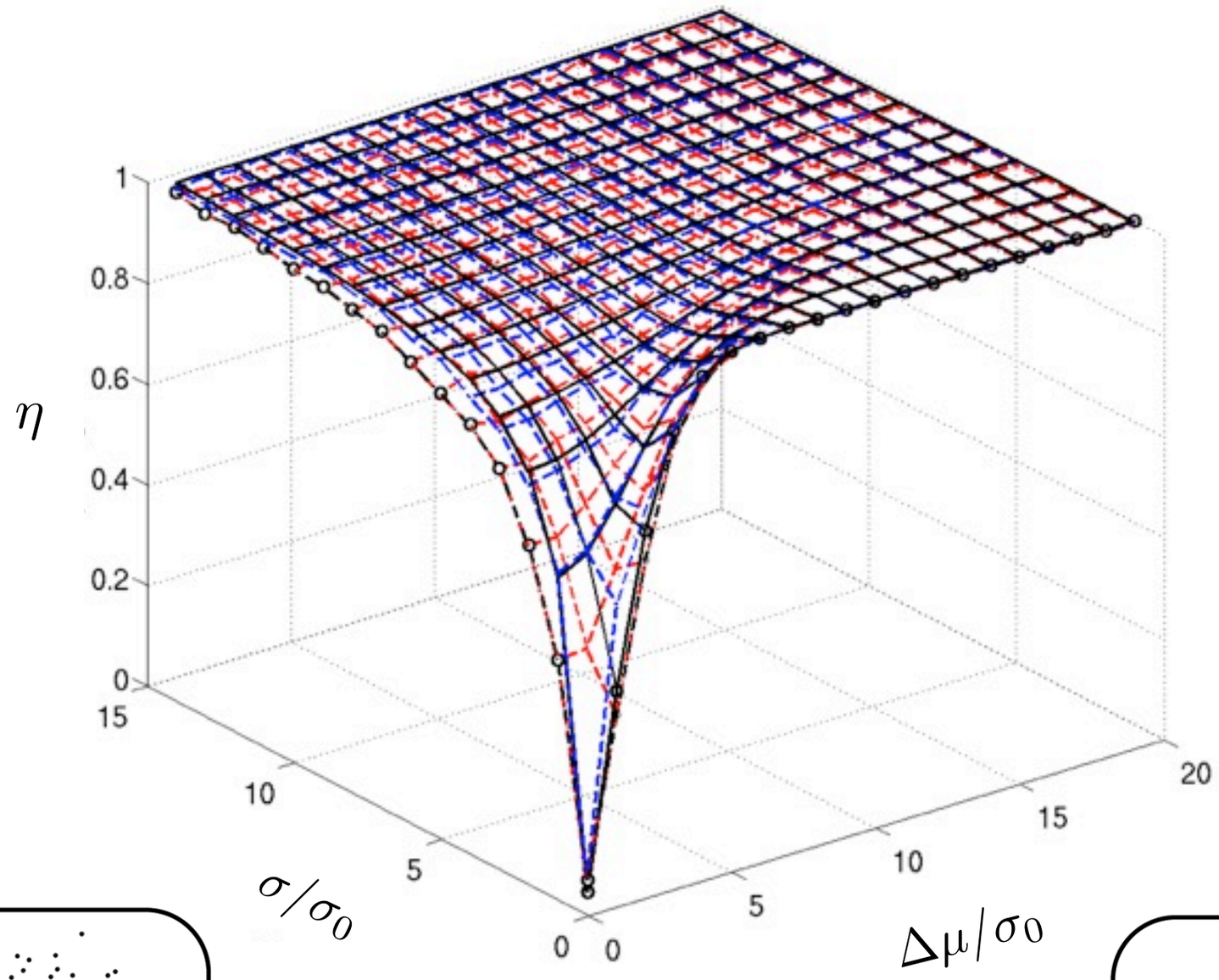

(b)



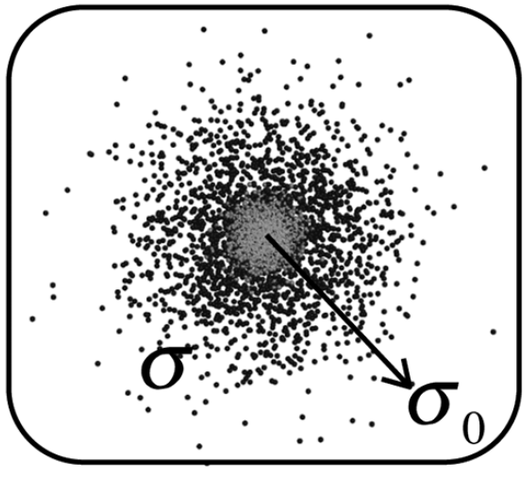

Leighty and Varma, JCTC 2013

Dutta, Siddiqui, Botlani, Varma. BJ 2016

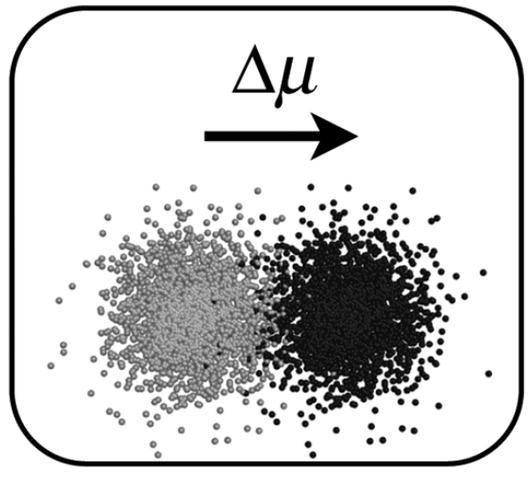
# Performance




Gaussian Distributions with different **widths**



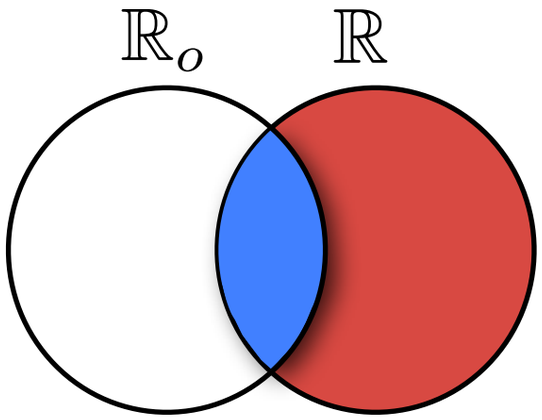

Gaussian Distributions with different **means**



# Performance

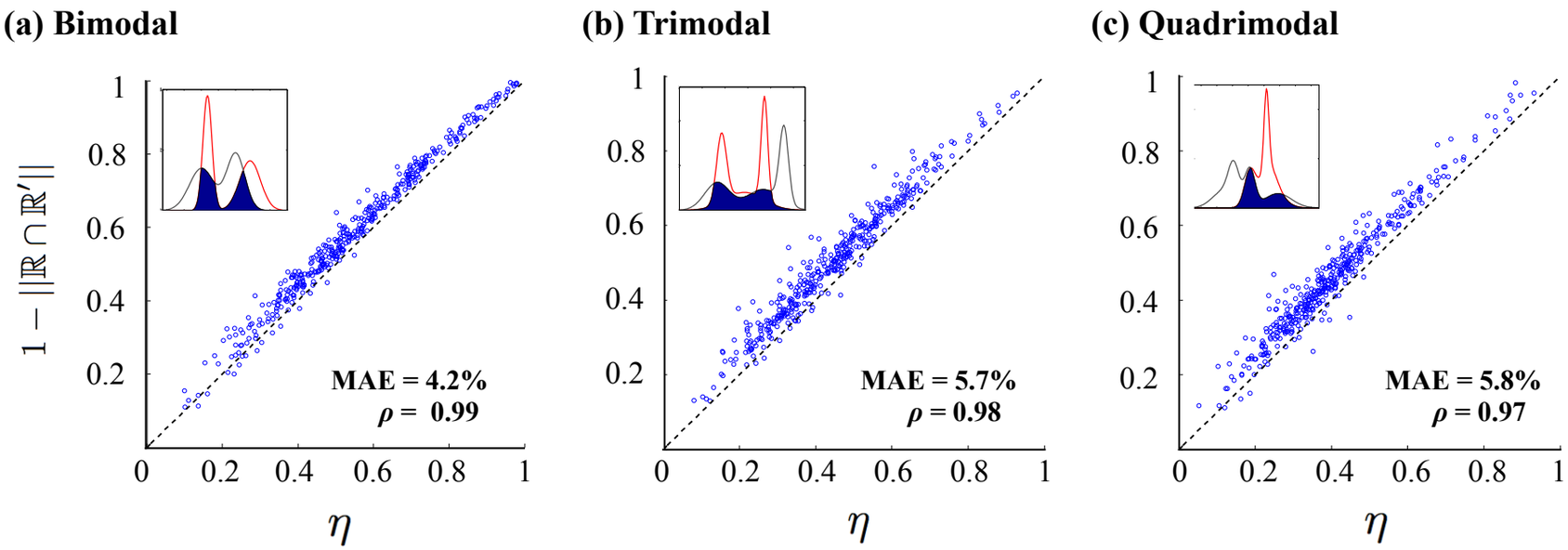
## Works for multi-modal distributions

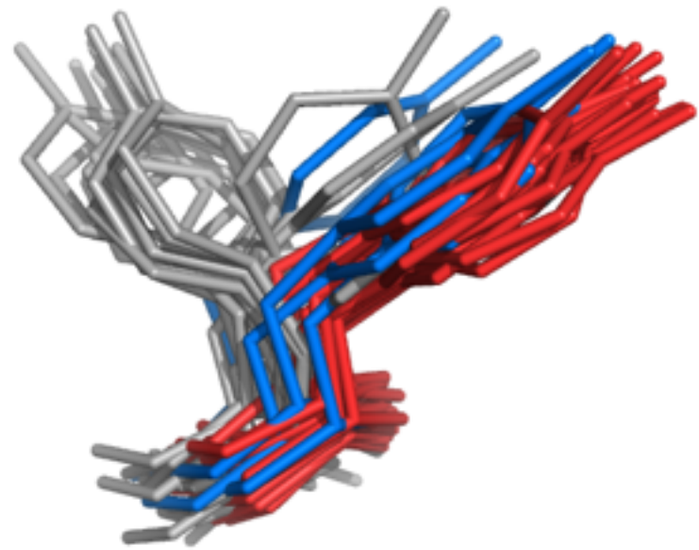
$$\eta = 1 - ||\sum_{i=1}^{n} c_i f_i \cap \sum_{j=1}^{n} c'_j f'_j|| = 1 - ||(\sum_{i,j=1}^{n} c_i f_i \cap c'_j f'_j||$$




$$\mathbb{R}_I = \mathbb{R} \cap \mathbb{R}_o$$

$$\mathbb{R}_\Delta = \mathbb{R} \setminus \mathbb{R}_o$$



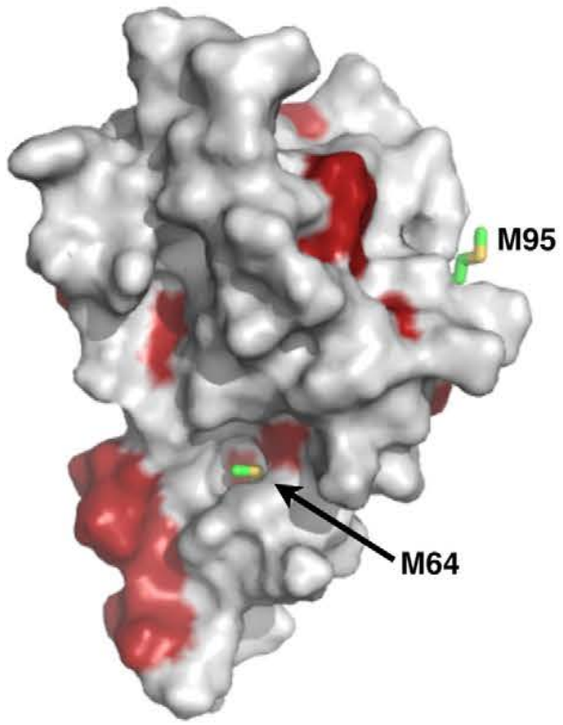



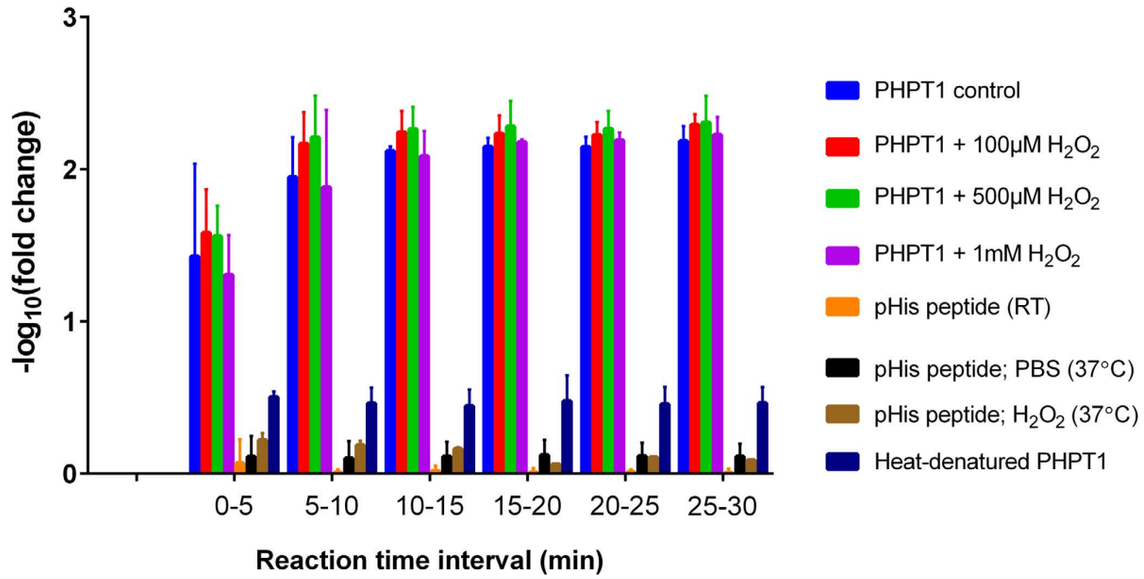
Dutta, Siddiqui, Botlani, Varma. BJ 2016

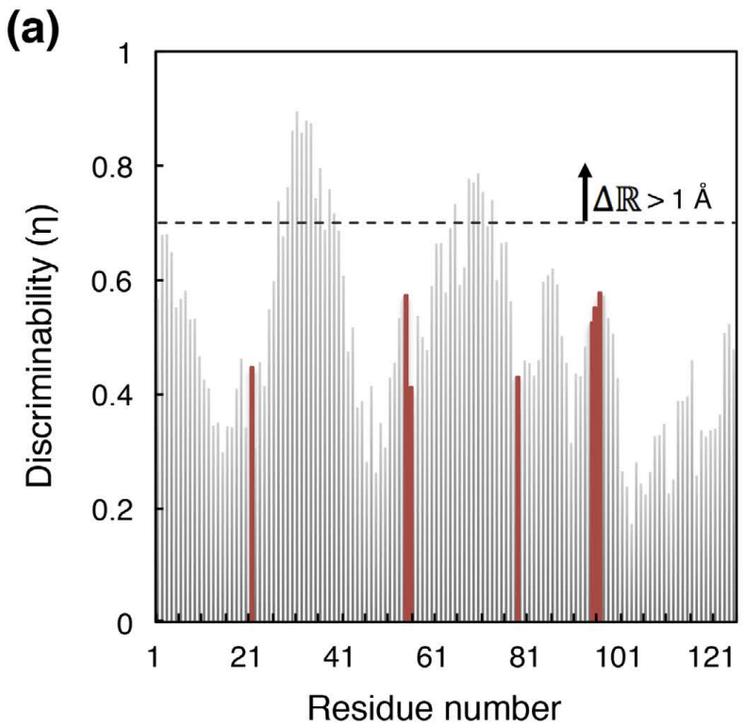
# New applications enabled

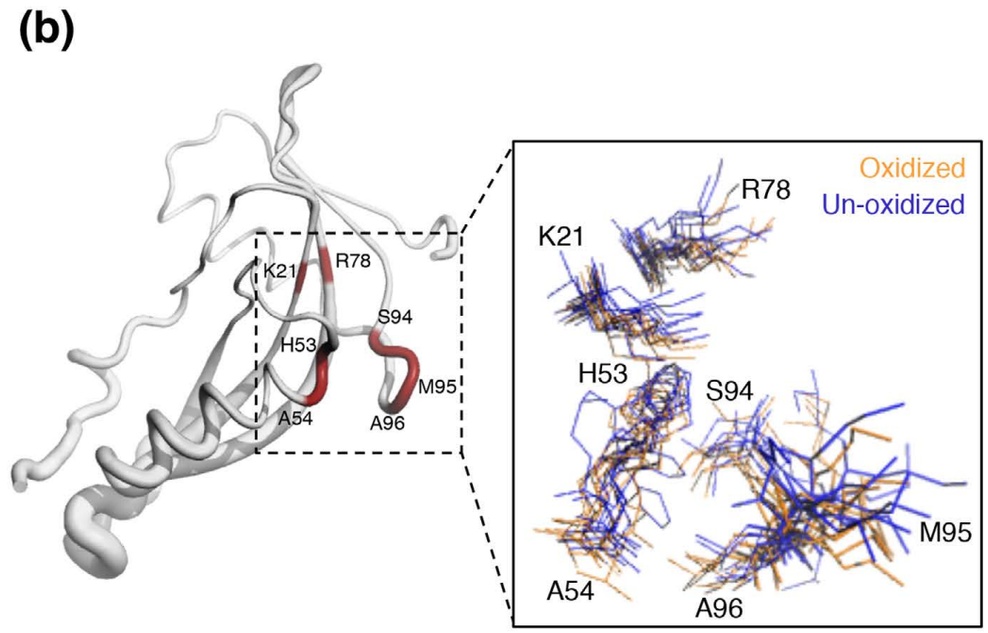
## Simplifies data representation and analysis

Effect of oxidation on phosphatase structure and dynamics



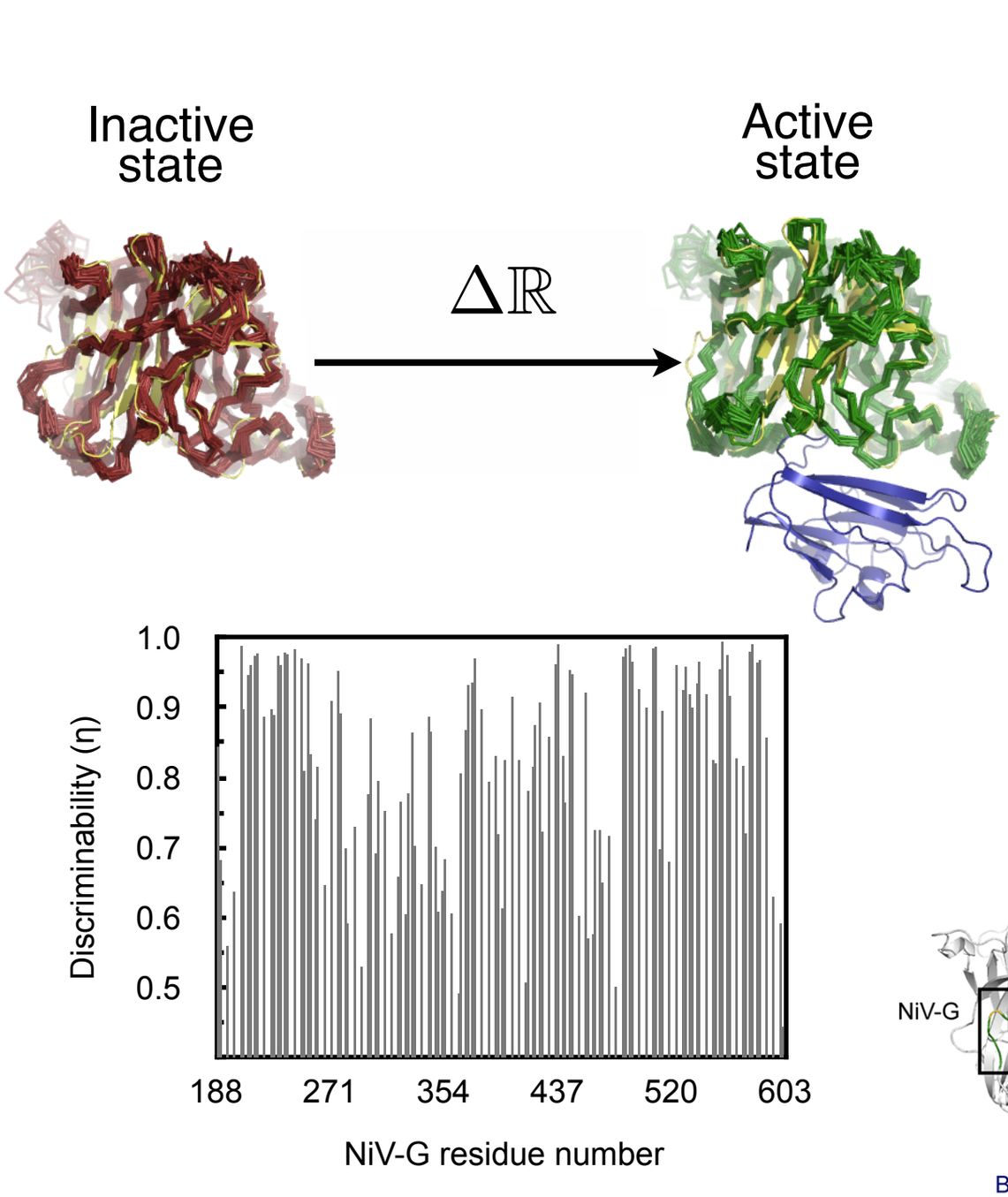



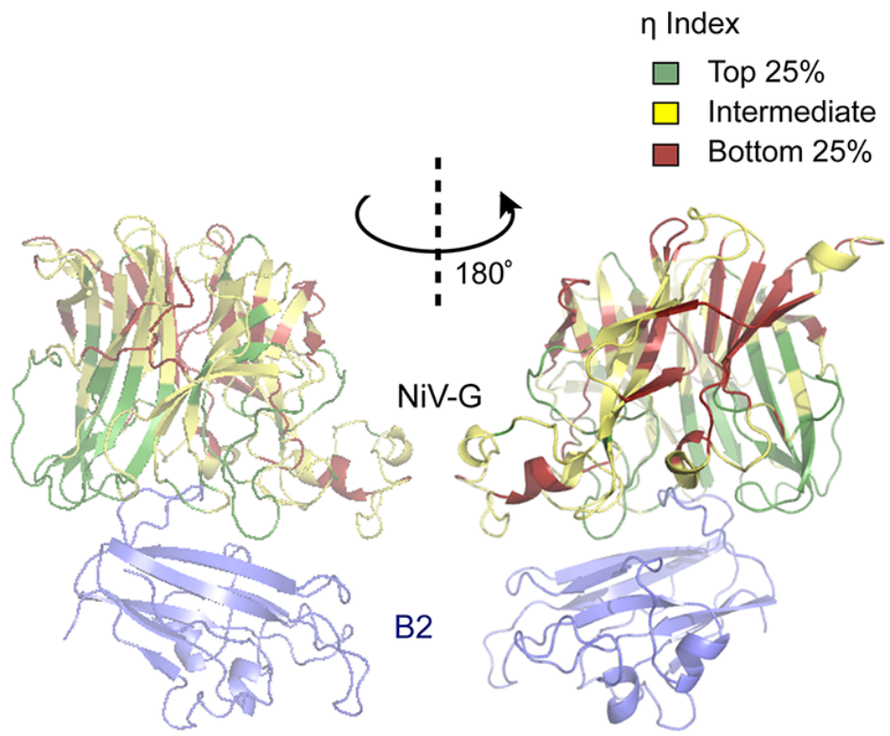



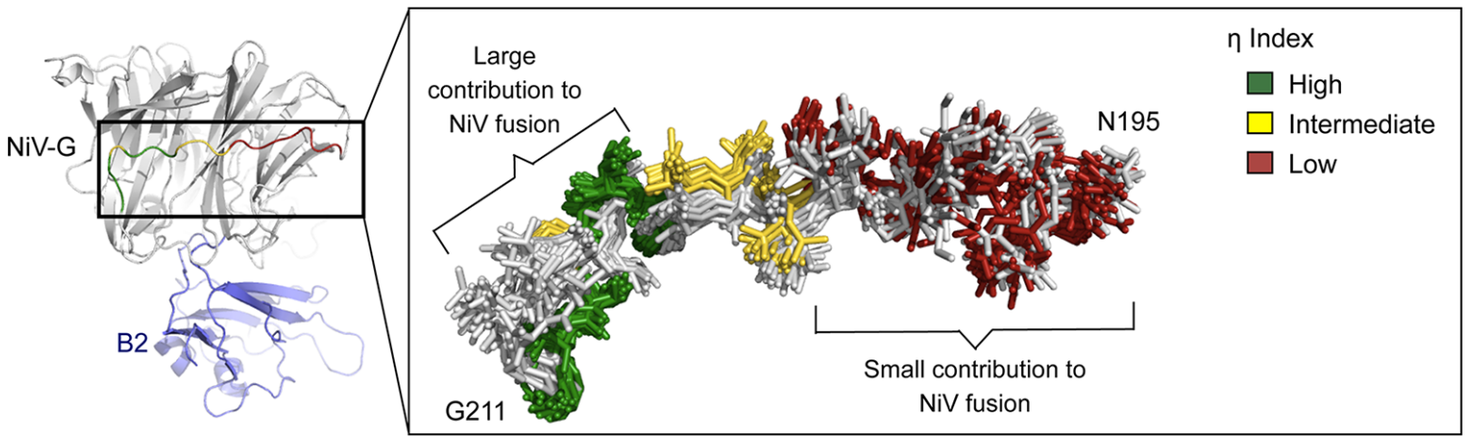

Martin et al. Sci Rep. 2016

# New applications enabled

## Permits ranking of population shifts



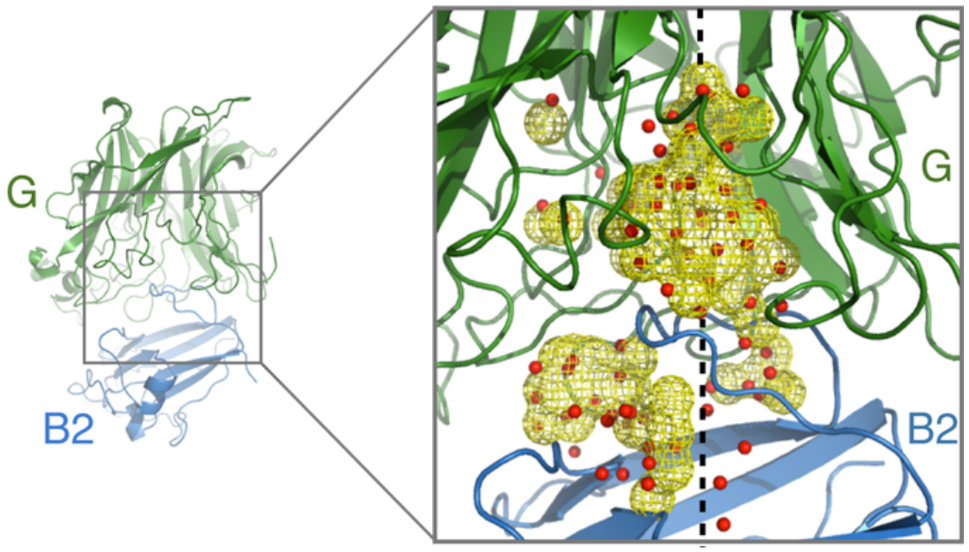




Leighty and Varma, JCTC 2013

# New applications enabled

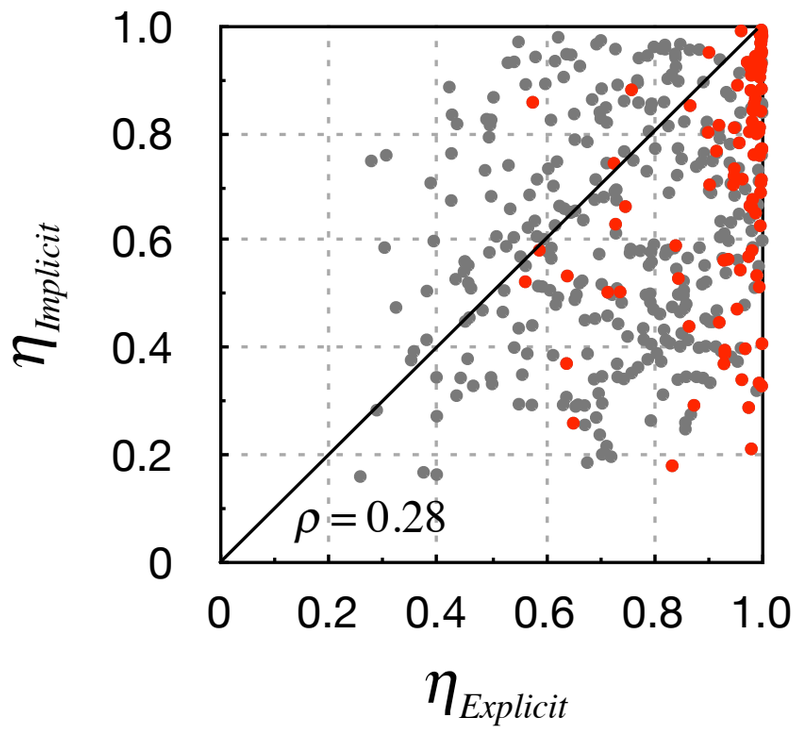
## Enables comparison of two population shifts

Determine role of the discrete nature of water on ligand-induced ensemble shifts




$$\mathbb{R}_{G}^{Explicit} \xleftrightarrow{\eta_{Explicit}} \mathbb{R}_{G\text{-}B2}^{Explicit}$$

$$\mathbb{R}_{G}^{Implicit} \xleftrightarrow{\eta_{Implicit}} \mathbb{R}_{G\text{-}B2}^{Implicit}$$




Dutta, Botlani, Varma. JPCB 2014

# New applications enabled

## Enables comparison of multiple population shifts

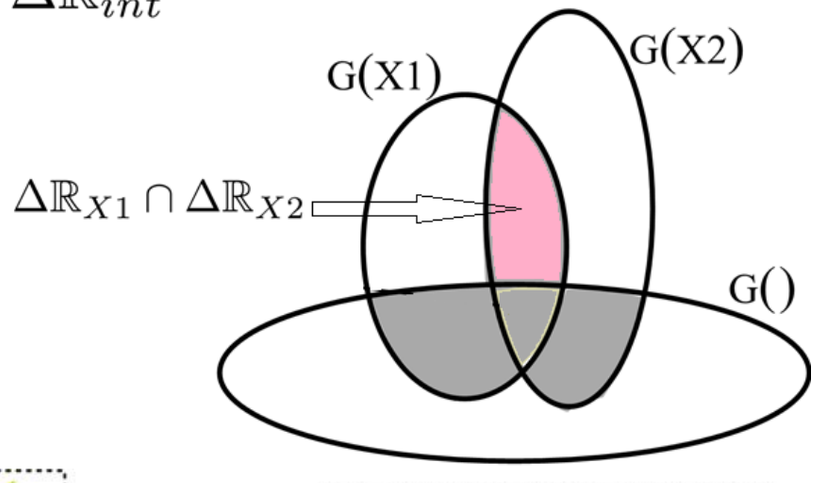
**Intersecting signaling pathways:** $\Delta\mathbb{R}_{signal} \subset \Delta\mathbb{R}_{int}$

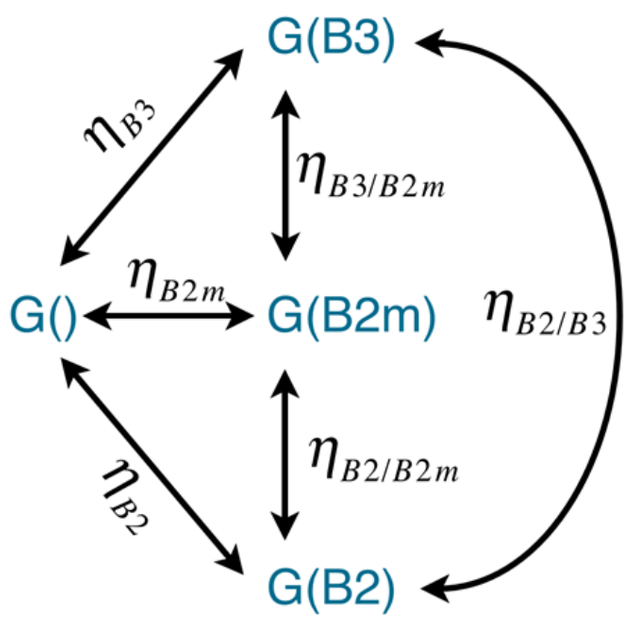
$$\Delta\mathbb{R}_{int} := \Delta\mathbb{R}_{B2} \cap \Delta\mathbb{R}_{B3} \cap \Delta\mathbb{R}_{B2m}$$

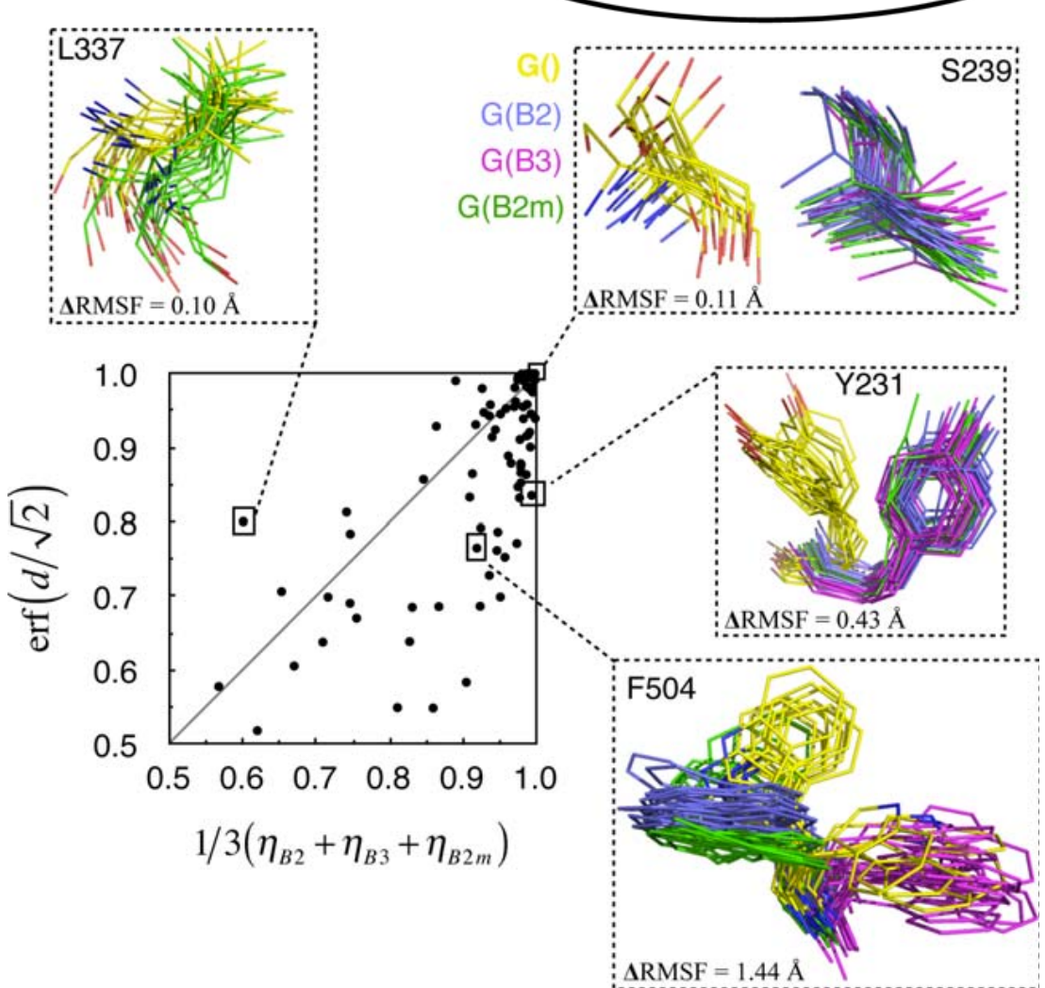
$\Delta\mathbb{R}_{X1} \cap \Delta\mathbb{R}_{X2}$ is defined by the inequalities:

1) $\min\{\eta_{X1}, \eta_{X2}\} > \eta_{X1/X2}$

2) $|\eta_{X1} - \eta_{X2}| < \langle \eta_{X1} - \eta_{X2} \rangle$



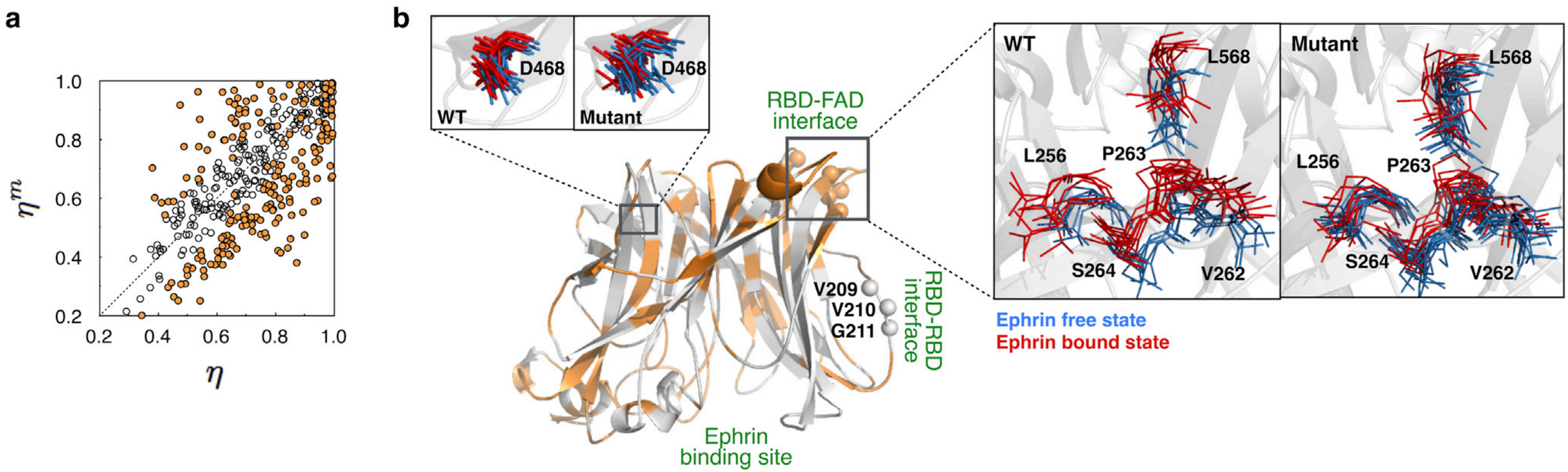




Varma, et al.
Proteins 2014

# New applications enabled

## Enables comparison of multiple population shifts

Effect of mutation on ligand-induced population shifts




Set of residues affected by mutation satisfy at least one of the three conditions

$$|\eta - \eta^m| > 2 \times \text{MAE},$$
$$\eta_{\text{apo}} > \text{erf}\left(1/\sqrt{2}\right), \text{ and}$$
$$\eta_{\text{bnd}} > \text{erf}\left(1/\sqrt{2}\right).$$

Dutta, et al. BJ 2016

# Source code and tutorials @ SimTK

simtk.org

SimTK | Search for projects | Projects | About | Sign Up | Log In

## Statistical analysis of conformational ensembles

Follow (0)

About | Downloads | Documents | Forums | Source Code | Issues

Provide user-friendly codes and algorithms written using GROMACS/CHARMM APIs for statistical analysis of conformational ensembles

This project provides computational tools and methods to analyze conformational ensembles of biomolecules, as well as their assemblies, such as those obtained from molecular simulations.

(A) PROTEINS: The molecular understanding of the functional regulation of proteins requires assessment of various states, including active and inactive states, as well as their interdependencies. For several proteins, their various states can be distinguished from each other on the basis of their minimum energy 3D structures. For many other proteins, like GPCRs, PDZ domains, nuclear transcription factors, heat shock proteins, T-cell receptors and viral attachment proteins, their states can be distinguished categorically from each other only when their finite-temperature conformational ensembles are considered alongside their minimum-energy structures. We are developing tools/methods for:
(A1) Direct comparison of conformational ensembles - The traditional approach to compare two or more conformational ensembles is to compare their respective summary statistics. This approach is, however, prone to artifactual bias, as data is compared after dimensionality reduction. The proper way to compare ensembles is to compare them directly with each other and prior to any dimensionality reduction. g_ensemble_comp is a tool we have developed that does just that and reports the difference between ensembles in terms of a true metric defined by the zeroth law of thermodynamics.
(A2) Prediction of allosteric signaling networks - method under development.

(B) LIPID MEMBRANES: The surface area of a lipid bilayer is related fundamentally to many other observables, such as thermal phase transitions and domain formation in mixed lipid bilayers. We have developed g_tessellate_area to compute the 3D surface area of a bilayer using Delunay tessellation.

111 downloads

1 forum posts

Last updated Jul 19, 2016

Project Statistics

Join Mailing Lists

Suggest Idea

Sameer Varma

SimTK is maintained through Grant R01 GM107340 from the National Institutes of Health (NIH). It was initially developed as part of the Simbios project funded by the NIH as part of the NIH Roadmap for Medical Research, Grant U54 GM072970.

Our Pledge and Your Responsibility

Feedback About Join



Dutta, Siddiqui, Botlani, Varma. BJ 2016

# Sincerely thankful to...

## Contributing lab members

Mohsen Botlani

Ralph Leighty

Ahnaf Siddiqui

Priyanka Dutta

## Collaborators

### Machine learning

Klaus Muller
(TU Berlin)

### Protein activity & regulation

Matteo Porotto
(Columbia U)

Stan Stevens
(USF)

## Funding

DARPA

NSF

NIH National Institutes of Health

University of South Florida 1956

# Thank you.