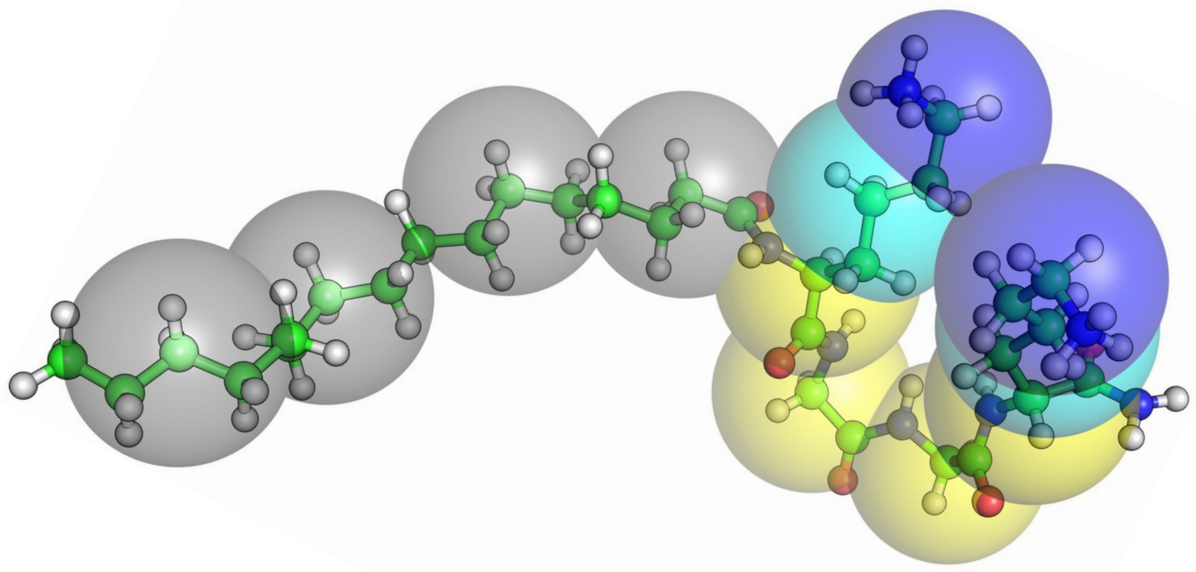
# COARSE GRAINING IN PHYSICAL SPACE

Lorenzo Boninsegna & Cecilia Clementi

Long Program EL2017



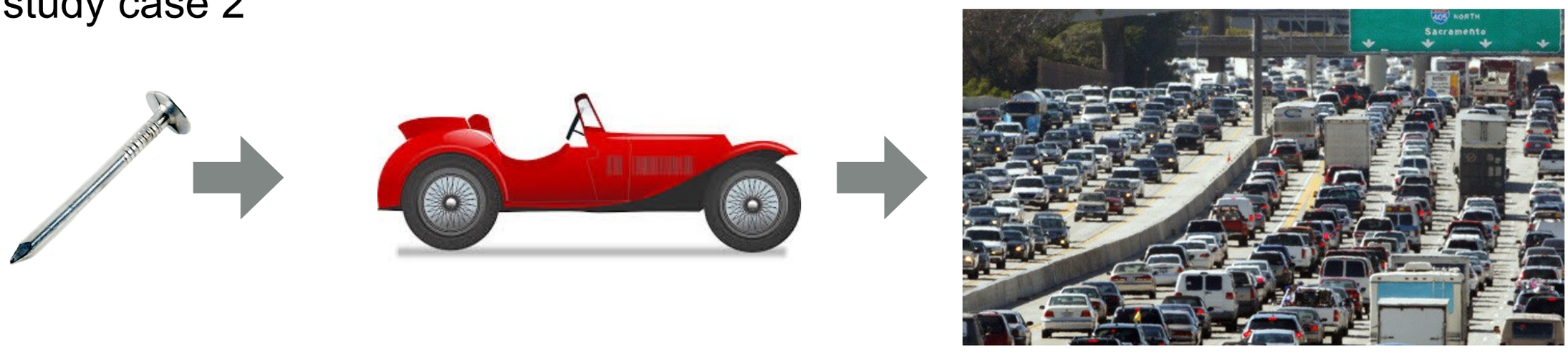
# OUTLINE

- MOTIVATION
- COARSE GRAINING (CG) INGREDIENTS
- COARSE GRAINING STRATEGIES
- CHALLENGES:
  - IDENTIFICATION OF CG BEADS
  - REPRODUCING THE CORRECT DYNAMICS
- INCORPORATING EXP DATA
- OUTLOOK

# HIERARCHY IN PHYSICAL SYSTEMS

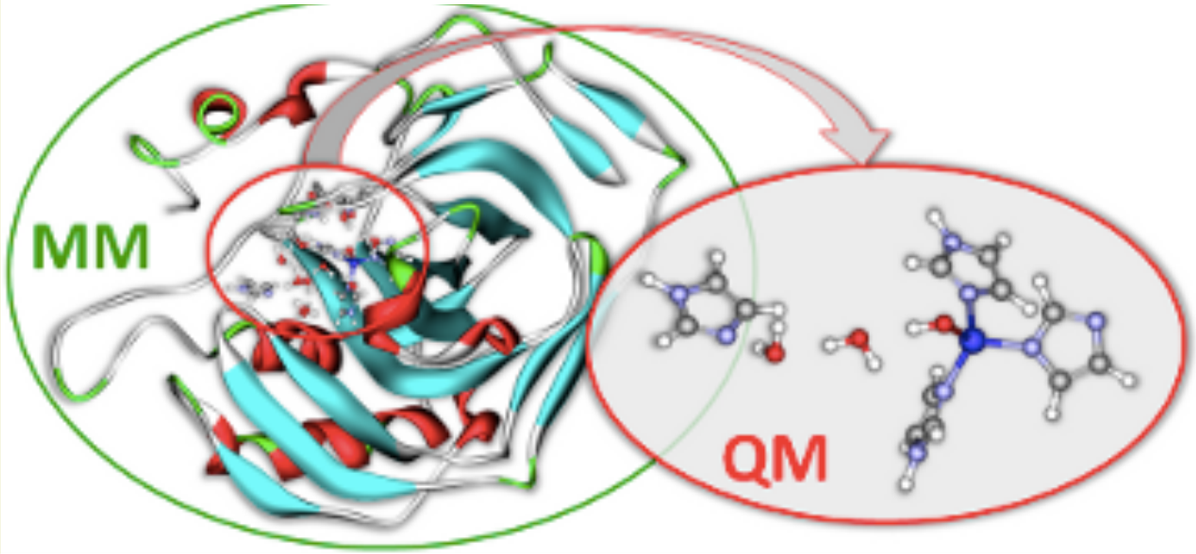
study case 1

study case 2




assembling of microscopic constituents impacts mesoscopic features

## QM/MM

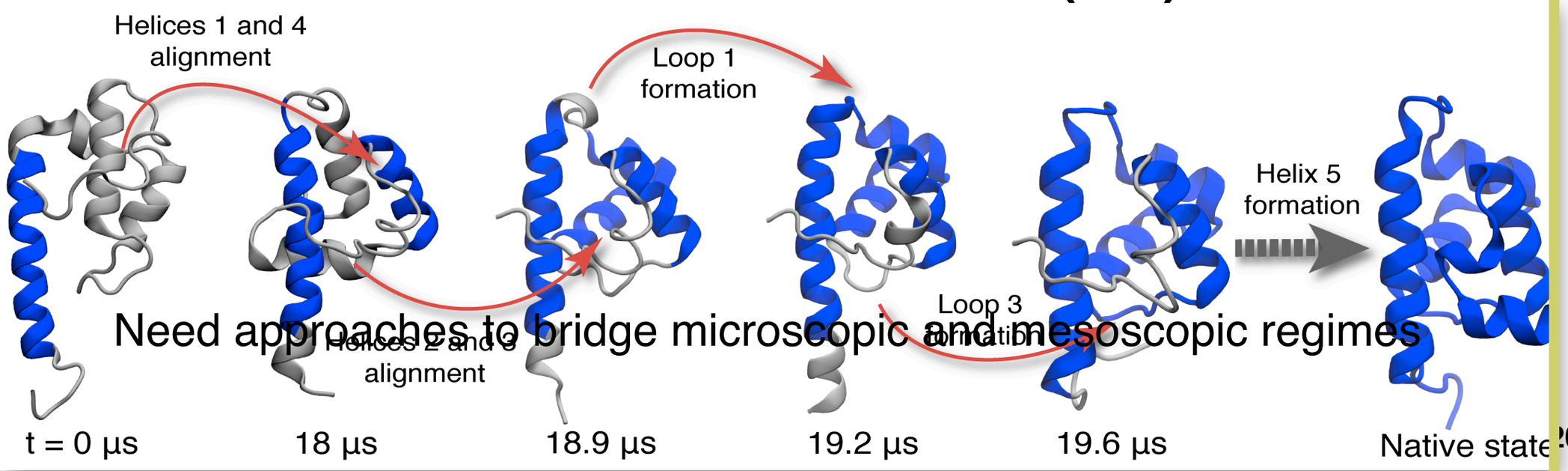



## RXN-DIFFUSION





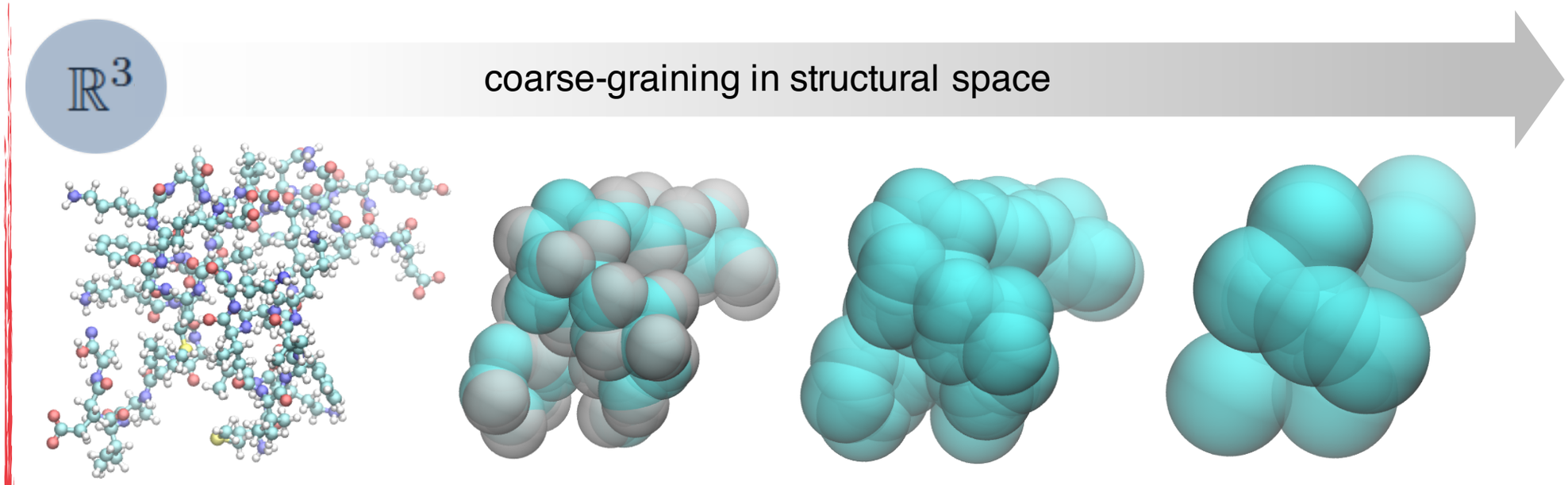
## MOLECULAR DYNAMICS (MD)




Need approaches to bridge microscopic and mesoscopic regimes

**2002**), 3, 754

# BIOMOLECULES AND COARSE-GRAINING




Instead of distilling information from complex data, use data to **make system** intrinsically **simpler**

- High dimensionality
- Sampling
- Interpretation

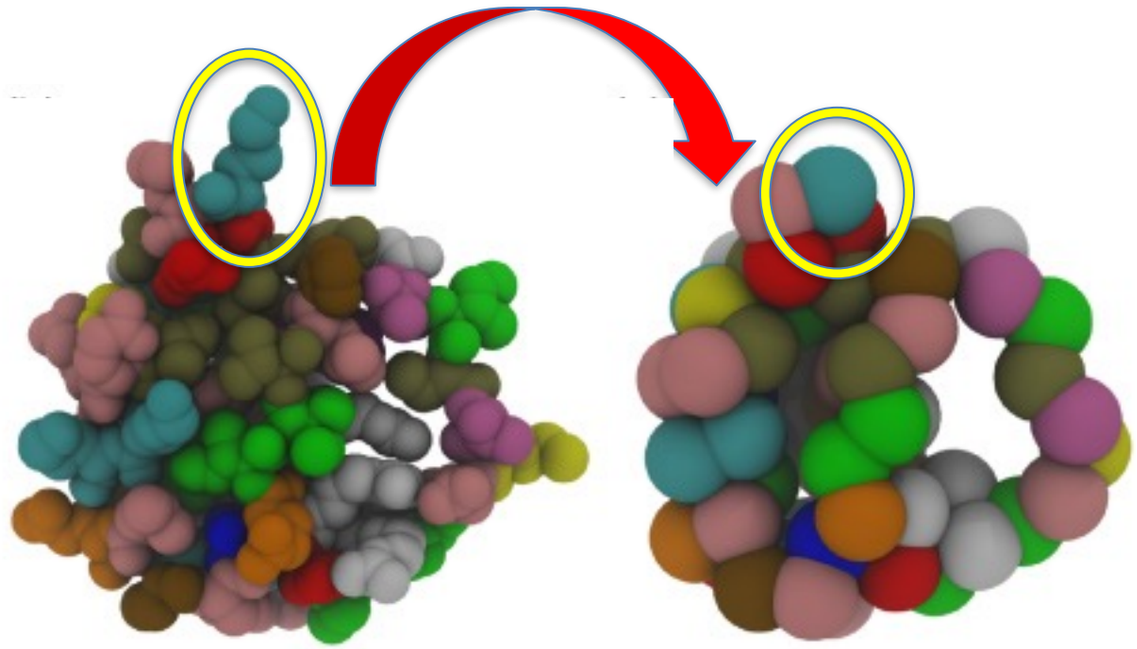
(courtesy of Cecilia Clementi)

# COARSE GRAINING IN  A NUTSHELL

Coarse-Graining (CG)

decimating degrees of freedom into effective ones

groups of atomic degrees of freedom…



all atom

coarse grained

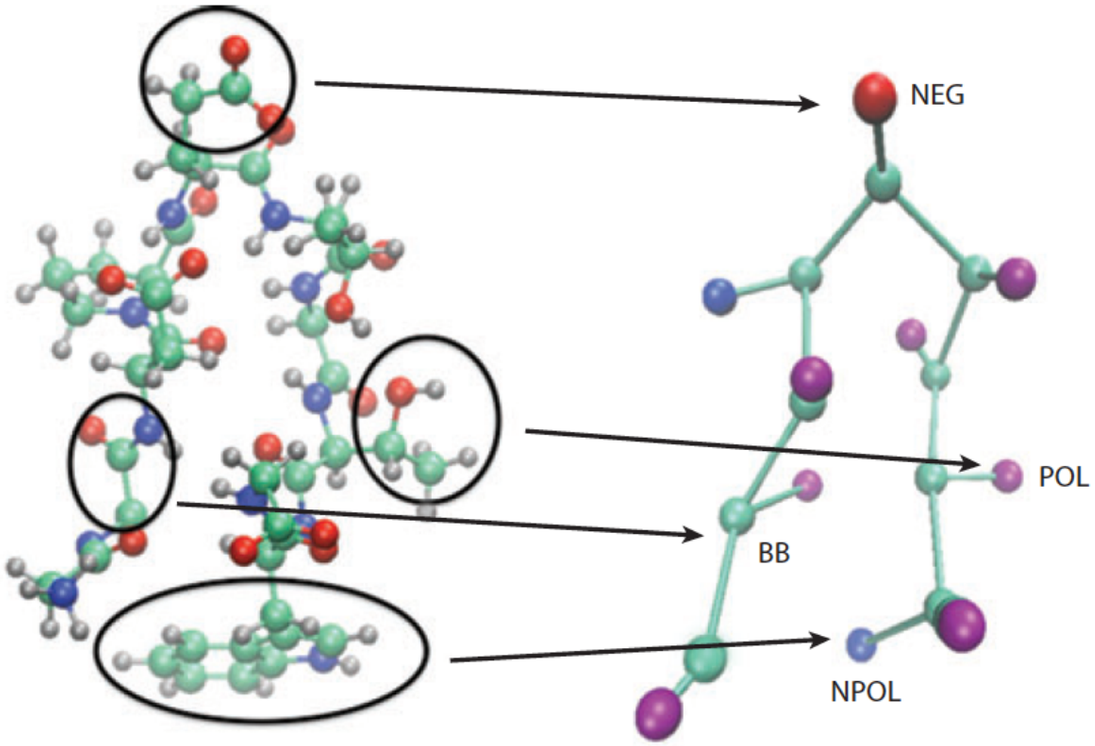
(courtesy of Fernando Yrazu)

…lumped into one set of effective degrees of freedom (‘bead’)

effective interactions have to be computed too

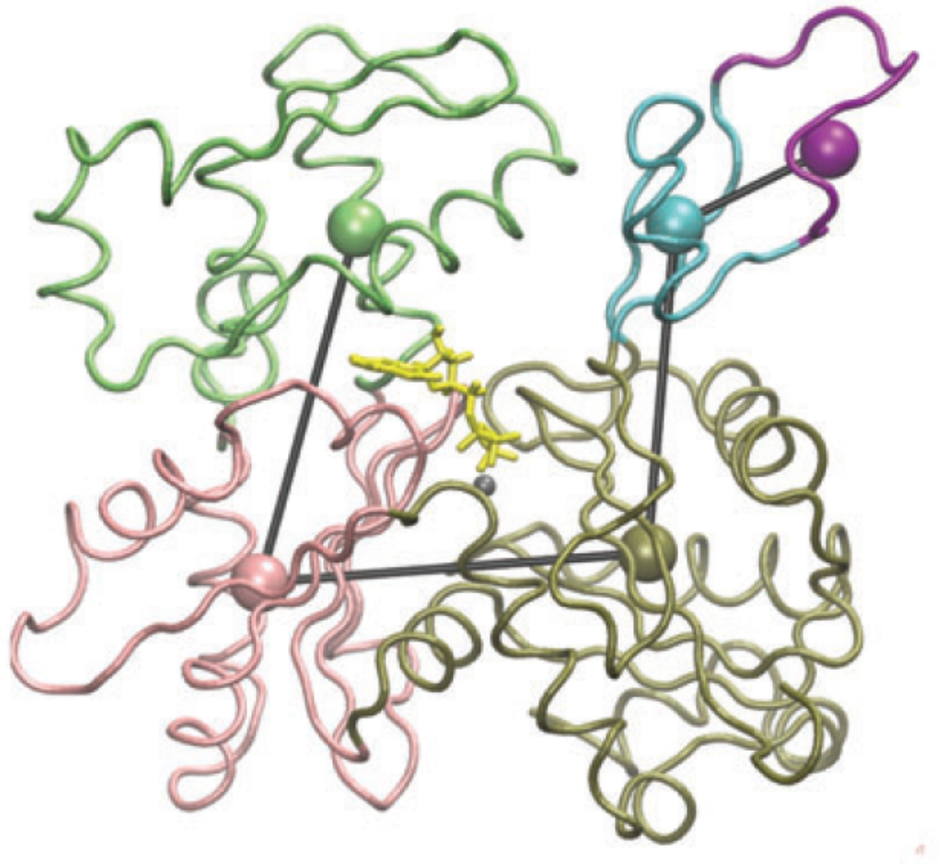
- Physical intuition or principles usually guide the choices
- Model needs to retain ‘important’ features
- Increasing interest in data-driven, automatic procedures

# RESOLUTION IS KEY




low resolution, ultra CG
many atoms/residues are grouped
into a single CG bead

high resolution, properties are
preserved by different CG sites



Saunders, *Annu. Rev. Biophys.* (**2013**) 42, 73

# THE CG INGREDIENTS

**…effective degrees of freedom** $\xi_i$

identify groups of atoms moving together in the dynamics

$$(\mathbf{x}_1, \cdots, \mathbf{x}_N) \xrightarrow{\Phi} (\xi_1, \cdots, \xi_M)$$

+

**…effective Hamiltonian** $\mathcal{H}$

project the all atom Hamiltonian onto reduced space

$$H(\mathbf{x}_1, \cdots, \mathbf{x}_N) \xrightarrow{\Gamma} \mathcal{H}(\xi_1, \cdots, \xi_M)$$

$(M < N)$

=

CG MODEL

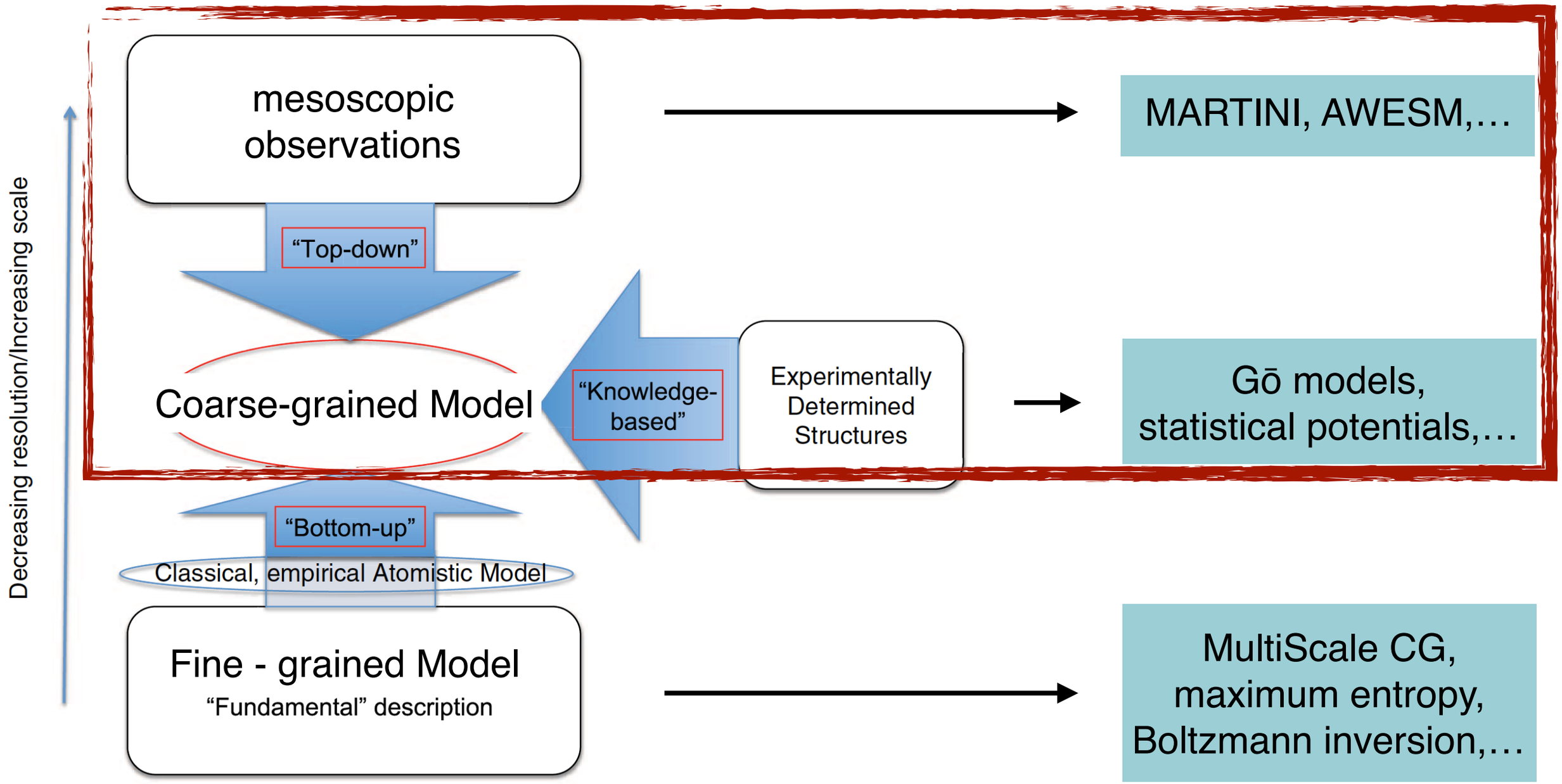
**CG thermodynamics**

**CG dynamics**

are these the actual low resolution description of the FG properties?

# COARSE GRAINING PHILOSOPHIES

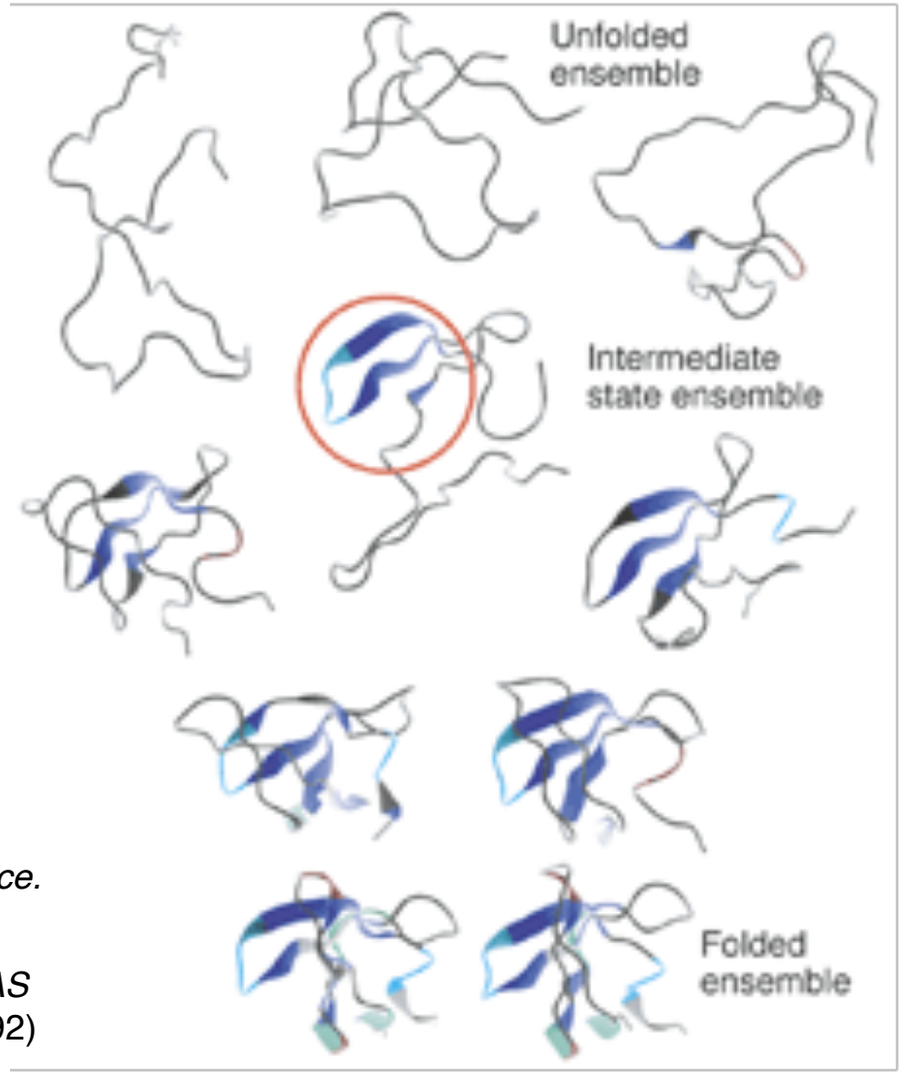
Problem is formidable and can be tackled from alternative points of view

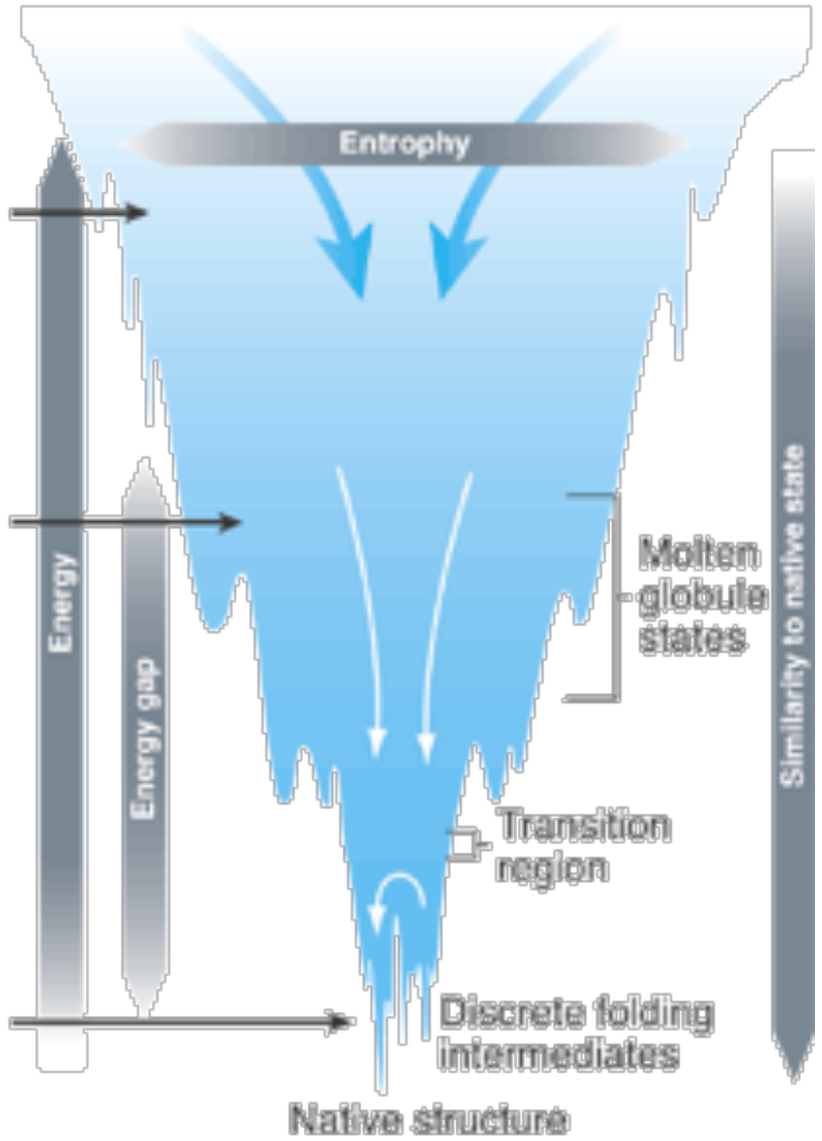




Noid, *J. Chem. Phys* (**2013**), 139, 090901
Saunders and Voth, *Ann. Rev. Biophys.* (**2013**), 42, 73

# STRUCTURE - BASED MODELS: PREQUEL




Brooks *et al.*, *Science.* **293**, 612 (2001)

Leopold *et al.*, *PNAS USA.* **89**, 8721 (1992)

Natural protein energy landscape resembles a **rugged funnel**

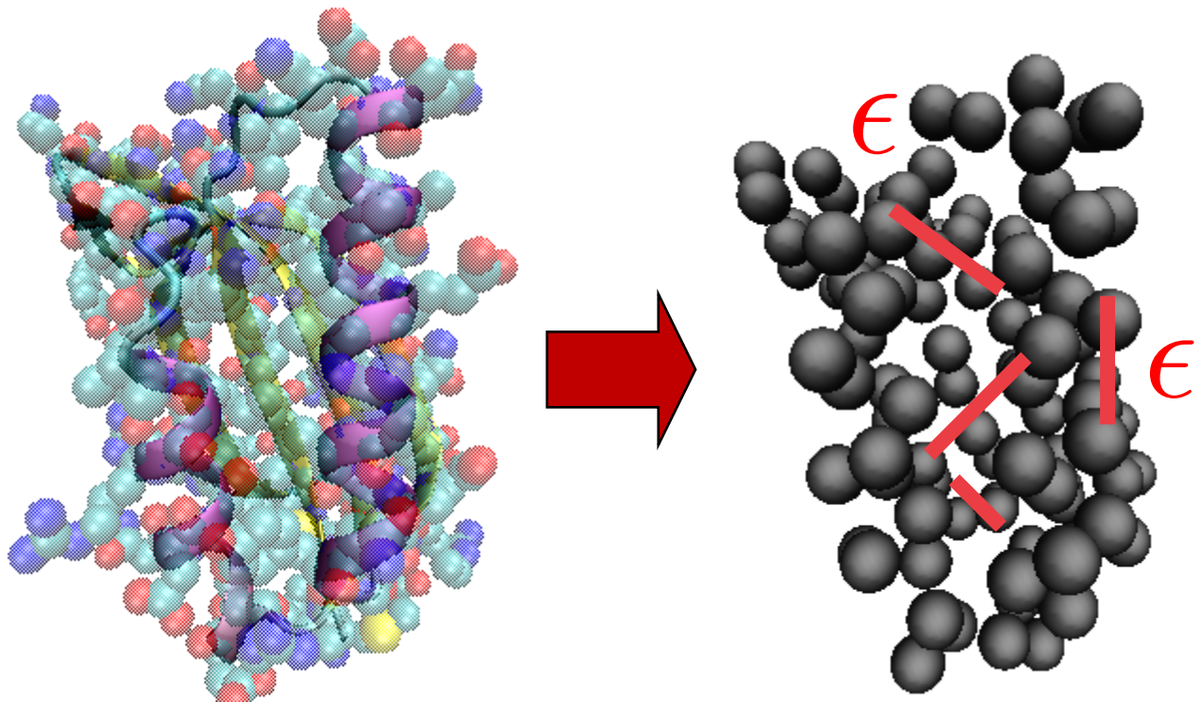
frustration, kinetic traps in the funnel envelope

Hp: native structures are mostly favored over all others

# STRUCTURE - BASED MODELS: ZEROTH ORDER

Theoretical justification in **principle of minimal frustration**:

proteins have evolved their ability to fold by minimizing frustration




**native contacts**

one CG bead per residue

$$\mathcal{H}_{Go} = -\sum_{native} \epsilon$$

smooth, funneled landscape
(ideal gas)

Despite simplicity, many successful studies:

- Das et al., PNAS (**2005**), **102**, 14569
- Hyeon et al., PNAS (**2007**), **104**, 2175
- Koga et al., PNAS (**2006**), **103**, 5367

....

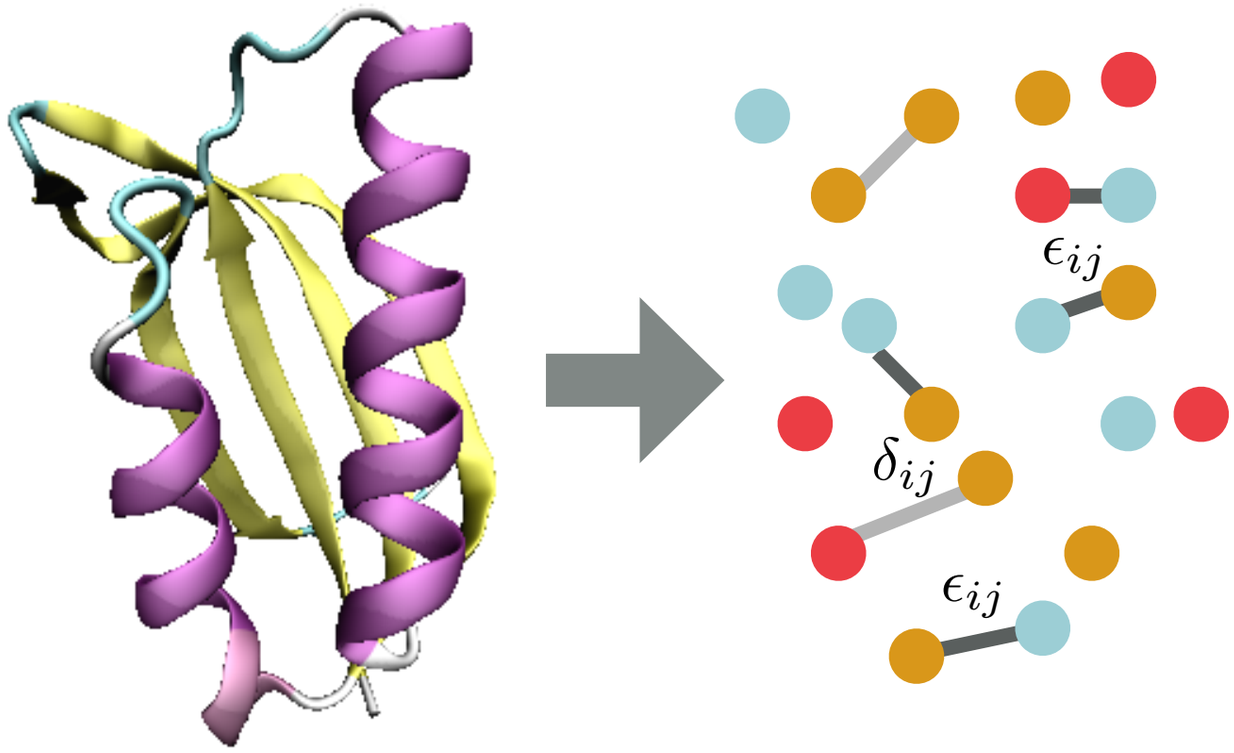
Clementi, Curr. Opin. Struct. Biol. (**2008**),18,10
Clementi, Nymeyer & Onuchic (JMB **2000**)

# STRUCTURE - BASED MODELS: CORRECTIONS

Beyond zeroth order: $\mathcal{H}_{SBM} = \mathcal{H}_{Go} + \sum_{native} \delta\epsilon_{ij} + \sum_{non-native} \delta_{ij} + \cdots$

**heterogeneity** **frustration**

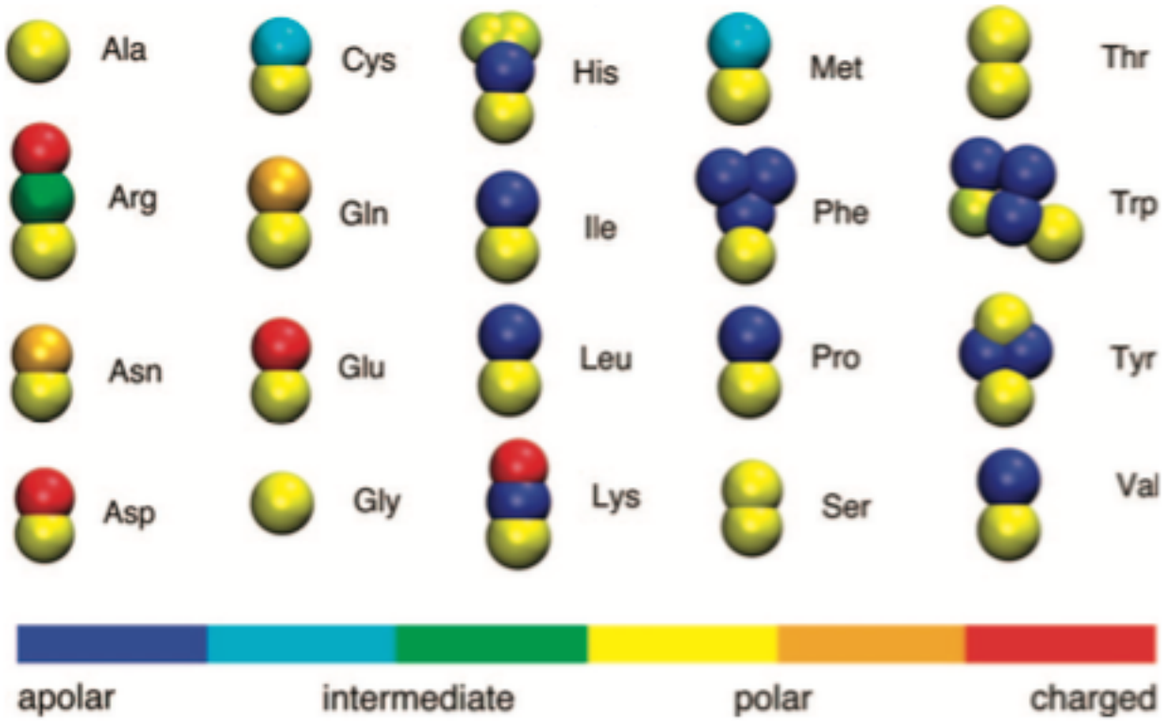
This is necessary to account for compact non native states or misfolded structures



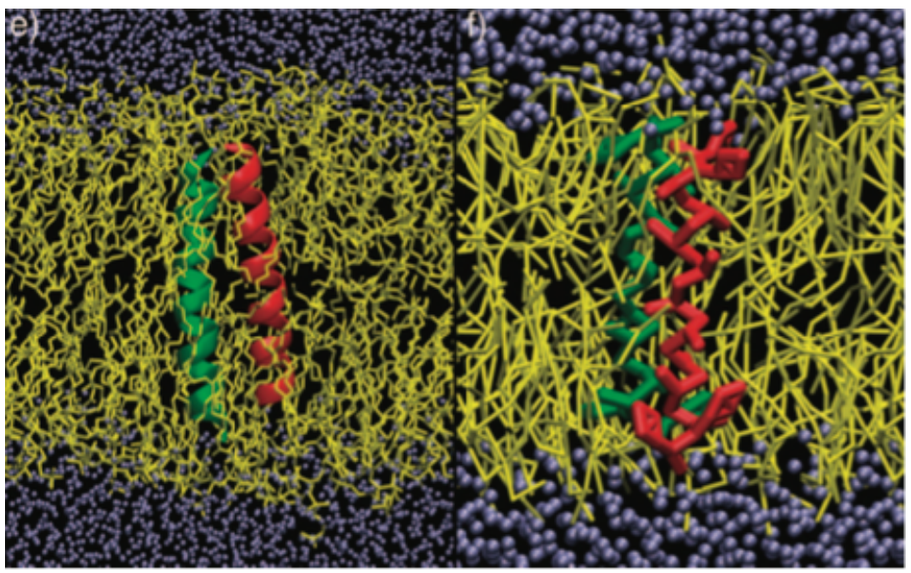

Performance improving upon incorporating experimental data (more later)

Effect of non-native interactions on protein folding
Clementi & Plotkin (Protein Science 2004)
Quantification of structural frustration
Chavez, Onuchic, & Clementi (JACS 2004)
Balance between energetic and entropic contributions
Das, Matysiak & Clementi (PNAS 2005)
Role of protein topology in modulating non-native frustration
Kluber, Burt, Clementi (2017, submitted)

# MARTINI FORCE FIELD




FG representation of the amino acids. Different colors represent different particle types



CG representation of a WALP23 peptide in a double lipid layer

- four-to-one mapping
- four main types of interaction sites, different levels
- integrated in GROMACS

non bonded potential

$$\mathrm{V}_{\text{Lennard-Jones}}(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

bonded potential

$$V_{\mathrm{b}} = \frac{1}{2}K_{\mathrm{b}}(d_{ij} - d_{\mathrm{b}})^2$$

$$V_{\mathrm{a}} = \frac{1}{2}K_{\mathrm{a}}[\cos(\varphi_{ijk}) - \cos(\varphi_{\mathrm{a}})]^2$$

$$V_{\mathrm{d}} = K_{\mathrm{d}}[1 + \cos(n\psi_{ijkl} - \psi_{\mathrm{d}})]$$

$$V_{\mathrm{id}} = K_{\mathrm{id}}(\psi_{ijkl} - \psi_{\mathrm{id}})^2$$


Marrink *et al.*, *J. Phys. Chem. B* (**2004**) 108, 750
Monticelli *et al.*, *JCTC* (**2008**) 4, 819

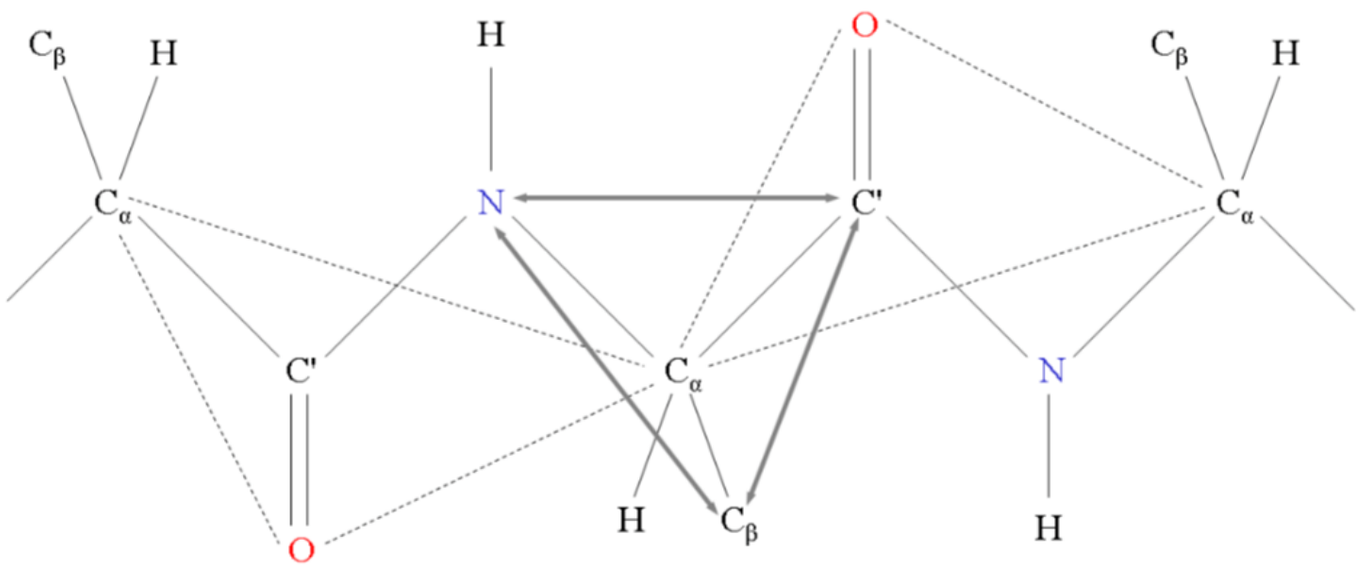
# AWSEM FORCE FIELD

**AWSEM** (Associative memory Water Mediated Structure and Energy Model)

both knowledge and physics based approaches, with a bioinformatically justified term which correspond to memories of proteins structures with similar sequences

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{helical} + V_{FM}$$

| backbone structure | tertiary contact interactions | burial preference | hydrogen bonds | local structure bias |
|---|---|---|---|---|



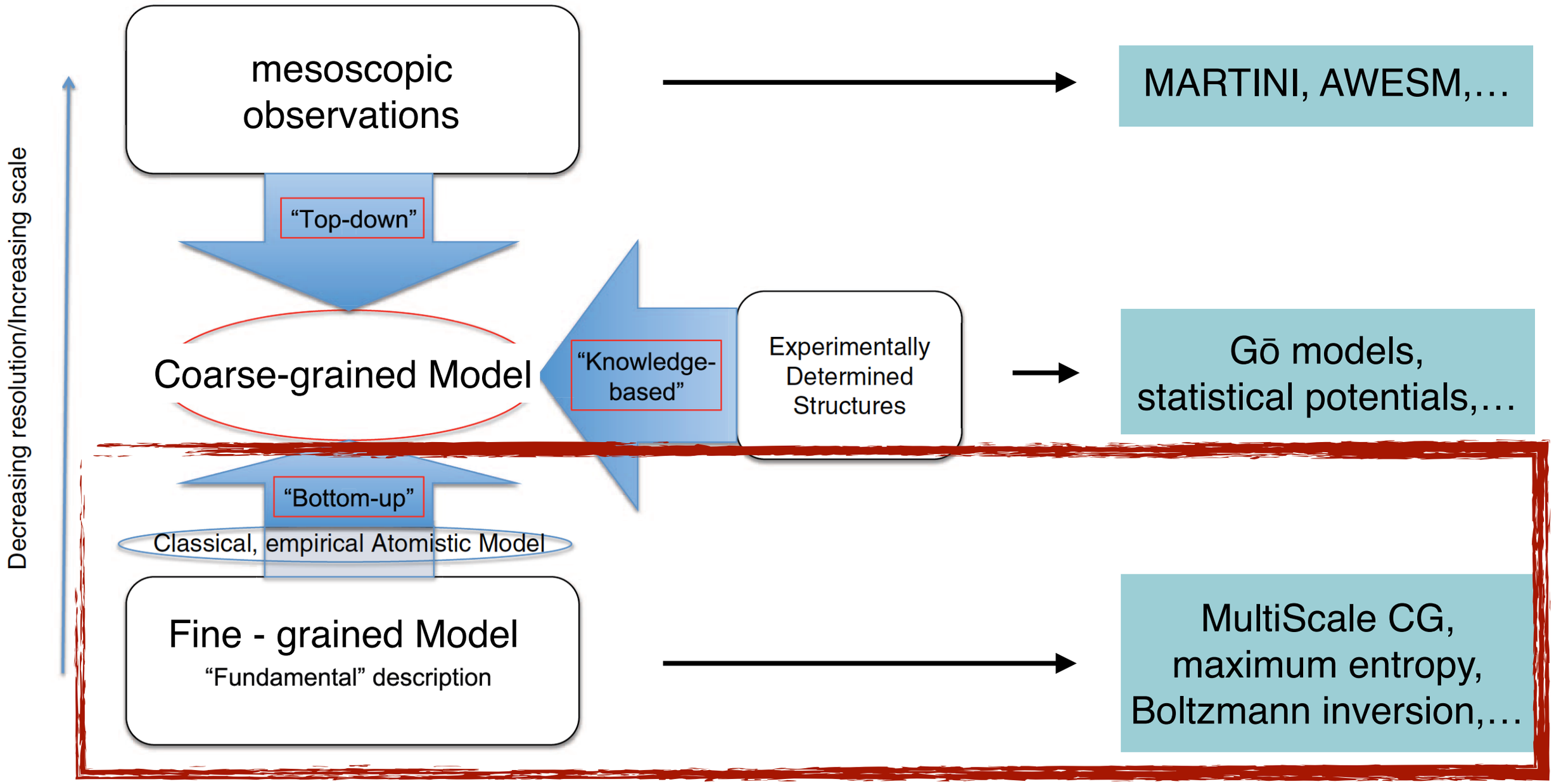

Successful at performing *de novo* protein structure prediction

Parameters optimization by energy landscape theory

Davtyan *et al.*, J. Chem. Phys. B (**2012**), 116, 8494

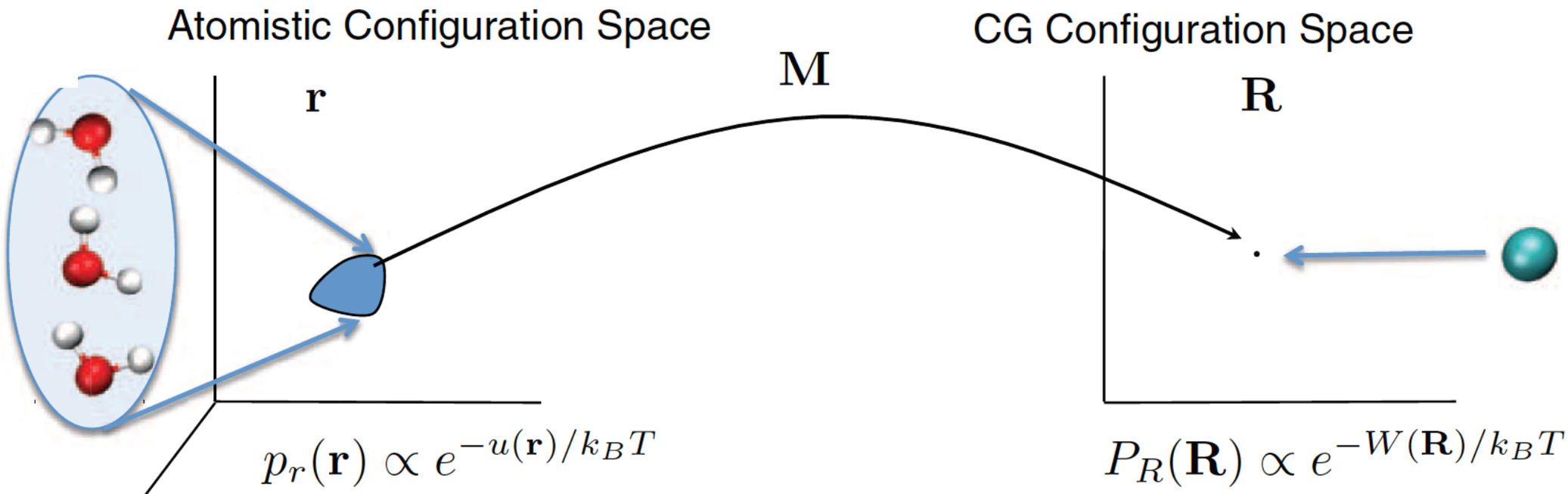
# COARSE GRAINING PHILOSOPHIES

Problem is formidable and can be tackled from alternative points of view




Noid, *J. Chem. Phys* (**2013**), 139, 090901
Saunders and Voth, *Ann. Rev. Biophys.* (**2013**), 42, 73

# CG THERMODYNAMICS




Let $\mathbf{M} : \mathbf{x}_i \rightarrow \mathbf{X}_i$ be the **CG mapping**. Then CG thermodynamics given (in principle)

$$e^{-\beta W(\mathbf{X})} \propto \int d\mathbf{x} \delta\left(\mathbf{X} - \mathbf{M}(\mathbf{x})\right) e^{-\beta U(\mathbf{x})}$$

reproducing PMF is one of major CG goals

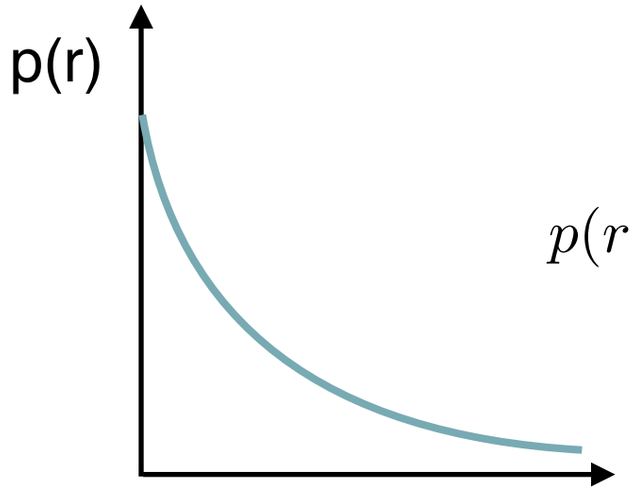
→ integrating on the mapping hyperplane

→ ‘renormalization’ of interactions

→ CG potential of mean force (PMF)

Noid, J. Chem. Phys (**2013**), 139, 090901

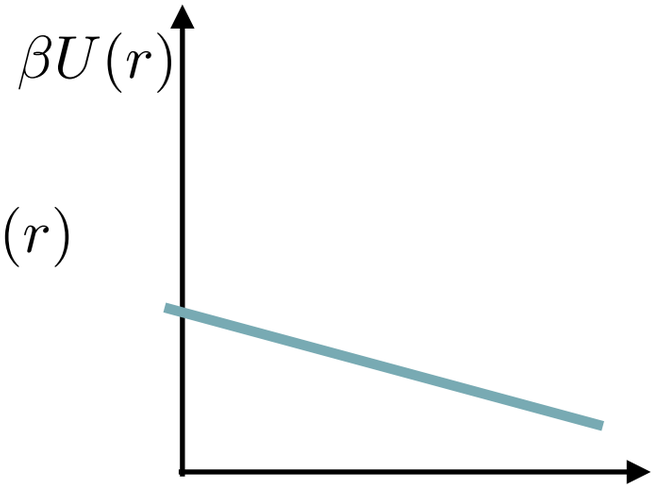
# INVERSION METHODS




e.g., pair correlation function

$$p(r) \propto e^{-\beta U(r)} \Rightarrow U(r) = \frac{1}{\beta} \ln p(r)$$

Boltzmann inversion




effective potential

## Different formulations

(iterative) Boltzmann inversion

$$V_{n+1}(r) = V_n(r) + kT \ln \frac{\mathrm{RDF}_n(r)}{\mathrm{RDF}_{\mathrm{target}}(r)}$$

Monte Carlo

$$V_{n+1}(r_i) = V_n(r_i) + \Delta V_n(r_i)$$

$$\Delta \mathrm{RDF}_n(r_i) = \sum_{j=1}^{N} \frac{\partial \langle \mathrm{RDF}_n(r_i) \rangle}{\partial V_n(r_j)} \Delta V_n(r_j) + O(\Delta V_n^2)$$

RDF = (any) radial distribution function


Lyubartsev *et al, Phys. Rev. E* (**1995**) 52, 3730
Müller-Plathe , *Chemphyschem* (**2002**) 3, 734
Soper, *Chem. Phys.* (**1996**) 202, 295

# MULTISCALE COARSE GRAINING (MS-CG)

**FG**

$$H_{FG} = \sum_i \frac{\mathbf{p}_i^2}{2m_i} + U_{FG}(\mathbf{x}_i)$$

**CG**

$$H_{CG} = \sum_i \frac{\mathbf{P}_i^2}{2M_i} + U_{FG}(\mathbf{X}_i)$$

$$\exp\left(-\beta U(\mathbf{X}^N)\right) \propto \int d\mathbf{x} \exp\left(-\beta u(\mathbf{x})\right) \delta(\mathbf{M}(\mathbf{x}) - \mathbf{X}^\mathbf{N})$$

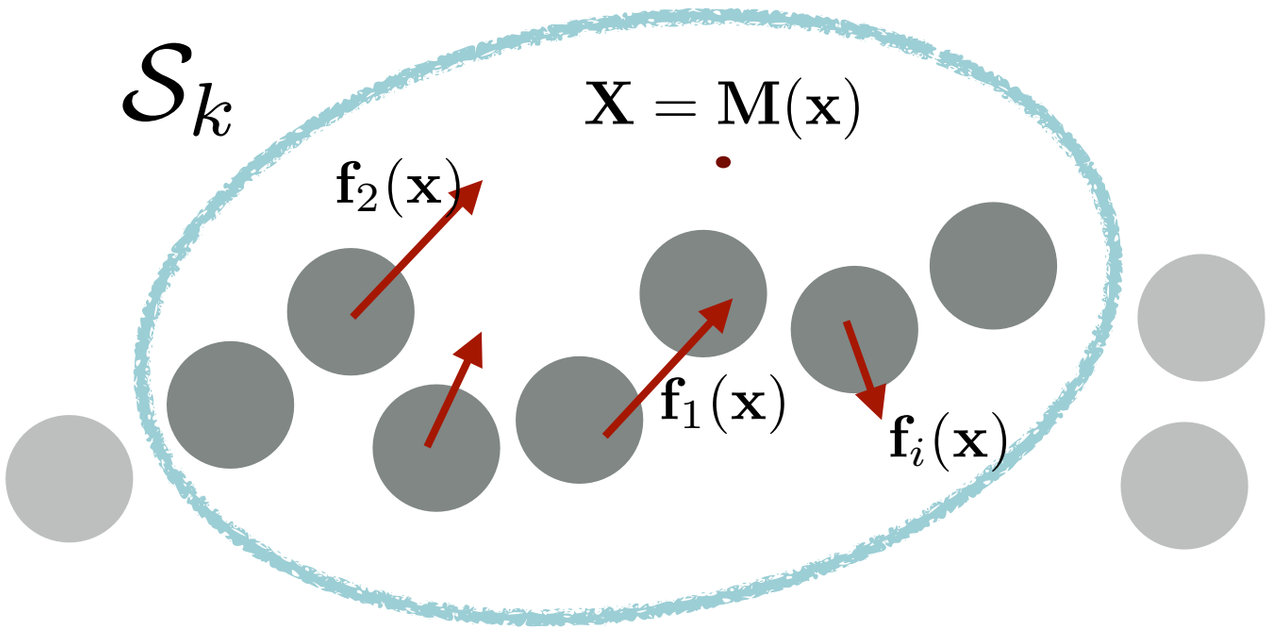
free energy mapping

Assume CG mapping is known (crucial) and cast as follows

$$\mathbf{X}_k = \mathbf{M}_k(\mathbf{x}) = \sum_i c_{ik}\mathbf{x}_i, \qquad \forall k = 1, \cdots, N$$

then

$$\mathbf{F}_k(\mathbf{X}^N) = -\frac{\partial U(\mathbf{X}^N)}{\partial \mathbf{X}_k} = \cdots = \left\langle \sum_{j \in \mathcal{S}_k} \eta_{kj} \mathbf{f}_j(\mathbf{x}) \right\rangle_{\delta(\mathbf{X}^N - \mathbf{M}(\mathbf{x}))}$$




conditional average of local atomistic forces, given than the system is in an allowed configuration

Noid et al., *J. Chem. Phys.* (**2008**) 128, 244114

Introduce a set of vectors $\mathbf{G}(\mathbf{X}^N) = (\mathbf{G}_1, \cdots, \mathbf{G}_{N_D})(\mathbf{X}^N)$

Define functional

$$\chi^2[\mathbf{G}] = \frac{1}{3N}\langle \sum_{k=1}^{N} | \sum_{q \in \mathcal{S}_k} \eta_{kj} \mathbf{f}_j(\mathbf{x}) - \mathbf{G}_k(\mathbf{M}(\mathbf{x})) |^2 \rangle_\beta$$

**effective atomistic force** **value taken by vectors**

It can be proved that

$$\chi^2[\mathbf{G}] \geq \chi^2[\mathbf{F}] = \chi^2[-\frac{\partial U(\mathbf{X}^N)}{\partial \mathbf{X}}]$$ CG force from PMF

amenable to a systematic variational optimization (see QM also)

basis functions

$$\mathcal{G}_1(\mathbf{X}^N) = -\nabla U_1(\mathbf{X}^N)$$
$$\cdots$$
$$\mathcal{G}_p(\mathbf{X}^N) = -\nabla U_p(\mathbf{X}^N)$$

+

optimization of linear combinations

$$\mathbf{G}_q = \sum_{i=1}^{N_D} \phi_D \mathcal{G}_D(\mathbf{X}^N)$$

=

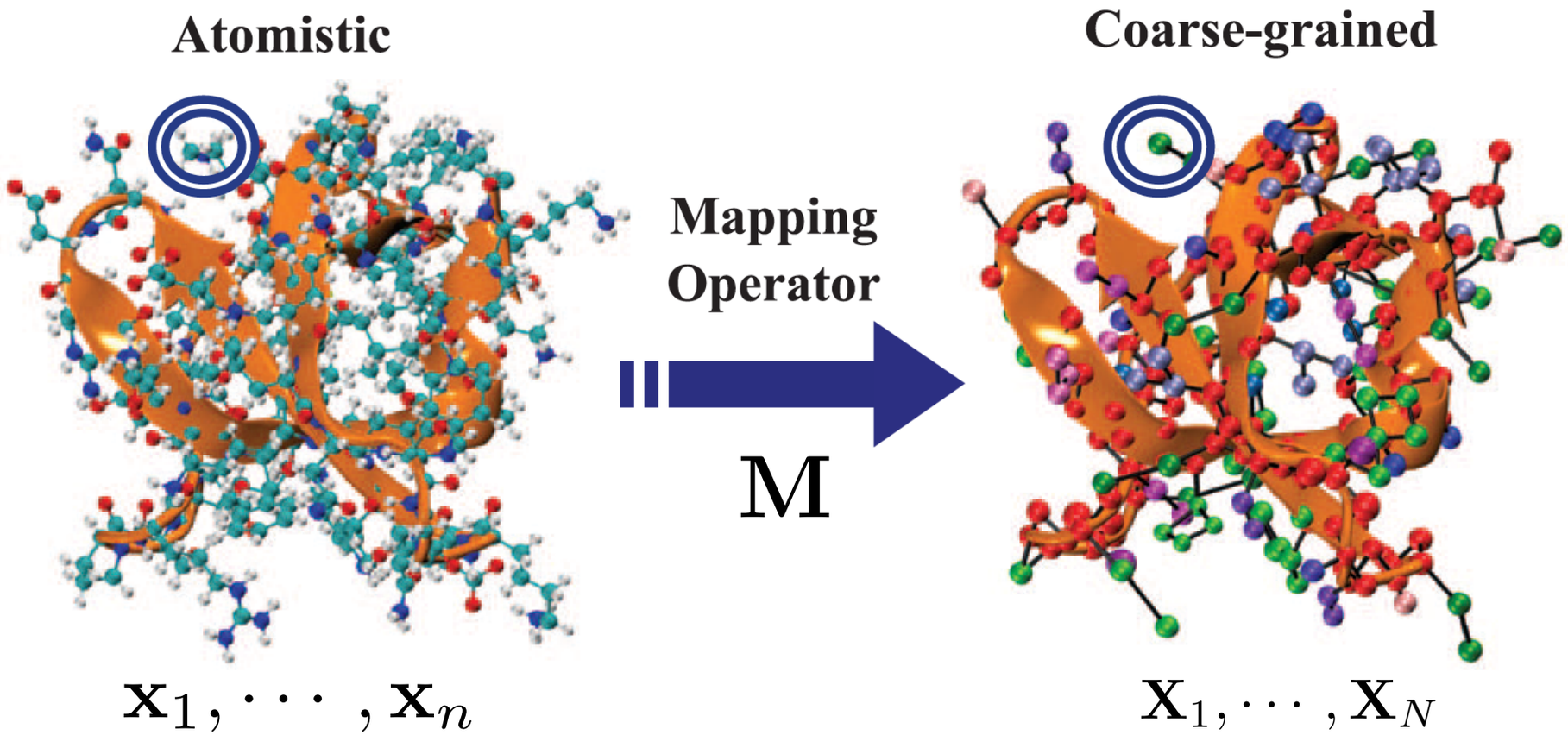
variational PMF

$$U^*(\mathbf{X}^N) = \sum_i \phi_i^* U_i(\mathbf{X}^N)$$

Noid et al., *J. Chem. Phys.* (**2008**) 128, 244114
Ritz variational principle

# CHALLENGES

# MAPPING AND REPRESENTATIONS



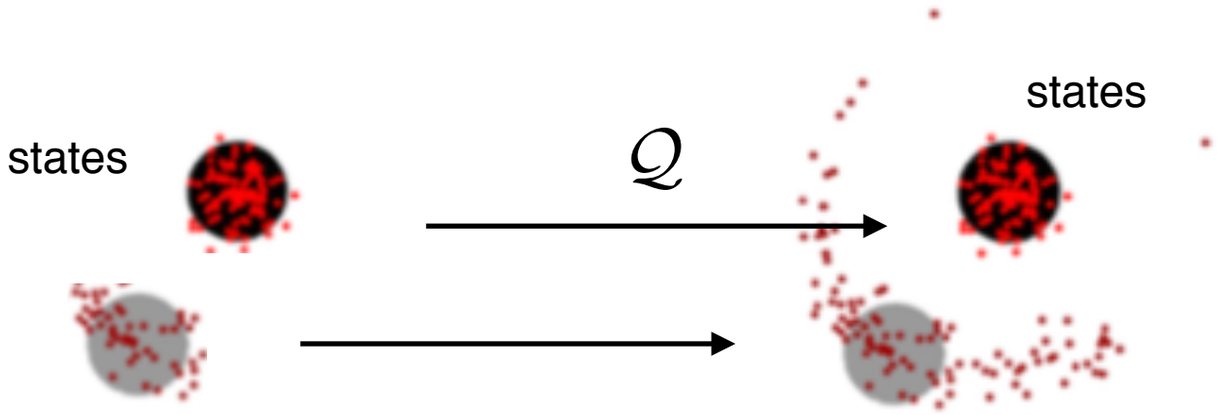

Definition of CG beads

**structural modularity**
(e.g., amino acids, nucleotides)

**dynamical proximity**
(e.g., quasi-rigid domains)

Noid et al., *J. Chem. Phys.* (**2008**) 128, 244114

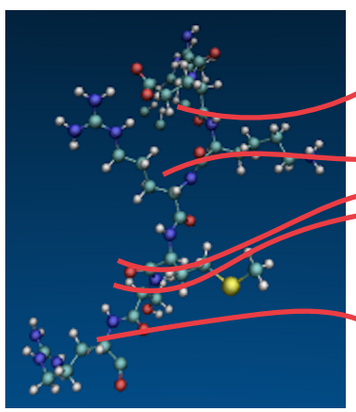
# DYNAMICAL PROXIMITY




**coherent states**

regions in state space that keep their geometric integrity, allowing very little transport in and out of themselves (i.e. atmospheric vortices)

each atom has its own trajectory



approximate Q

**Space Time Diffusion Map**

diagonalize

$\hat{Q} \approx \mathcal{Q}$

$\hat{Q}|q\rangle = \lambda|q\rangle$

clustering

each cluster is a group of heavy atoms

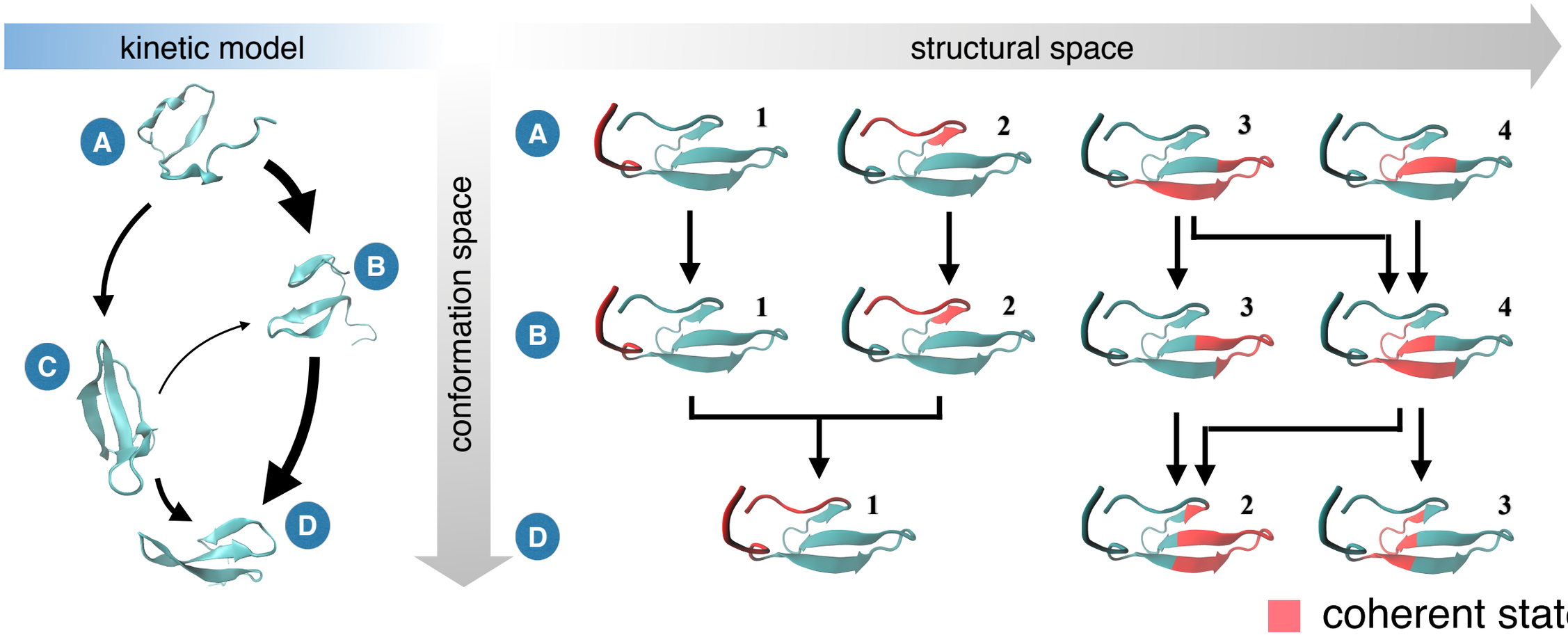
$(\xi_1, \cdots \xi_M)$
effective degrees of freedom

Banisch & Koltai (**2017**) Chaos **27**, 035804

# COHERENT DOMAINS IN FIP35 KINETICS

**S3D** (Structural and State Space Decomposition) algorithm:

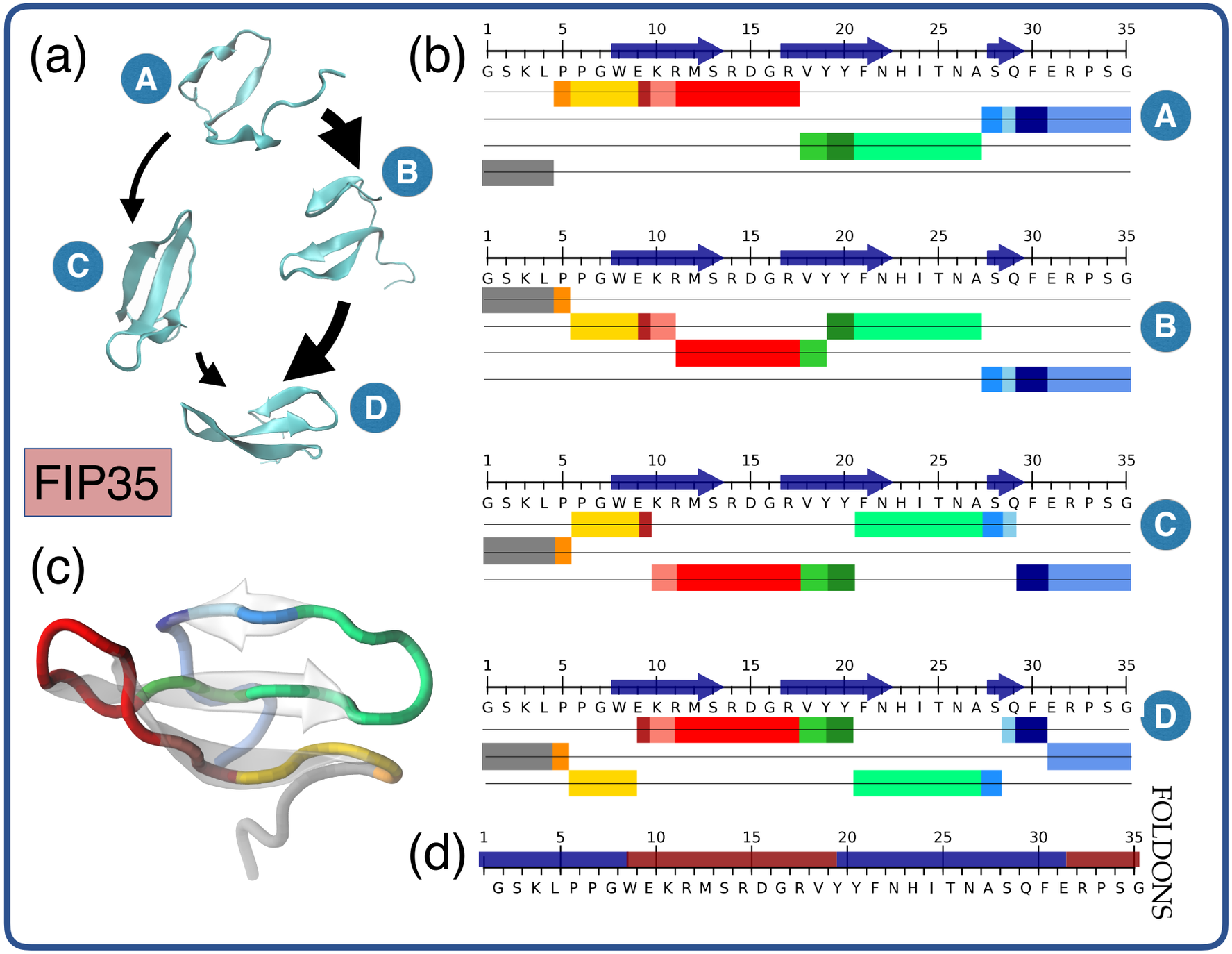
1. Kinetic model using Markov State Model analysis
2. Find state-specific coherent domains using SPTDMap




- state - dependent resolutions/decompositions
- interstate transitions as clusters' ***splitting and merging***
- clusters' hierarchy

Boninsegna, Banisch, Clementi (**2017**) *submitted*

# MINIMAL ASSEMBLY UNITS




notes-on-the-staff representation

dynamical building blocks

heterogeneity of minimal units

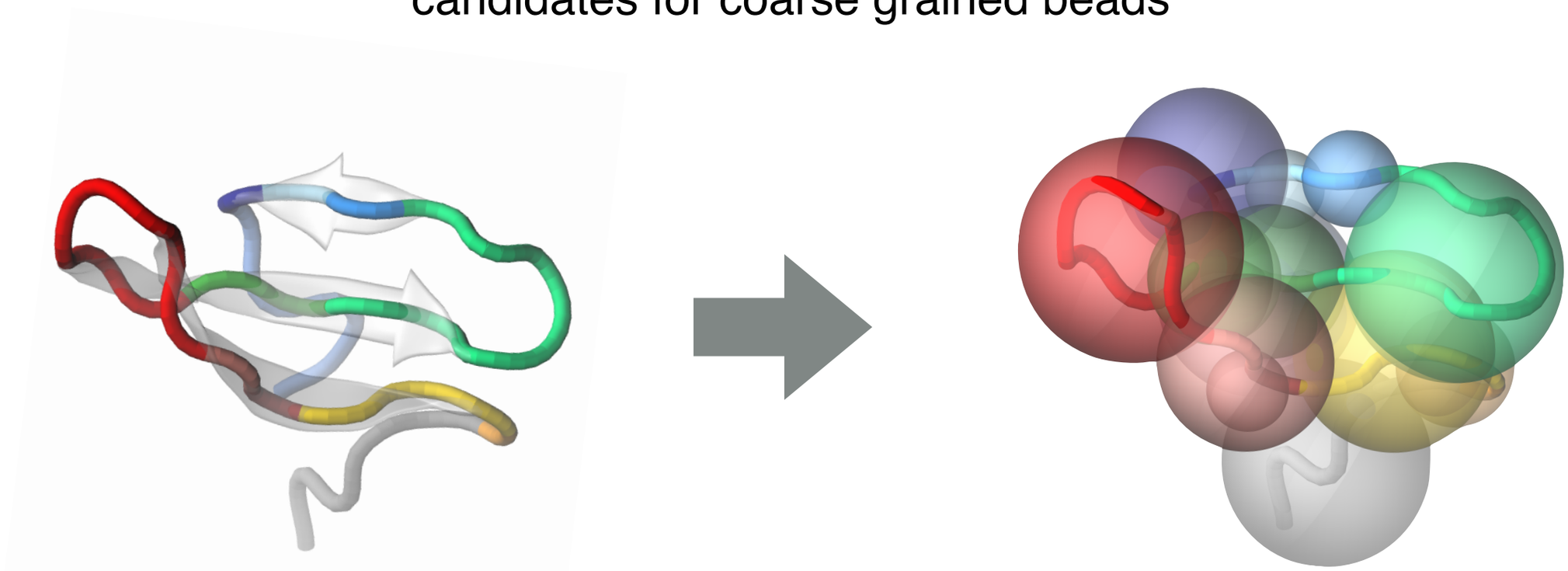
Panchenko *et al*, PNAS 330 (**1996**), 341

multi scale characterization → minimal assembly units as *Lego* logs

Boninsegna, Banisch, Clementi (**2017**) *submitted*

# TOWARDS A RIGOROUS CG MAPPING

Minimal assembly blocks are then optimal
candidates for coarse grained beads



We know the ‘beads’, but actual mapping: $\mathbf{M} : \mathbf{x}_i \rightarrow \mathbf{X}_i$ still missing

Could use the same suggested in MS-CG?

Boninsegna, Banisch, Clementi (**2017**) *submitted*

# COARSE GRAINED DYNAMICS

# COARSE GRAINED DYNAMICS

Mori - Zwanzig projection theory → CG and FG dynamics may be different!

↓

Approximating PMF does not guarantee that dynamical features are also reproduced

$$dx_t = -\nabla V(x)dt + \sqrt{2\beta^{-1}}\sigma dB_t \quad \not\longrightarrow \quad dz_t = -A'(z)dt + \sqrt{2\beta^{-1}}\sigma_z dB_t$$

Intuitively:

Degrees of freedom to be integrated out upon CG exert friction on those to be retained in the CG model

Of course compromise is needed, otherwise computational speedup due to degrees of f freedom decimation is lost

Legal & Lelièvre, Nonlinearity (2010)

# AN EXAMPLE: DYNAMICAL FORCE MATCHING

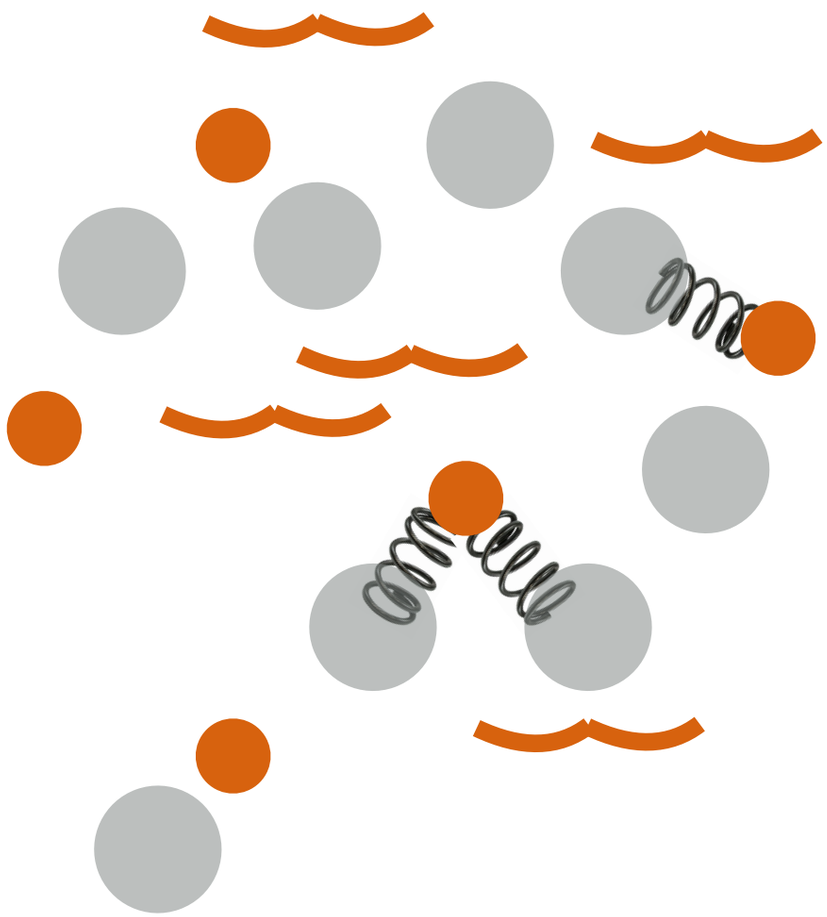
MS-CG shows that: $$\mathcal{F}_I(\mathbf{x}) = -\nabla_I V(\mathbf{R}^N) + \delta\mathcal{F}_I(\mathbf{x})$$

FG force on atoms in set **I**

*mean force* on bead **I**

**fluctuation force**

Introduce *ad hoc* fictional particles which reproduces 'missing' FG fluctuations neglected in Newtonian simulation!



MS-CG beads

'ghost' particles $(m, \omega, \gamma)$

oscillators

interactions

Langevin forces

Parametrization of 'ghosts' is based on atomistic information only

Zwanzig, J. Stat. Phys. (**1973**)
Davtyan et al., J. Chem. Pays. (**2015**)

# SINDy algorithm

Brunton *et al.* (**2016**), PNAS

- Collect all positions and 'velocities'
- Define a **function database**, i.e. space of polynomials

$$\Theta(\mathbf{X}) = \begin{bmatrix} | & | & | & | & & | & | & \\ 1 & \mathbf{X} & \mathbf{X}^{P_2} & \mathbf{X}^{P_3} & \cdots & \sin(\mathbf{X}) & \cos(\mathbf{X}) & \cdots \\ | & | & | & | & & | & | & \end{bmatrix}.$$

- Linear combinations of these functions with appropriate coefficients are going to approximate the components of the dynamic constraints $f$

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t))$$

- Coefficients $\Xi = [\boldsymbol{\xi}_1 \; \boldsymbol{\xi}_2 \; \cdots \; \boldsymbol{\xi}_n]$ are found by solving the linear regression

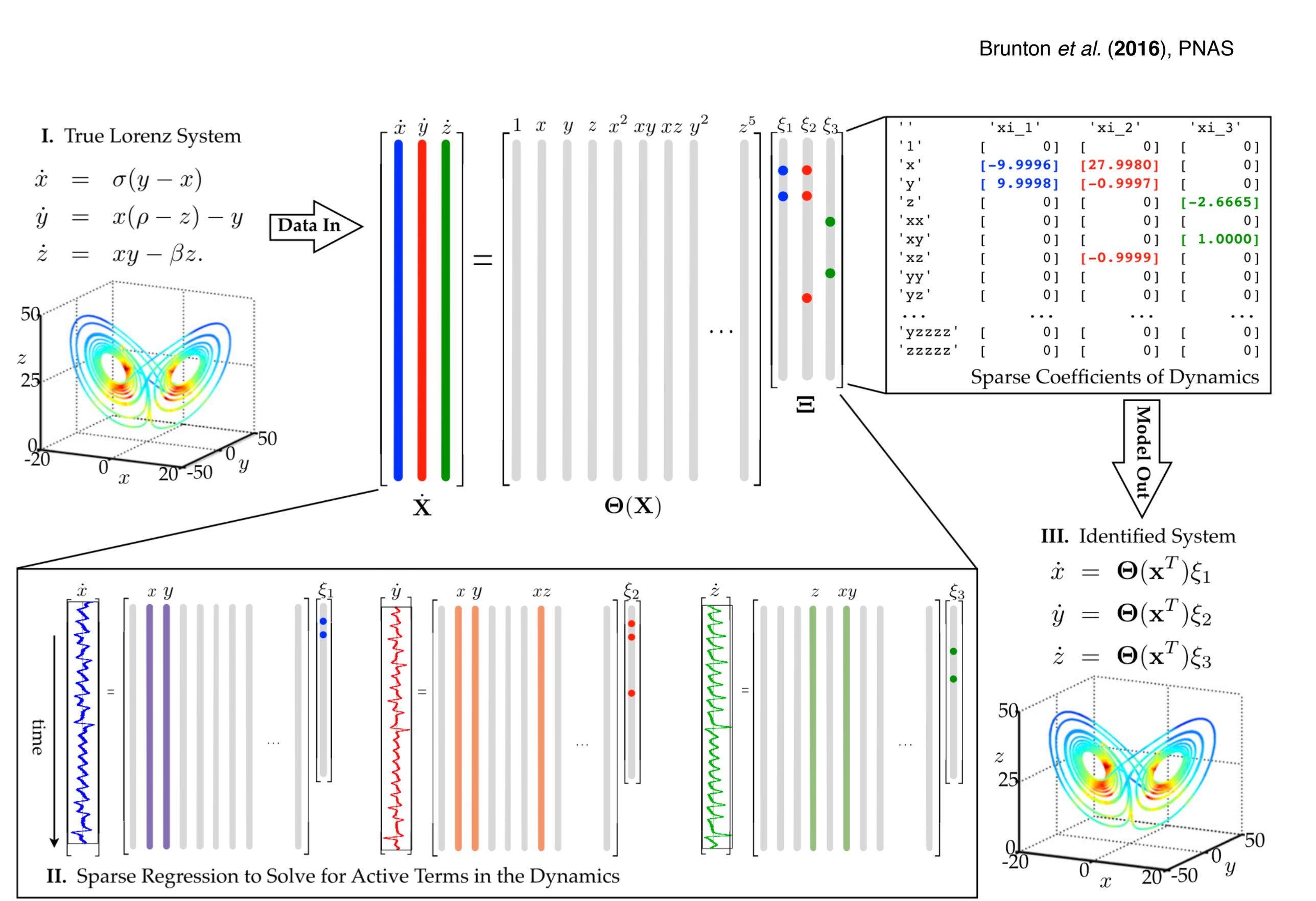
$$\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi.$$

- **Coefficient sparsity** is imposed on-the-fly, i.e. each entry is a sparse vector
- The machine-learned dynamics is then obtained by

$$\dot{\mathbf{x}}_k = \mathbf{f}_k(\mathbf{x}) = \Theta\left(\mathbf{x}^T\right)\boldsymbol{\xi}_k$$

Q: How do original and machine learned dynamics compare?

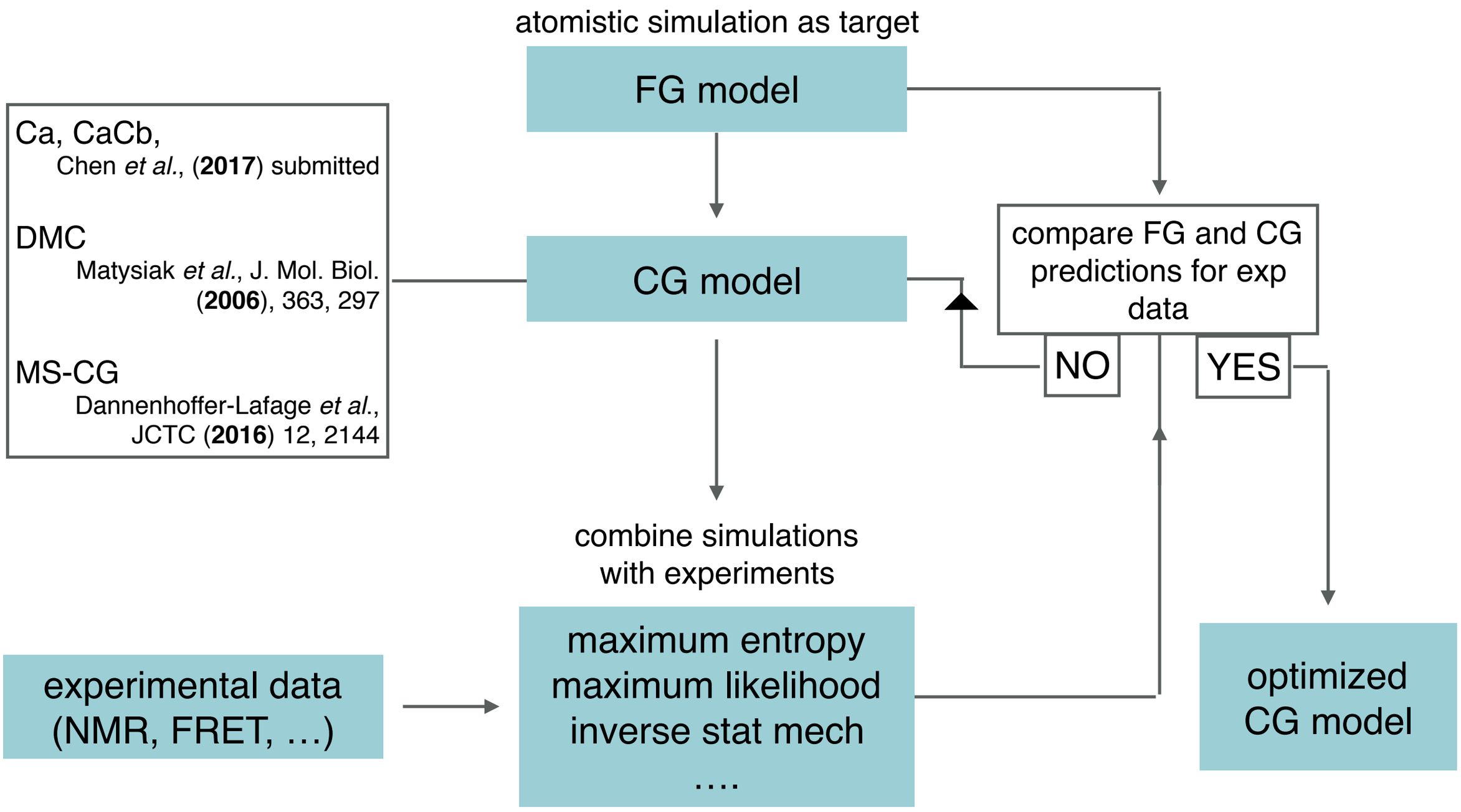
Brunton *et al.* (**2016**), PNAS




Adapt this to **high dimensional stochastic** complex dynamics to coarse grain on-the-fly?

# INCORPORATE EXPERIMENTAL DATA IN DESIGNING COARSE GRAINED MODELS

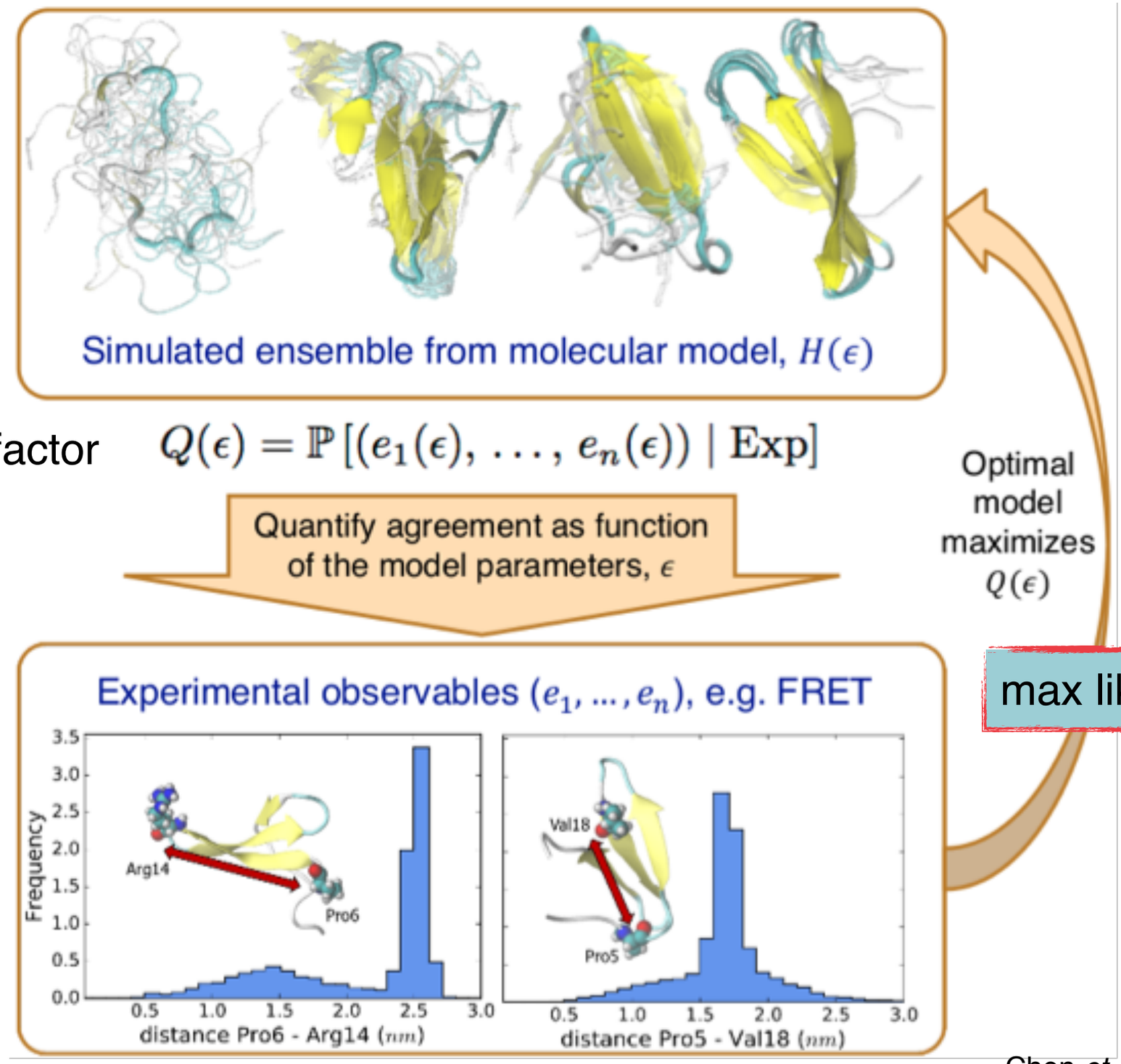
# STRATEGY




Pitera & Chodera, JCTC (**2012**)
Boomsma, Plos. Comput. Biol. (**2014**)

GC Hamiltonian: $$\mathcal{H}_{CG} = \mathcal{H}_{bond} + \sum_{contacts} K_c V_q(r_{ij}, r_0, \epsilon_q)$$

heterogeneous & non-native interactions

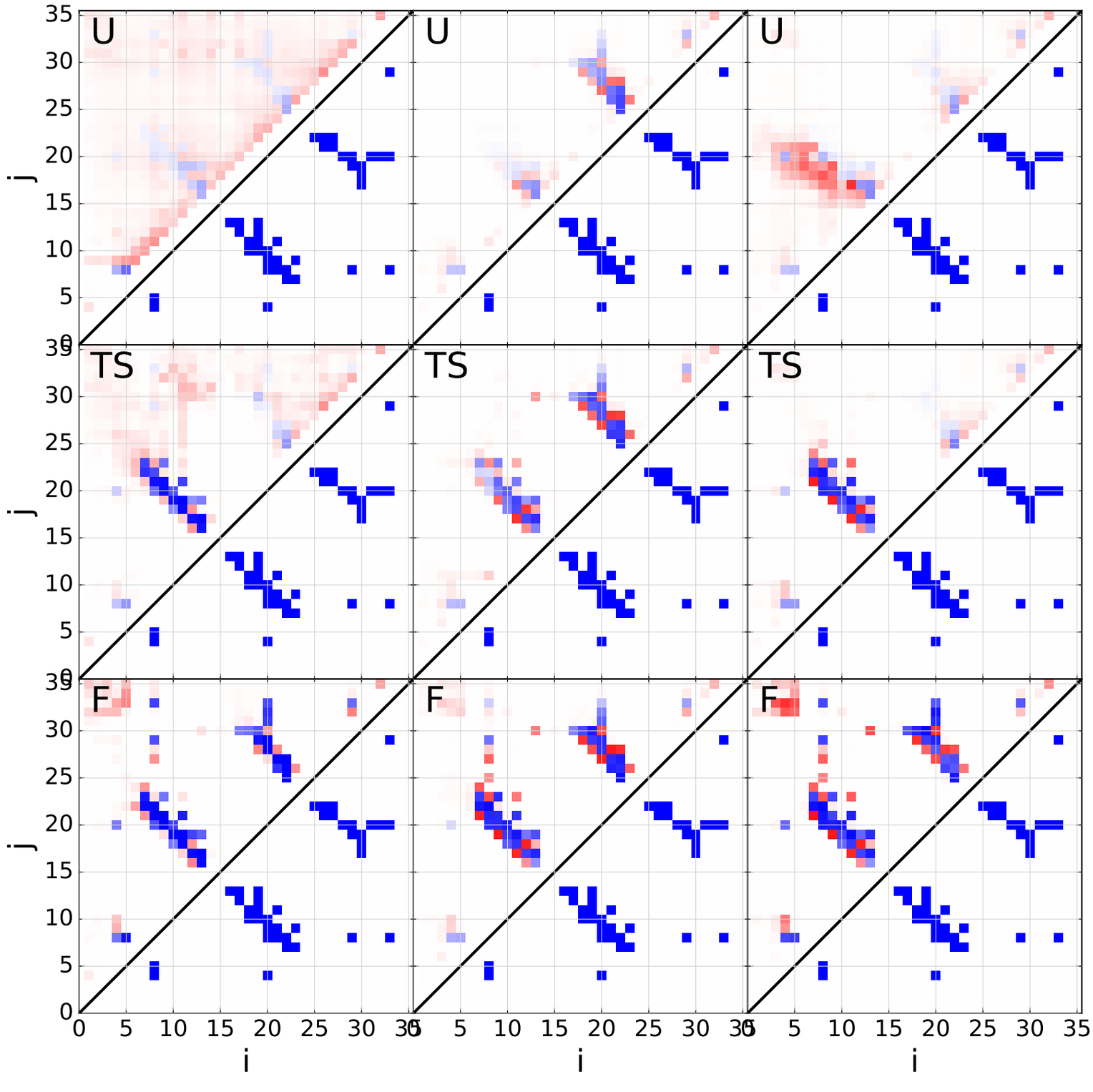



model quality factor

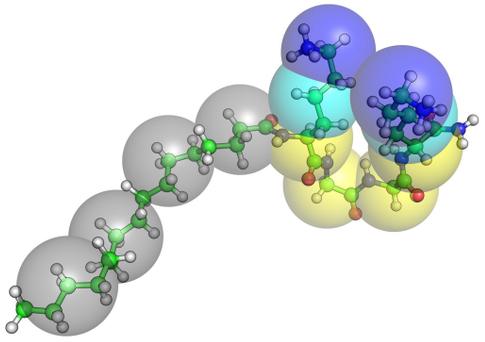
max likelihood

Chen *et al.*, (**2017**) submitted

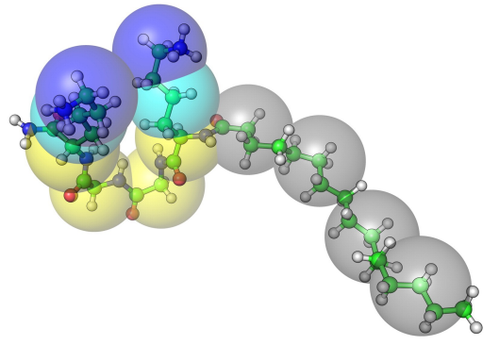
# ODEM MODELS




Chen *et al.*, (**2017**) submitted



# OUTLOOK & PERSPECTIVES



- CG is inevitable and a formidable problem

  e.g. CG degree of freedom, Hamiltonian, FG consistency

- ‘Ultimate CG’ would combine exp and theory

- Optimality, approximation control

- Thought provoking:

  - State dependent CG beads
  - ‘Parallel tempering’ CG simulations with on-the fly adapted resolution

  ….

Noid *et al.*, *J. Chem. Phys.* (**2008**) 128, 244114
Clementi, *Curr. Opin. Struct. Biol.* (**2008**),18,10

# ACKNOWLEDGMENTS

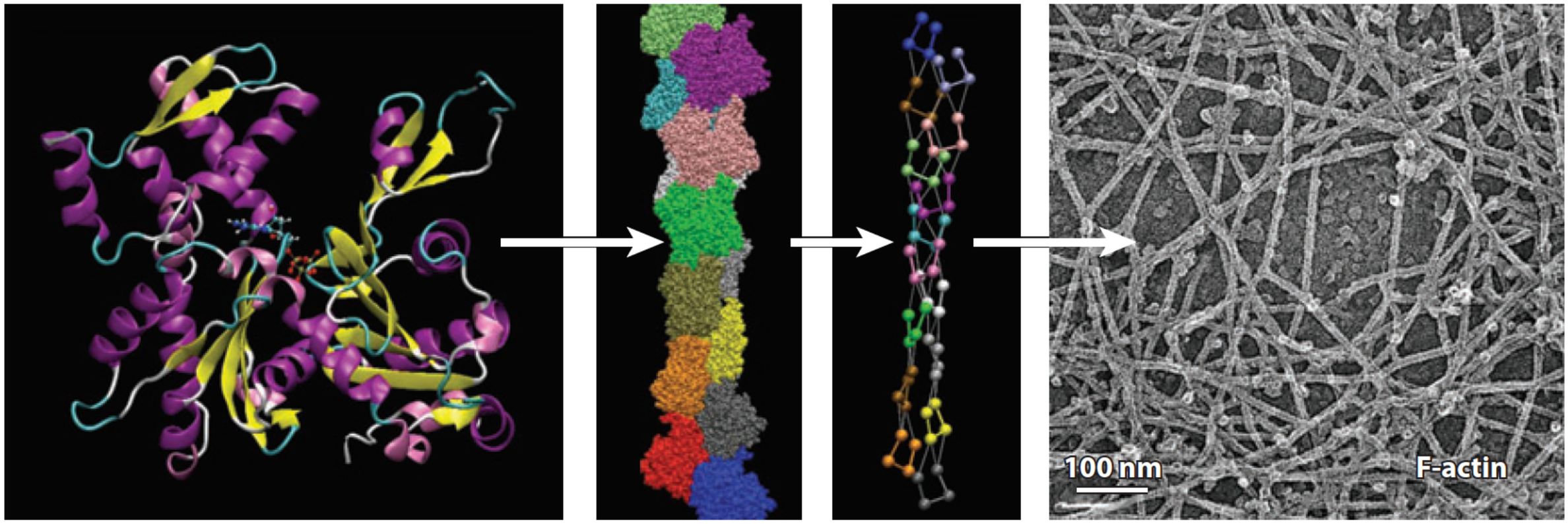

Prof. Cecilia Clementi

Dr. Feliks Nüske
Dr. Giovanni Pinamonti
Alexander Kluber
Justin Chen
Eugen Hruska

Our collaborators:

Prof. Frank Noé (Freie Universität Berlin)
Prof. Peter Koltai (Freie Universität Berlin)
Dr. Ralf Banisch (Freie Universität Berlin)

# MULTISCALE CHALLENGE: PARADIGMATIC EXAMPLE




acting subunit | actin filament | cytoskeleton network

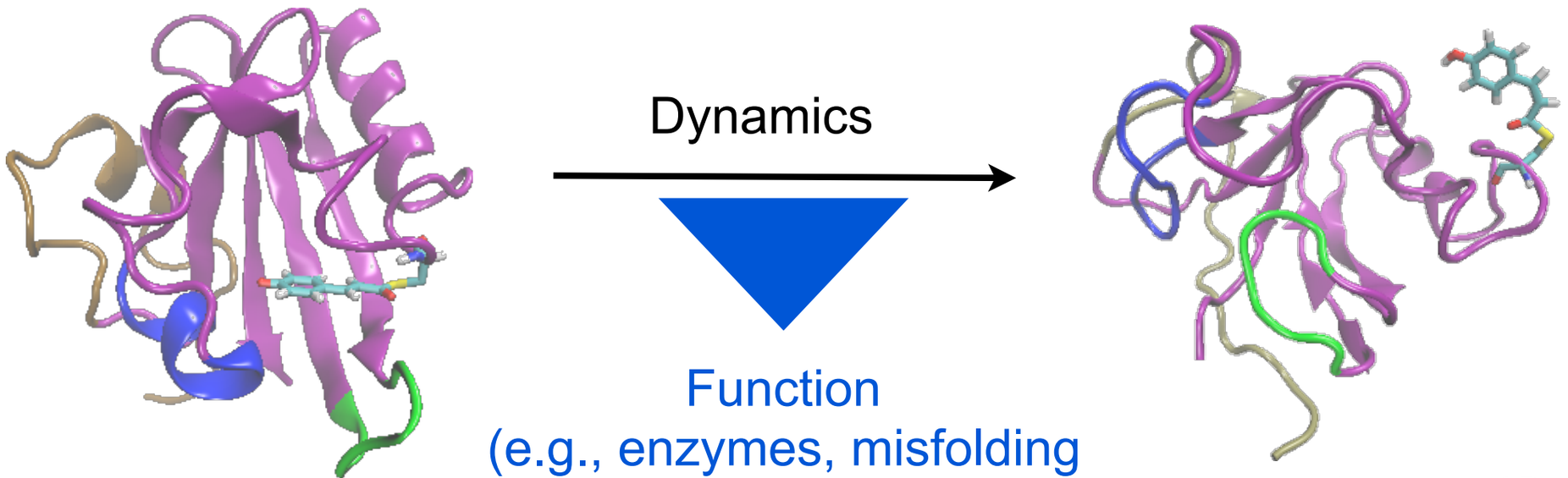
complex interactions

many degrees of freedom

**multi scale coupling**

Saunders, *Annu. Rev. Biophys.* (**2013**) 42, 73

# BIOMOLECULES: STRUCTURE, FUNCTION AND EQUILIBRIUM




(courtesy of Dr. Wenwei Zheng)

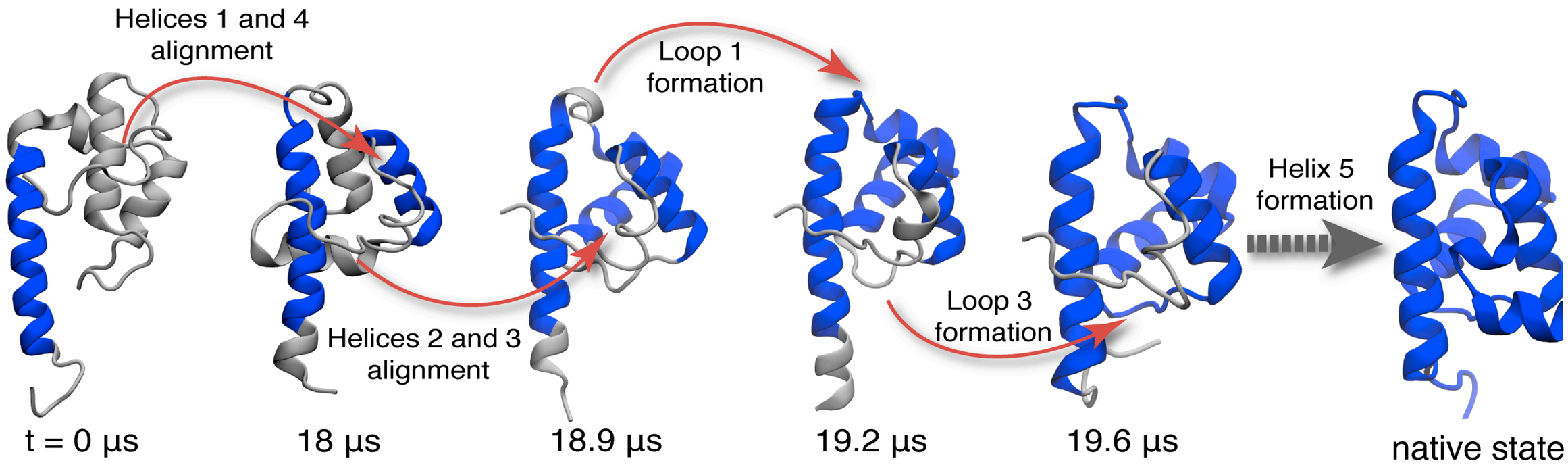
**dynamical model**
(Hamiltonian, eqs of motion)

**equilibrium trajectory**
(Boltzmann ensemble)

$$\langle \cdots \rangle_\beta = \langle \cdots \rangle_t$$

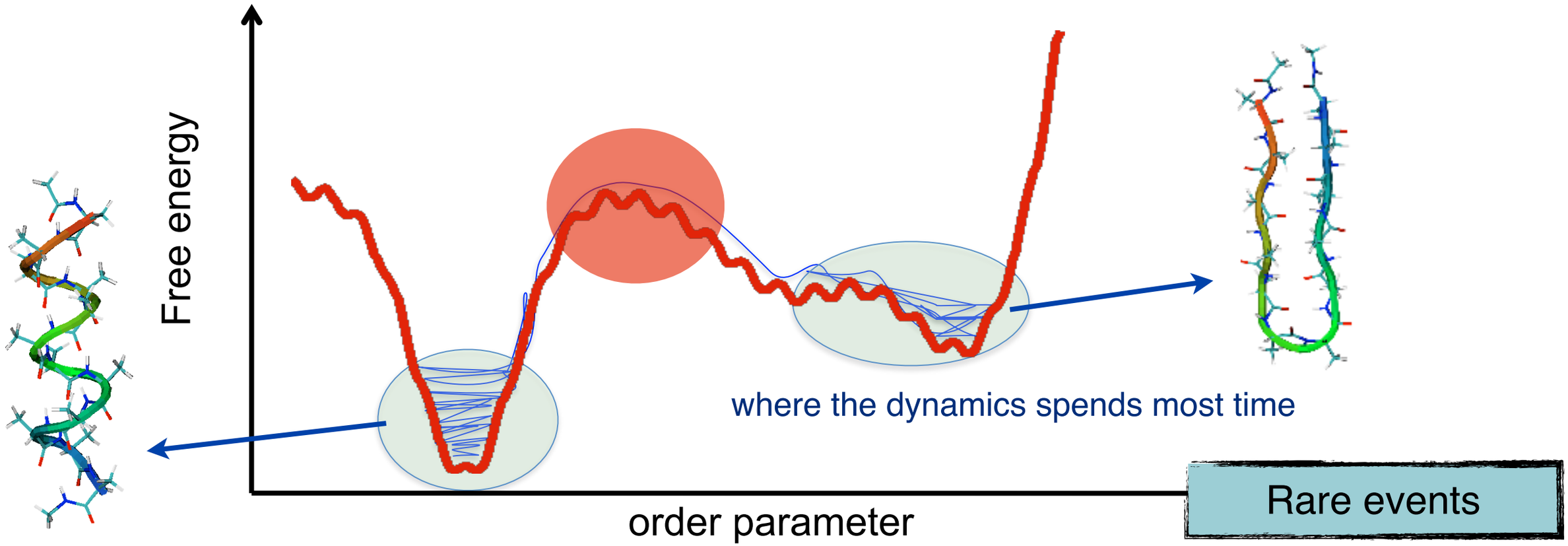
- observables (exp. data)
- structure - function
- predictions

BUT landscape is metastable, complex, frustrated



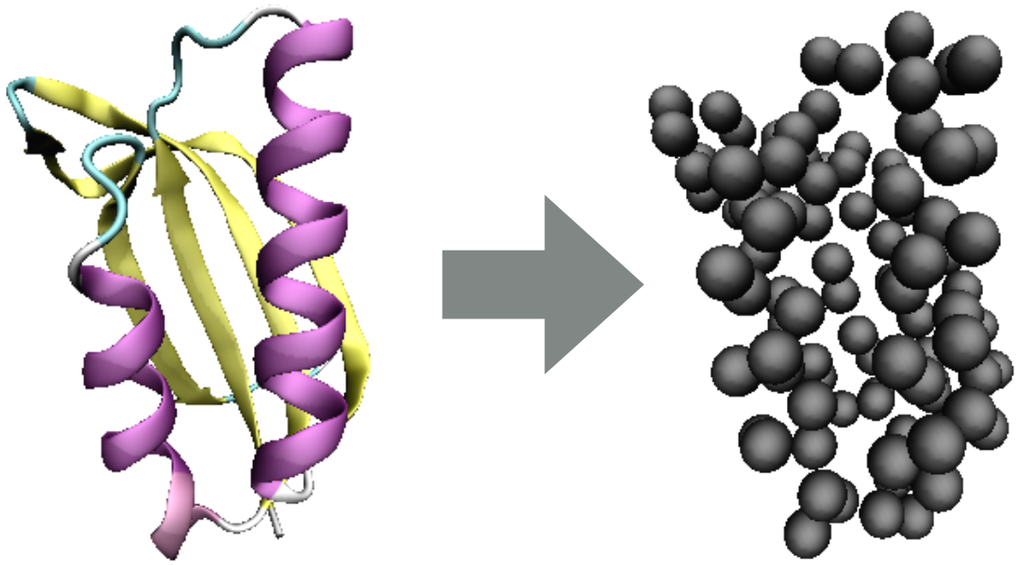

Collective motions

‘Big picture’



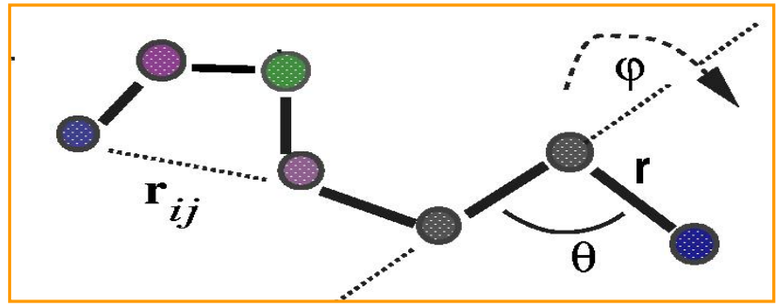

Rare events

# SAMPLE ATOMISTIC HAMILTONIAN

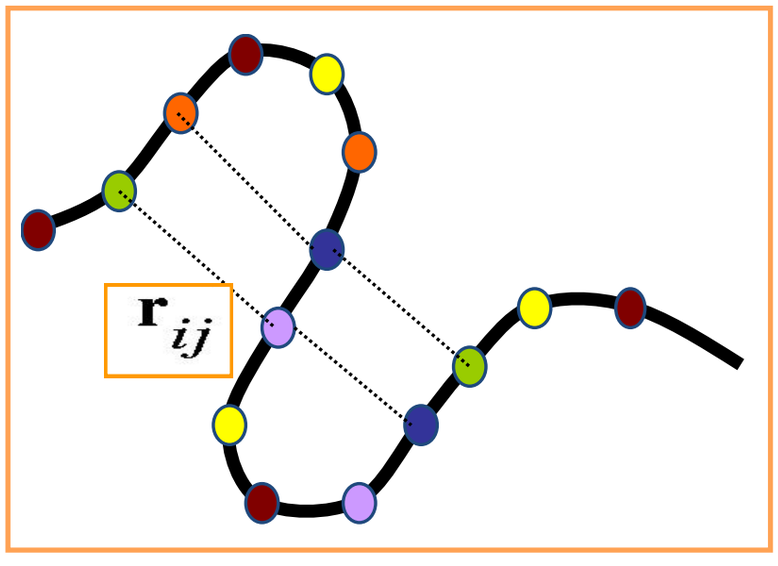


$$H = \sum_{bonds} K_d(d_i - d_0)^2 + \sum_{angles} K_\theta(\theta_i - \theta_0)^2$$
$$+ \sum_{Dihedrals} K_\phi(n)[1 + cos(n(\phi_i - \phi_0))]$$
$$+ \sum_{I-Dihedrals} K'_\phi(\phi'_i - \phi'_0)^2$$
$$+ \sum_{contacts} K_c V_q(r_{ij}, r_0, \epsilon_q)$$

bonded terms



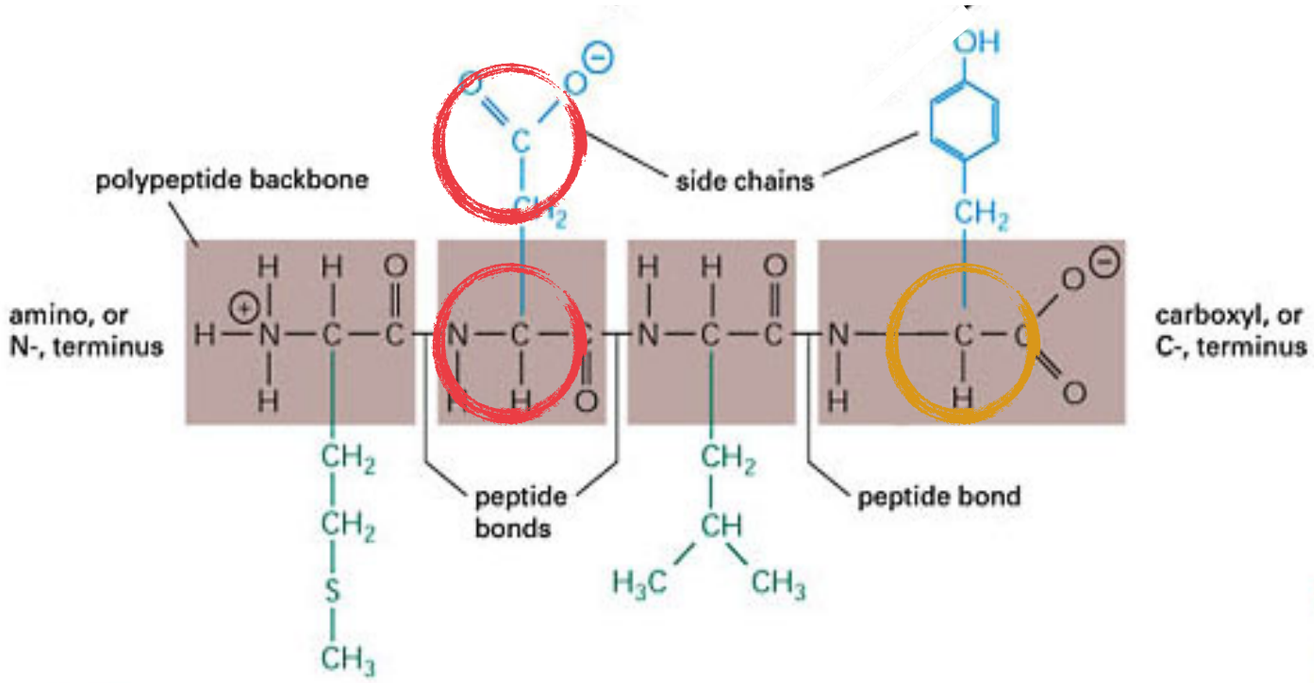

non bonded terms




Clementi, Nymeyer & Onuchic (JMB 2000)

(courtesy of Cecilia Clementi)

# MODULARITY

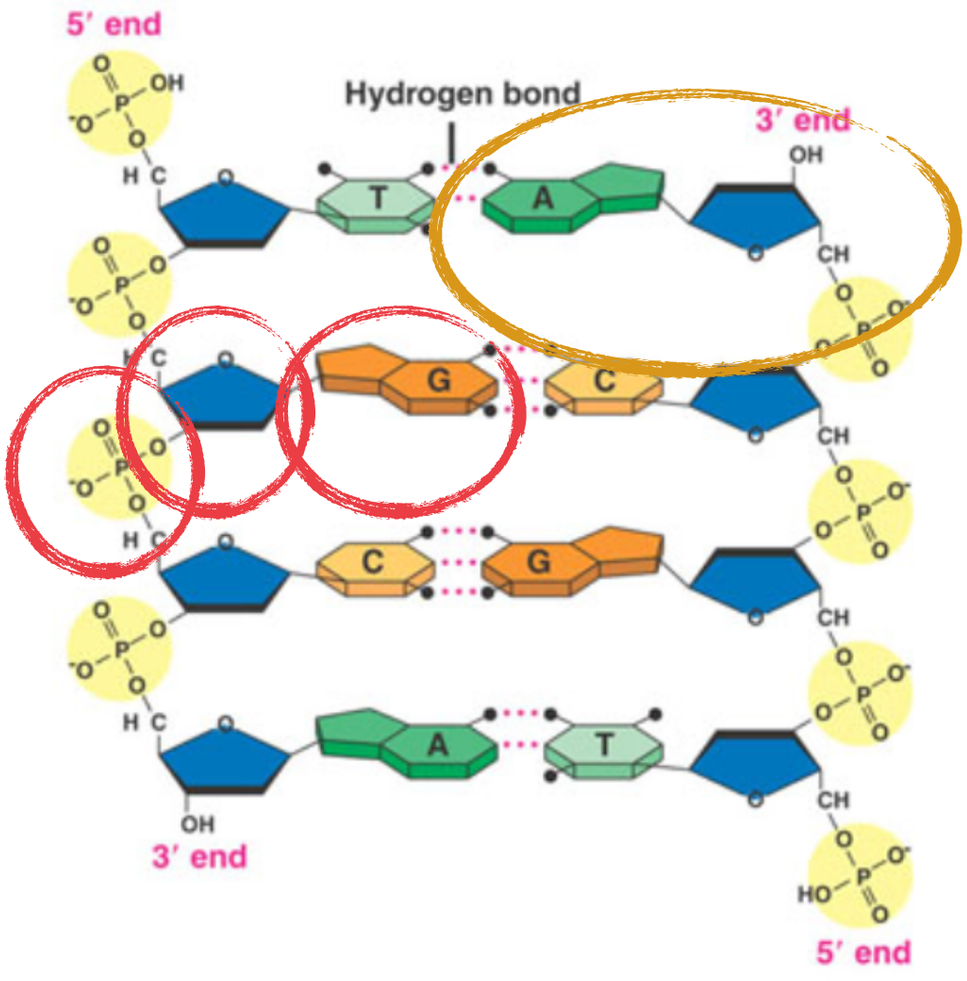



## NUCLEID ACIDS

## PROTEINS

$C_\alpha$ model

$C_\alpha C_\beta$ model



# PRINCIPLE OF MAXIMUM ENTROPY

Link between Shannon's information theory and stat. mech.

Q: what is the probability distribution which best represents the state of knowledge after observing a set of quantity?

entropy

$$S(p) = -\sum_{i}^{n} p(x_i) \ln p(x_i)$$

observations (constraints)

$$\sum_k p(x_k) = 1 \qquad \sum_k p(x_k) f(x_k) = F_k$$

Solve constrained optimization to get:

$$p(x_k) = \frac{1}{Z} e^{-\sum_j \lambda_j f_j(x_k)} \qquad Z = \sum_k e^{-\sum_j \lambda_j f_j(x_k)}$$

Enforcing averages on MD distribution is equivalent to introducing a biasing potential

$$p(x) = \frac{1}{Z} e^{-\beta U(x) - \sum_k \lambda_k f_k(x)}$$

CG potential

observable

Boomsma, *Plos. Comput. Biol.* (**2014**), 10, e1003406
Jaynes, *Proc. IEEE* (**1982**) 70, 939

# OBSERVABLE-driven DESIGN OF EFFECTIVE MOLECULAR MODELS (ODEM)

- Experimental measurements : $(f_1, \ldots, f_K)$ uncertainties $(\Delta f_1, \ldots, \Delta f_K)$
- Start from an initial guess for the parameters, $\epsilon^{(0)}$ ($0^{th}$ order model)
  Run MD with $H^{(0)}(\mathbf{x})$
- Estimate maximum likelihood equilibrium distribution $\pi^{(0)}(\mathbf{x})$ from MSM
- Compute value of observable $f_k$ from MSM as:

$$e_k[\boldsymbol{\pi}] = \sum_i \frac{\pi_i}{n_i} \sum_{j \in S_i} g_k(\mathbf{x}_{ij})$$

- Measure "goodness" $Q$ of the model:

$$Q = \mathbb{P}\left[(e_1[\boldsymbol{\pi}], \ldots, e_K[\boldsymbol{\pi}]) \mid \mathrm{Exp}\right] = \prod_k \mathcal{N}(e_k[\boldsymbol{\pi}];\ f_k, \Delta f_k)$$

- Upon infinitesimal parameter change $\epsilon^{(1)} = \epsilon^{(0)} + \delta\epsilon$ equilibrium distribution change is:

$$\pi_i^{(1)} = \frac{\pi_i^{(0)}}{n_i} \sum_{x \in S_i} \exp\left[-\beta(H^{(0)}(\mathbf{x}) - H^{(1)}(\mathbf{x}))\right]$$

- Compute expression of $Q^{(1)}$ as a function of $\epsilon^{(1)}$
- Pick $\epsilon^{(1)}$ to maximize $Q^{(1)}$

Chen, Pinamonti and Clementi, (**2017**) *submitted*