Learning molecular model from simulation and experimental data

Cecilia Clementi

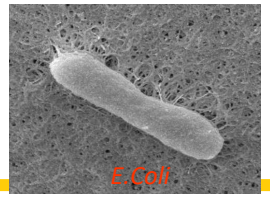
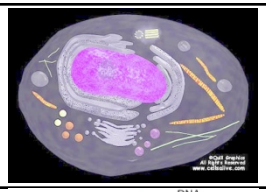
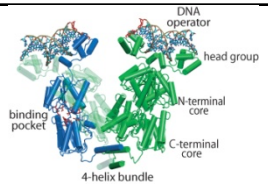
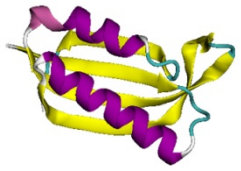
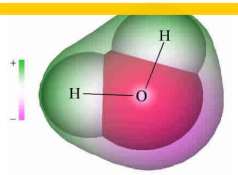
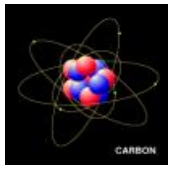
Center for Theoretical Biological Physics
Department of Chemistry & Department of Physics
Rice University, Houston, TX USA

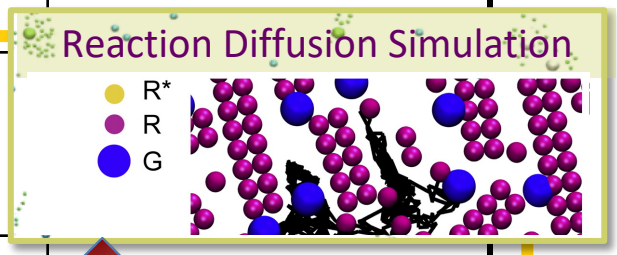
Einstein Visiting Professor
Multiscale Modeling of Biophysical Systems
Department of Mathematics and Computer Science
Freie Universität, Berlin, Germany

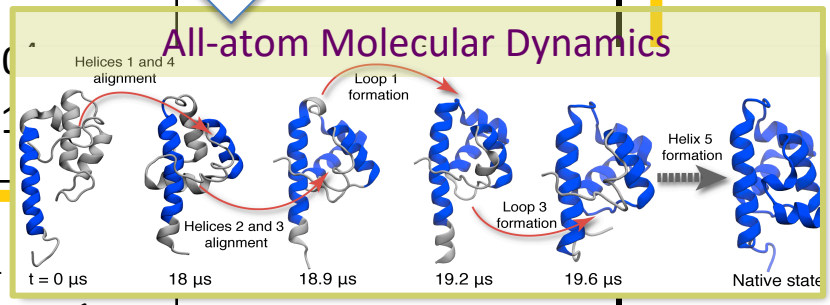
Einstein Stiftung Berlin
Einstein Foundation Berlin

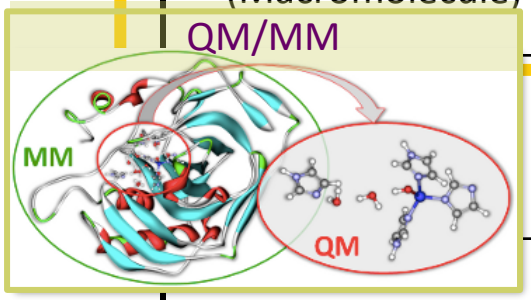
A main challenge in biophysics

broad range of interconnected length and time scales

Organism		~10^{20} atoms	Thermodynamics Macroscale
Cell		~10^{10} atoms ~1-10 μm	
System		~10^{4}-10^{5} atoms ~ 10-100 nm	Mesoscale Multiscale
Biomolecule (Macromolecule)		~10^{3}-10 ~ 1-	
		~10^{1} ~ 1-10 Å	
Atom		~1 atom ~ 1 Å	Quantum Physics



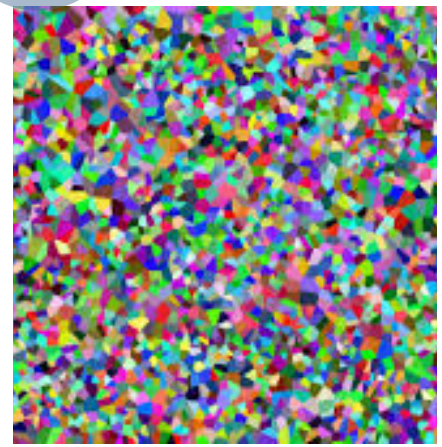


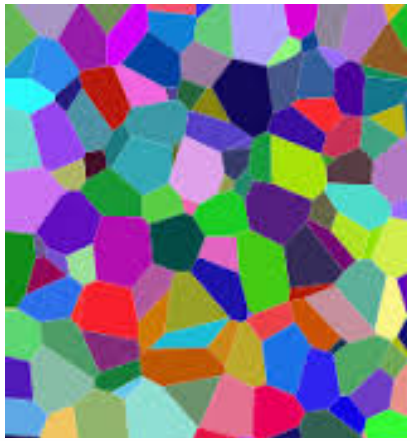


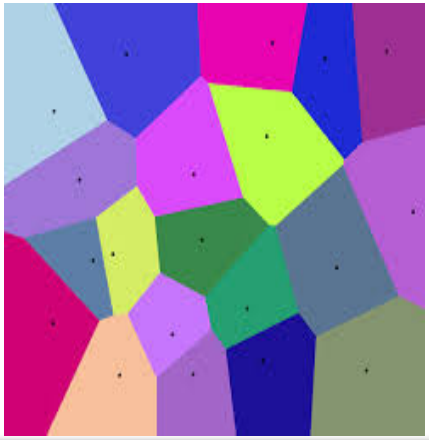
Coarse-Graining

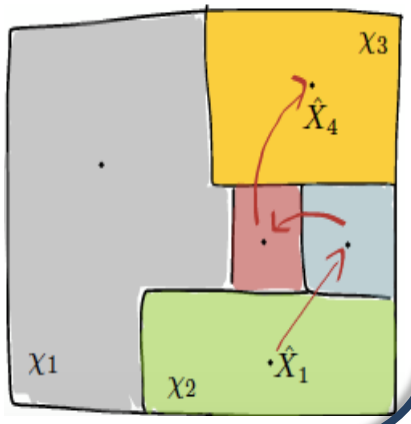
$\mathbb{R}^{3N}$

coarse-graining in conformation space



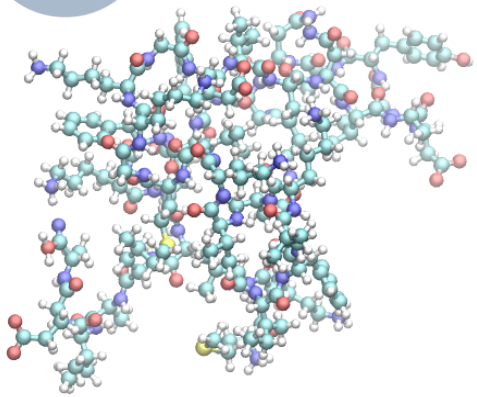


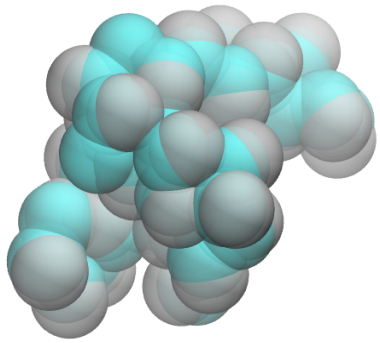


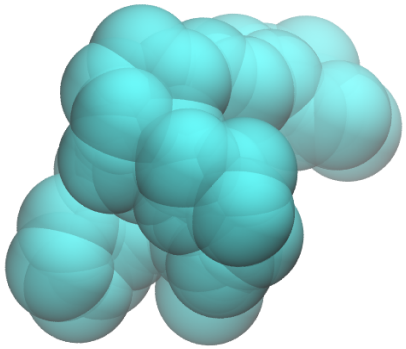


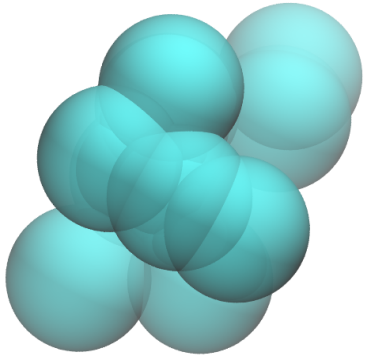
$\mathbb{R}^{3}$

coarse-graining in structural space









Coarse-graining in conformation space

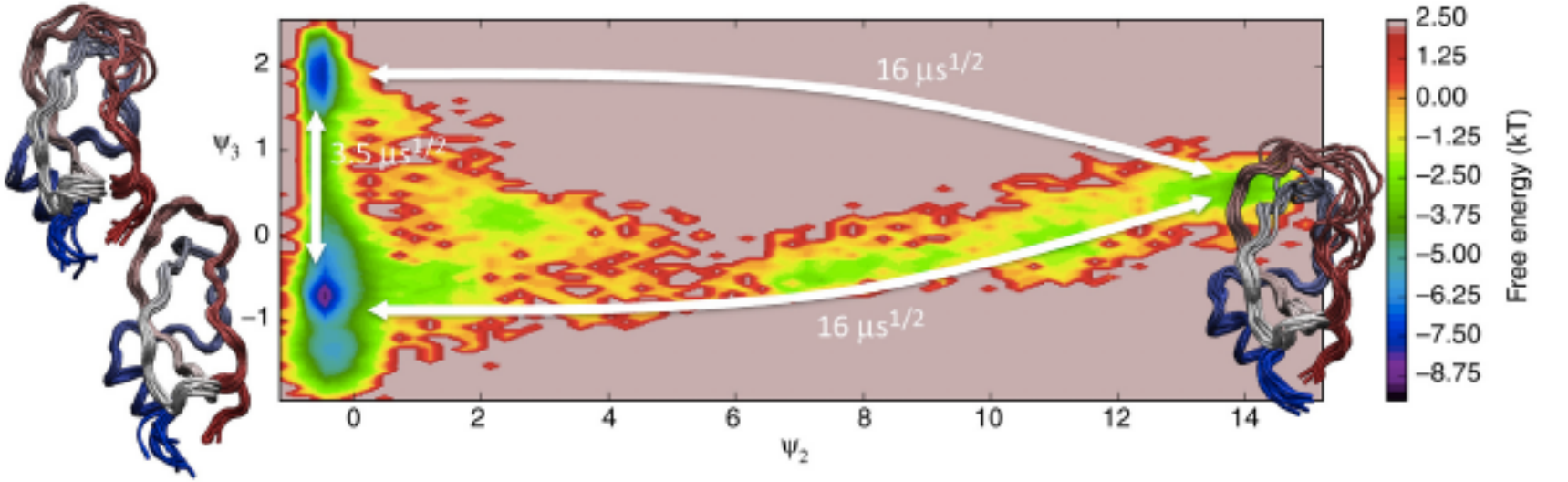
Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods

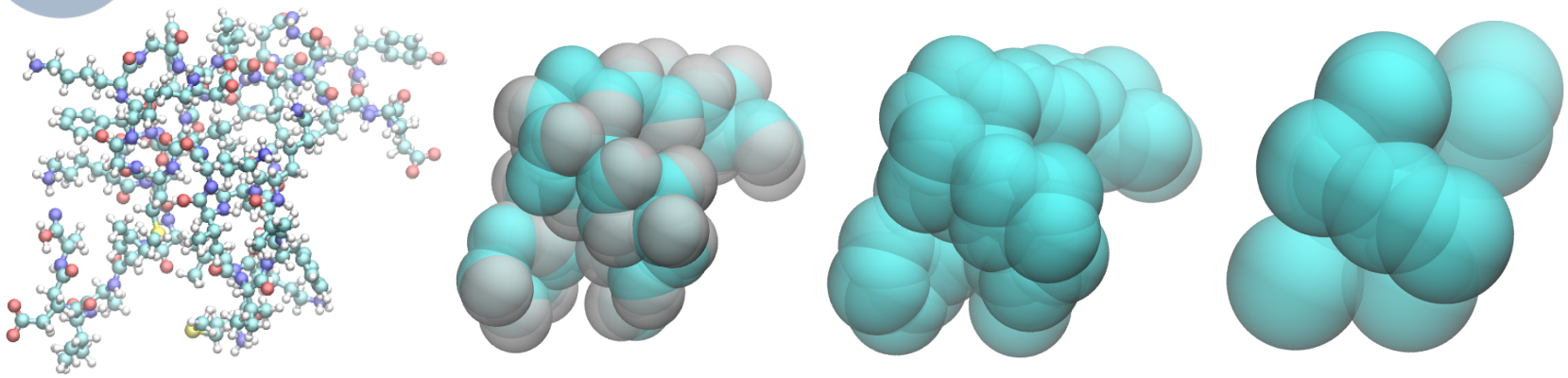
Frank Noé[1] and Cecilia Clementi[2]



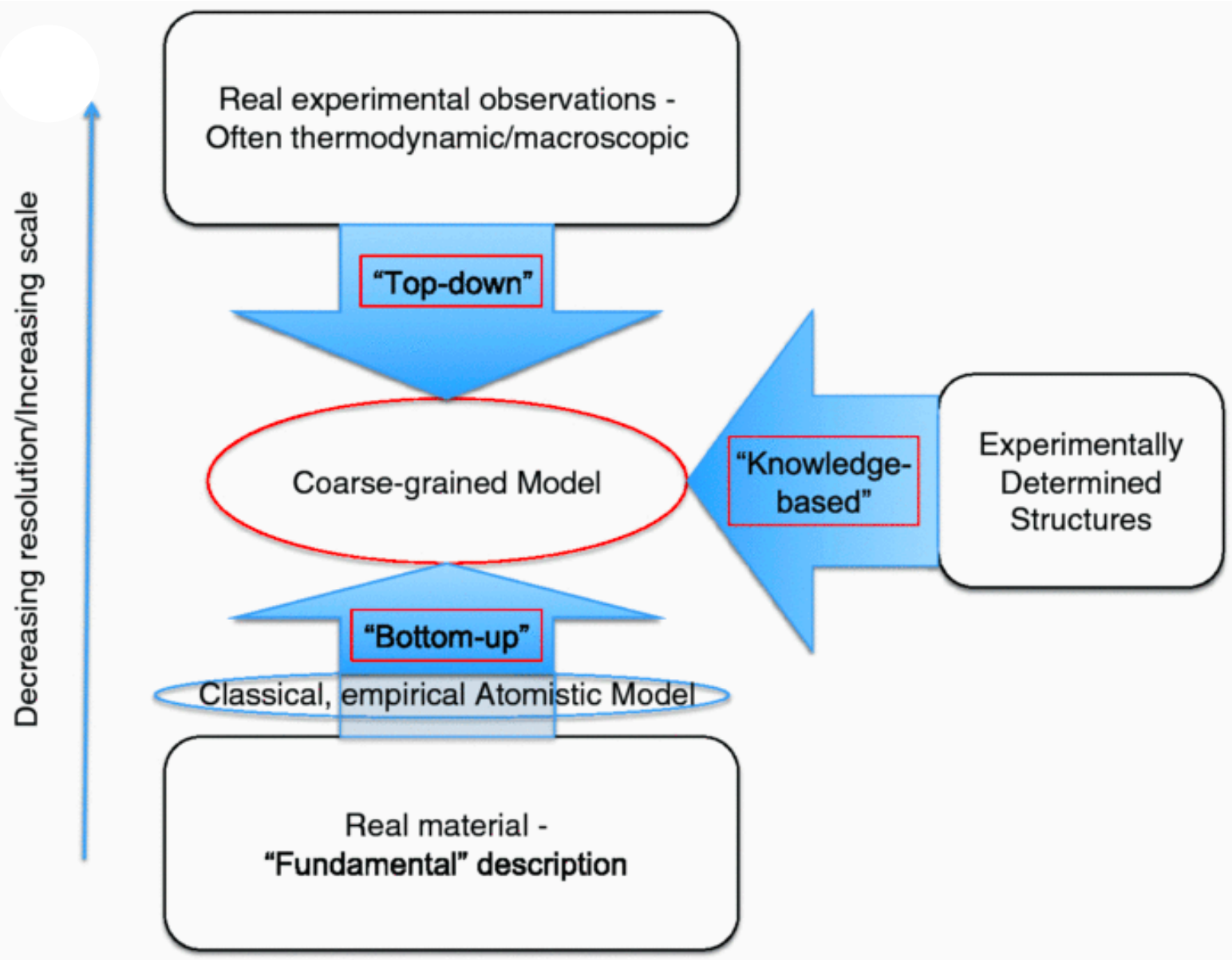
Coarse-Graining

$\mathbb{R}^3$

coarse-graining in structural space

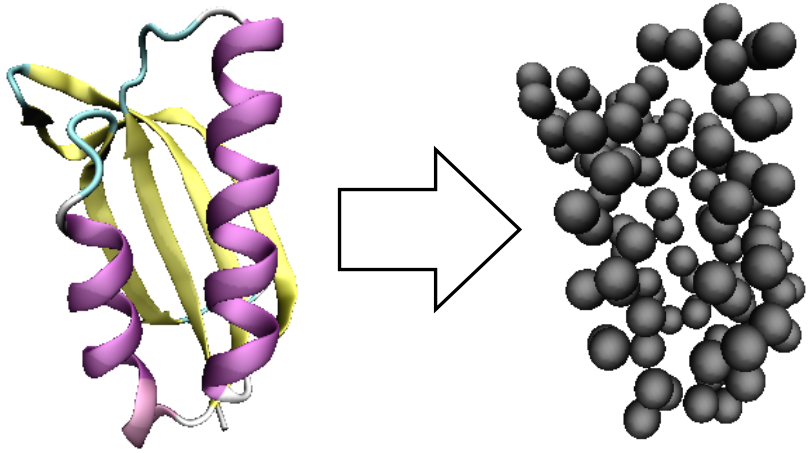


Coarse-graining: what properties should be preserved?



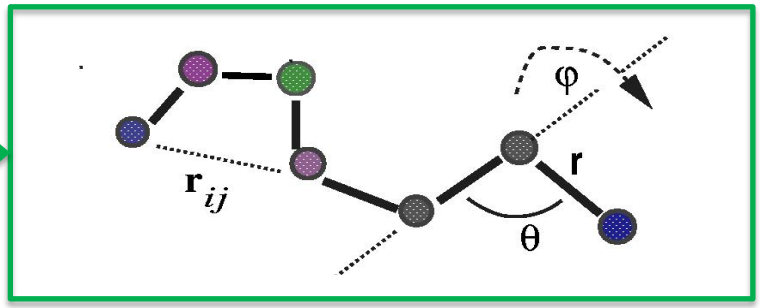
W.G. Noid, J. Chem. Phys. (2013) 139:9, 090901

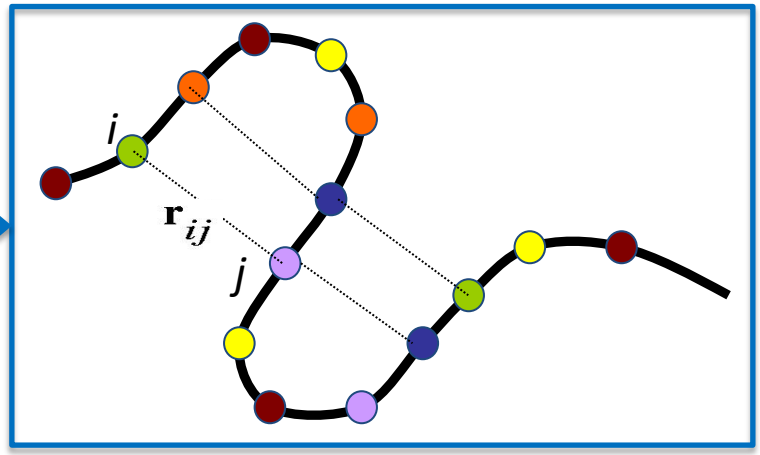
Coarse-graining in structural space



$$\mathbf{H} = \sum_{\text{bonds}} K_r(r - r_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_d(1 + \cos(n(\phi - \phi_0)) +$$

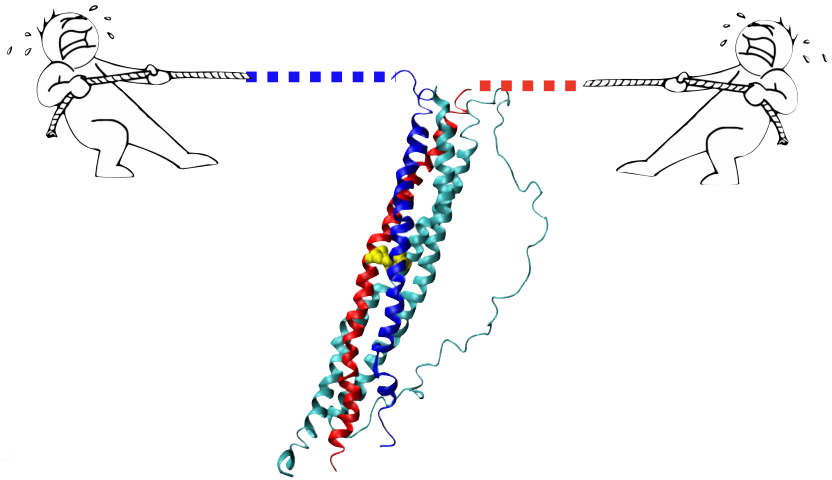
$$\sum_{i,j} \epsilon_{ij} \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right]$$

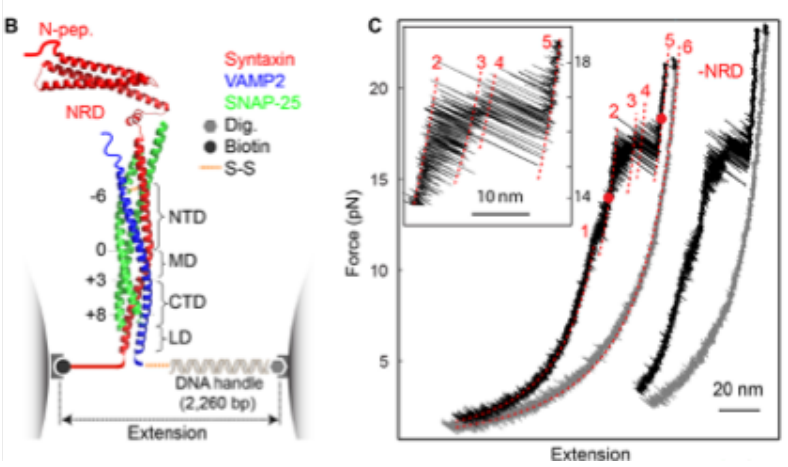


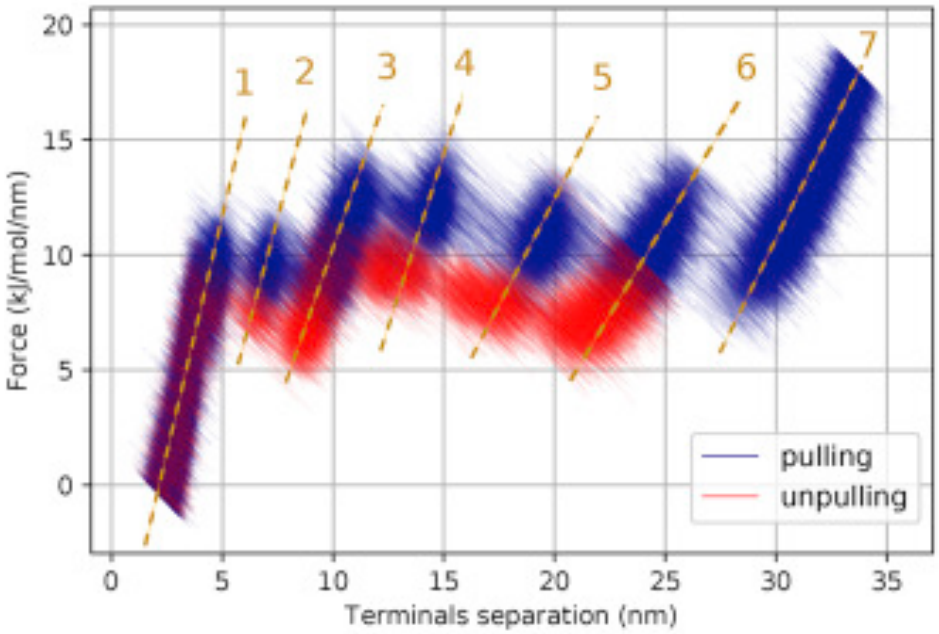


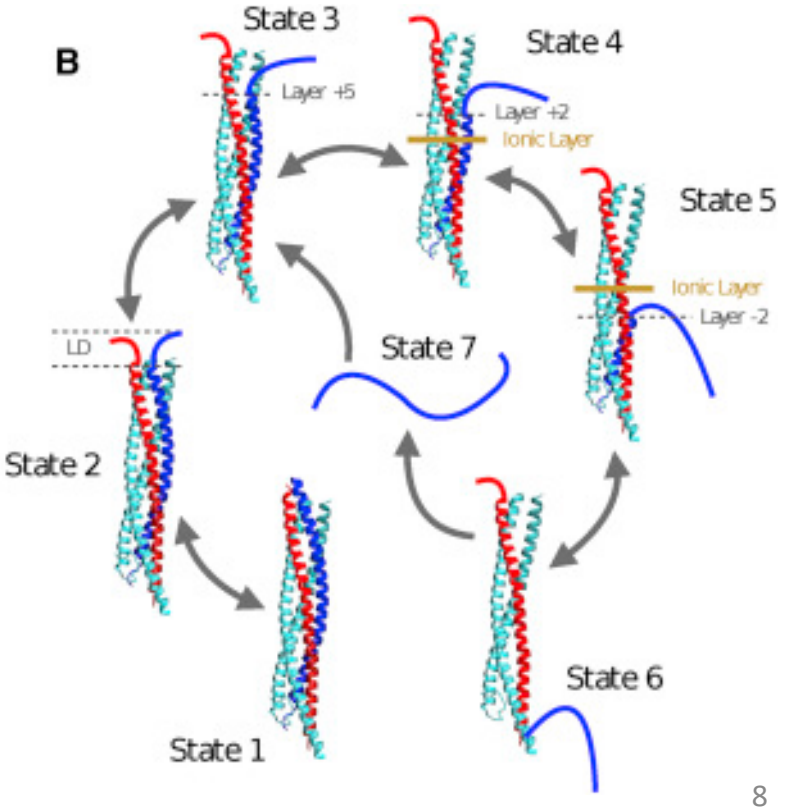
C. Clementi, H. Nymeyer, J.N. Onuchic, *J. Mol. Biol.* 298, 937-953 (2000)

SNARE complex: Results



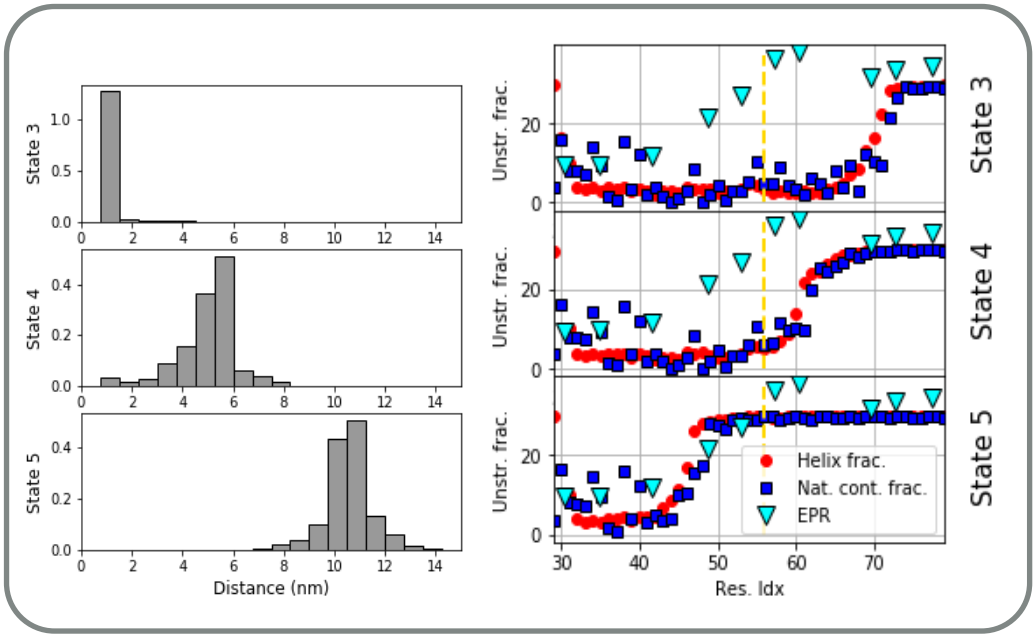


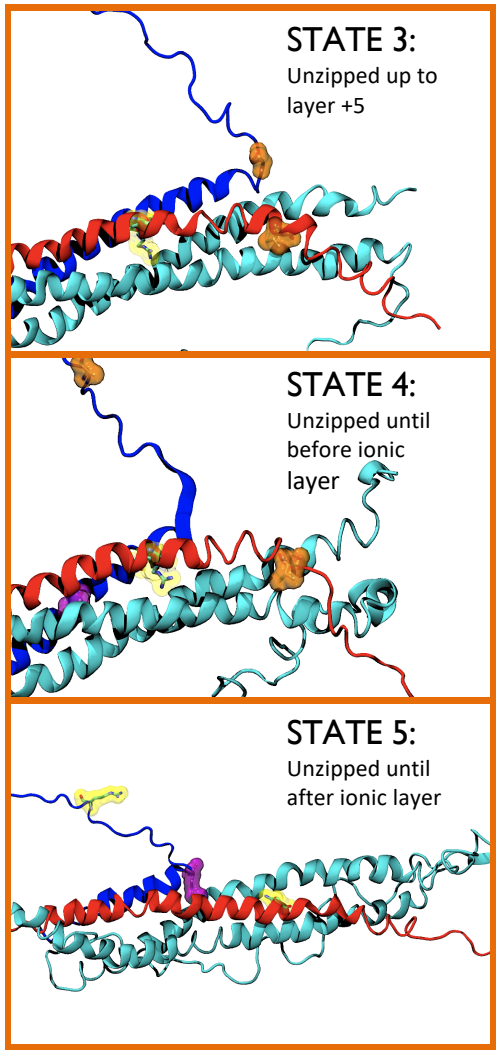




G. Pinamonti, G. Campo, J. Chen, A. Kluber & C.Clementi
Biophys. J. 115, 1470-1480 (2018)

Results: comparison with experiments





NEW STATE

compatible with FRET and optical tweezers experiments

compatible with EPR and optical tweezers experiments

G. Pinamonti, G. Campo, J. Chen, A. Kluber & C.Clementi
Biophys. J. 115, 1470-1480 (2018)

Revisiting coarse-graining:
outstanding challenges

- Definition of coarse variables
- Definition of effective energy function (and dynamic equations)
- Incorporation of experimental data

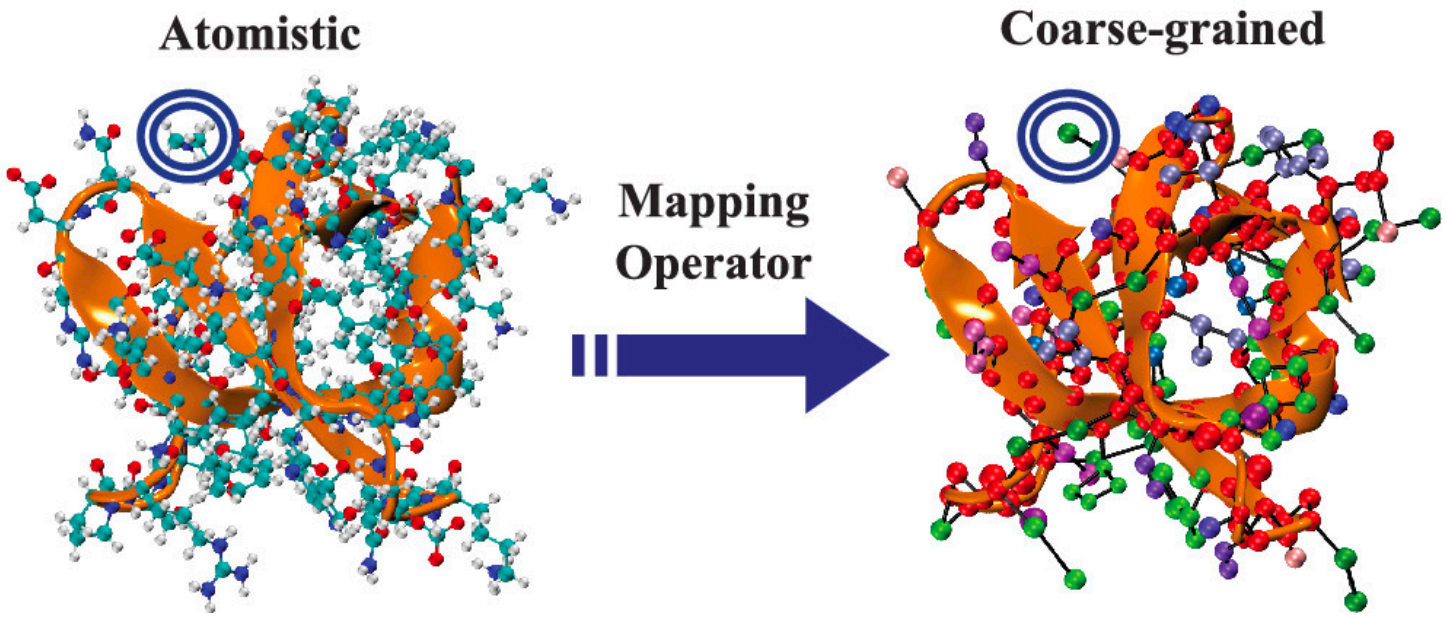
Coarse-grained mapping

atomistic representation $\mathbf{r} \in \mathbb{R}^{3N}$

CG representation $\mathbf{x} = \xi(\mathbf{r}) \in \mathbb{R}^{3n}$

atomistic potential $V(\mathbf{r})$

CG potential $U(\mathbf{x})$



W.G. Noid, J. Chem. Phys. (2013) 139:9, 090901

Coarse-graining with thermodynamic consistency

atomistic representation $\mathbf{r} \in \mathbb{R}^{3N}$

CG representation $\mathbf{x} = \xi(\mathbf{r}) \in \mathbb{R}^{3n}$

atomistic potential $V(\mathbf{r})$

CG potential $U(\mathbf{x})$

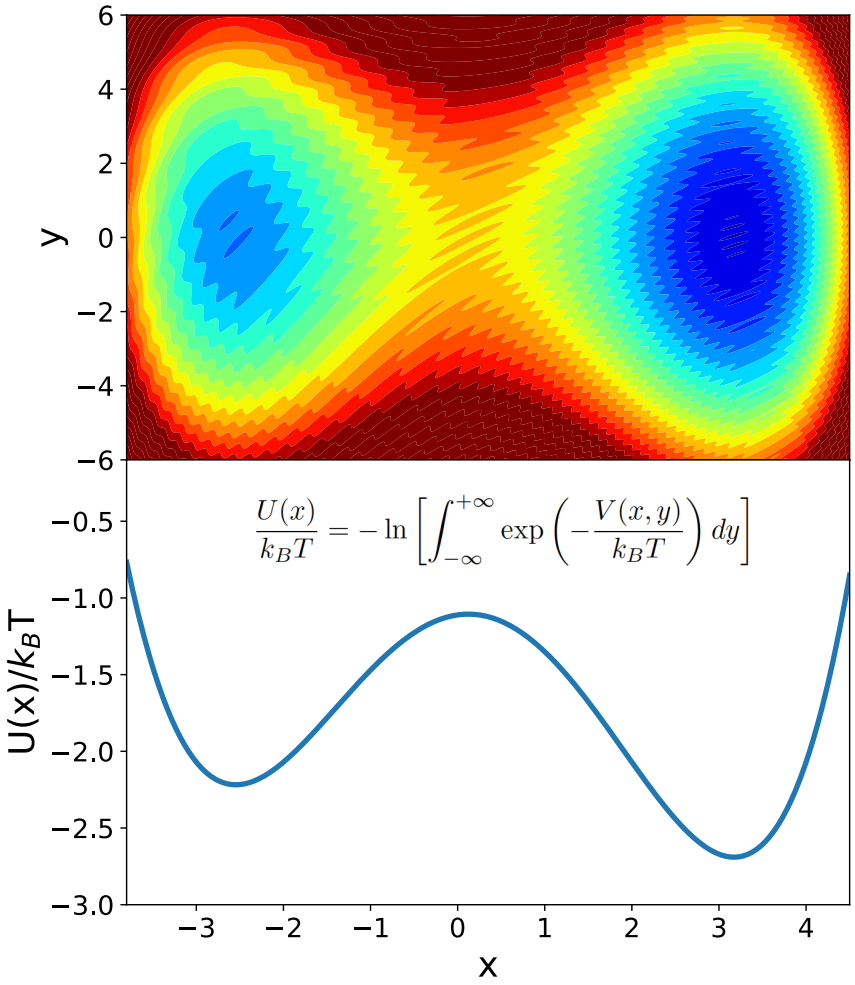
Thermodynamic consistency

$$U(\mathbf{x}) = -k_B T \ln p^{CG}(\mathbf{x}) + \text{const}$$

$$p^{CG}(\mathbf{x}) = \frac{\int \exp\left(-\frac{V(\mathbf{r})}{k_B T}\right) \delta\left(x - \xi(\mathbf{r})\right) d\mathbf{r}}{\int \exp\left(-\frac{V(\mathbf{r})}{k_B T}\right) d\mathbf{r}}$$

GOAL:

Optimize the parameters of a CG model $U(\mathbf{x};\boldsymbol{\theta})$ to satisfy this consistency as best as possible



W.G. Noid, J. Chem. Phys. (2013) 139:9, 090901

Force matching

Enforcing the thermodynamic consistency is equivalent to minimize the force matching error:

$$\chi^2(\boldsymbol{\theta}) = \left\langle \left\| \xi(\mathbf{F}(\mathbf{r})) + \nabla U(\xi(\mathbf{r}); \boldsymbol{\theta}) \right\|^2 \right\rangle_{\mathbf{r}}$$

instantaneous atomistic forces

The force matching error is always > 0 and can be decomposed in two parts:

$$\chi^2(\boldsymbol{\theta}) = \text{PMF error}(\boldsymbol{\theta}) + \text{Noise}$$

The PMF error is what should be as close to zero as possible

$$\text{PMF error}(\boldsymbol{\theta}) = \langle \left\| \mathbf{f}(\xi(\mathbf{r})) + \nabla U(\xi(\mathbf{r}); \boldsymbol{\theta}) \right\|^2 \rangle_{\mathbf{r}}$$

$$\text{Noise} = \langle \left\| \xi(\mathbf{F}(\mathbf{r})) - \mathbf{f}(\xi(\mathbf{r})) \right\rangle_{\mathbf{r}}$$

The noise term is determined solely by the coarse-graining mapping

$$\mathbf{f}(\mathbf{x}) = \langle \xi(\mathbf{F}(\mathbf{r})) \rangle_{\mathbf{r}|\mathbf{x}}$$

mean force

the gradient of the CG potential should be as close as possible to the “mean force”

Coarse-graining as a machine learning problem

Enforcing the thermodynamic consistency is equivalent to minimize the force matching error:

$$\chi^2(\boldsymbol{\theta}) = \left\langle \left\| \xi(\mathbf{F}(\mathbf{r})) + \nabla U(\xi(\mathbf{r}); \boldsymbol{\theta}) \right\|^2 \right\rangle_{\mathbf{r}}$$

Define a loss function to minimize over a Boltzmann-distributed sample:

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{R}) &= \frac{1}{M} \sum_{i=1}^{M} \left\| \xi(\mathbf{F}(\mathbf{r}_i)) + \nabla U(\xi(\mathbf{r}_i); \boldsymbol{\theta}) \right\|^2 \\ &= \left\| \xi(\mathbf{F}(\mathbf{R})) + \nabla U(\xi(\mathbf{R}); \boldsymbol{\theta}) \right\|_F^2 \end{aligned}$$

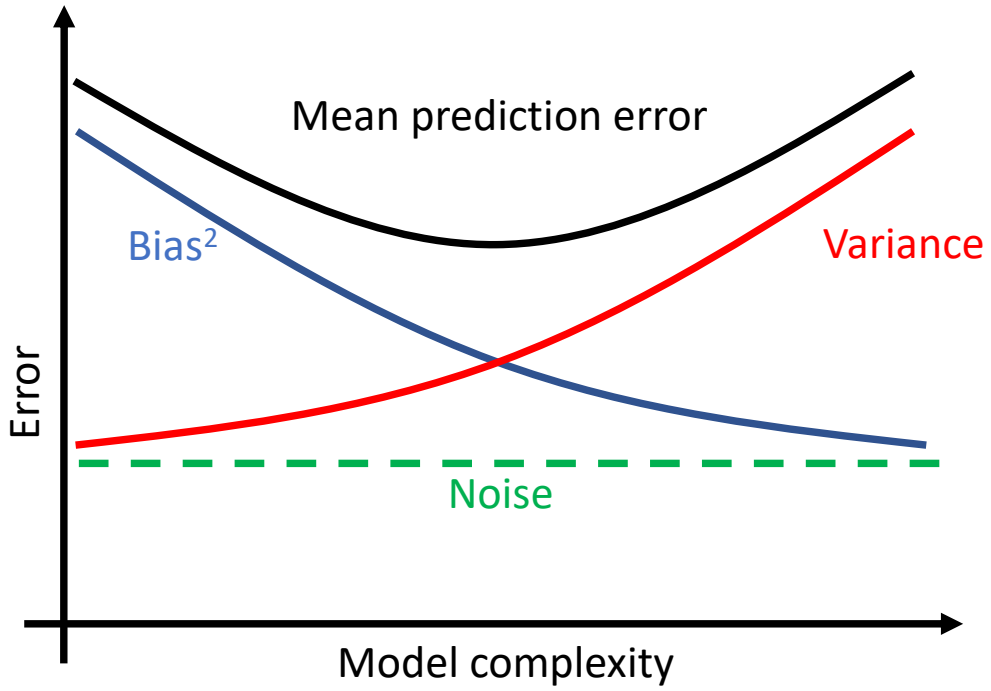
Coarse-graining as a machine learning problem

$$\mathbb{E}\left[L(\boldsymbol{\theta};\mathbf{R})\right] = \mathrm{Bias}^2 + \mathrm{Var} + \mathrm{Noise}$$

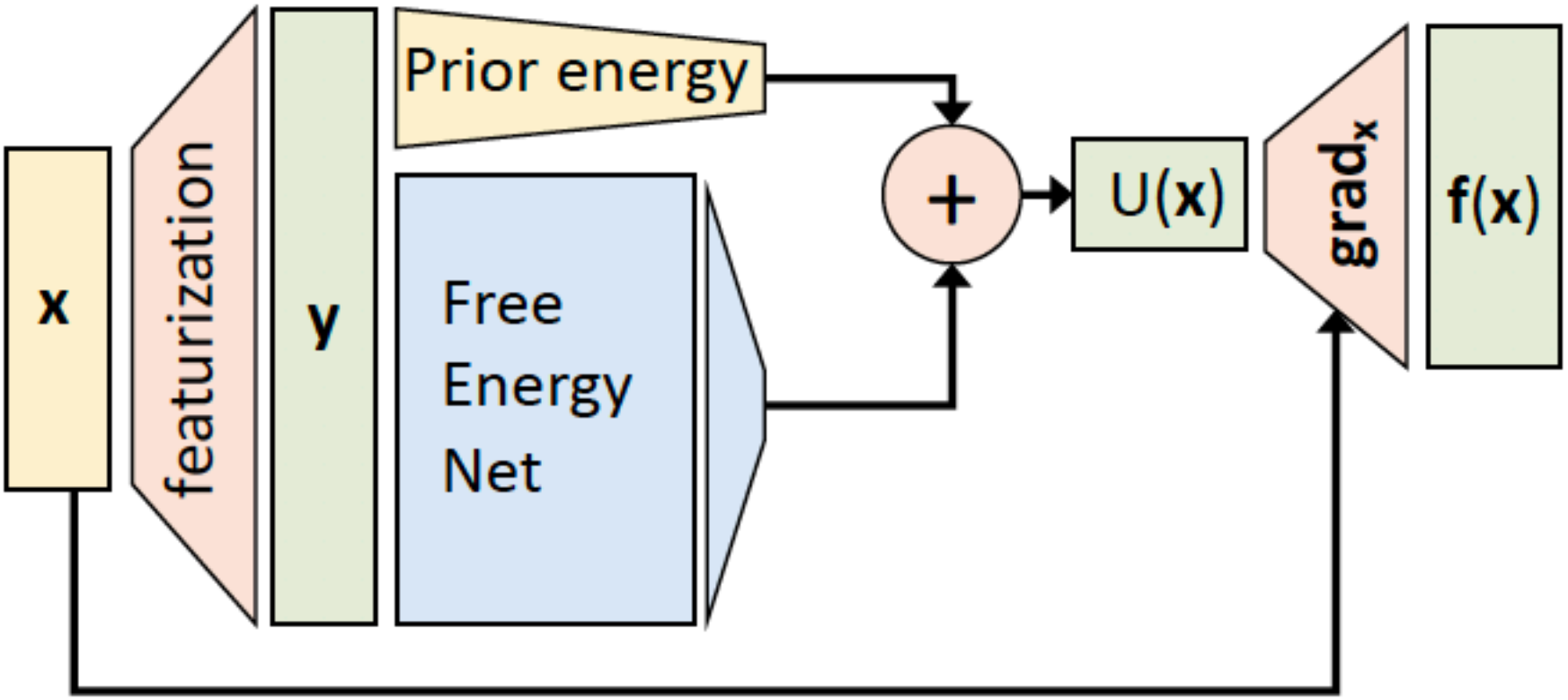
$$\mathrm{Bias}^2 = \left\|\mathbf{f}(\mathbf{X}) - \bar{\mathbf{f}}(\mathbf{X})\right\|_F^2$$

$$\mathrm{Var} = \mathbb{E}\left[\left\|\bar{\mathbf{f}}(\mathbf{X}) + \nabla U(\mathbf{X})\right\|_F^2\right]$$

$$\mathbb{E}\left[L(\boldsymbol{\theta};\mathbf{R})\right] = \mathrm{Bias}^2 + \mathrm{Var} + \mathrm{Noise}$$



CGnets

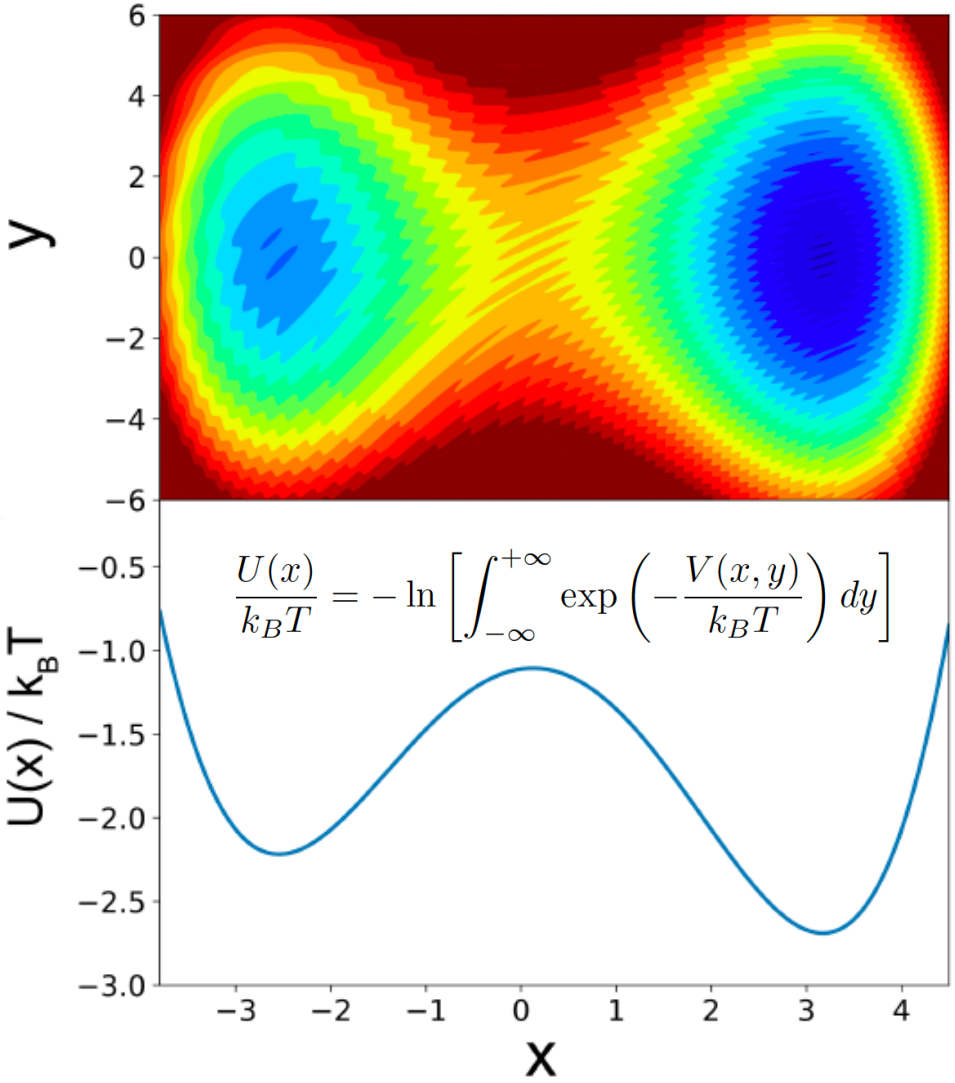


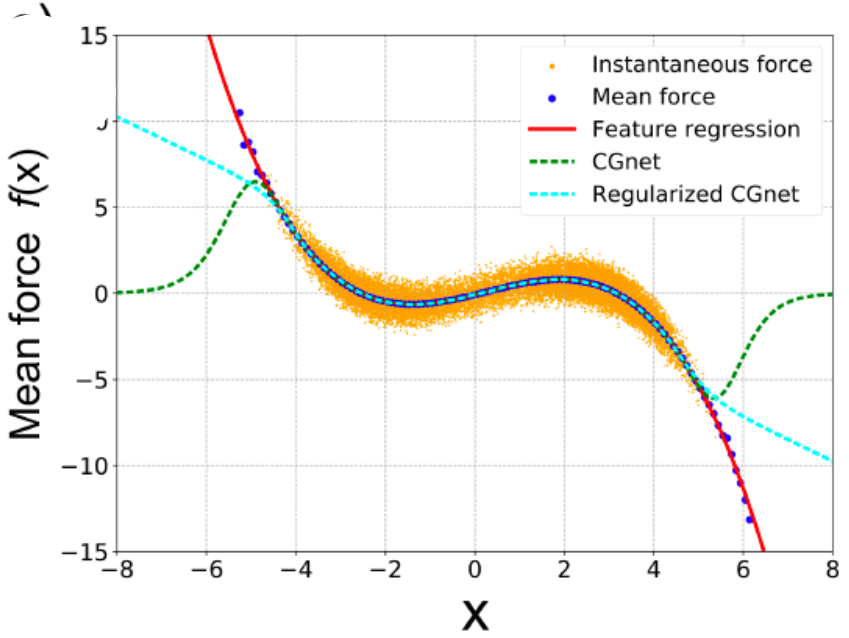
$$L(\boldsymbol{\theta};\mathbf{R}) = \frac{1}{M}\sum_{i=1}^{M}\left\|\xi(\mathbf{F}(\mathbf{r}_i)) + \nabla U(\xi(\mathbf{r}_i);\boldsymbol{\theta})\right\|^2$$

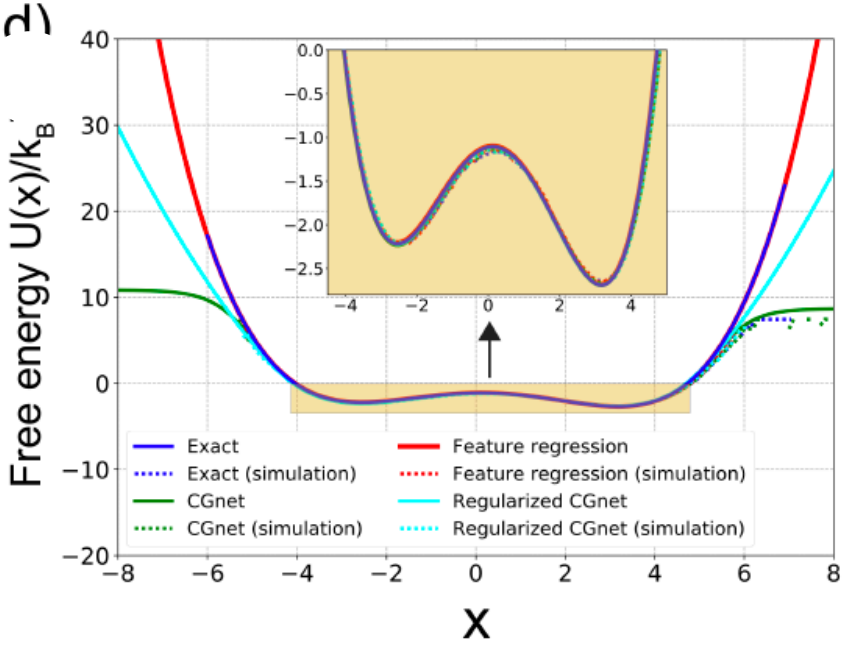
$$= \left\|\xi(\mathbf{F}(\mathbf{R})) + \nabla U(\xi(\mathbf{R});\boldsymbol{\theta})\right\|_F^2$$

J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron,
G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

CGnet for a 2d toy model

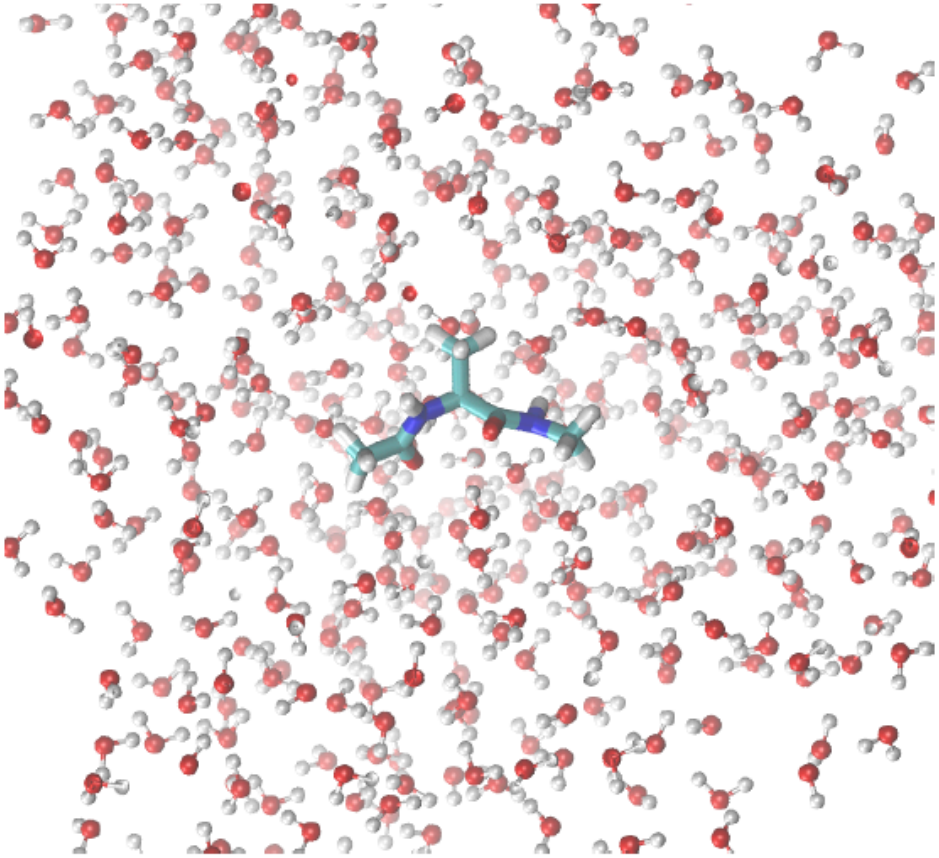




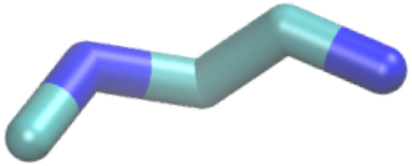


J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron,
G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

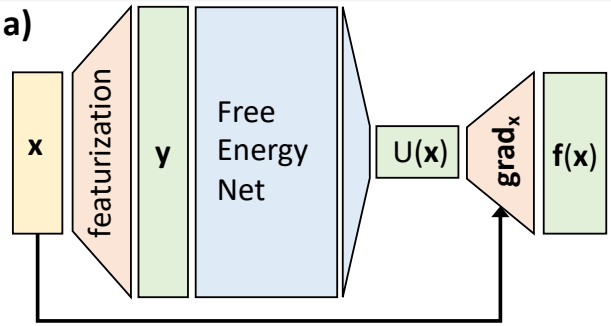
CGnet for Alanine Dipeptide



CG



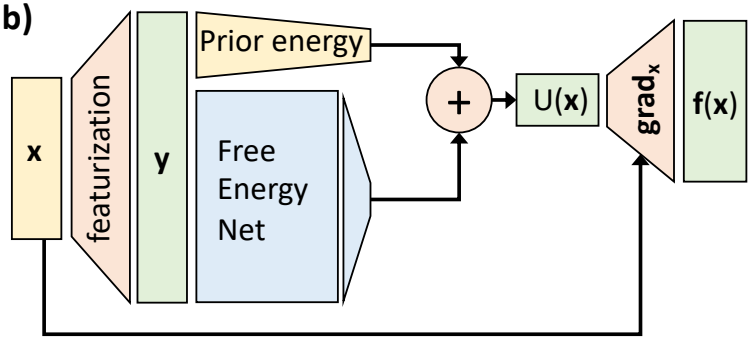
CGnets



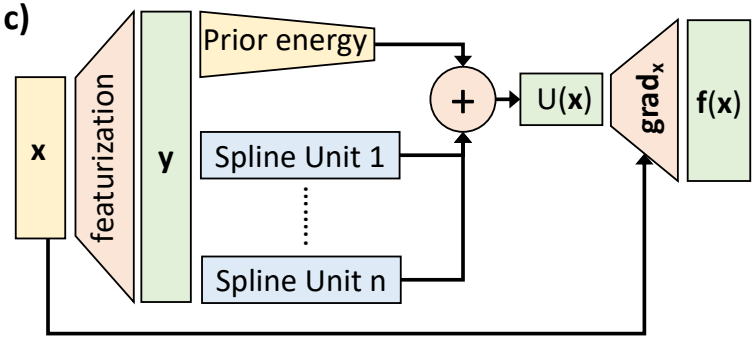
Unregularized CGnet

$$p(r_i) \propto \exp\left(-\frac{k_{b,i}(r_i - r_{i0})^2}{2k_BT}\right)$$

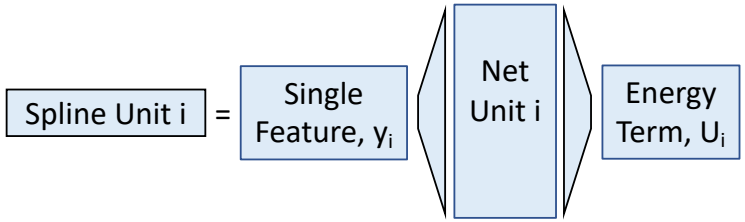
$$p(\theta_j) \propto \exp\left(-\frac{k_{a,j}(\theta_j - \theta_{j0})^2}{2k_BT}\right)$$



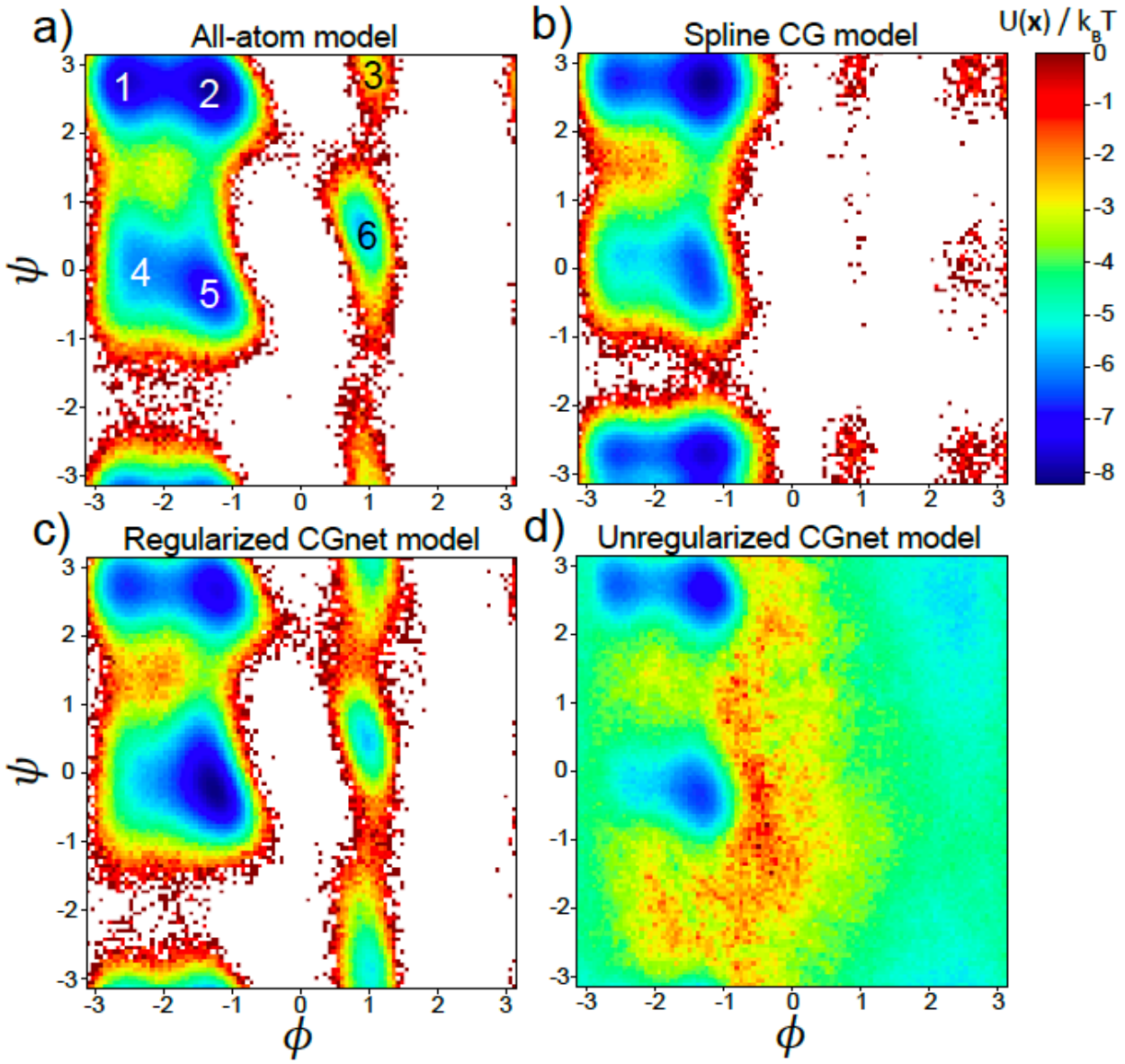
Regularized CGnet



"Spline" model

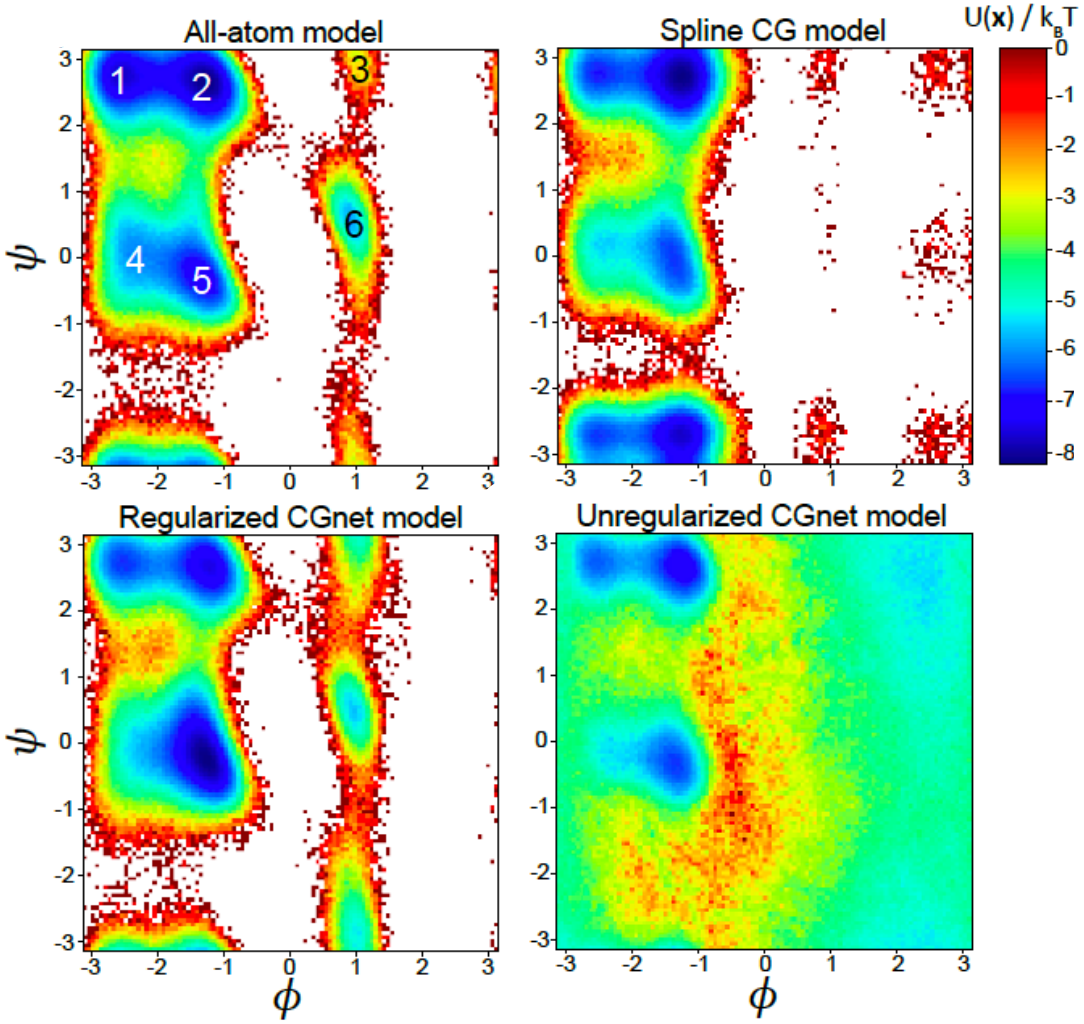


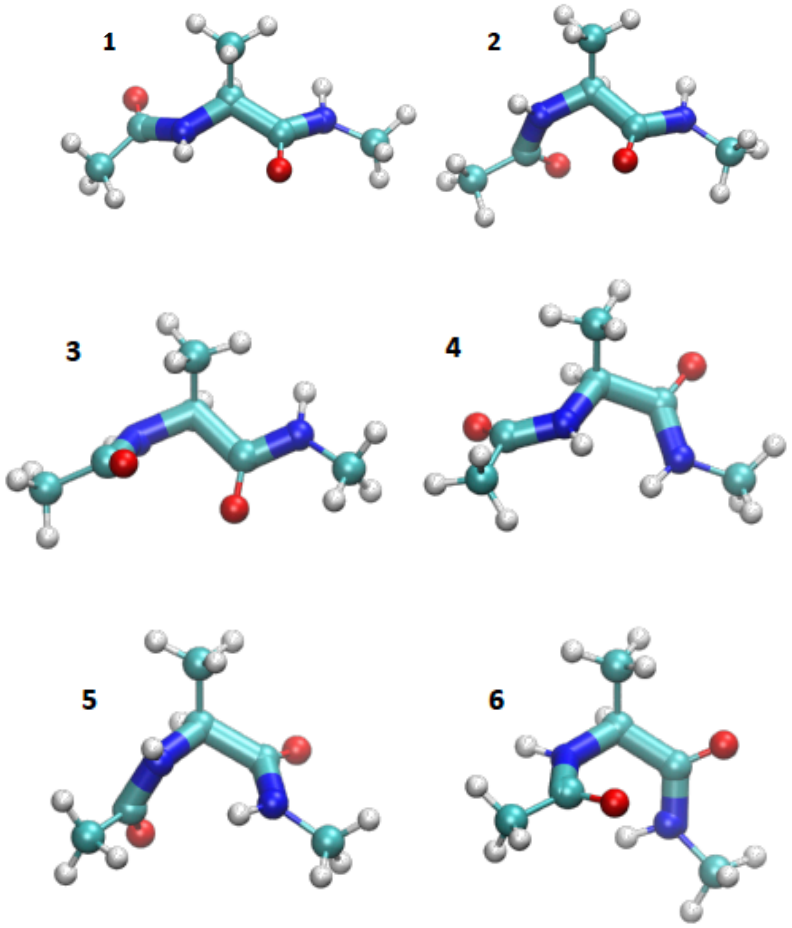
CGnets for Alanine Dipeptide: results



J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron,
G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

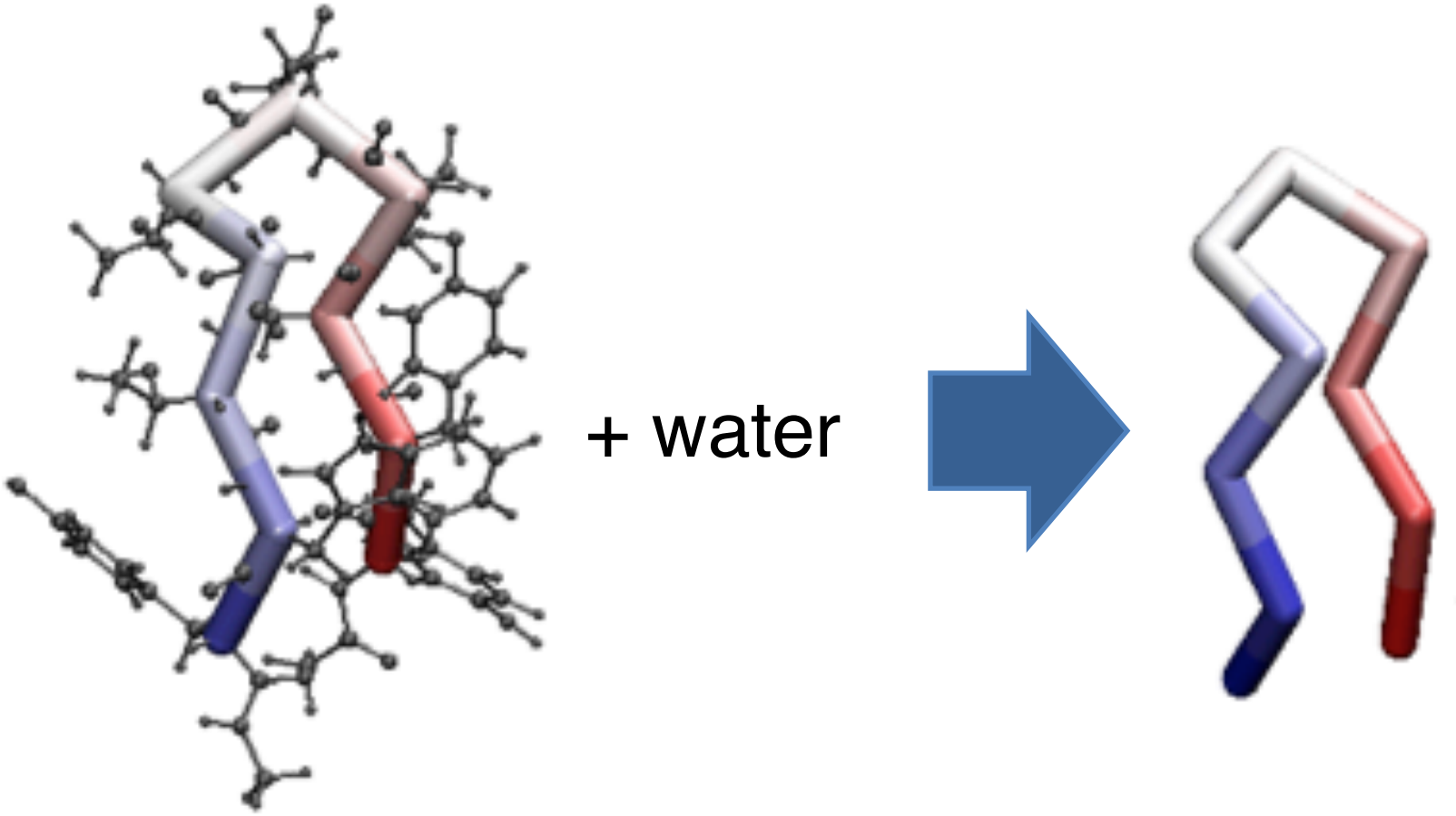
CGnets for Alanine Dipeptide: results





J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron,
G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

Coarse-graining of Chignolin folding/unfolding

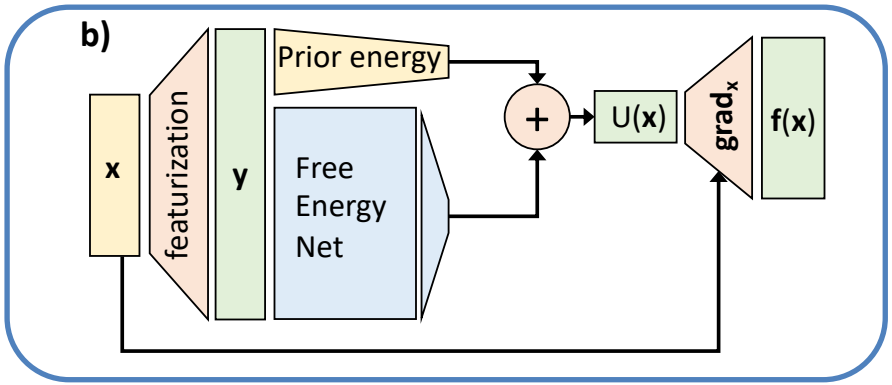


CGnets

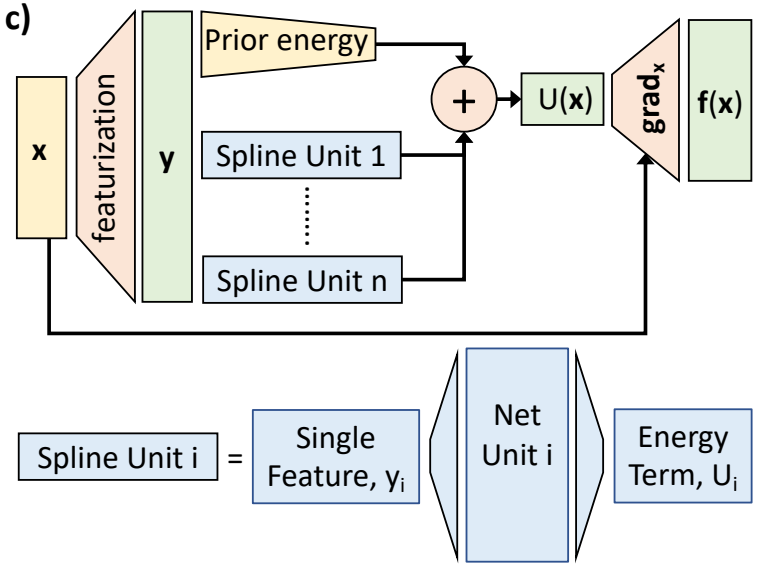
$$p(r_i) \propto \exp\left(-\frac{k_{b,i}(r_i - r_{i0})^2}{2k_BT}\right)$$

$$p(\theta_j) \propto \exp\left(-\frac{k_{a,j}(\theta_j - \theta_{j0})^2}{2k_BT}\right)$$

$$U_{rep}(r) = \left(\frac{\sigma}{r}\right)^c$$



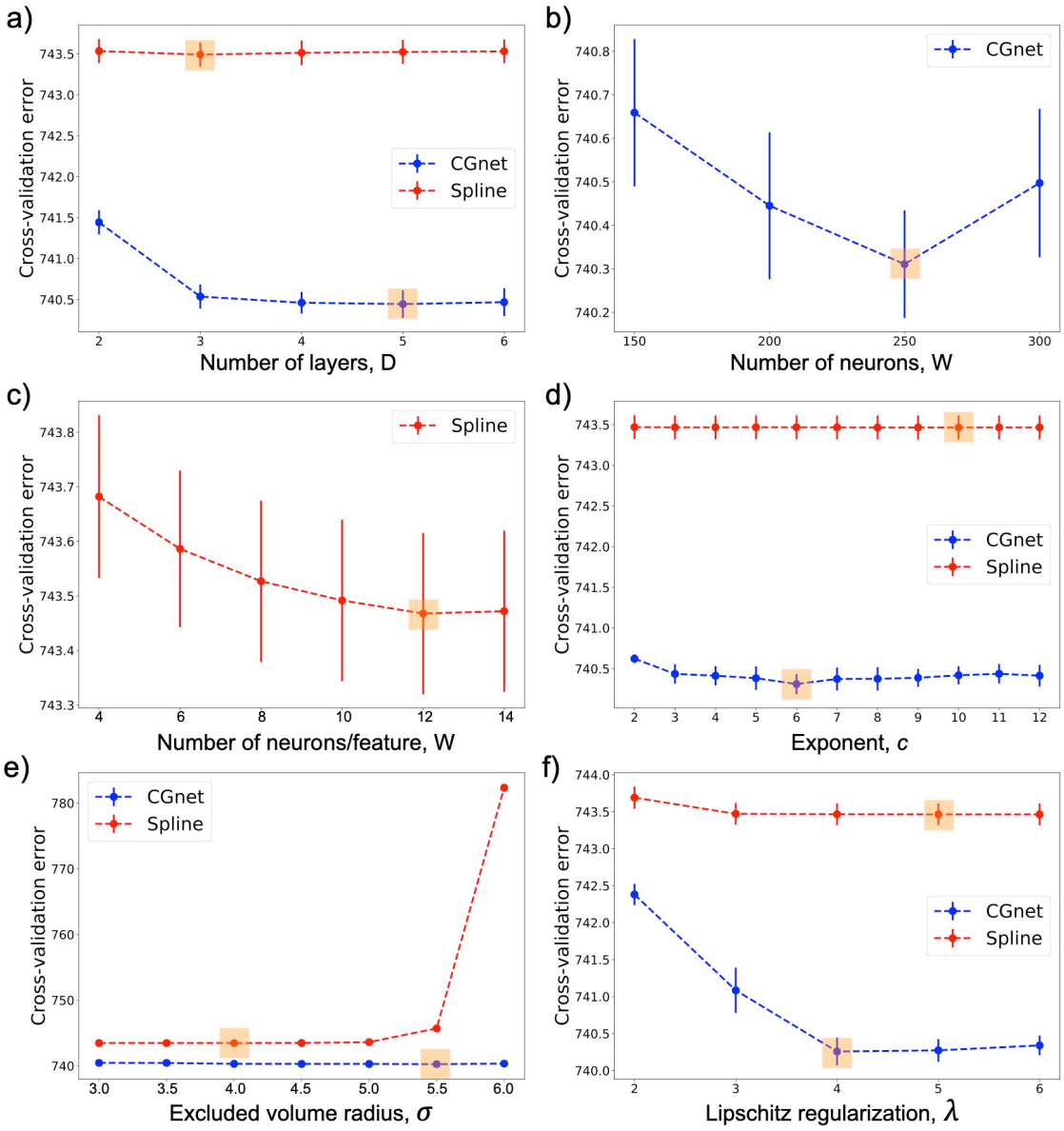
Regularized CGnet



"Spline" model

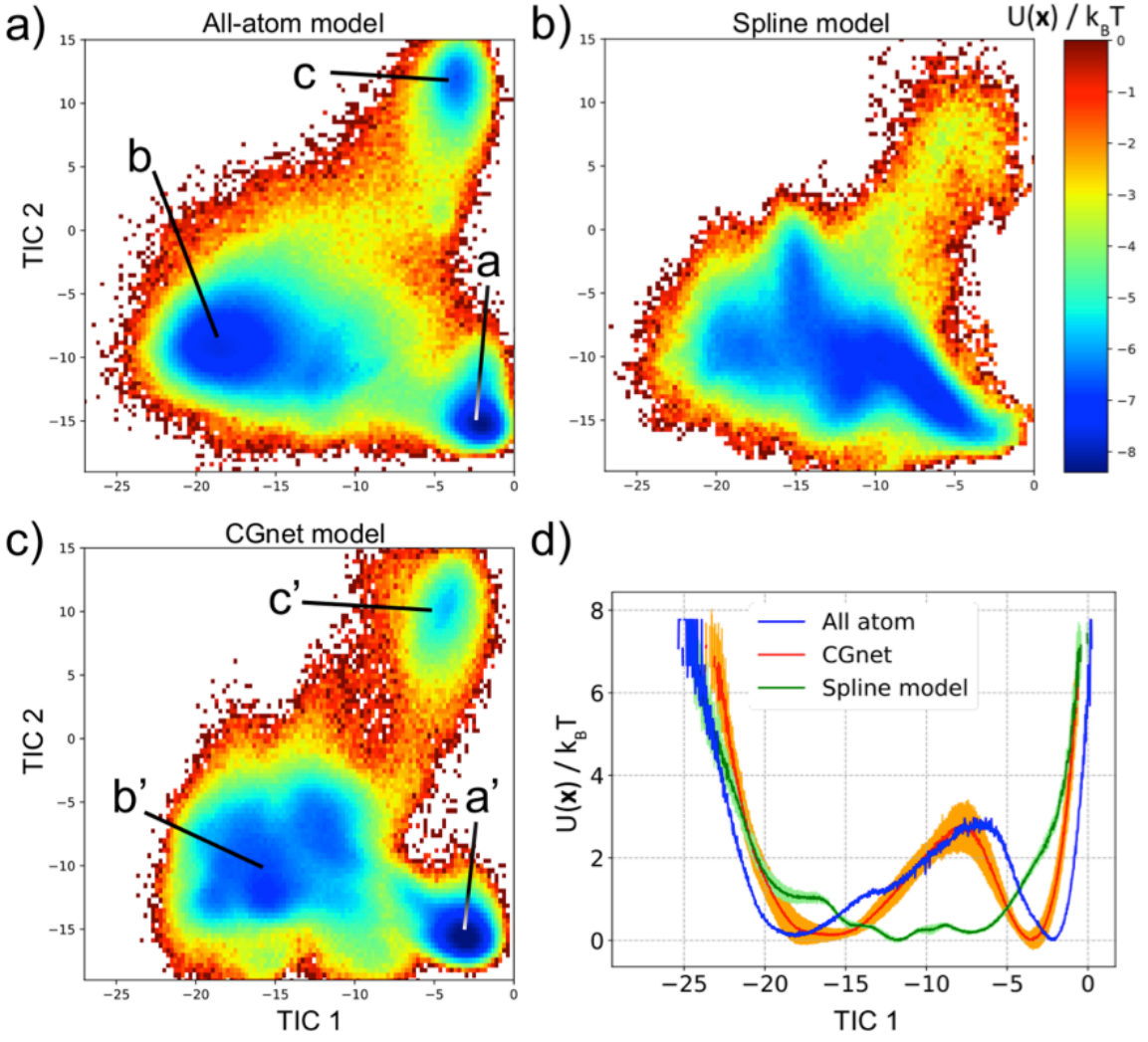
J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron, G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

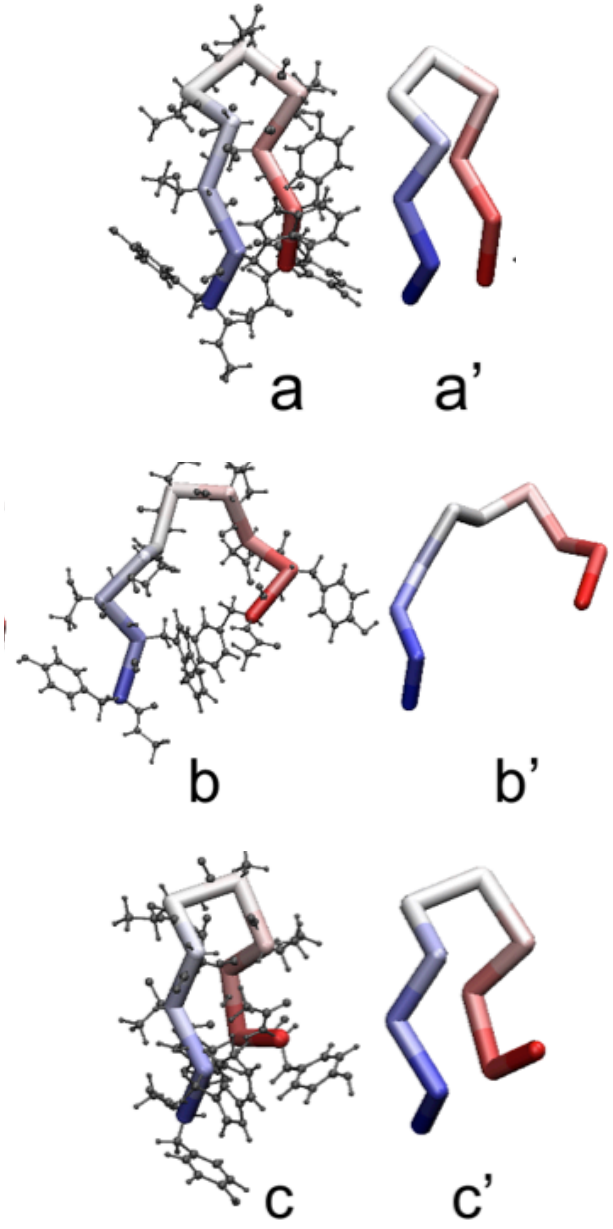
Coarse-graining of Chignolin folding/unfolding



J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron,
G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

Coarse-graining of Chignolin folding/unfolding





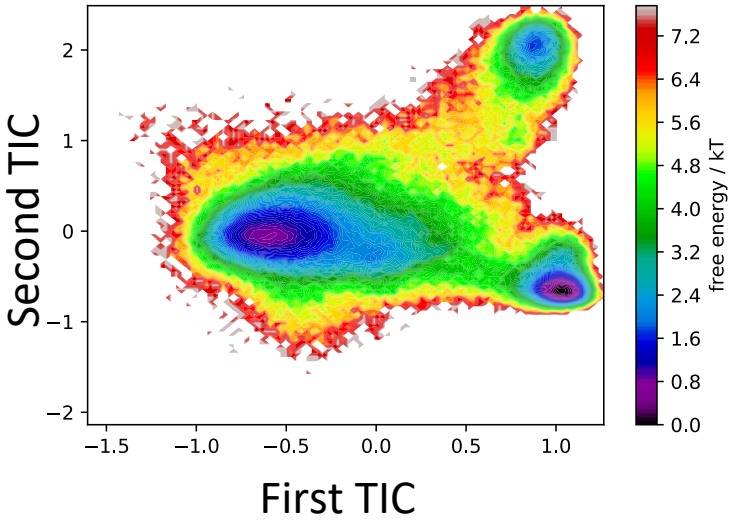
J. Wang, S. Olsson, C. Wehmeyer, A. Perez, N.E. Charron,
G. de Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci, 5, 755-767 (2019)

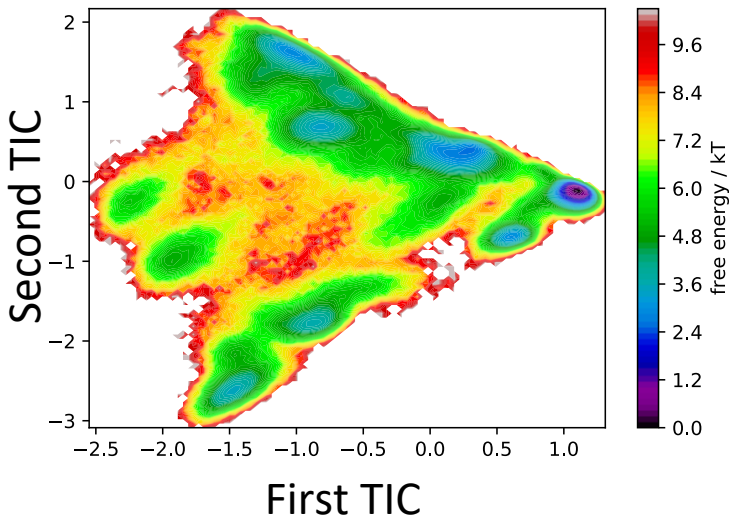
Machine learning for coarse-grained protein models

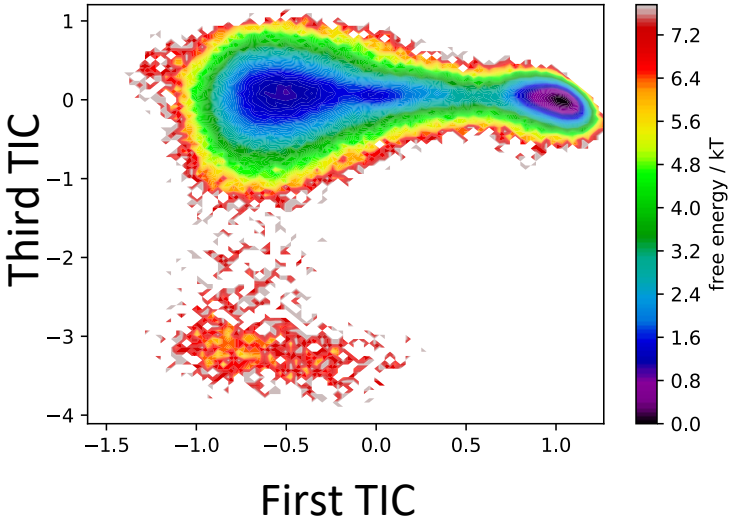
- Coarse-graining can be formulated as a machine learning problem
- A neural net naturally captures crucial multi-body effects
- Applications to more complex systems may require additional regularization (more physical constraints)
- Is it possible to design transferable CG models?

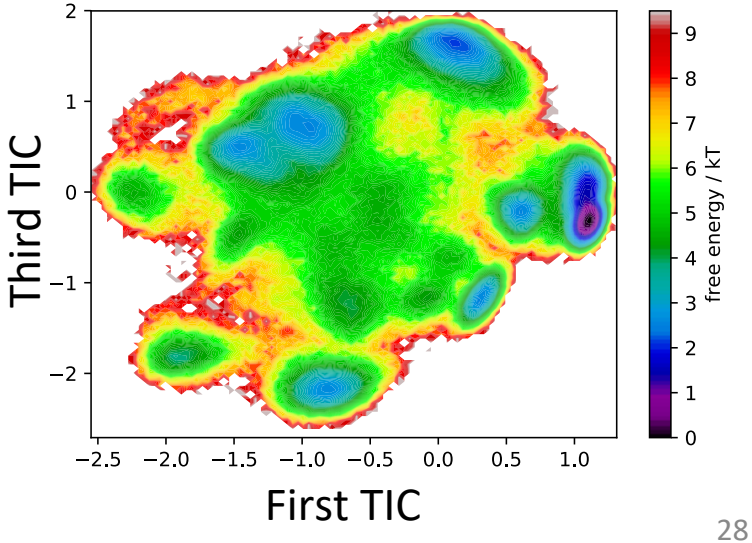
What about the dynamics?

Original all-atom Dynamics vs. CGnet Dynamics









Consistency

Thermodynamic Consistency:
potential recovers potential of mean force:

$$F^{\xi}(z) = -\log \int_{\Sigma_z} \mu(x) J^{-1/2}(x)\, \mathrm{d}x$$

Kinetic Consistency:
first few eigenvalues of original dynamics are restored:

$$\mathcal{L}^{\xi}\psi_i^{\xi} = -\kappa_i^{\xi}\psi_i^{\xi}$$

$$\kappa_i \approx \kappa_i^{\xi}$$

D. Crommelin and E. Vanden-Eijnden, Commun. Math. Sci. 4, 651 (2006)
D. Crommelin and E. Vanden-Eijnden, SIAM Multiscale Model. Simul. 9, 1588 (2011)

Overdamped Langevin dynamics with constant diffusion

$$d\mathbf{x}_t = -D\frac{\nabla U}{k_B T}dt + \sqrt{2D}dW_t$$

Generator:

$$\mathcal{L} = -\frac{D}{k_B T}\nabla U \cdot \nabla + D\Delta$$

Spectral matching:

$$\sum_{i=1}^{M} \|\mathcal{L}\hat{\psi}_i - \hat{\kappa}_i\hat{\psi}_i\|^2 = 0$$

$$\sum_{i=1}^{M} \| - \frac{\nabla U}{k_B T} \cdot \nabla\hat{\psi}_i + \Delta\hat{\psi}_i - \frac{\hat{\kappa}_i}{D}\hat{\psi}_i\|^2 = 0$$

Spectral matching

Test first few eigenvalue equations for effective generator in a weak sense:

$$\theta^* = \mathrm{argmin}_\theta \sum_{i=1}^{M} \sum_{j=1}^{P} \langle \mathcal{L}_\theta^\xi \tilde{\psi}_i^\xi + \tilde{\kappa}_i^\xi \tilde{\psi}_i^\xi, f_j \rangle_\nu^2$$

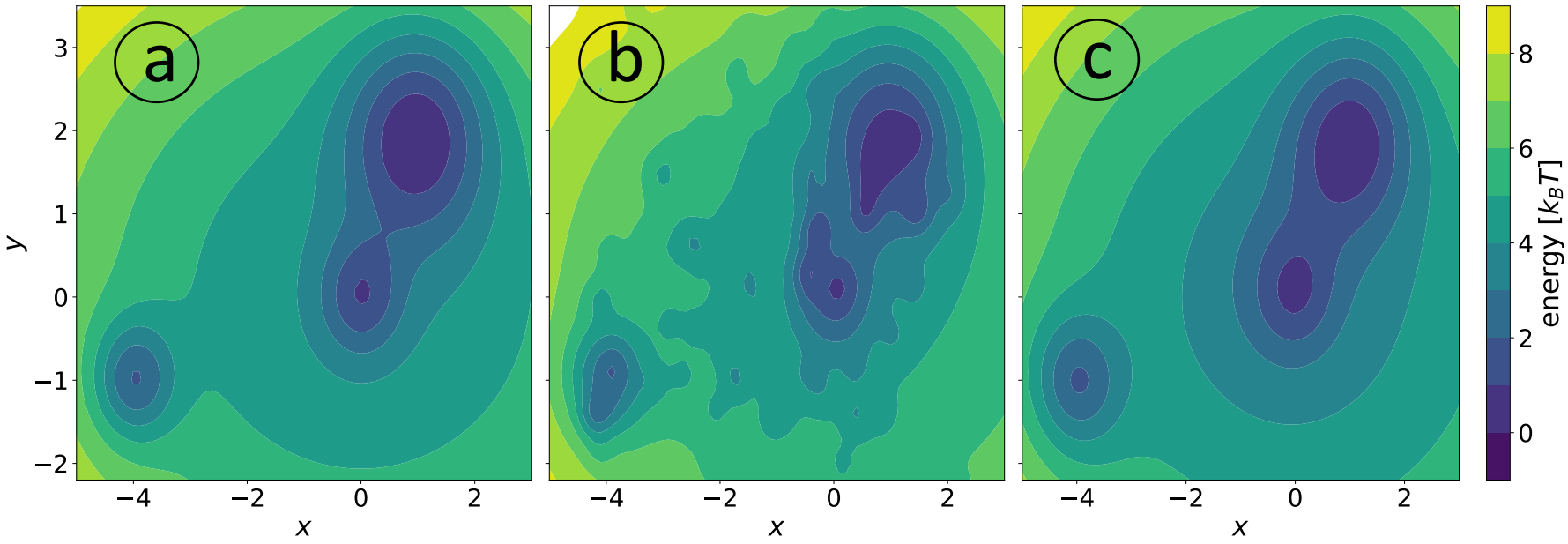
Requires:

- spectral data $\tilde{\kappa}_i^\xi,\ \tilde{\psi}_i^\xi$ (TICA, MSM, …)
- test functions f_j (user selection)
- parametric model

Output: optimal parameters.

Nüske, Boninsigna, Clementi, J. Chem. Phys. 151, 044116 (2019)

Toy example



The addition of 100 small random Gaussians to the smooth potential in (a) creates the noisy three-well potential in (b).

The potential learned by using the slow eigenfunctions of the latter is shown in (c):
It successfully encodes the long timescale features as it recovers the position and depth of the main energy minima while smoothing out the local and faster motions.

Nüske, Boninsigna, Clementi, J. Chem. Phys. 151, 044116 (2019)

Example: adjusting diffusion

Assume that thermodynamically consistent potential is available (e.g. from force matching).

Parametric model for the diffusion leads to symmetric generator:

$$\langle \mathcal{L}_\theta^\xi \tilde{\psi}_i, f_j \rangle_\nu = -\int A_\theta^\xi \cdot \nabla_z \tilde{\psi}_i \cdot \nabla_z f_j \, \mathrm{d}\nu$$

Using a linear model:

$$A_\theta^\xi = \left[\sum_{n=1}^{N} w_n g_n \right] \mathrm{Id}$$

We arrive at a linear regression:

$$E_1(w) = \|Xw - y\|^2,$$

$$X_{i,j;n} = -\langle g_n \nabla_z \tilde{\psi}_i \cdot \nabla_z f_j \rangle_\nu,$$

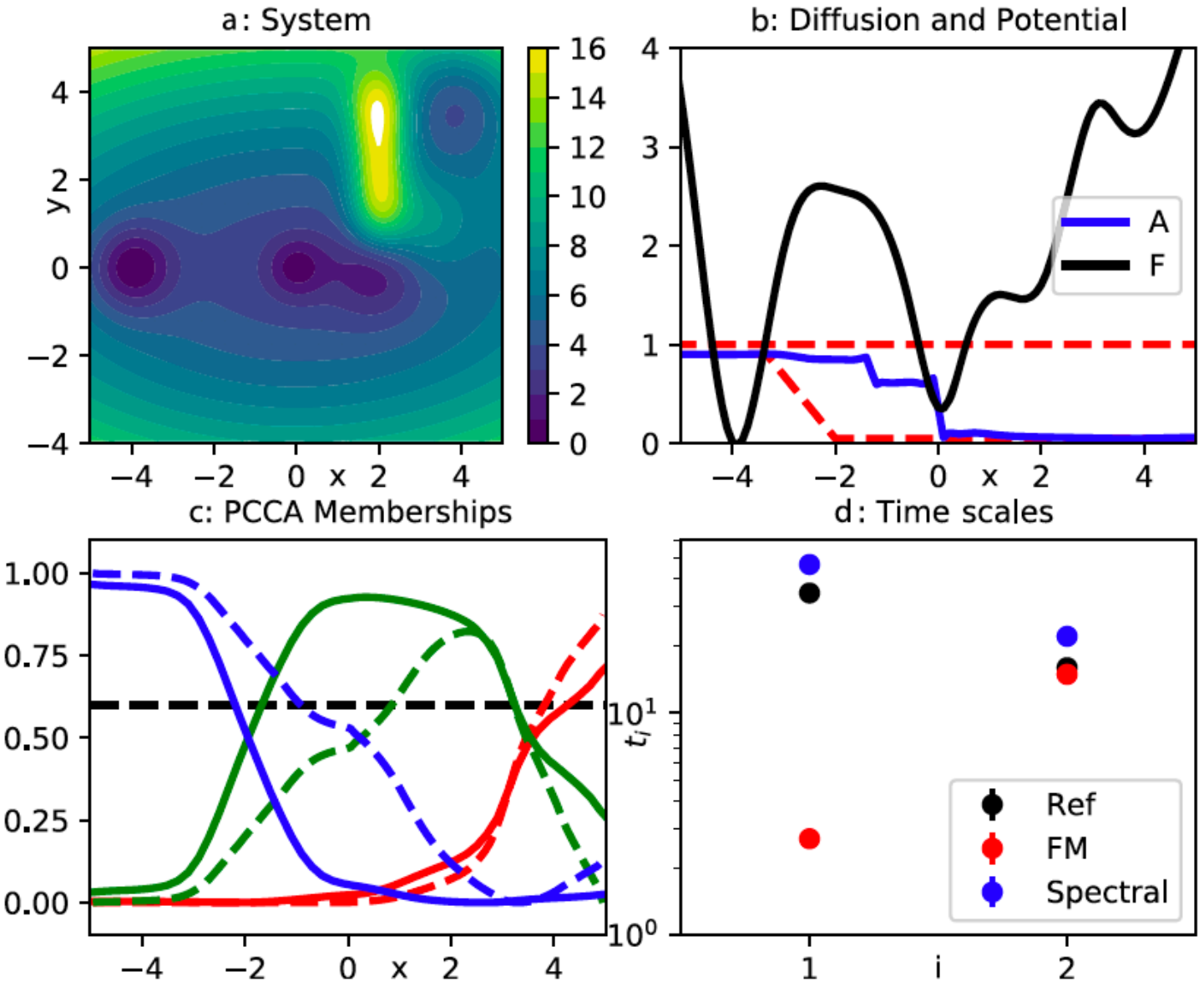
$$y_{i,j} = -\tilde{\kappa}_i \langle \tilde{\psi}_i, f_j \rangle_\nu.$$

Subject to positivity constraint:

$$\sum_{n=1}^{N} w_n g_n(Z_{t_k}) \geq a_{min} \geq 0$$

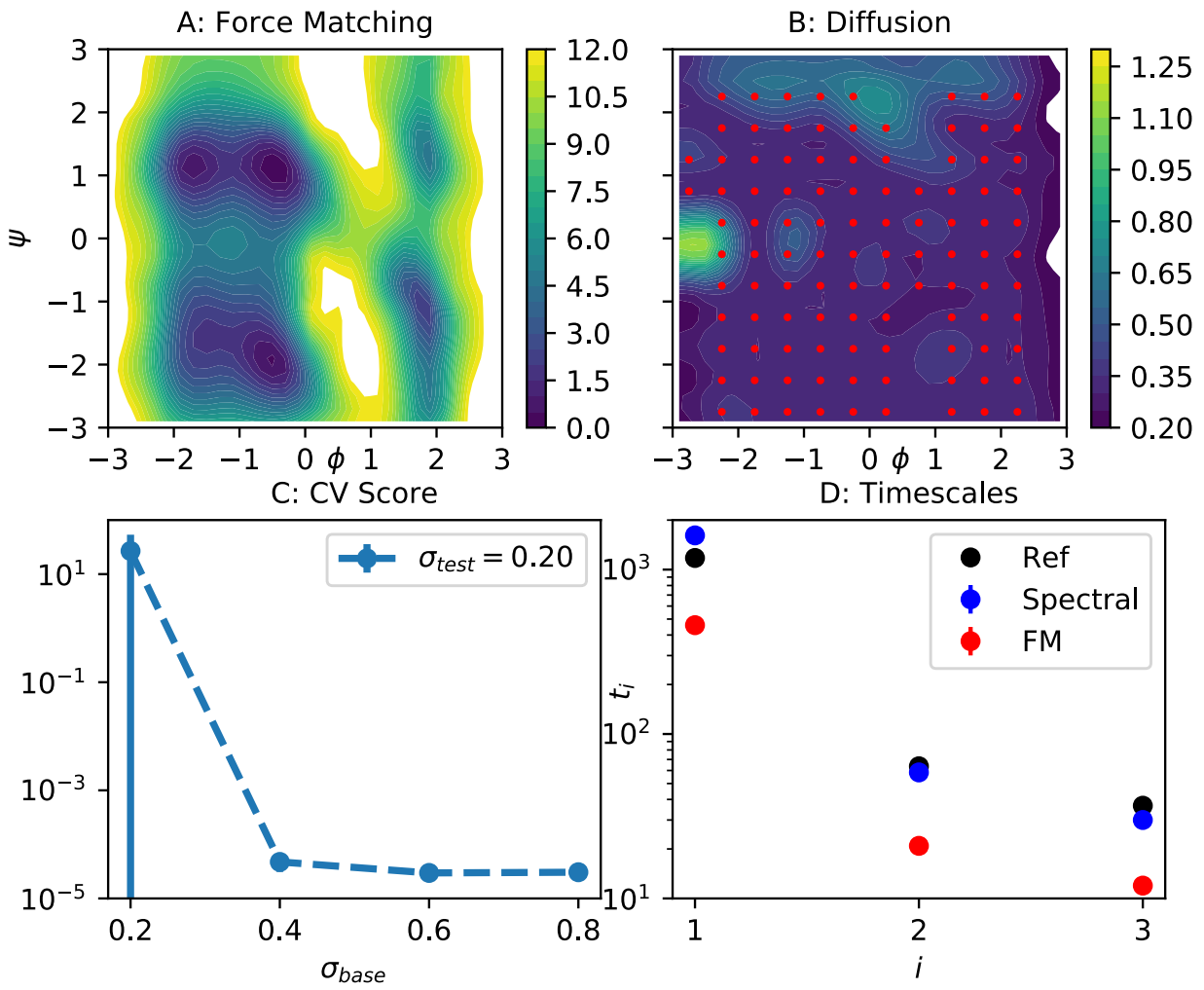
Nüske, Boninsigna, Clementi, J. Chem. Phys. 151, 044116 (2019)

Example: adjusting diffusion



Example: alanine dipeptide

- Projection to backbone dihedrals.
- Force Matching with 152 Gaussians.
- FM dynamics are too fast by factor 2-3.
- Spectral matching with 104 Gaussians.
- Diffusion is mostly constant.
- Timescales are restored.

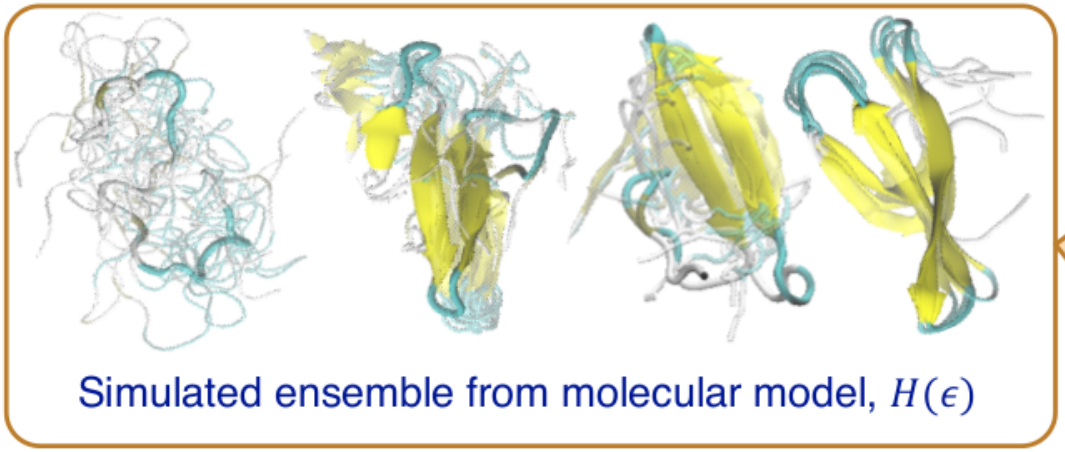


Nüske, Boninsigna, Clementi, J. Chem. Phys. 151, 044116 (2019)

Revisiting coarse-graining:
outstanding challenges

- Definition of coarse variables
- Definition of effective energy function (and dynamic equations)
- Incorporation of experimental data

Theoretical framework for optimal combination of simulation and experiment



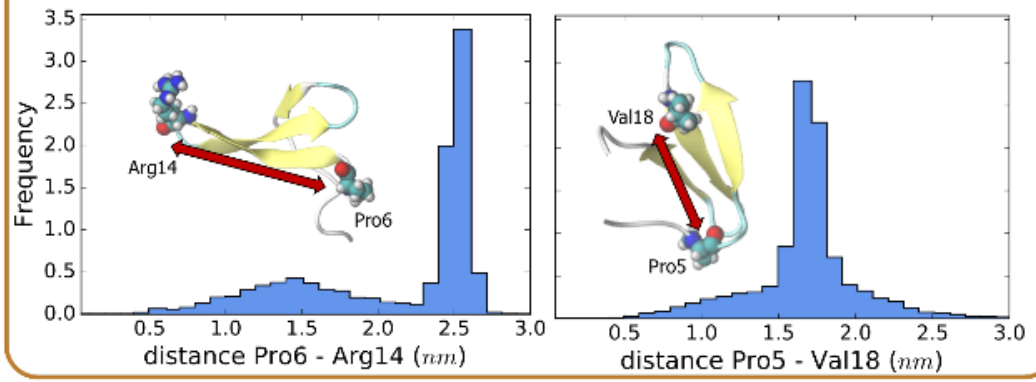
Simulated ensemble from molecular model, $H(\epsilon)$

$$Q(\epsilon) = \mathbb{P}\left[(e_1(\epsilon), \ldots, e_n(\epsilon)) \mid \text{Exp}\right]$$

Quantify agreement as function of the model parameters, ϵ

Optimal model maximizes $Q(\epsilon)$

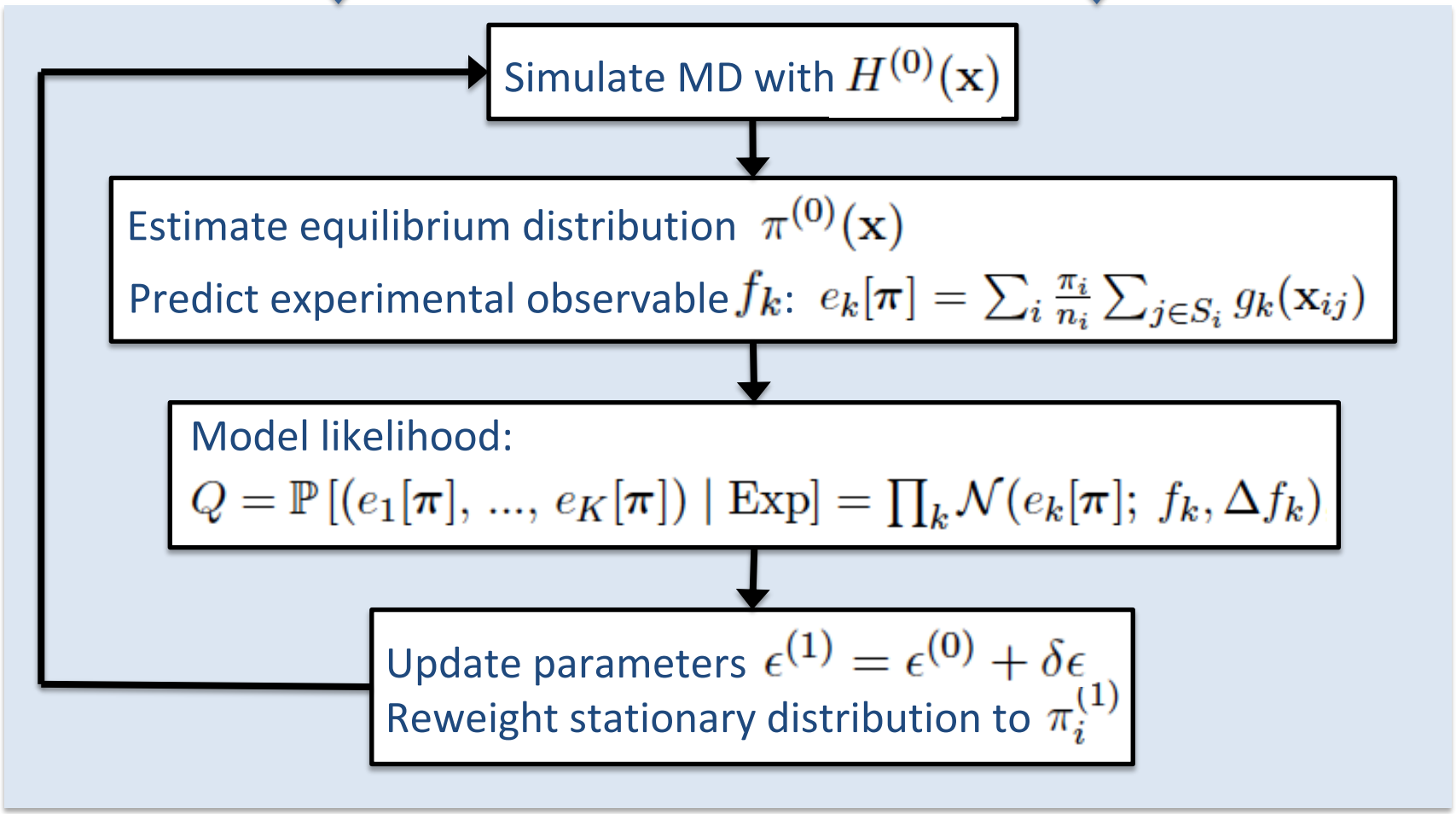
Experimental observables $(e_1, \ldots, e_n)$, e.g. FRET



Theoretical framework for optimal combination of simulation and experiment

Experimental measurements $(f_1, ..., f_K)$
& uncertainties $(\Delta f_1, ..., \Delta f_K)$

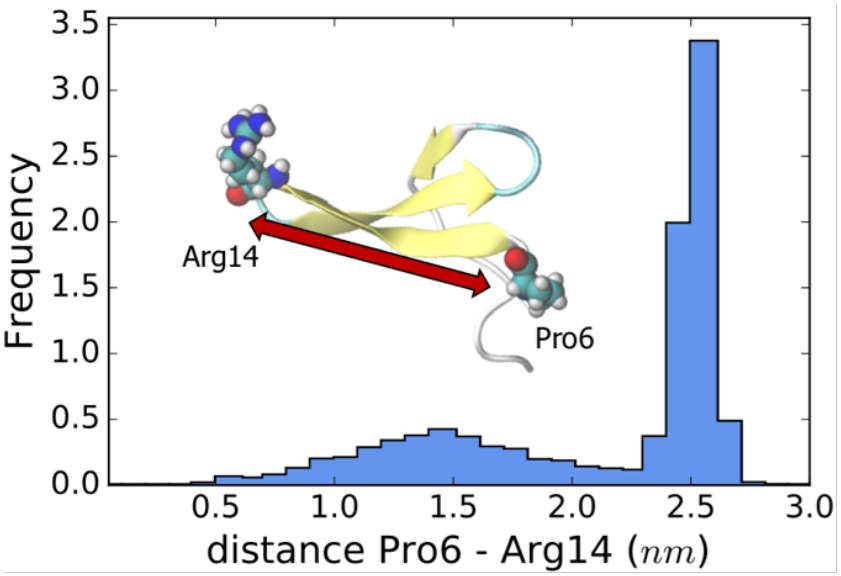
Initial guess for parameters $\epsilon^{(0)}$

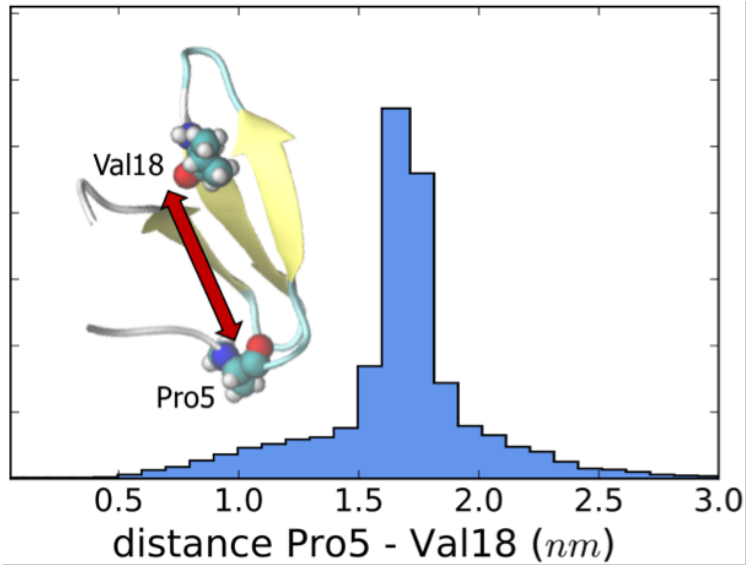


J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

Application: CG model of FIP35 from “synthetic FRET"

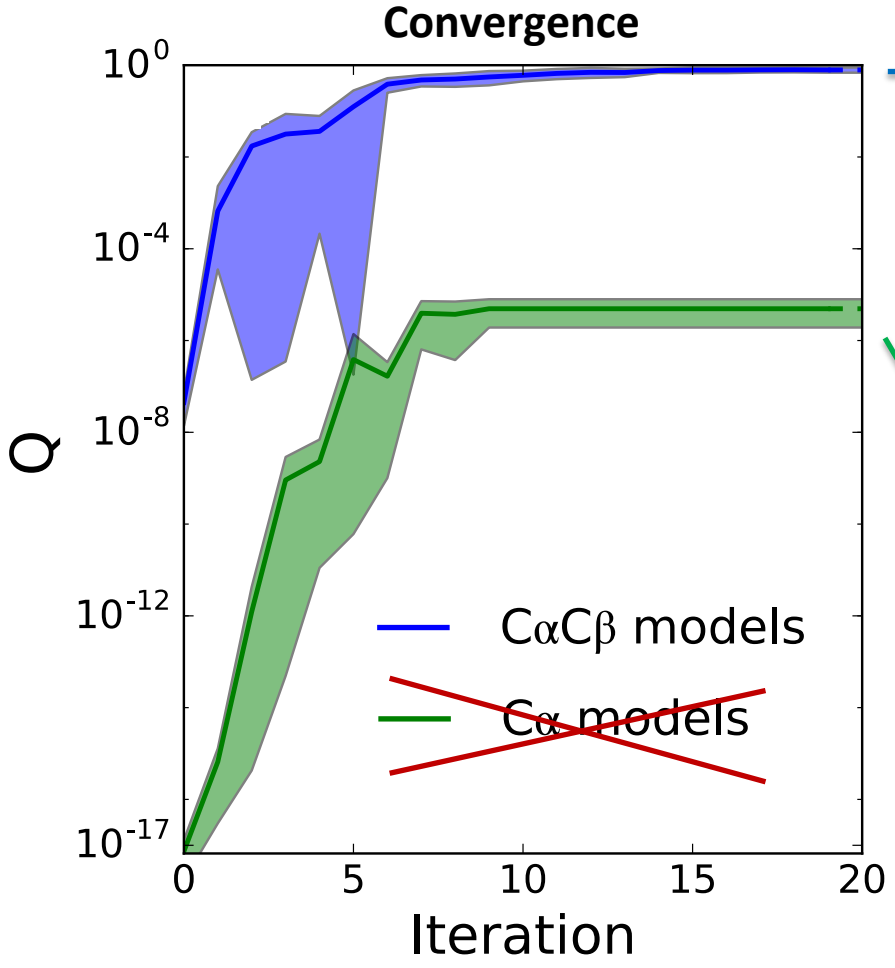
- Generate synthetic FRET distributions from all-atom (DESRES) equilibrium simulations of FIP35
- Learn "optimal" Coarse-Grained model from FRET data



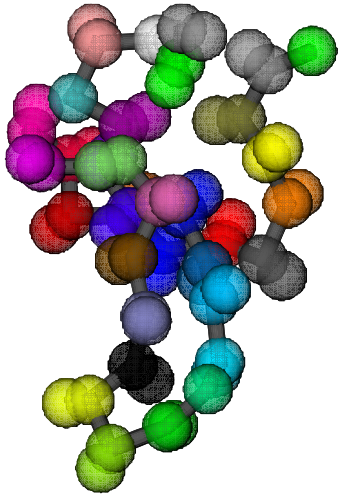


J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

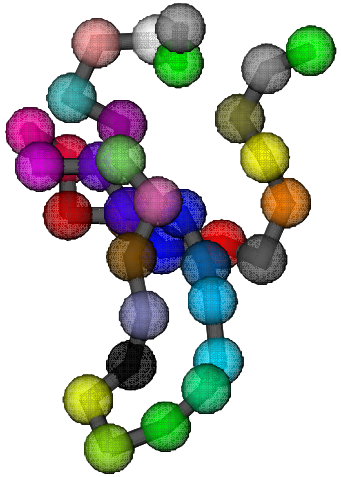
Application: CG model of FIP35 from "synthetic FRET"



CαCβ model

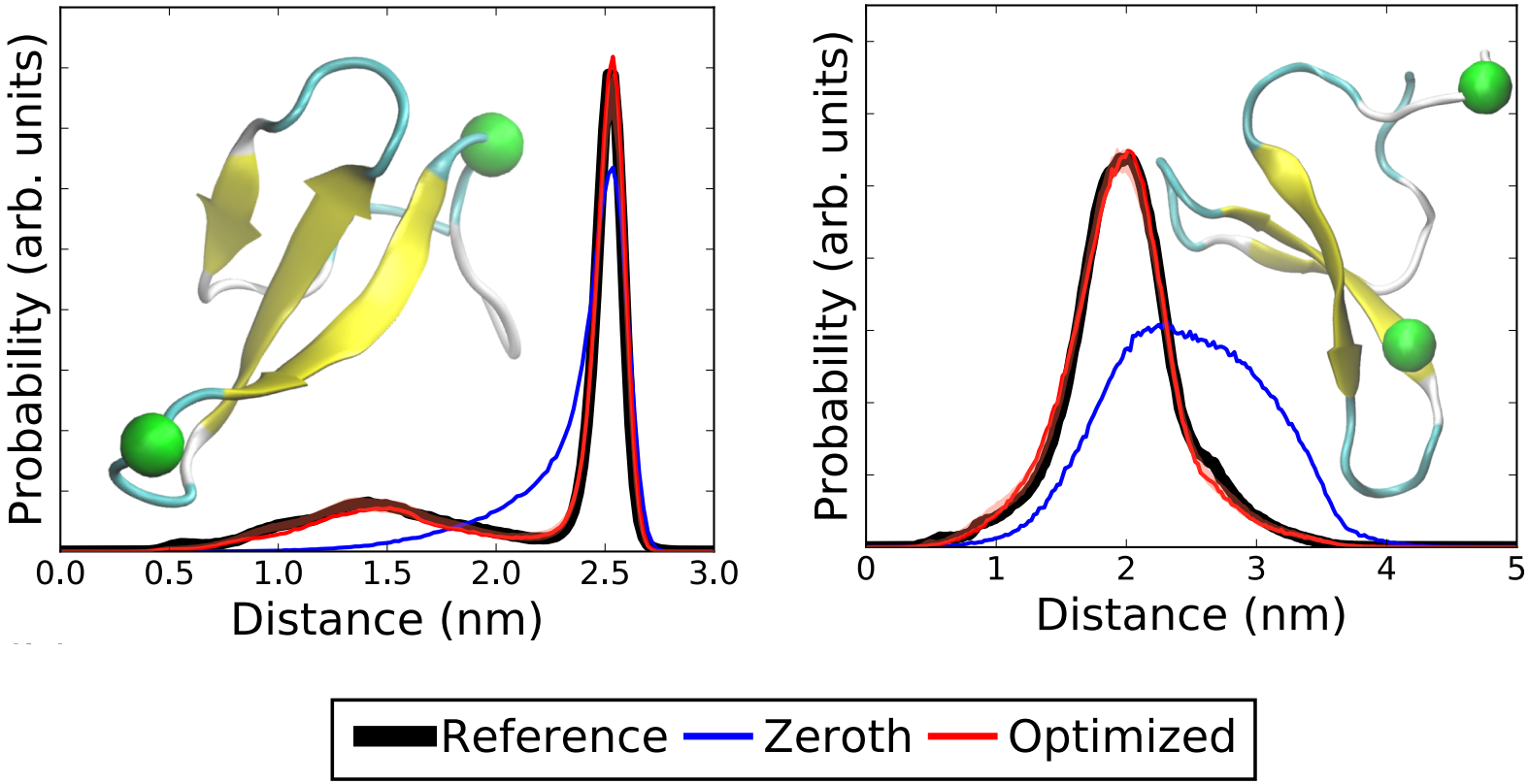


Cα model



J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

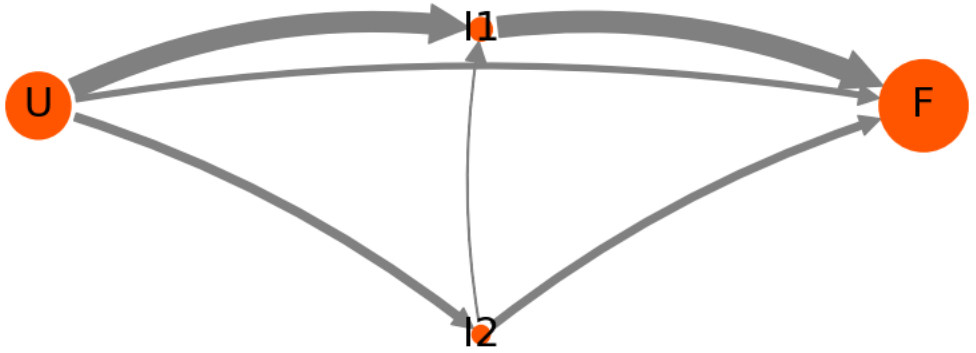
Application: CG model of FIP35 from "synthetic FRET"



J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

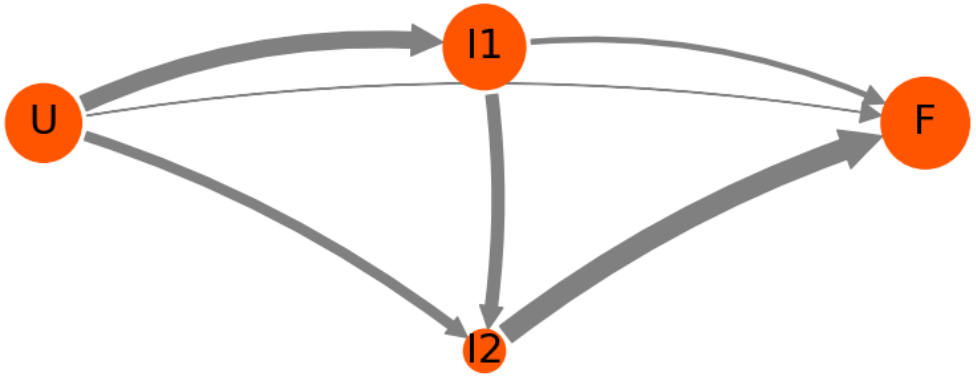
Application: CG model of FIP35 from "synthetic FRET"

"True" model



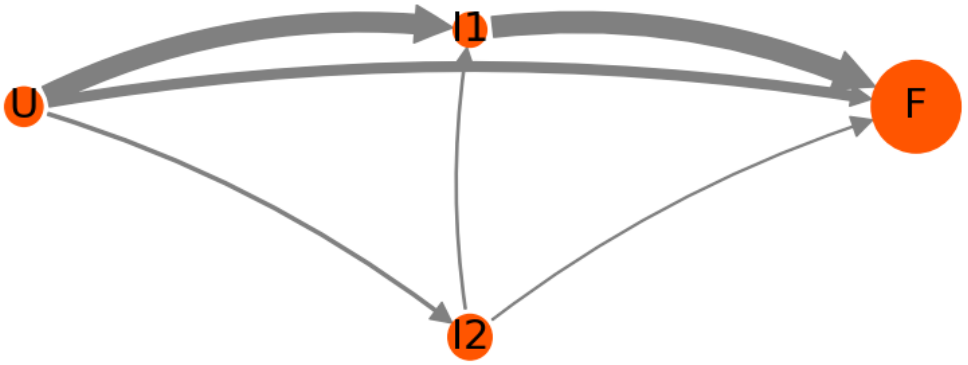
0^{th} order model

slowest timescale (folding)
5000 faster than all-atom



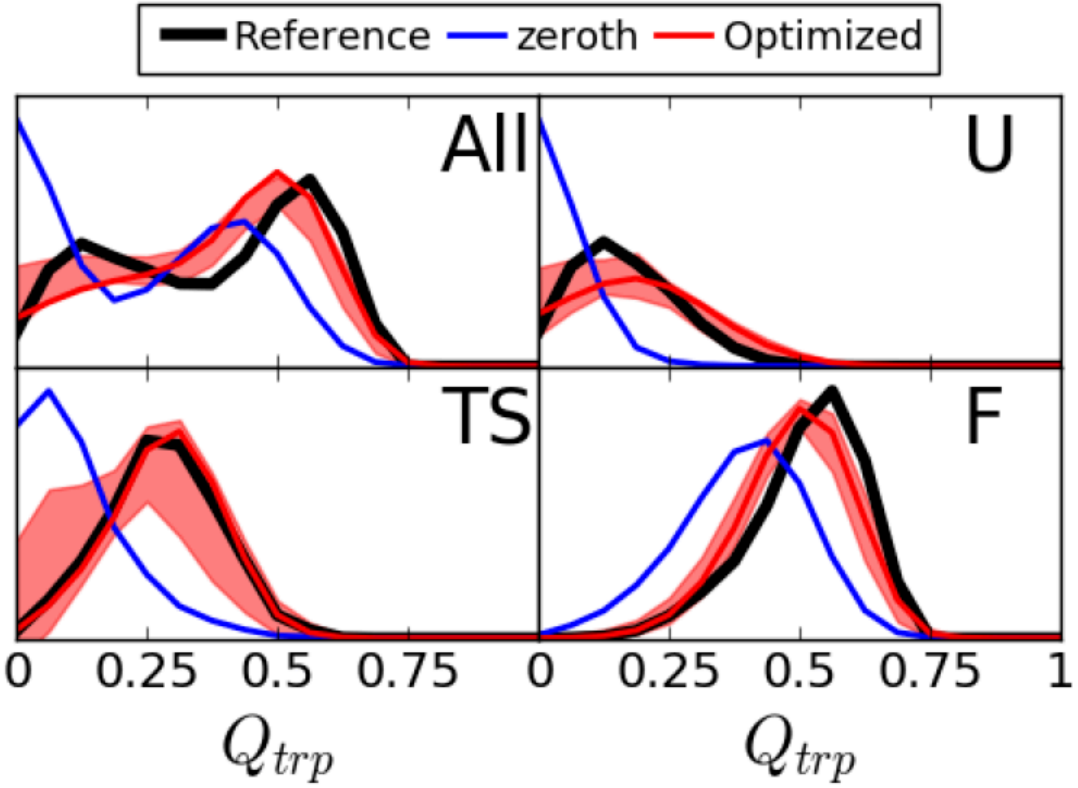
Trained CG model

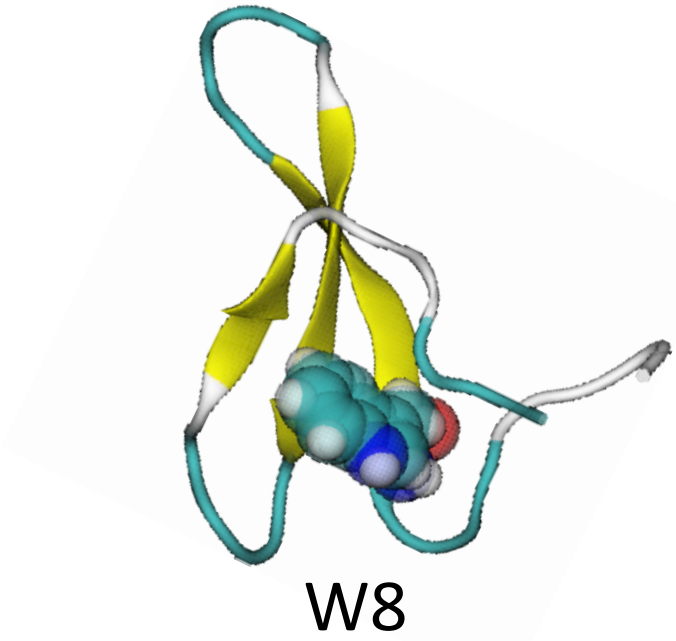
slowest timescale (folding)
500 faster than all-atom



J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

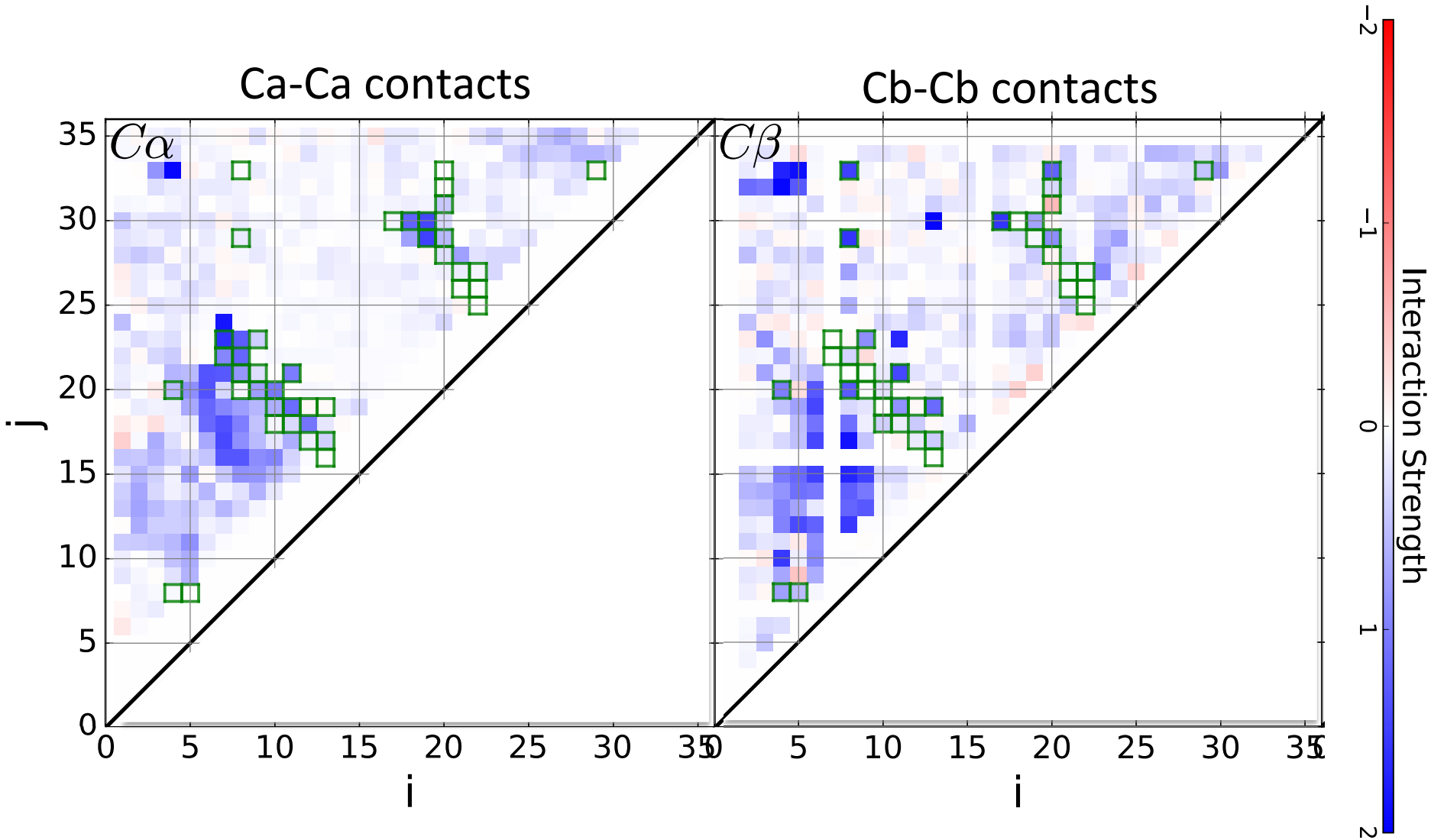
Cross validation: synthetic "Trp fluorescence"





J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

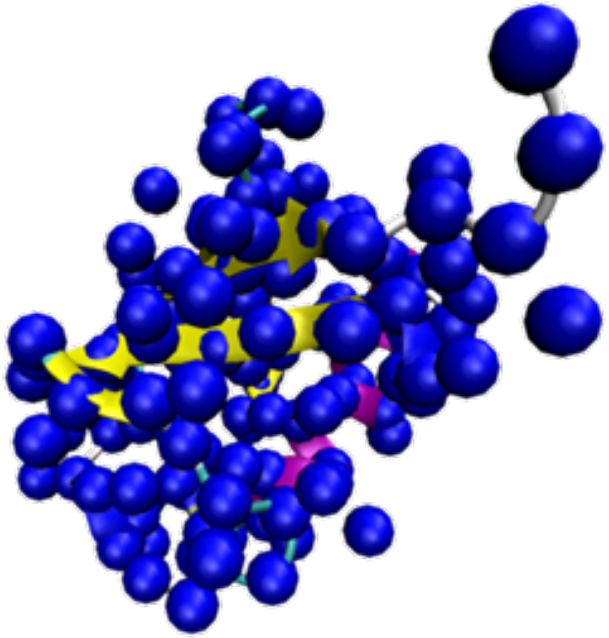
Analysis of parameter space



J. Chen, J. Chen, G. Pinamonti & C. Clementi, *JCTC* 14, 3849 (2018)

Ubiquitin as a model system

- C_α-C_β model
- Only native contacts included (219 parameters)
- ^3J ($H_{C\alpha}$-H_N) spin-couplings as observables (63 values)
- Δf_i = 0.25 Hz



I. Zaporozhets & C. Clementi, unpublished (2019)

Optimization

- Model likelihood:

$$Q = \prod_{i=1}^{N} \boldsymbol{N}(e_i(\epsilon), f_i, \Delta \mathrm{f}_i)$$

e_i calculated observable value

f_i experimental observable value

Δf_i standard deviation

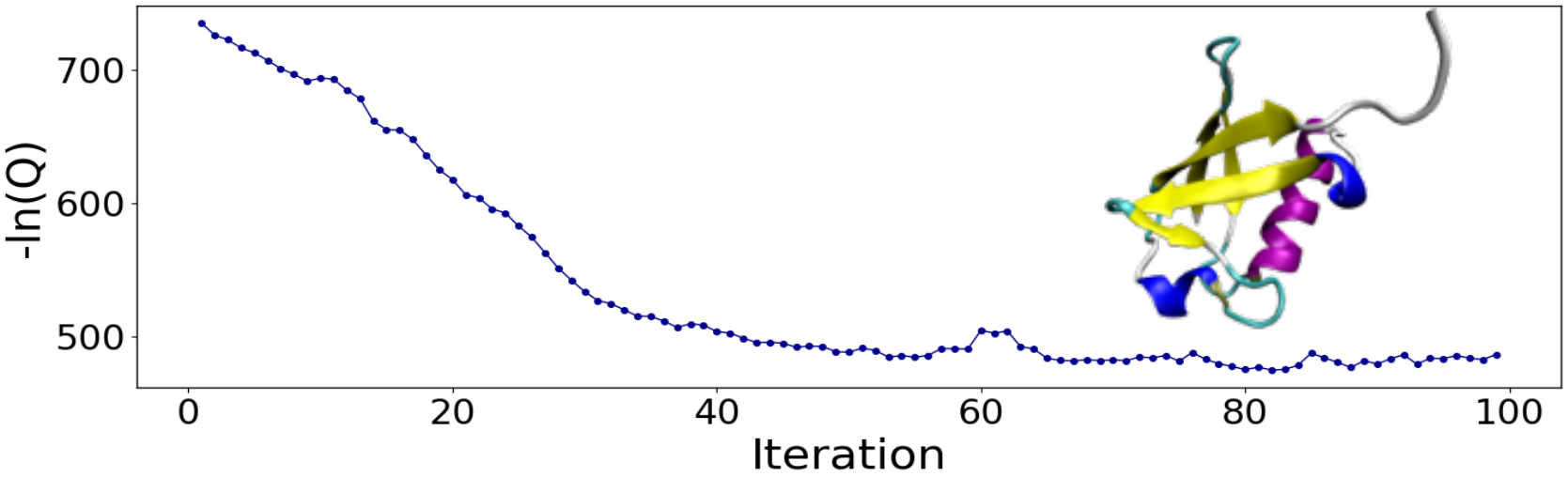
$\{\epsilon\}$ model parameters (interaction strength)

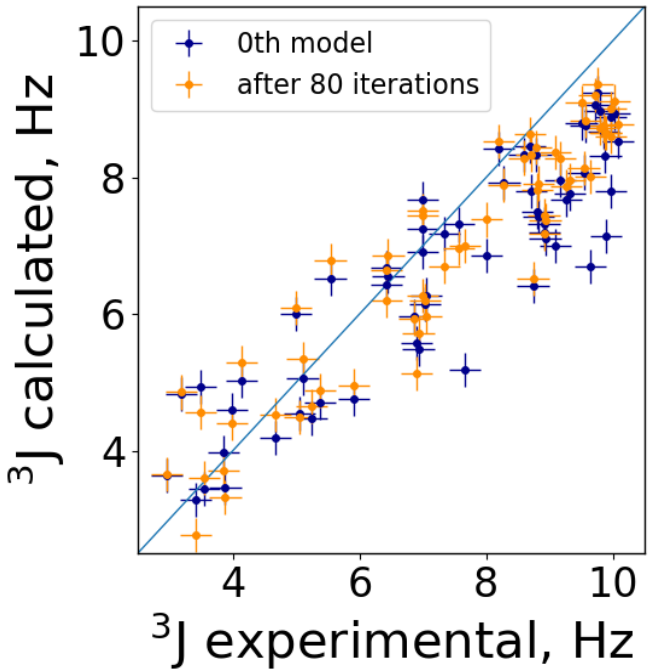
- Model loss-function:

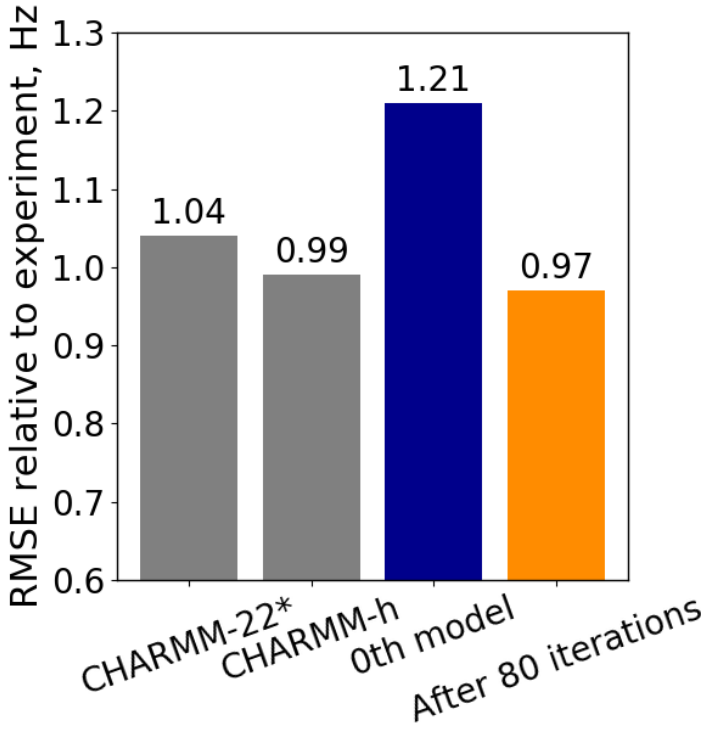
$$L^{(k+1)} = -lnQ + \alpha \sum_{j=1}^{M} (\epsilon_j^{(k+1)} - \epsilon_j^{(k)})^2$$

I. Zaporozhets & C. Clementi, unpublished (2019)

Results: Ubiquitin with NMR data



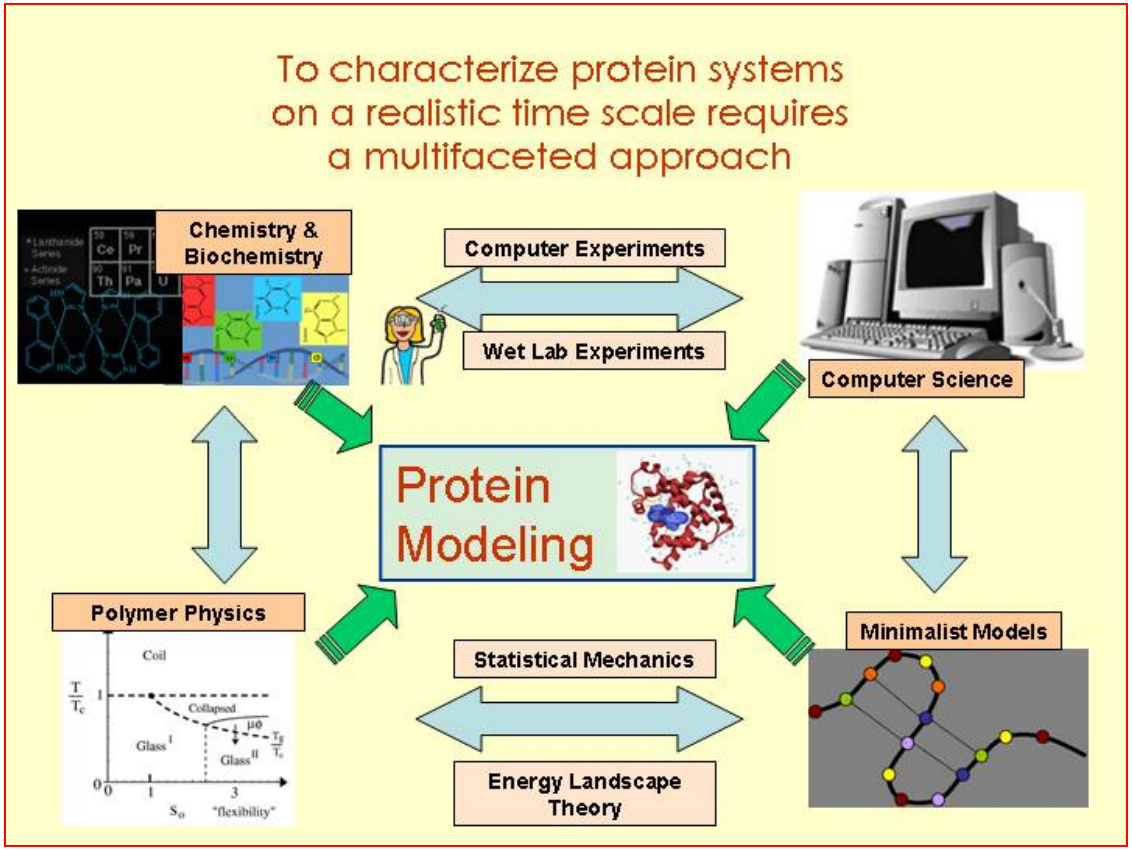




I. Zaporozhets & C. Clementi, unpublished (2019)

Cecilia Clementi's research group

http://clementiresearch.rice.edu



Clementi's group

Dr. Feliks Nüske
Dr. Jiang Wang
Dr. Giovanni Pinamonti
Dr. Fabio Trovato
Eugen Hruska
Wangfei Yang
Nick Charron
Iryna Zaporozhets

previous:

Alex Kluber
Dr. Justin Chen
Dr. Lorenzo Boninsegna (UCLA)
Dr. Fernando Yrazu (Rice)
Dr. Jordane Preto (U Alberta)
Dr. Mary Rohrdanz (MD Anderson)
Dr. Wenwei Zheng (U Arizona)
Dr. Amarda Shehu (GMU)
Dr. Payel Das (IBM)
Dr. Silvina Matysiak (U. Maryland)
Dr. Brad Lambeth (Shell)

$$ NSF

$$ Welch Foundation

$$ Einstein Foundation