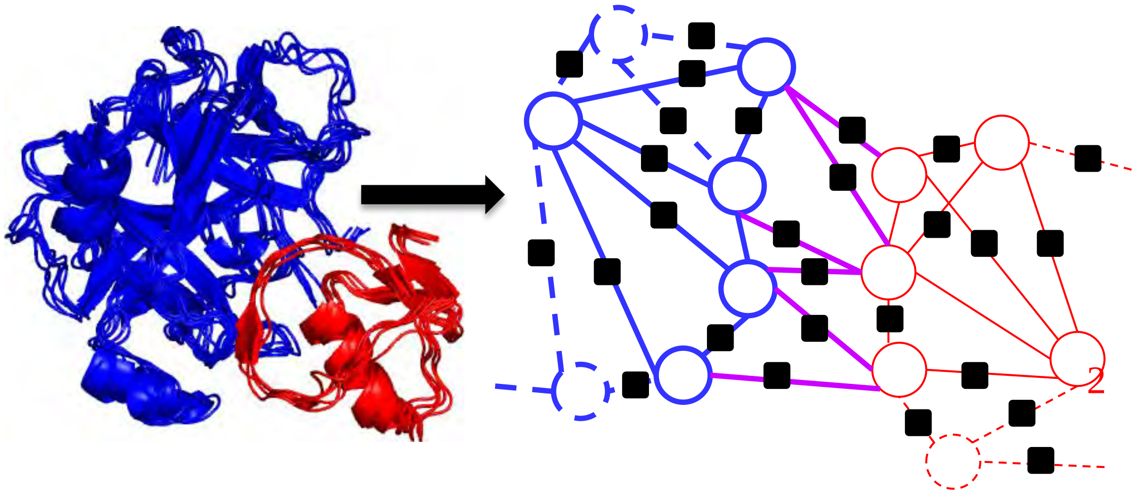
Probabilistic Graphical Model Based Analysis and Modeling of Ensembles of Conformers



Christopher James Langmead

School of Computer Science

Carnegie Mellon University

Carnegie Mellon
School of Computer Science

Context within the BTRR

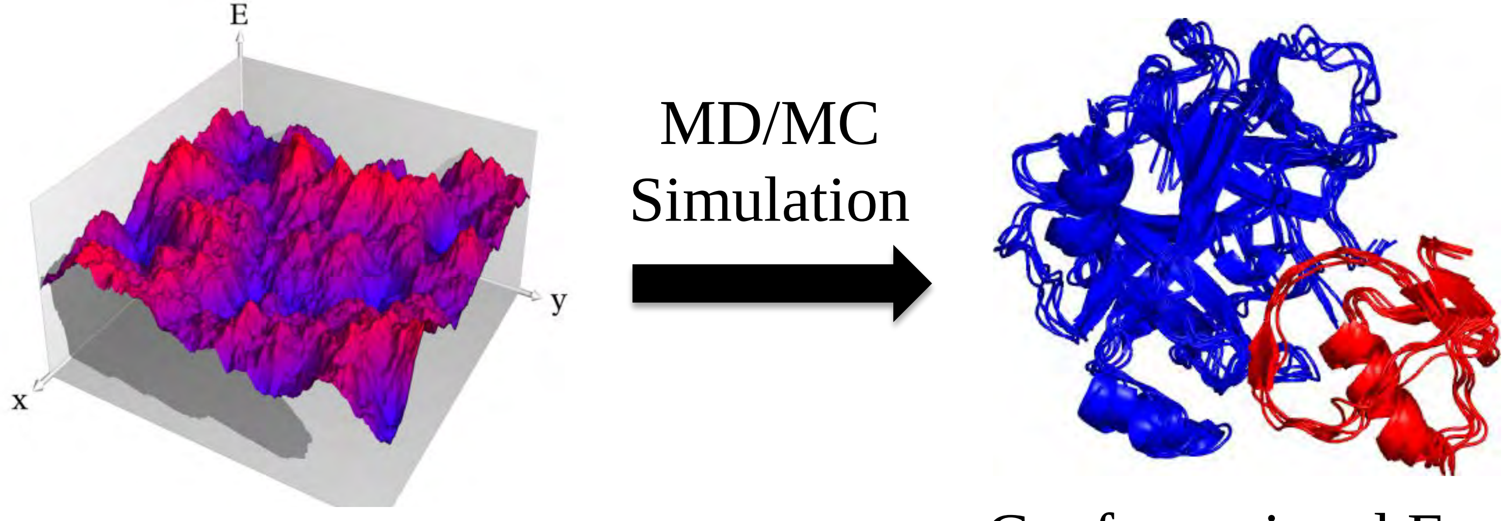
- TR&D1: Molecular Modeling
  - Specific Aims 1 – 3
  - Today's presentation is most relevant to
    - Subaim 1.4: "PGM-based analysis and modeling of ensembles of conformers"
    - Subaim 2.2: "Binding geometry and affinity computations for protein-protein and protein-ligand interactions using novel methods based on PGMs and/or mixed-resolution models with LBMC"

Context within the BTRR

- Relevant C&SPs & DBP
  - C&SP3; DBP1
- Our methods are mostly scale and data agnostic, and so they can also be used for TR&Ds 2 and 3
  - Analysis of trajectories
  - Generative Models
  - Parameter Estimation

Conformational Ensembles

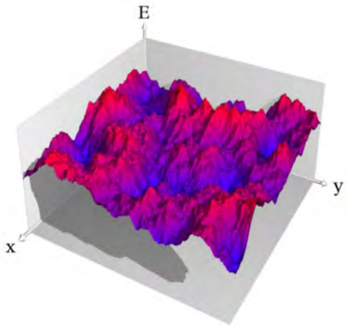
- Molecular Dynamics and Monte Carlo Simulation trajectories consist of molecular conformations sampled from an energy landscape



Energy Landscape

Conformational Ensemble

Motivation



- Conformational Ensembles contain important information relevant to function
- Unfortunately, extracting information from large ensembles (i.e., Big Data) can be challenging
- Our goals are to:
  - Learn generative models from ensembles
  - Use those models to analyze, simulate, and (re)engineer molecular motions

From Conformational Ensembles to Generative Models

- Each conformation corresponds to a point in a high-dimensional space; i.e., $x \in \mathbb{R}^n$
  - One dimension for each degree of freedom
- Examples
  - Internal degrees of freedom
  - Cartesian coordinates
  - Atomic fluctuations from a mean conformation
  - Inter-atomic distance matrices

From Conformational Ensembles to Generative Models

- Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a set of random variables corresponding to the degrees of freedom for some system
- A generative model is an encoding of $P(\mathbf{X})$
  - i.e., an encoding of the joint distribution
  - Thus, the ensemble is a sample from $P(\mathbf{X})$

From Conformational Ensembles to Generative Models

- Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a set of random variables corresponding to the degrees of freedom for some system
- A generative model is an encoding of $P(\mathbf{X})$
  - i.e., an encoding of the joint distribution
- Question: how can we compactly represent $P(\mathbf{X})$?
- Answer: Probabilistic Graphical Models (PGM)

Probabilistic Graphical Models

- A PGM, (G, Φ), is a factored encoding of a joint probability distribution $P(\mathbf{X})$ over a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, in terms of a graph $G = (V,E)$ and a set of non-negative functions $\Phi = \{\phi_1, \ldots, \phi_m\}$

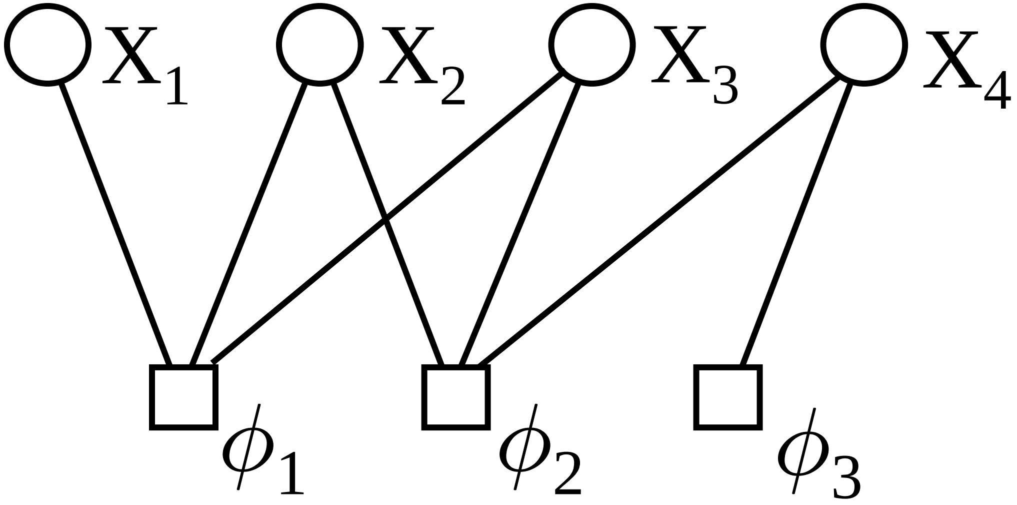
The Graphical Model Zoo

- Bayes Nets
- Hidden Markov Models
- Kalman Filters
- Dynamic Bayesian Networks

- Ising Model
- Potts Model
- Markov Random Fields
- Factor Graphs

Etc

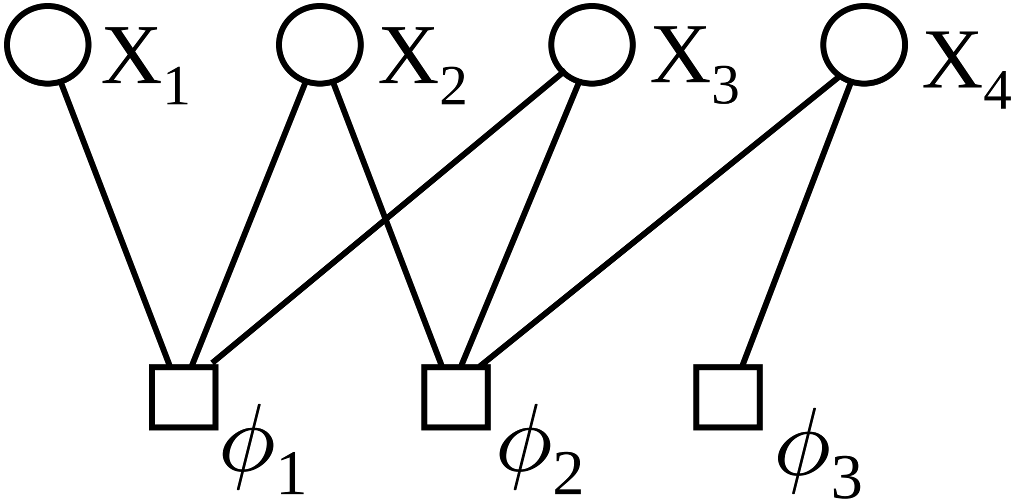
Factor Graphs



Circles correspond to random variables

Squares correspond to factors (functions) over the variables

Factor Graphs



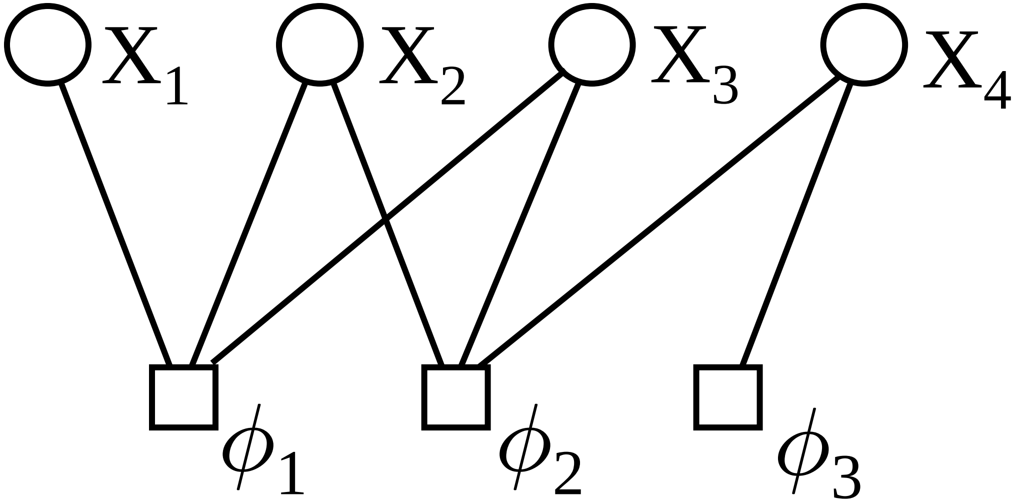
If each ϕ_i is a positive function …

Theorem (Hammersely and Clifford)

$$\boxed{P(x) = \frac{1}{Z}\prod_{a\in\Phi}\phi_a(x_a)}$$

$$\boxed{Z = \sum_x \prod_{a\in\Phi}\phi_a(x_a)}$$

Example

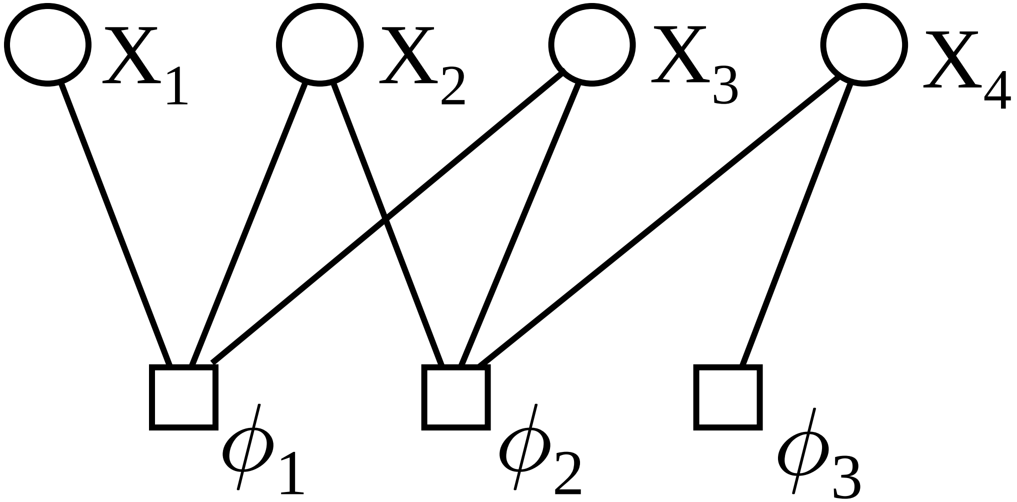


$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_1(x_1, x_2, x_3)\phi_2(x_2, x_3, x_4)\phi_3(x_4)$$

Conditional Independencies

- The topology of the graph defines a set of conditional independencies (CI)
  - Variables A and B are conditionally independent, given C (denoted $A \perp B \mid C$) iff $P(A,B|C) = P(A|C)P(B|C)$ or, equivalently, $P(A|B,C) = P(A|C)$
- Informally, CIs let us use ‘simpler’ functions to encode the joint distribution

Example



$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \phi_1(x_1, x_2, x_3) \phi_2(x_2, x_3, x_4) \phi_3(x_4)$$

In this graph:

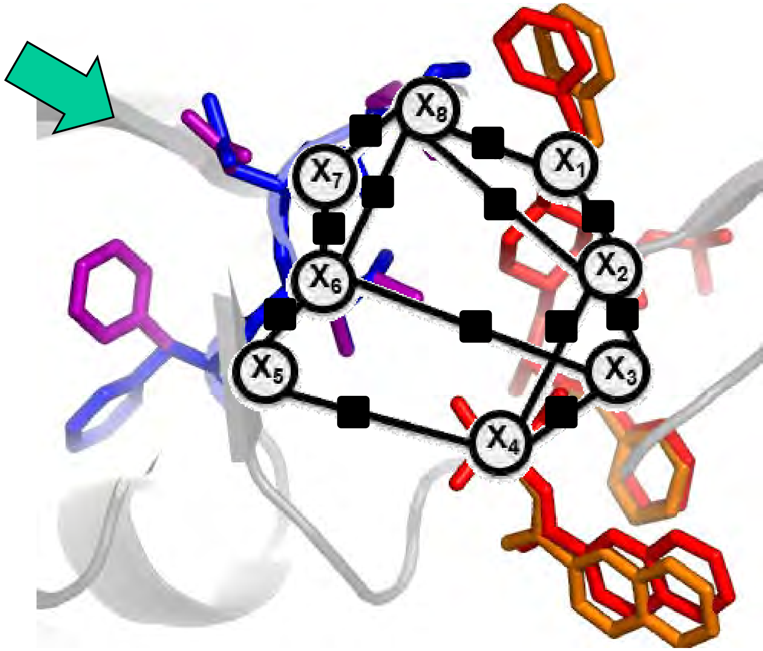
$X_1 \perp X_4 | X_2 , X_3$

Key Point

- *Any* joint probability distribution over N variables can be represented via a suitably defined factor graph
- User must specify (or learn from data):
  1. Topology of the graph
  2. Functional form and parameters of the factors

PGMs of Molecular Structures

- The user gets to decide which degrees of freedom they wish to model
  - Internal degrees of freedom
  - Cartesian coordinates
  - Atomic fluctuations
  - Inter-atomic distances



Learning PGMs from Ensembles

GAMELAN (GrAphical Models of Energy LANdscapes)

- Input
  - Ensemble encoded as an $n \times t$ matrix, D
    - n is the number of covariates $\mathbf{X} = \{X_1, \ldots, X_n\}$
    - t is the number of conformations in the ensemble
- Output : PGM (G, Φ) over $\mathbf{X}$ that "fits" D
  - Algorithmic subtasks:
    1. Learn topology of the graphical mode, $G = (V,E)$
    2. Learn model parameters (i.e., Φ), given G

Learning PGMs from Ensembles

GAMELAN (GrAphical Models of Energy LANdscapes)

- Optimization problem

$$(G,\mathit{\Phi})^* = \mathrm{argmax}_{G,\mathit{\Phi}}\ f(G, \mathit{\Phi}; D) = \sum_t \log P_{G,\mathit{\Phi}}(d_t) - \lambda R(G, \mathit{\Phi})$$

Learning PGMs from Ensembles

GAMELAN (GrAphical Models of Energy LANdscapes)

- Optimization problem

$$(G,\Phi)^* = \mathrm{argmax}_{G,\Phi}\ f(G, \Phi; D) = \underbrace{\textstyle\sum_t \log P_{G,\Phi}(d_t)}_{\text{1st term reflects the PGM's fit to the data}} - \lambda R(G, \Phi)$$

Learning PGMs from Ensembles

GAMELAN (GrAphical Models of Energy LANdscapes)

- Optimization problem

$$(G,\Phi)^* = \mathrm{argmax}_{G,\Phi}\ f(G,\Phi; D) = \sum_t \log P_{G,\Phi}(d_t) - \underbrace{\lambda R(G,\Phi)}$$

2nd term penalizes complex PGMs by counting the number of edges (and thus parameters)

Learning PGMs from Ensembles

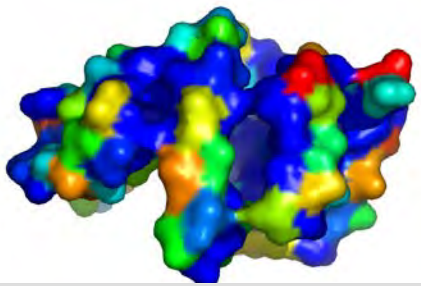
GAMELAN (GrAphical Models of Energy LANdscapes)

- Algorithms for solving optimization problem
  - Discrete Random Variables: BKLCL11
  - Continuous Random Variables
    - Angular Data (von Mises distribution): RKL11
    - Unimodal distributions: RKL12
    - Multi-modal Distributions: RL12; L14
  - Time-varying models: RMKL10; L14

*P41 acknowledged

Using PGMs of Molecular Structures

- Given a PGM, there are algorithms for:
  - Computing (approximate) free energies
    - KXL07; KL08; KBL09; KXL11; KGLB14
  - Visualizing entropic contributions to the free energy
    - KXL11
  - Sampling new configurations
    - RKL11



Heatmap of Configurational Entropy for Lysozyme

*P41 acknowledged

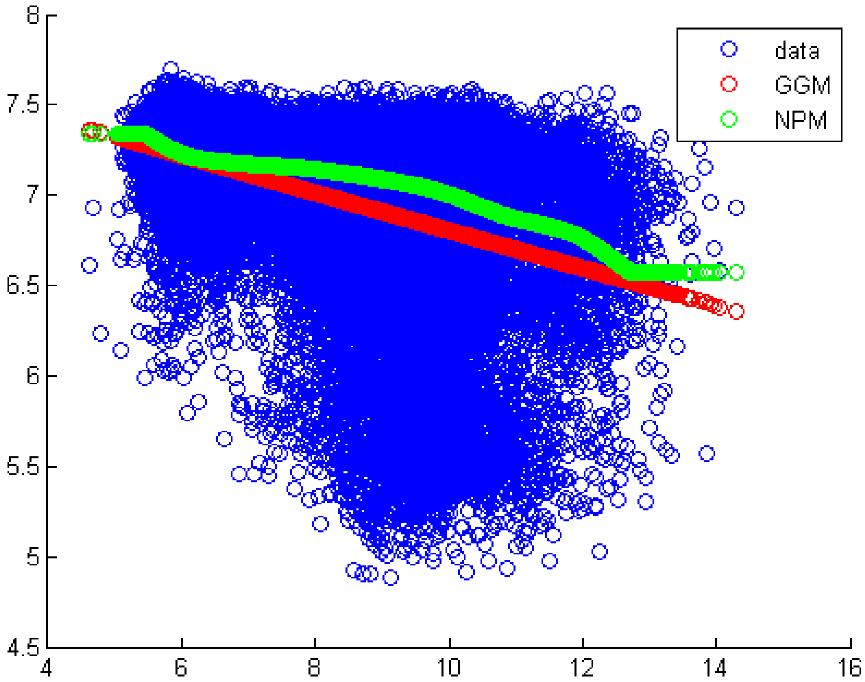
Using PGMs of Molecular Structures

- Given a PGM, there are algorithms for:
  - (re)Designing Proteins
    - KGBL09; KGLB14
  - Predicting how the distribution changes under perturbations [i.e., P(**X** | Y)]
    - Examples: allosteric regulation; effects of mutations
    - KXL11; RKL12; L14

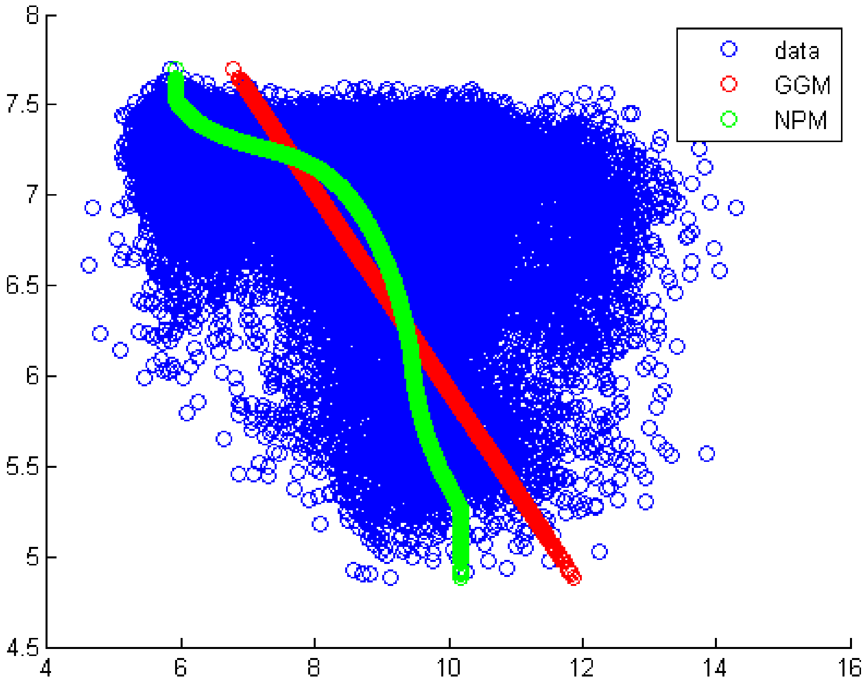
*P41 acknowledged

Example: Inference in GGM vs NPN [L14]

- Data: 50 μs simulation of the engrailed homeodomain
- Conditioned model on one variable, computed MLE of remaining variable



argmax $_y$ P(y|x)



argmax $_x$ P(x|y)

Ongoing Work

- Distribution GAMELAN Software
  - Custom-version for Anton Trajectories
- PGM-based Markov-State Models
- Writing manuscripts for semi- and non parametric models
- Rory and Dan are integrating Dan’s high-resolution rotamer libraries into our framework

Potential Applications to other areas of the BTRR

- Analyzing MCELL/BNG trajectories
- Alternative algorithms for learning PGMs from image data
- Parameter estimation
  - Specifically, learning PGMs over model parameters

Thank you!

- Students & Post Docs
  - – Dr. Hetunandan Kamisetty
  - – Dr. Narges Sharif Razavian
  - – Subhodeep Moitra
- Collaborators
  - – Dr. Chris Bailey-Kellogg
  - – Dr. Jaime Carbonell

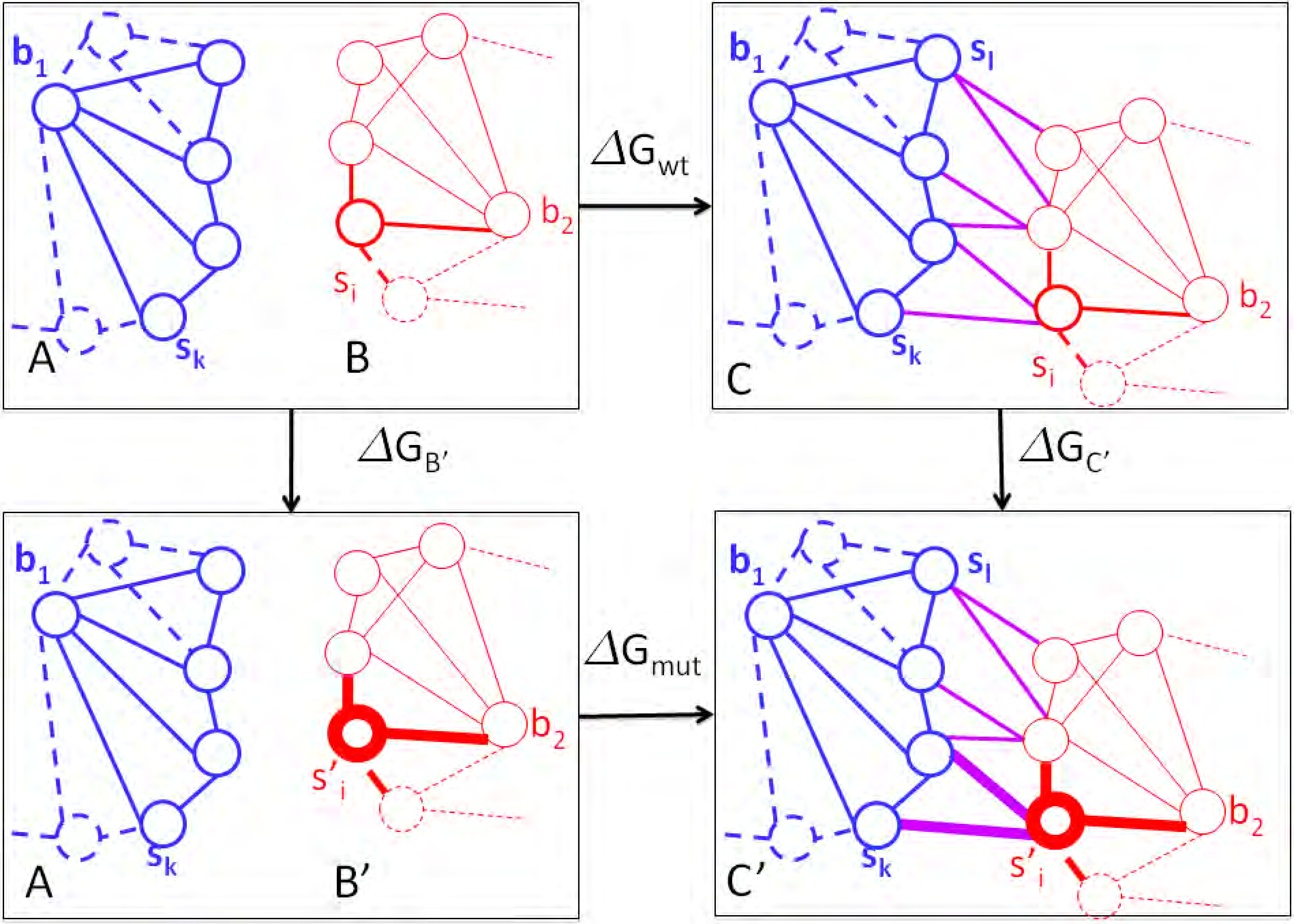
References

- [L14] C.J. Langmead “*Generative Models of Conformational Dynamics*“ in “Protein Conformational Dynamics” Springer International Publishing, Eds. Han, Keli; Zhang, Xin; Yang, Mingjun, 2014, pp. 87-105
- [KGLB14] H. Kamisetty, B. Ghosh, C.J. Langmead, C. Bailey-Kellogg “*Learning Sequence Determinants of Protein:Protein Interaction Specificity with Sparse Graphical Models*”. RECOMB pp129-143, 2014

Timeline

	Timeline	Year1		Year2		Year3		Year4		Y5
Aim 1	1. Inclusion of lipid bilayer into network models	🟥	🟥	🟥	🟦	🟦	🟩	🟩	🟧	🟧
	2. Dev of a hybrid methodology that integrates ENMs, MD & MC	🟥	🟥	🟦	🟩	🟩	🟧	🟧	🟧	🟧
	3. Analysis suite for WE sim & application to neurosignaling proteins	🟥	🟦	🟦	🟩	🟧	🟧	🟧	🟧	🟧
	4. Critical assessment and sampling quality	🟦	🟦	🟩	🟧	🟧	🟧	🟧	🟧	🟧
	5. PGM-based analysis modeling of ensembles of conformers	🟥	🟥	🟦	🟩	🟧	🟧	🟧	🟧	🟧
	6. Combined use of ENM-, WE- and PGM-based methods				🟥	🟥	🟦	🟩	🟧	🟧
Aim 2	1. Improving QC methods in hybrid QC/MM	🟦	🟦	🟩	🟧	🟧	🟧	🟧	🟧	🟧
	2. Affinity calculations using mixed resolution models with LBMC			🟥	🟥	🟦	🟩	🟩	🟧	🟧
	3. PGM-based binding affinity calculations					🟥	🟦	🟩	🟧	🟧
	4. Combining PGMs with statistical mechanical libraries					🟥	🟥	🟦	🟩	🟧
	5. Elucidating allosteric signaling mechanisms & multimerization effects	🟥	🟦	🟩	🟧	🟧	🟧	🟧	🟧	🟧
Aim 3	1. Information transfer across scales - scale integration		🟥	🟥	🟦	🟦	🟦	🟩	🟩	🟧
	2. Application of WE methods to accelerate MCell simulations			🟥	🟥	🟦	🟦	🟩	🟩	🟧
	3. Analysis of MCell trajectories using PGMs and PCA-based methods		🟥	🟥	🟦	🟦	🟩	🟩	🟧	🟧
	4. Software optimization and parallelization		🟦	🟦	🟩	🟩	🟧	🟧	🟧	🟧
	5. PGM-based software and API				🟥	🟥	🟦	🟩	🟩	🟧
	6. Development of interfaces for easy access and interoperation					🟥	🟥	🟦	🟩	🟧
	7. Alternative strategies:ENMs & resolution exchange applied to MCell						🟥	🟥	🟦	🟩

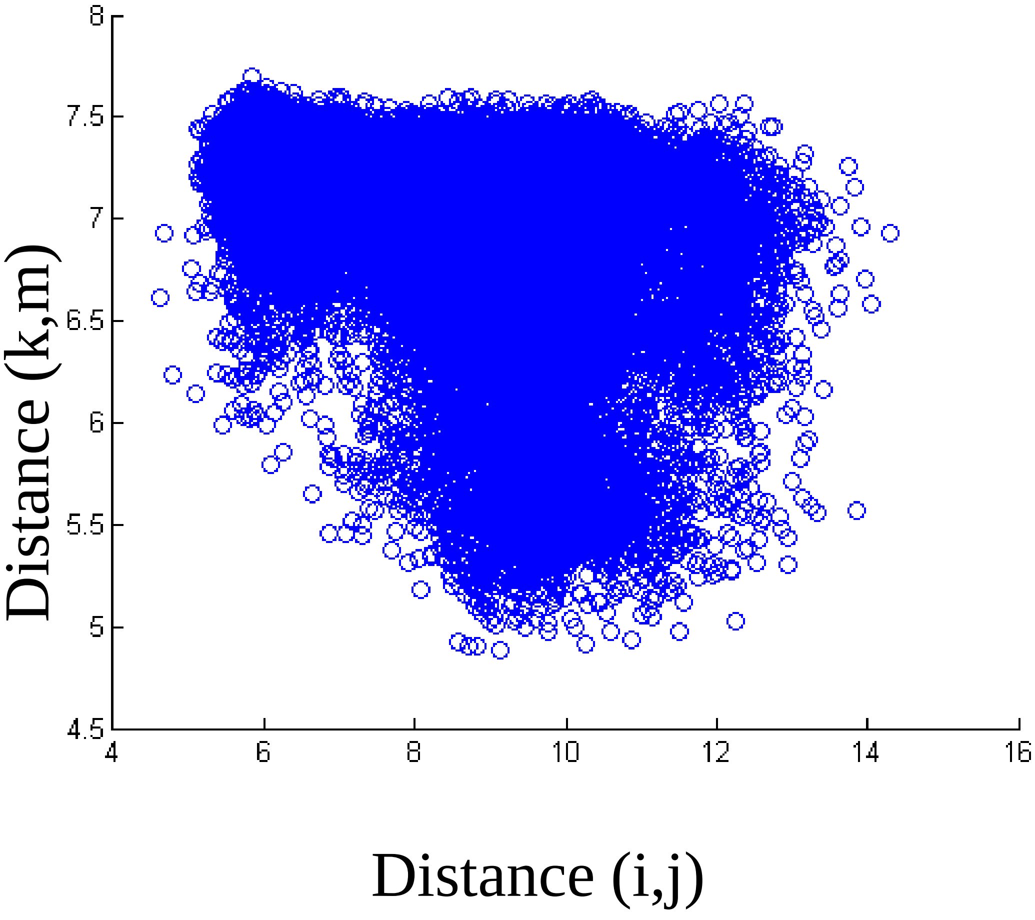
🟥 Design	🟦 Implementation and improvements	🟩 Alpha testing	🟧 User evaluation and refinements (beta testing)



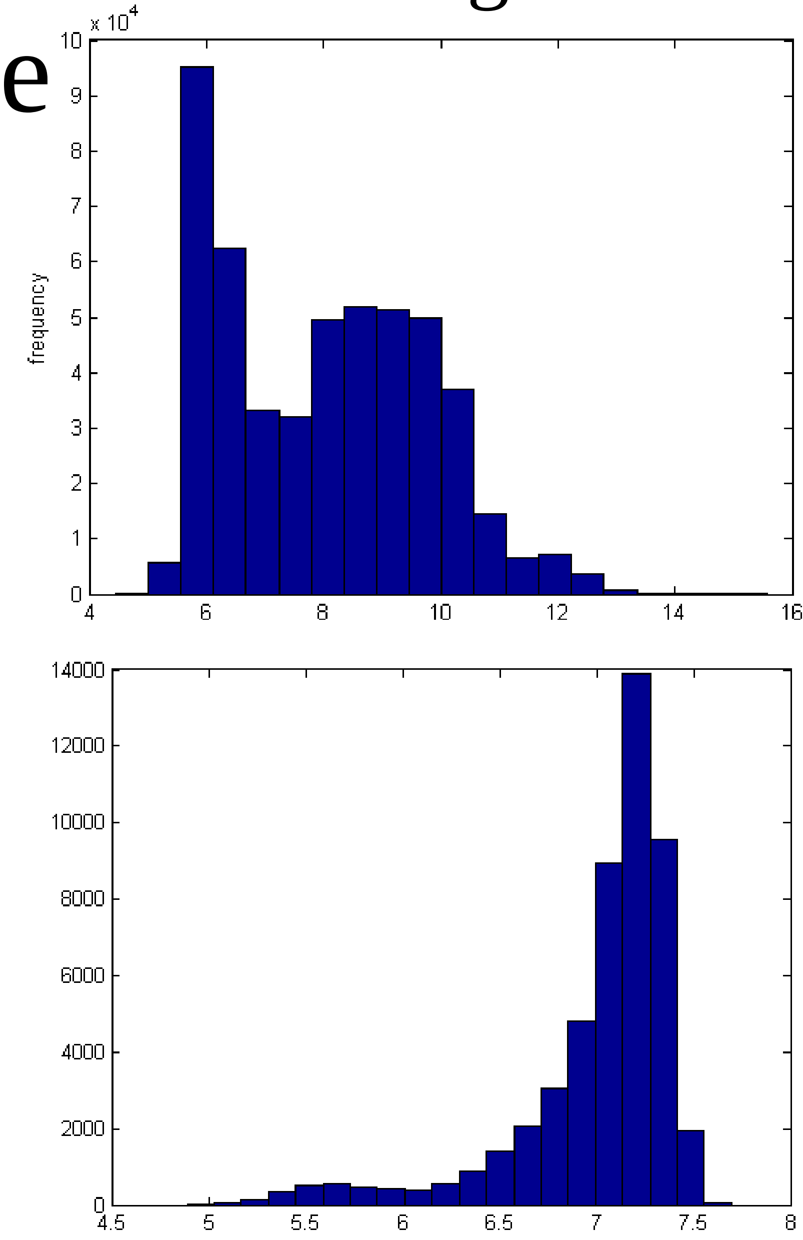
b1
A
sk
B
si
b2
ΔGwt
C
sl
ΔGB′
ΔGC′
ΔGmut
B′
s′i
C′

Example

Joint Distribution



Marginals


x 10^4
frequency