Argonne NATIONAL LABORATORY

REFINING LOW-RESOLUTION CRYO-EM STRUCTURES WITH BAYESIAN INFERENCE DRIVEN INTEGRATION OF MULTISCALE SIMULATIONS

AARON HAGSTROM[1], HENG MA[1], DEBSINDHU BHOWMIK[1], CHRISTOPHER B. STANLEY[1], JULIE C. MITCHELL[2], HUGH O'NEILL[3], ARVIND RAMANATHAN[1,4]

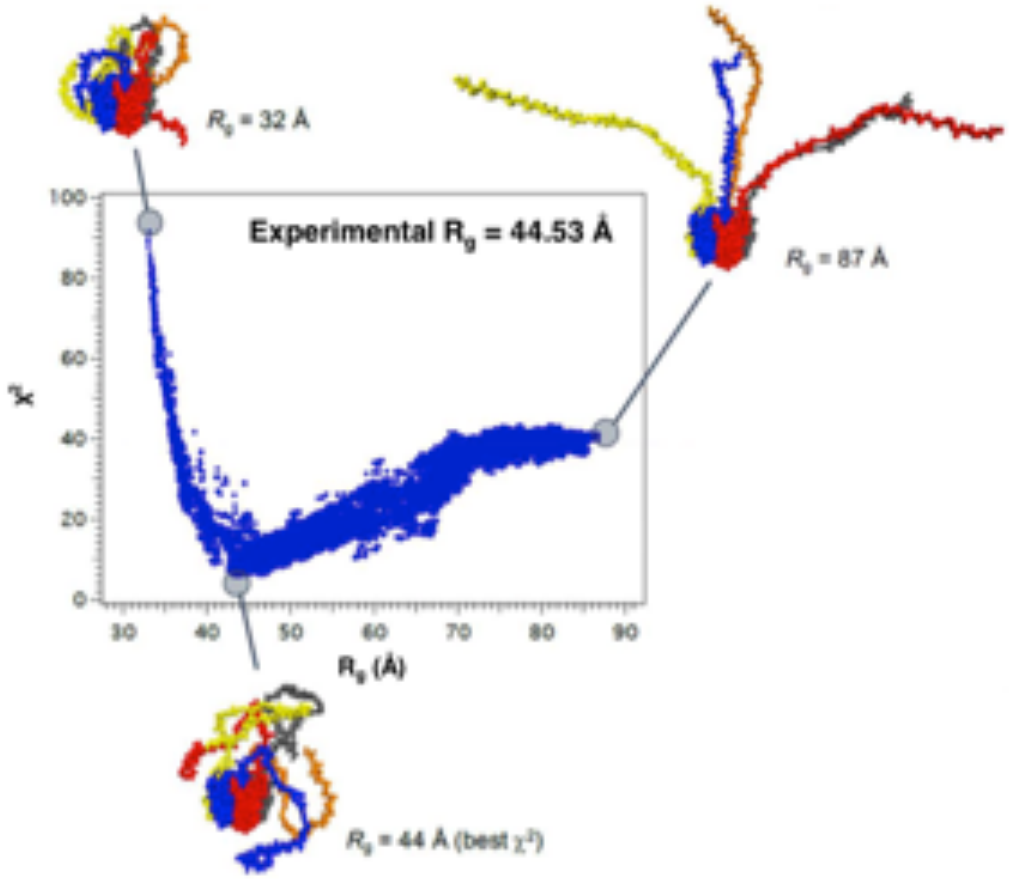
[1]Computational Sciences and Engineering Division, [2]Biosciences Division, [3]Neutron Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[4]Data Science & Learning Division, Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL, USA.
http://ramanathanlab.org

U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Comp-bio-med 2019

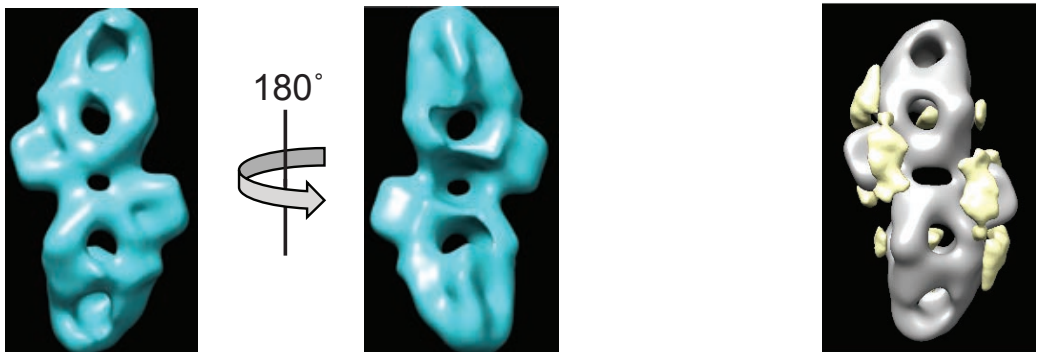
MOTIVATION (1): CRYO-EM – AN ADVANCED TOOL FOR STRUCTURAL BIOLOGY

- Larger bio-assemblies are easier to probe with cryo-EM
  - Atomic resolution structures for large proteins from frozen hydrated samples
  - Fills a gap for structures that are difficult/recalcitrant to crystallography and NMR (esp. membrane proteins)
- Rapid progress enabled by digital electron detector technology, new algorithms for image analysis
- Flexible regions in proteins are often challenging to characterize



Merk et al. Cell 165, 1698–1707, June 16, 2016
Nogales, Nature Methods, 2016, 13, 24
Mitrea, Kriwacki et al. Nat. Chem. Biol. (2019)

MOTIVATION (2): LOW RESOLUTION CRYO-EM CONSEQUENCE OF NOT USING ALL AVAILABLE DATA?

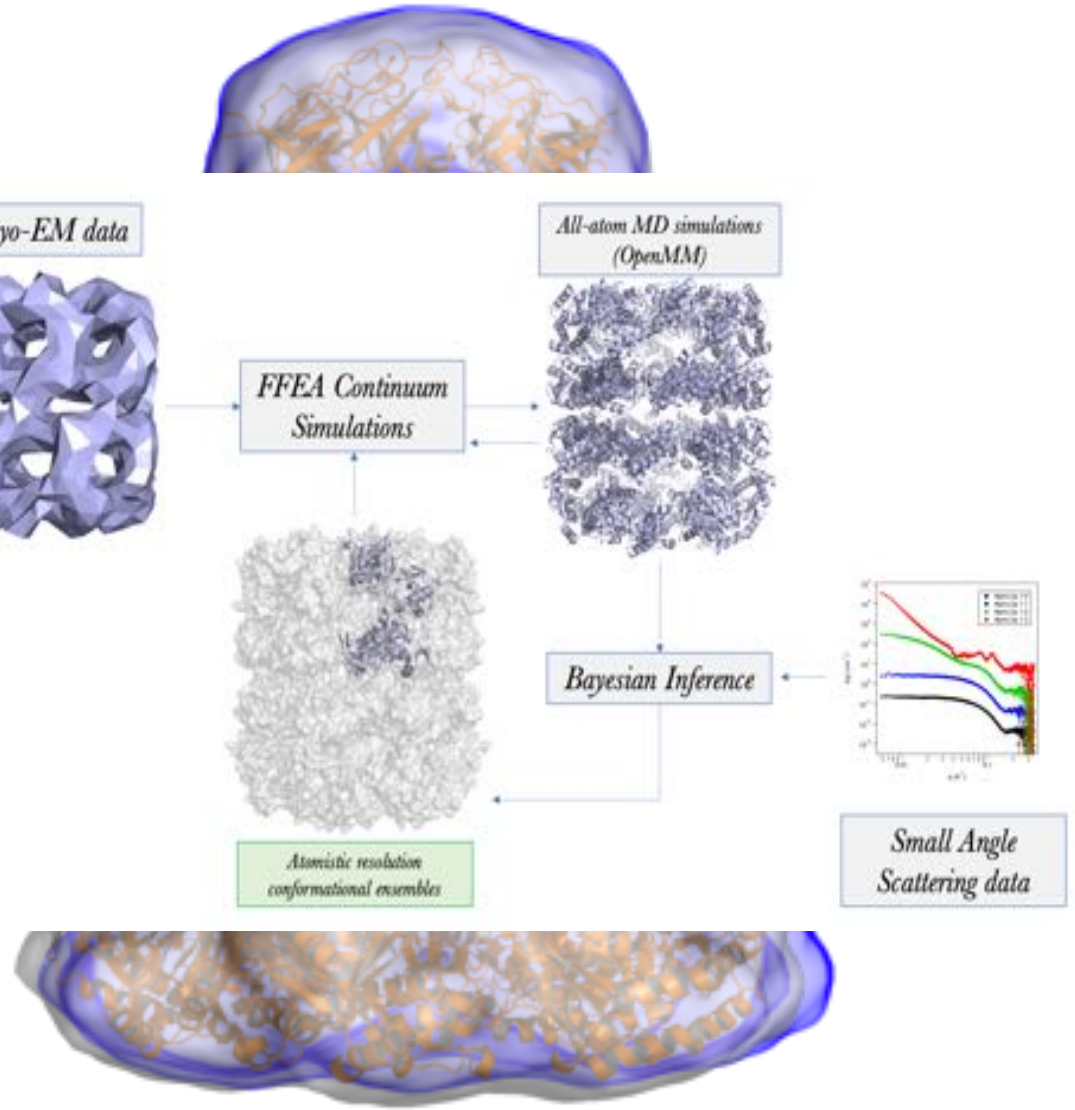


M. Sherekar, et al, J. Biol Chem (2019; in review)

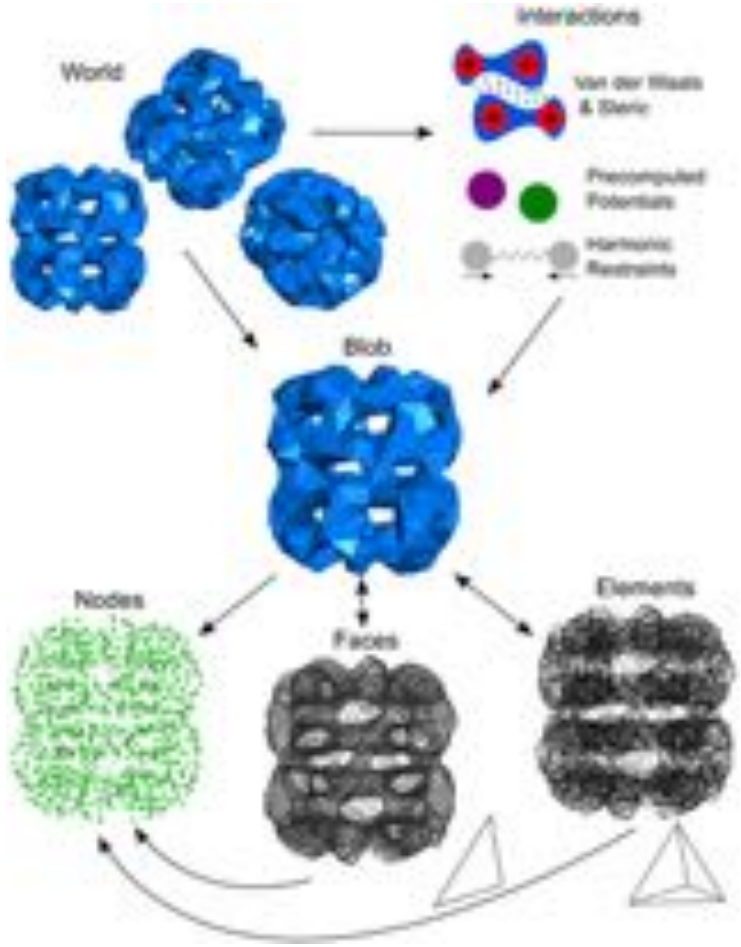
- From collected data, only a partial set is used for cryo-EM reconstruction:
  - 20-40% of data can be “ignored” for lack of clarity in images
- Is this usable data?
  - image processing ideas have been able to use both model based and model free techniques to reconstruct
- Lacking key tools for “building in” resolution:
  - Generative models for cryo-EM data from volumetric representations

OUTLINE

- Building a generative model for volumetric data from cryo-EM:
  - Fluctuating finite element analysis (FFEA)
  - Generating 2D projections from FFEA simulations
- Automatically convert between all-atom and volumetric representations:
  - Machine learning determination of high flexibility regions
- Bayesian fits to 2D projections – and evaluate with cryoEM data



FLUCTUATING FINITE ELEMENT ANALYSIS (FFEA)



A. Solberneu, et al PLoS Computational Biology (2018)

- FFEA simulations represent proteins as visco-elastic continuum solids
  - a 6-12 LJ potential with repulsive potential that is proportional to the overlapping volume
  - specific interactions can be defined w/ precomputed potentials

Density

External force

$$\rho \frac{D\vec{u}}{Dt} = \nabla.(\vec{\sigma}^{v} + \vec{\sigma}^{e} + \vec{\pi}) + \vec{f}$$

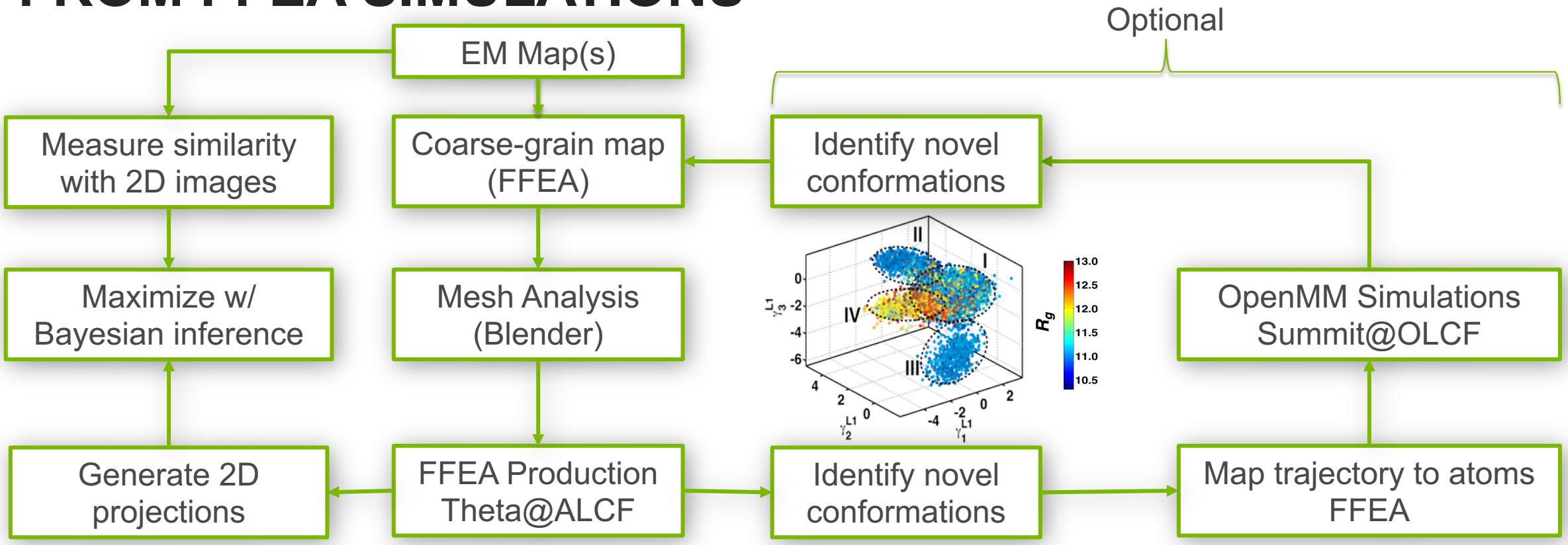
Velocity

Viscous stress

Elastic stress

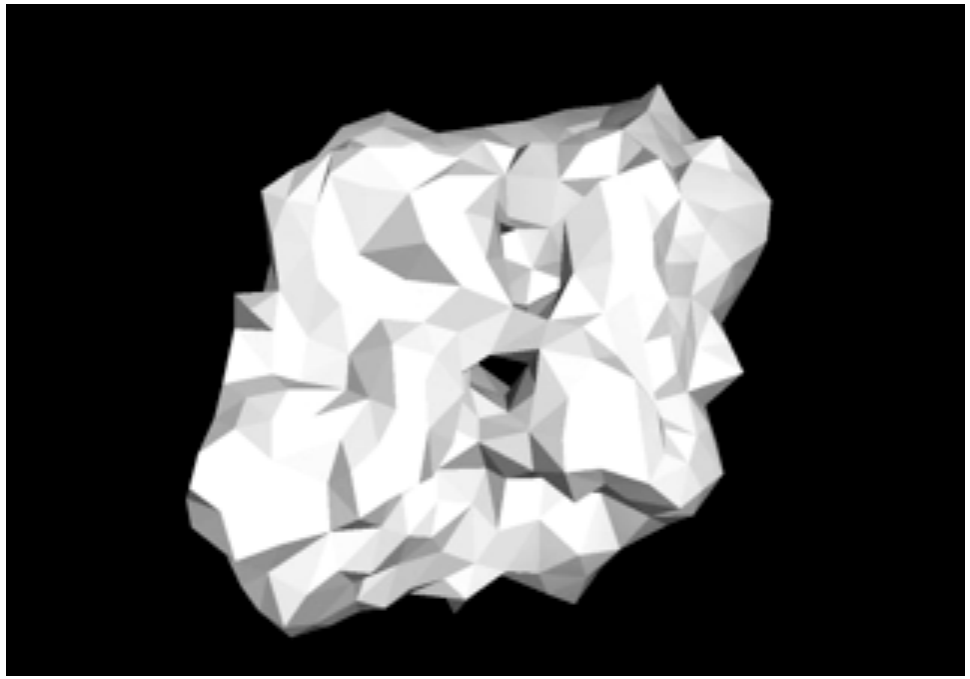
Thermal noise

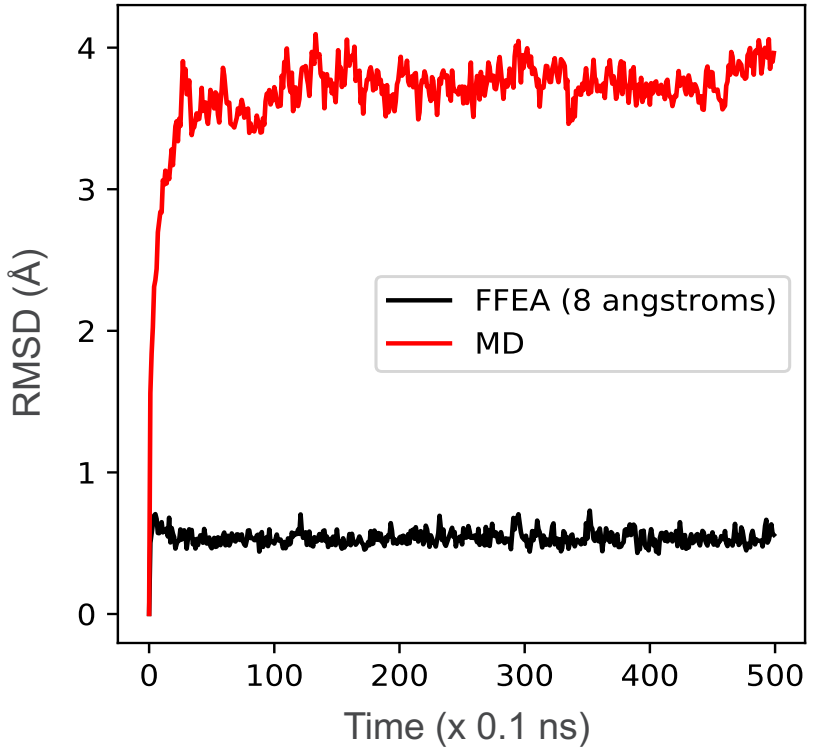
WORKFLOW FOR GENERATING 2D PROJECTIONS FROM FFEA SIMULATIONS



- A reconstructed EM model with improved resolution
- Subset of conformations (from MD) → mapped to EM data

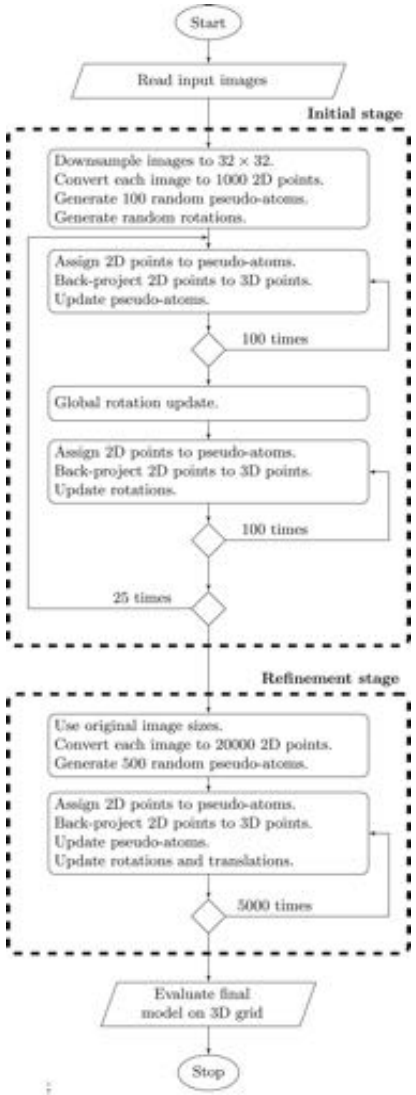
FFEA GENERATED CONFORMATIONS ARE STABLE AT SHORTER TIMESCALES

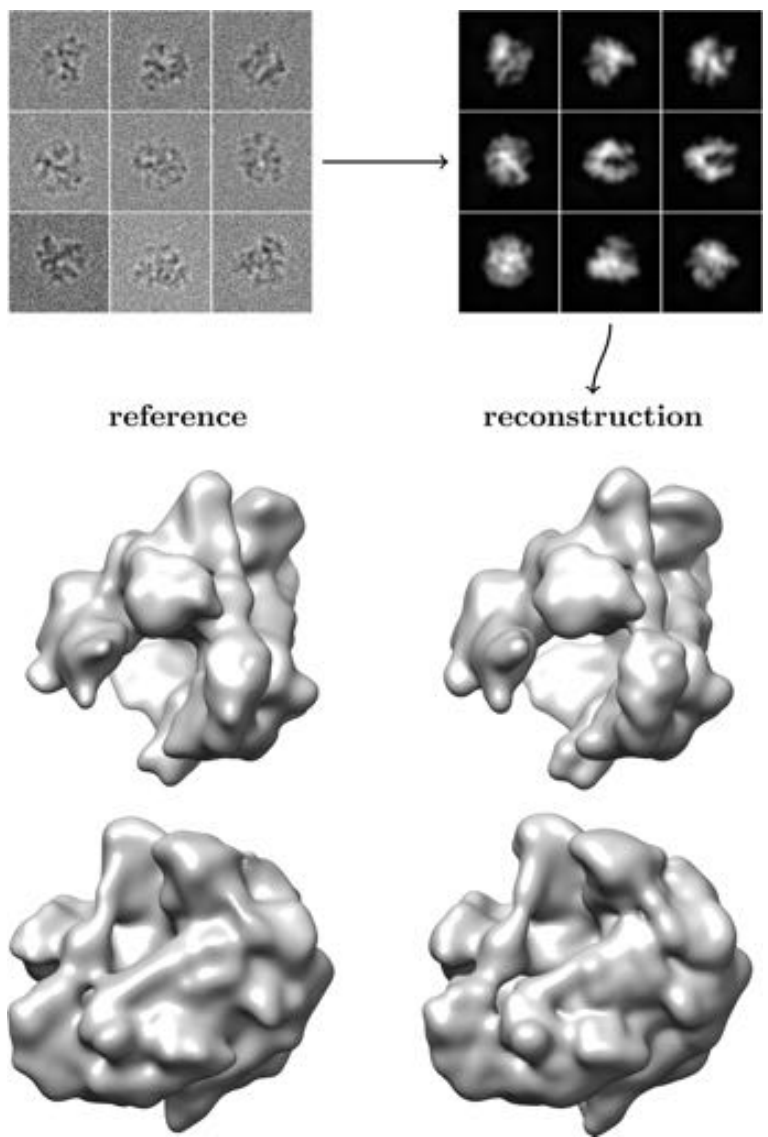




- FFEA Simulations of beta-galactosidase cryo-EM map derived at 8 Å resolution
  - Additional resolutions also give similar results (although with higher RMSD)
- All-atom MD simulations also show stable RMSD (shorter timescales)

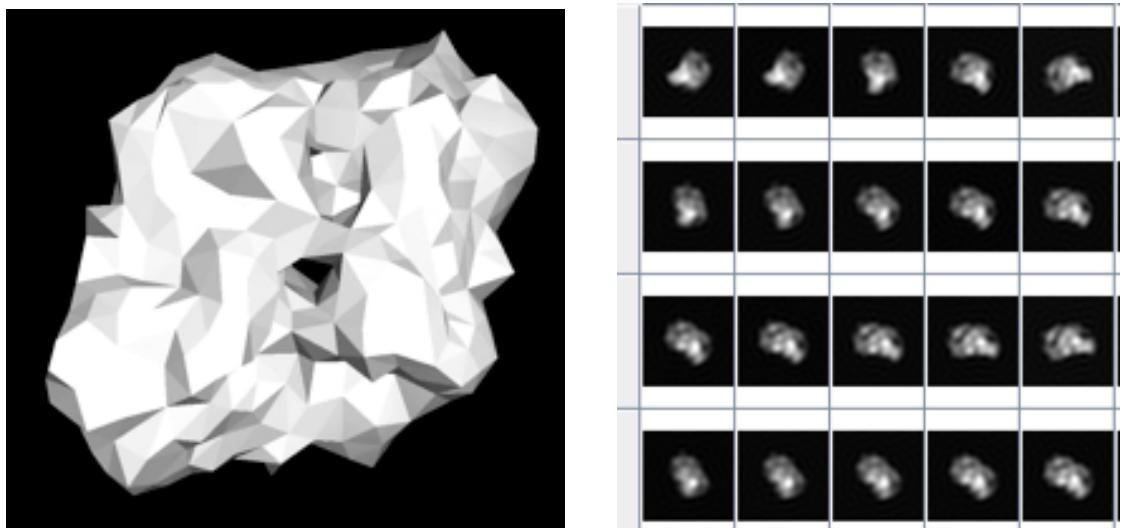
BAYESIAN INFERENCE (PSEUDO-ATOM METHOD)





- Joubert and Habeck used 3D Gaussian mixture models (GMMs) to a 2D mixture model:
  - evaluate this two-dimensional mixture model on a two-dimensional grid
  - Two stages: initial coarse-atom generation stage, refinement stage to “add” more atoms
- Our approach uses FFEA generated conformations to replace the initial and refinement stages
  - Bayesian inference used only for model evaluation on 3D grid

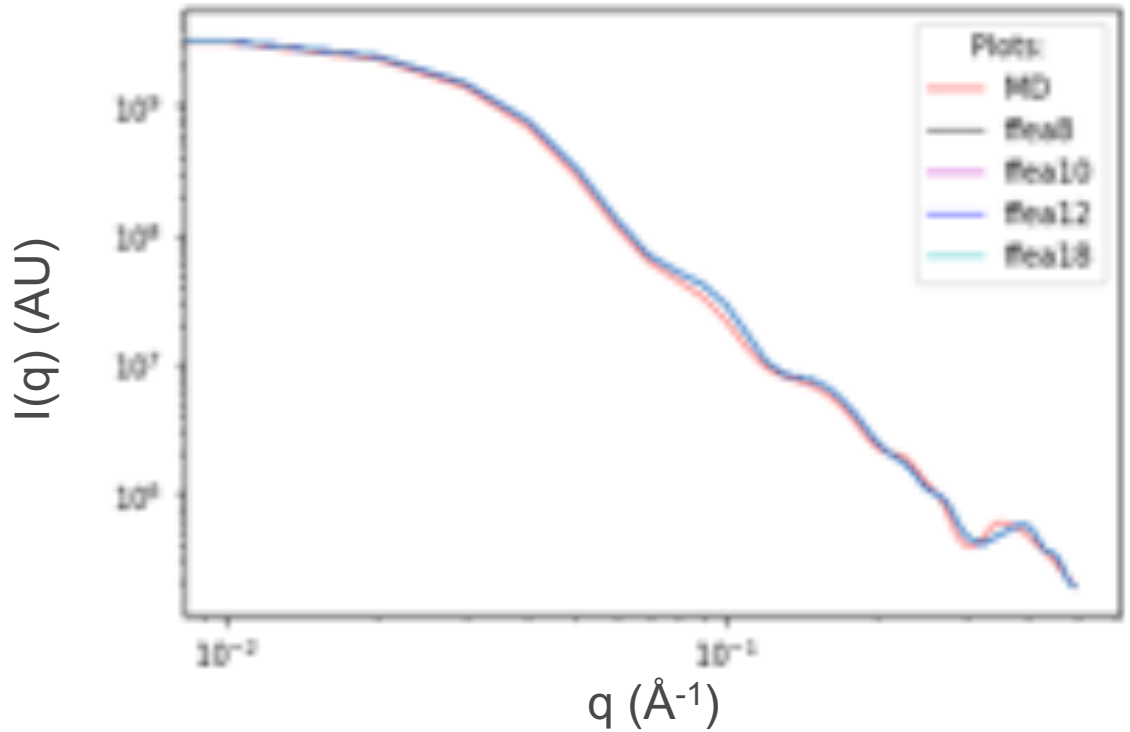
FFEA GENERATED CONFORMATIONS PROVIDE CLASS REPRESENTATIVES …

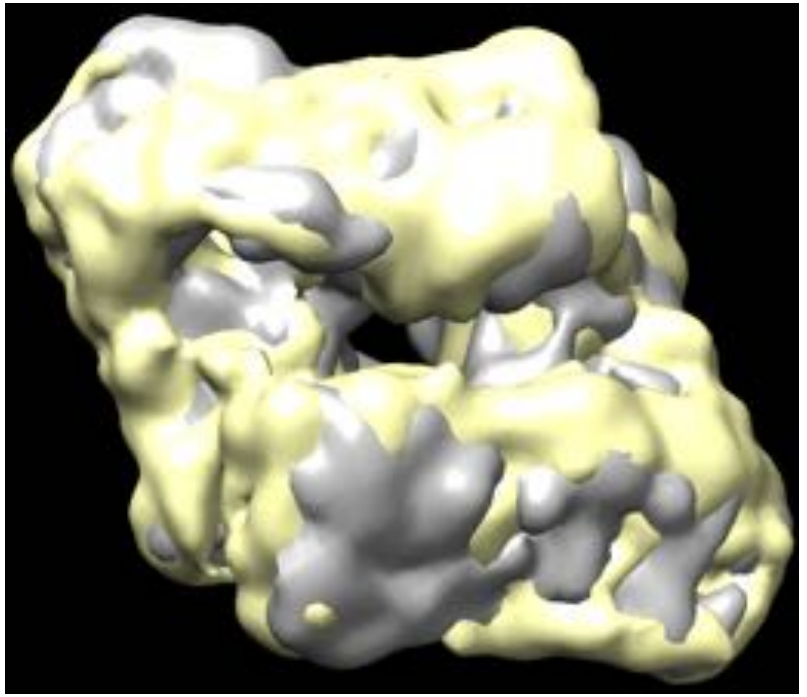


20 largest populated, 2D class averages from the FFEA simulated ensembles. Each of these represents an average from at least 100 conformations FFEA-generated 3D volumes. (Summary of largest variations are shown next.)

- FFEA simulations are used to generate 2D images with different angle settings
  - currently generated using random angles (similar to pseudo-atom method)
  - experimental settings can be used to guide the generation process
- Noise (in pixels) – representative of data acquisition set to zero (for now)
  - could be modeled as perturbations to the generated data

… CLASS REPRESENTATIVES IMPROVE AGREEMENTS BETWEEN SAXS/SANS DATA





- fitting to SAXS data, both FFEA and MD show similar profiles
- Regions with higher flexibility in protein structure are also regions where fit to experimental cryo-EM is not best (yellow)

SUMMARY

- Low resolution in Cryo-EM data is an important problem
  - FFEA coupled to MD simulations allowed us to explore states that may be represented in the imaging data, but not necessarily explained
  - Initial results seem promising – improves agreement with experimental SAXS/SANS data
- Complementary to a number of techniques:
  - will have to quantify improvement within this context
  - improvement in simulation run times
- Existing proteins with well-defined structures help in modeling:
  - Nf1 protein (collaboration with D. Esposito, A. Stephen, M. Sherekar, S. Subramanyam, et al at Frederick National Lab)

FURTHER WORK

- Improve representation for proteins in FFEA:
  - electrostatics (from APBS or similar)
  - computing solvent interactions (important for SAXS/SANS)
- Switching between all-atom and continuum representations:
  - ML techniques (including deep learning)
  - local modeling of loops and other flexible regions
- Scalability and testing of FFEA:
  - Adaptive meshing/ areas which have less data/ certainty vs. more data
  - modeling large bio-molecular assemblies with FFEA
  - docking small molecules and other interactions

ACKNOWLEDGEMENTS

PEOPLE/ TEAM

- St. Jude Children's Hospital
  - Richard Kriwacki
- Frederick National Lab
  - Andrew Stephen
  - Dom Esposito
  - Dwight Nissley
  - Natalia de Val
- Many summer interns

FUNDING

- DOE Exascale computing – Cancer Deep learning environment (CANDLE) project
- Joint design of advanced computing solutions for cancer (JDACS4C)
- Computing support:
  - Summit computing time (Director's discretionary award)
  - DOE ALCC award
  - ALCF director's discretionary award

THANK YOU!!

RAMANATHANA@ANL.GOV