Dealing with heterogeneity in single particle cryo-EM.

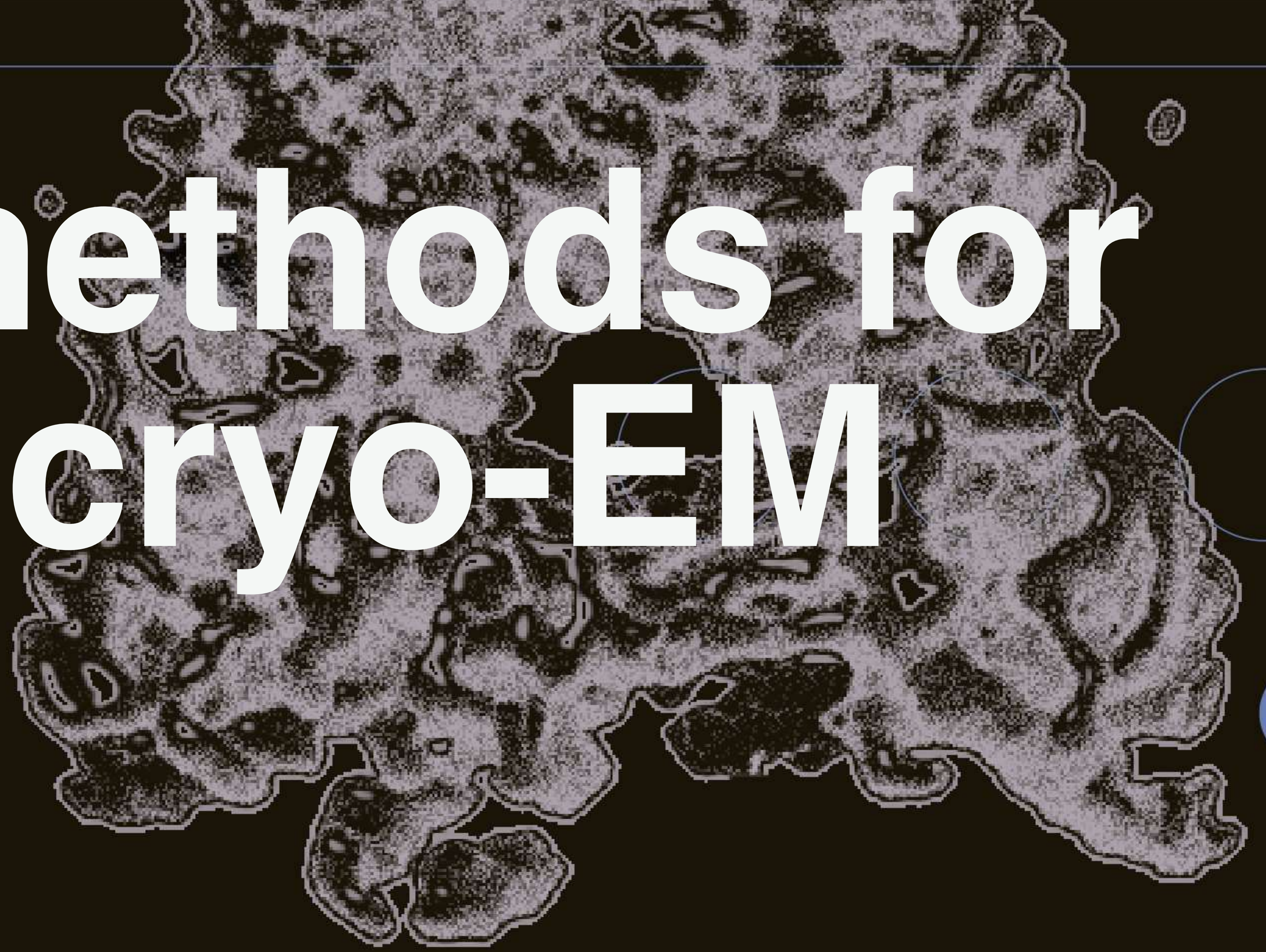
Sonya M. Hanson
Structural and Molecular Biophysics, CCB/CCM
Flatiron Institute, Simons Foundation, NYC

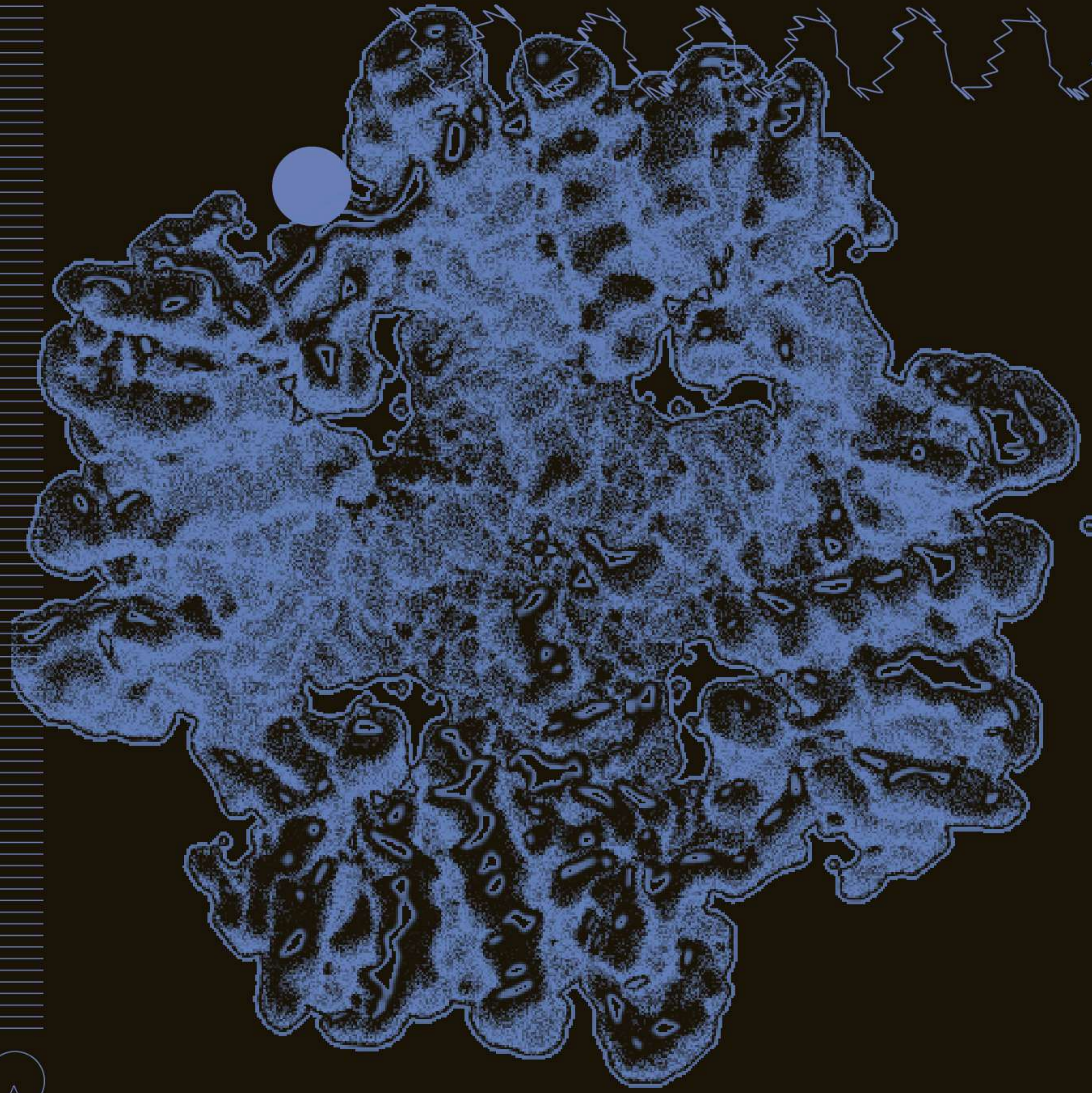
~~Tuesday, March 25, 2025~~ Monday, March 24, 2025

FLATIRON INSTITUTE

The Winter-Spring 2025 EM Course @ NYSBC

Computational methods for heterogeneity in cryo-EM





ML methods for heterogeneity in cryo-EM
- PCA
- ManifoldEM
- cryoDRGN
- GMM in eman2
- cryoSPARC/RELION
- … and more!

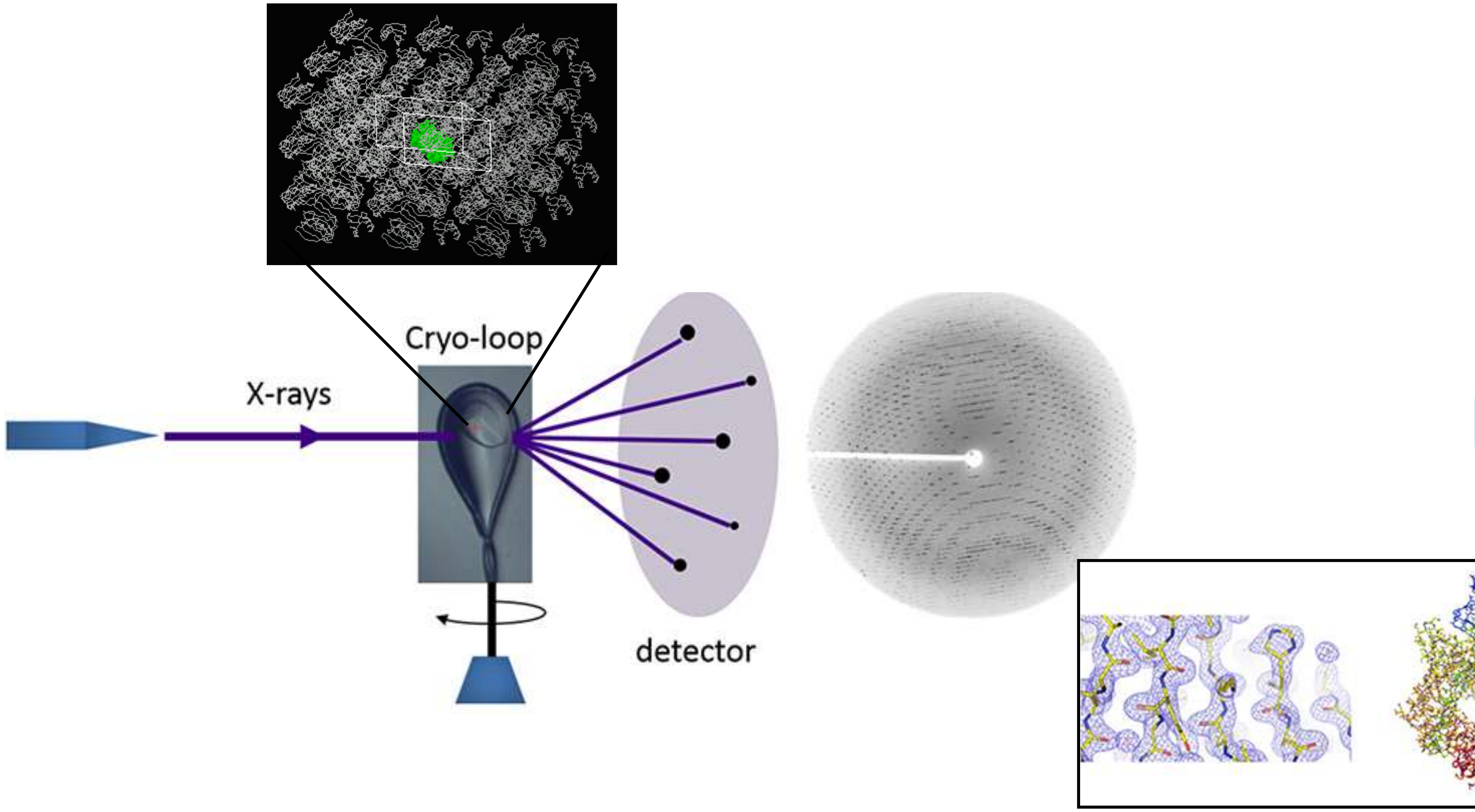
Practical considerations when using continuous heterogeneity methods.

Combining cryo-EM and Molecular Dynamics

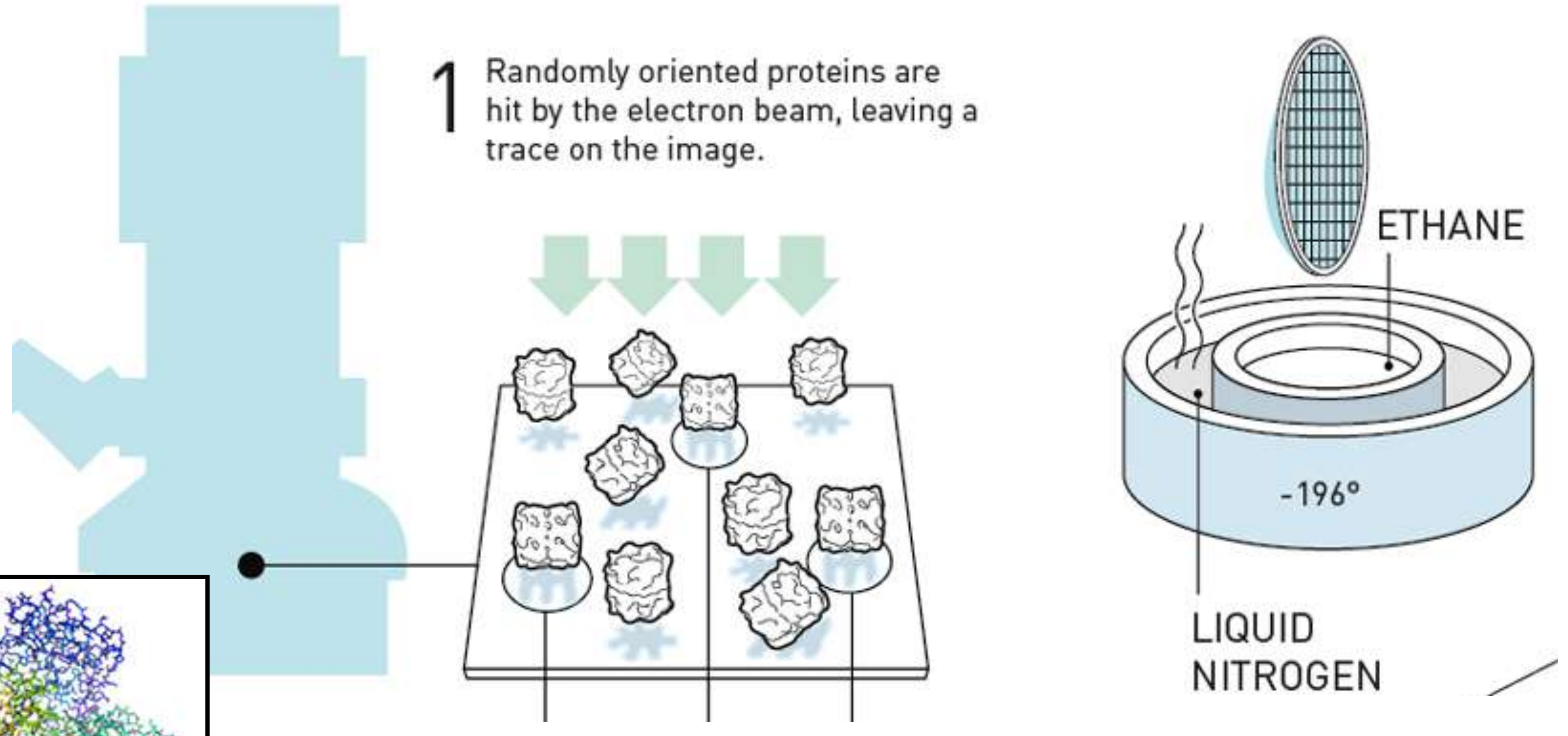
Assessing the performance of continuous heterogeneity methods.

Conformational landscapes of biomolecules are currently both experimentally and computational difficult to access.

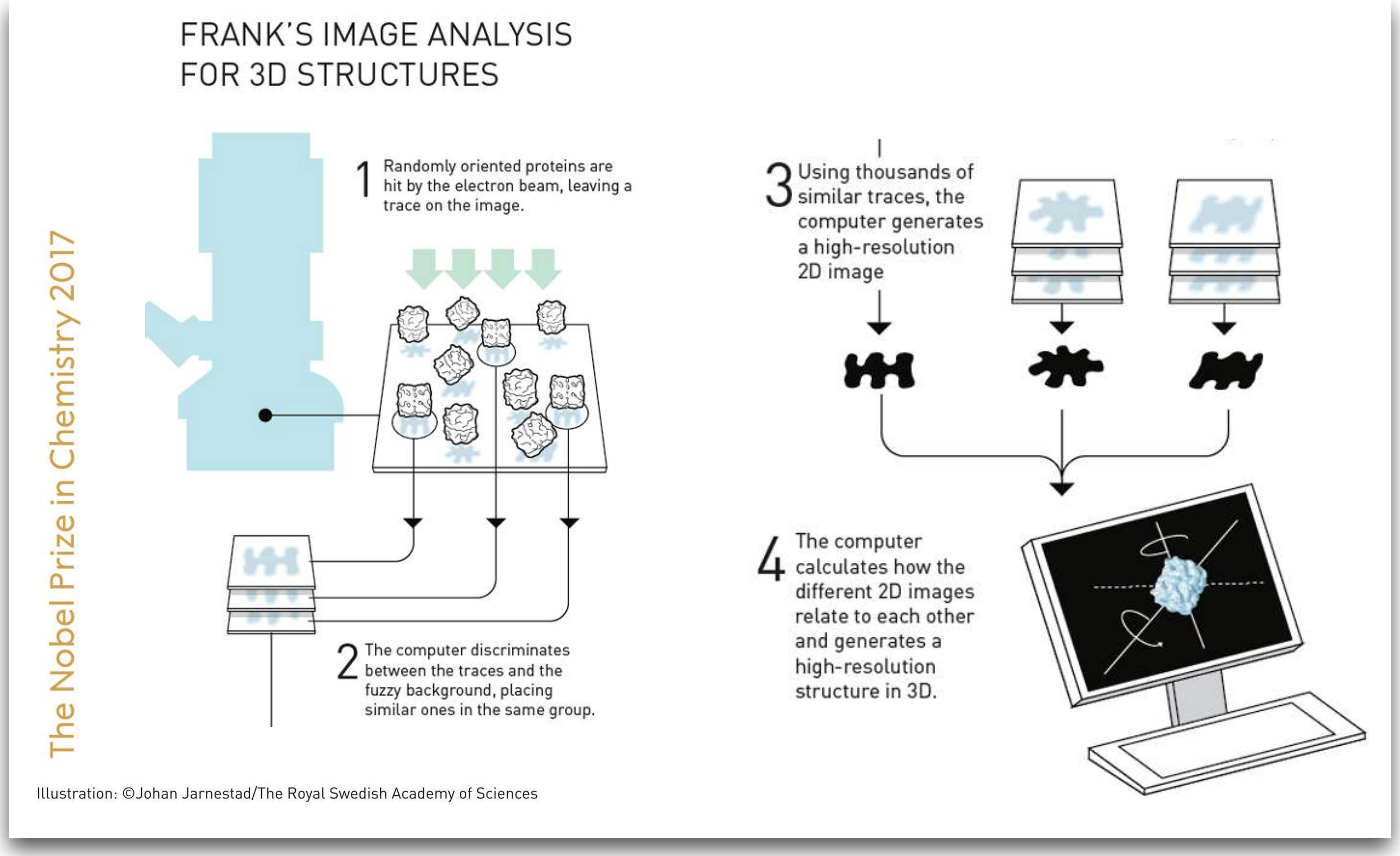
X-ray crystallography



cryo-EM


1 Randomly oriented proteins are hit by the electron beam, leaving a trace on the image.

Cryo-EM achieves atomic resolution 3D structures via 2D imaging of many differently oriented molecules.



FRANK'S IMAGE ANALYSIS
FOR 3D STRUCTURES

1 Randomly oriented proteins are hit by the electron beam, leaving a trace on the image.

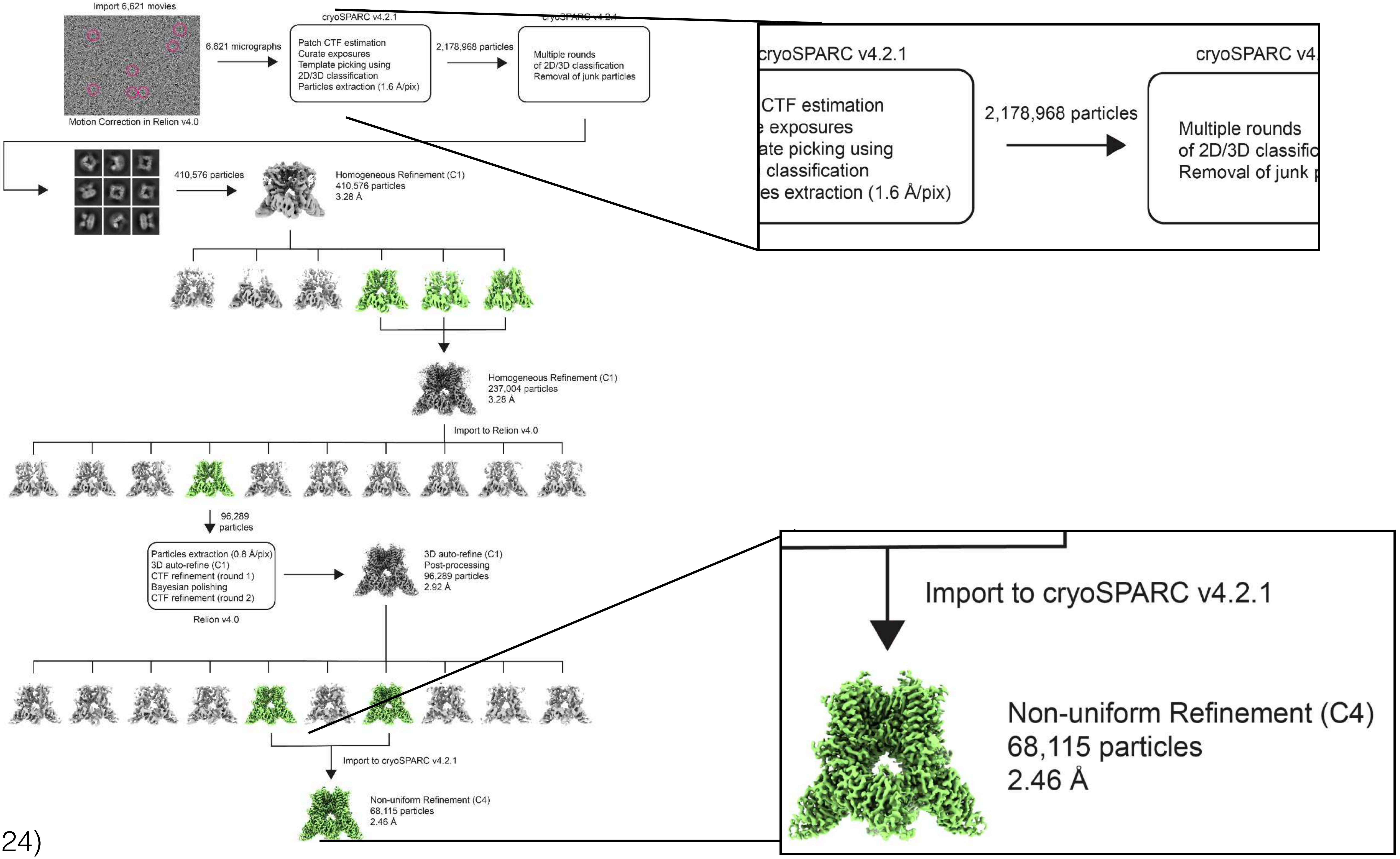
2 The computer discriminates between the traces and the fuzzy background, placing similar ones in the same group.

3 Using thousands of similar traces, the computer generates a high-resolution 2D image

4 The computer calculates how the different 2D images relate to each other and generates a high-resolution structure in 3D.

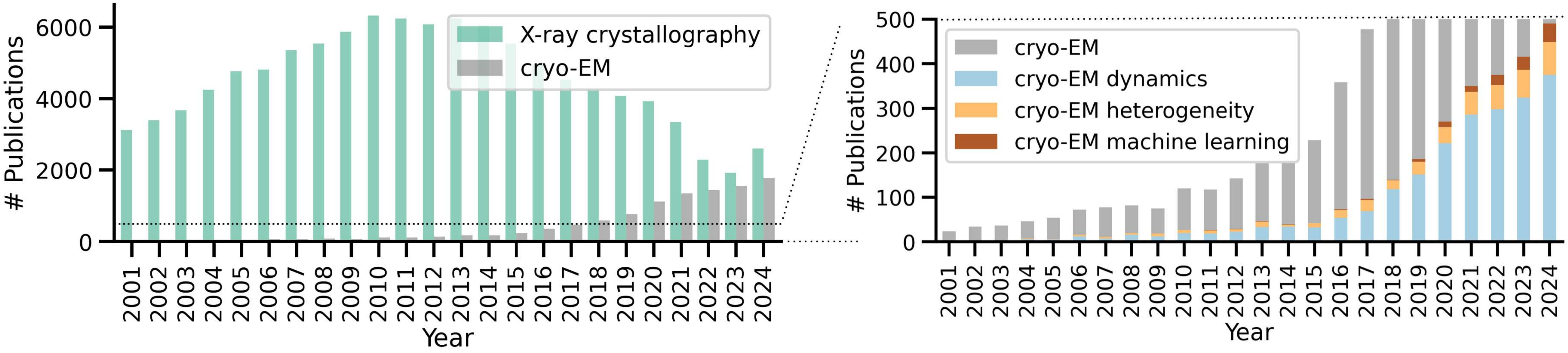
Illustration: ©Johan Jarnestad/The Royal Swedish Academy of Sciences

Discrete classes in single particle analysis restricts access to information.

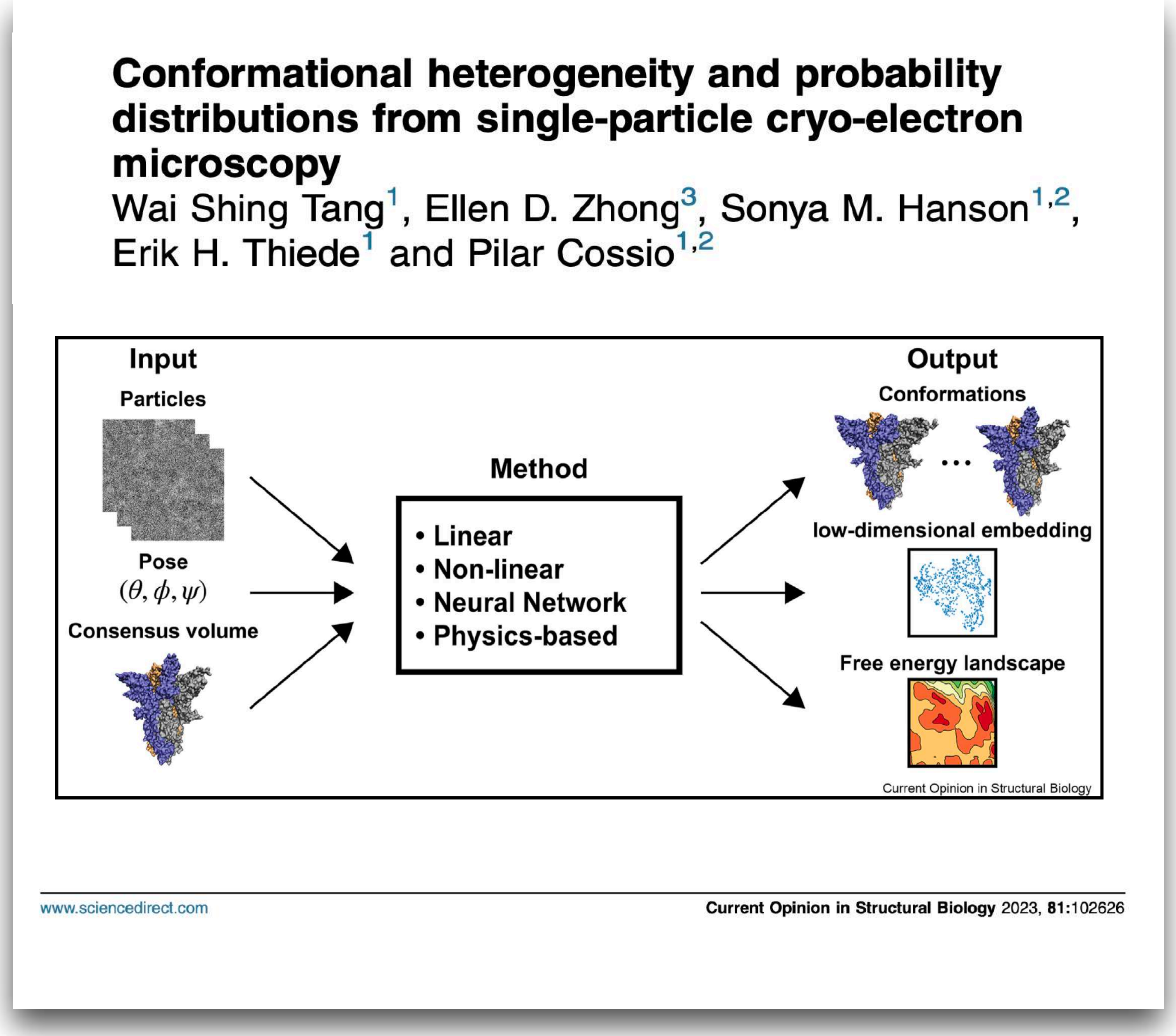


Nadezhdin et al, *Sci. Adv.* (2024)

Publication trends are in line with this vision.



A couple useful reviews to check out:

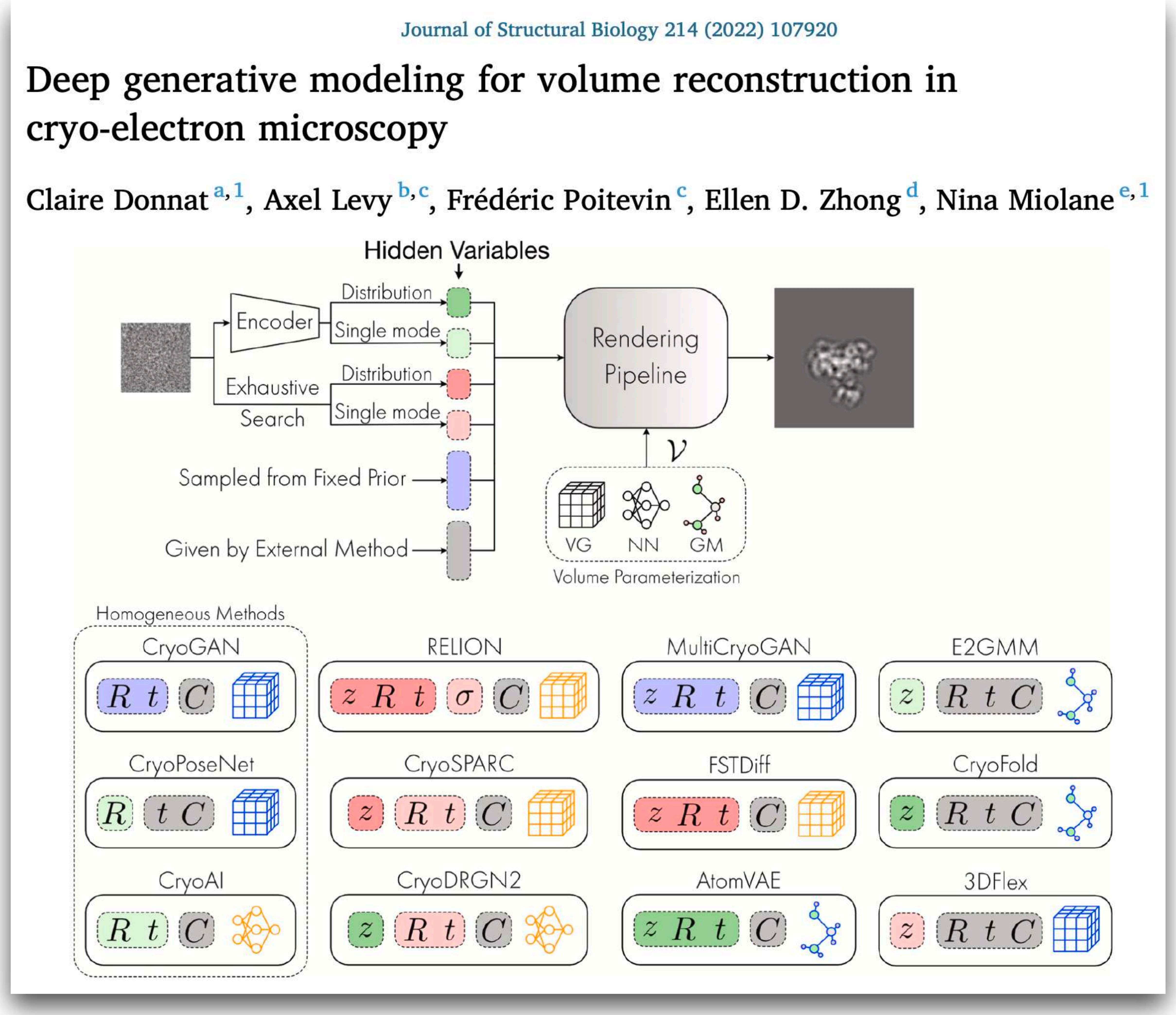


Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy

Wai Shing Tang[1], Ellen D. Zhong[3], Sonya M. Hanson[1,2], Erik H. Thiede[1] and Pilar Cossio[1,2]

www.sciencedirect.com

Current Opinion in Structural Biology 2023, **81**:102626

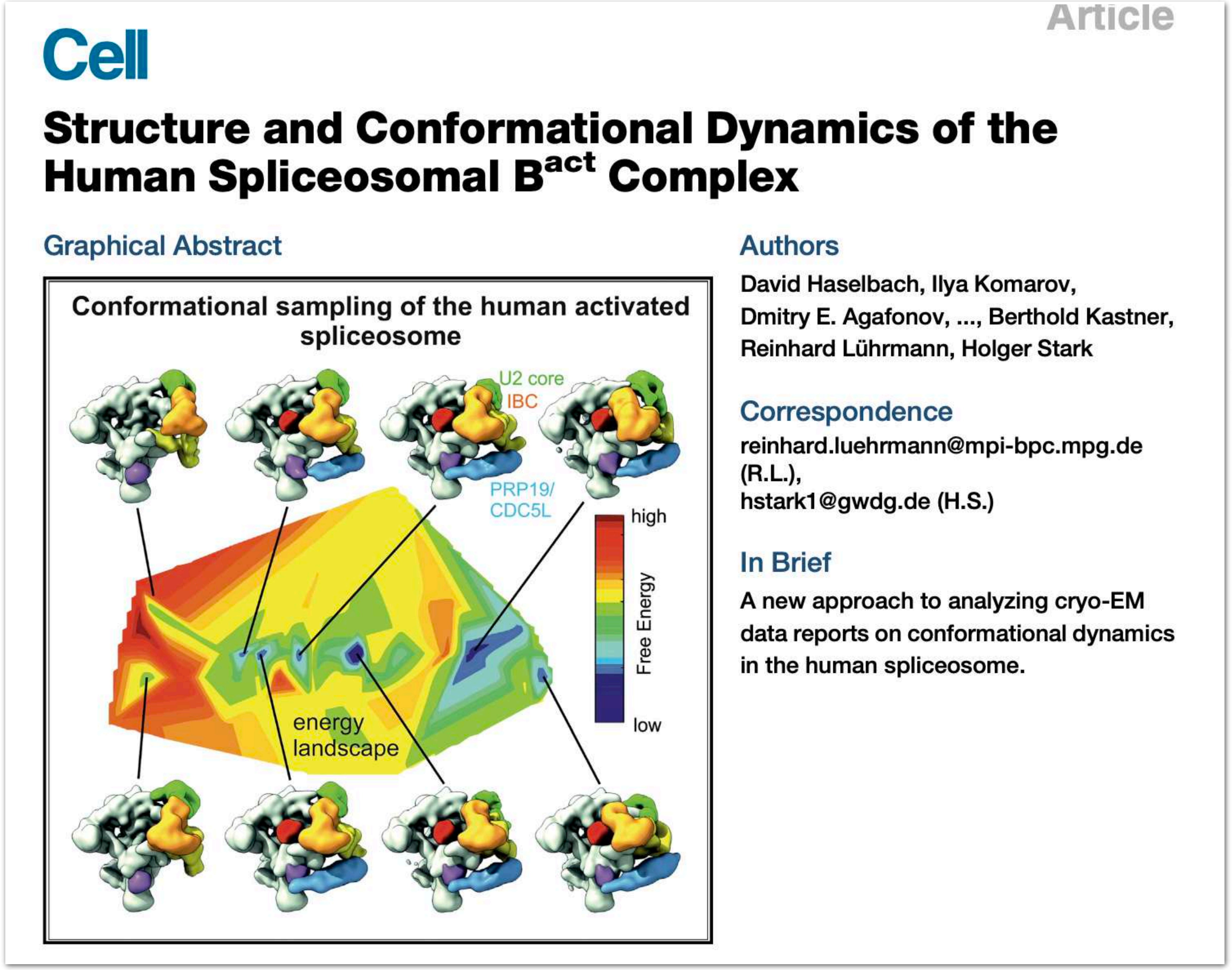


Journal of Structural Biology 214 (2022) 107920

Deep generative modeling for volume reconstruction in cryo-electron microscopy

Claire Donnat[a,1], Axel Levy[b,c], Frédéric Poitevin[c], Ellen D. Zhong[d], Nina Miolane[e,1]

Principal component analysis (PCA) for continuous heterogeneity in SPA.



Article

Cell

Structure and Conformational Dynamics of the Human Spliceosomal B^{act} Complex

Graphical Abstract

Authors

David Haselbach, Ilya Komarov,
Dmitry E. Agafonov, ..., Berthold Kastner,
Reinhard Lührmann, Holger Stark

Correspondence

reinhard.luehrmann@mpi-bpc.mpg.de
(R.L.),
hstark1@gwdg.de (H.S.)

In Brief

A new approach to analyzing cryo-EM data reports on conformational dynamics in the human spliceosome.

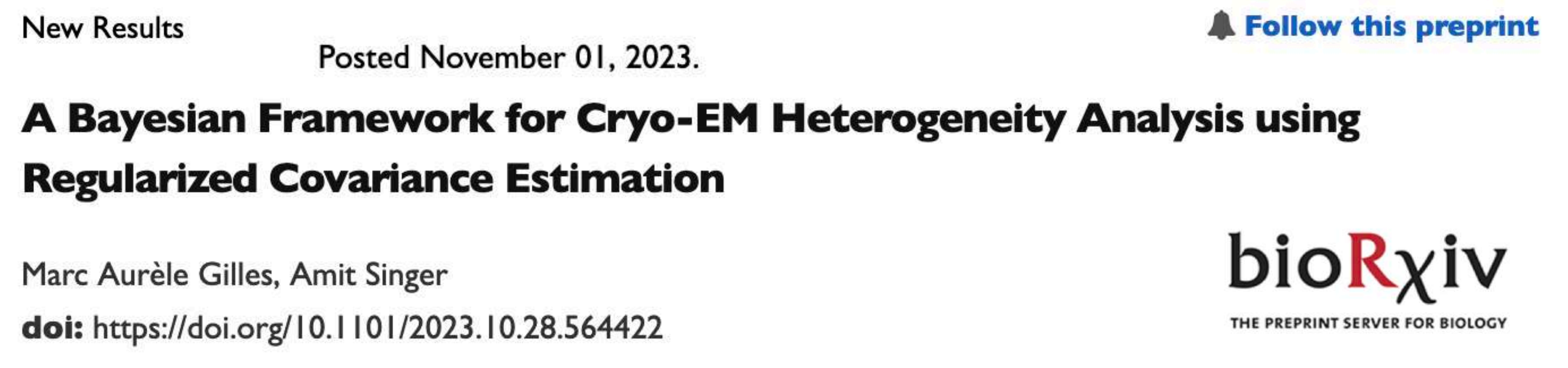
Acta Cryst D STRUCTURAL BIOLOGY

ISSN 2059-7983

Acta Cryst. (2021). D77, 835–839

Principal component analysis is limited to low-resolution analysis in cryoEM

Carlos Oscar S. Sorzano* and Jose Maria Carazo



New Results

Posted November 01, 2023.

🔔 **Follow this preprint**

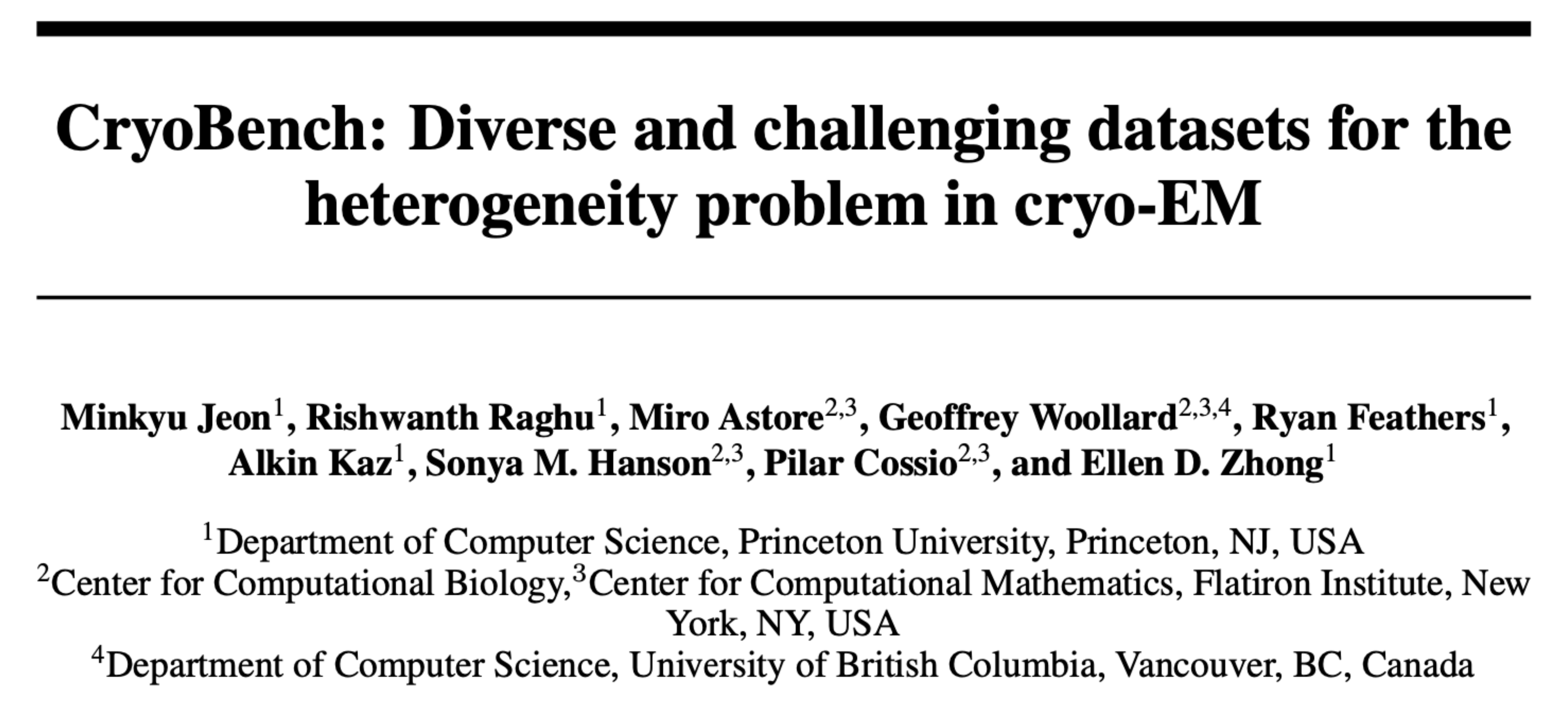
A Bayesian Framework for Cryo-EM Heterogeneity Analysis using Regularized Covariance Estimation

Marc Aurèle Gilles, Amit Singer

doi: https://doi.org/10.1101/2023.10.28.564422

bioRχiv THE PREPRINT SERVER FOR BIOLOGY

cryoBench ❄️🪑



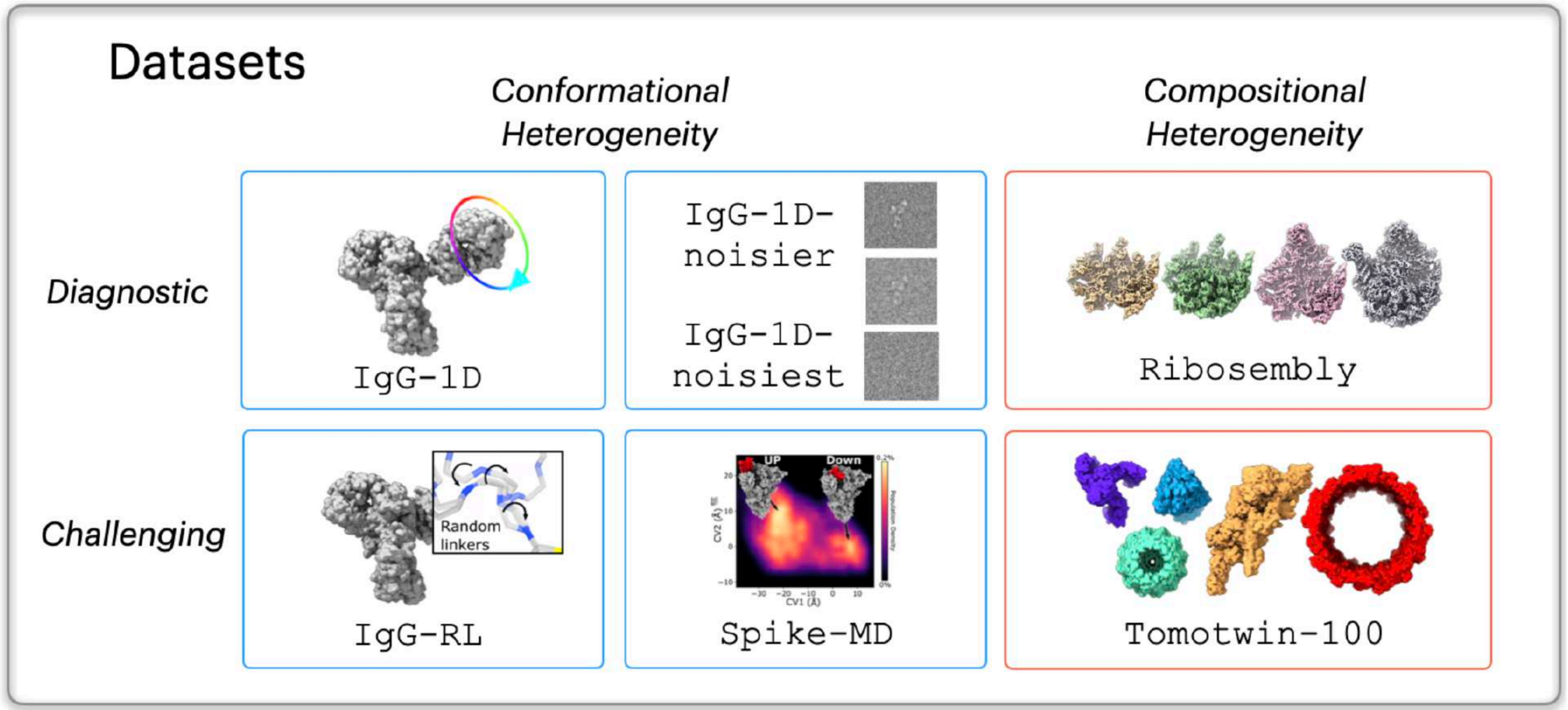
CryoBench: Diverse and challenging datasets for the heterogeneity problem in cryo-EM

Minkyu Jeon[1], Rishwanth Raghu[1], Miro Astore[2,3], Geoffrey Woollard[2,3,4], Ryan Feathers[1], Alkin Kaz[1], Sonya M. Hanson[2,3], Pilar Cossio[2,3], and Ellen D. Zhong[1]

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA
[2]Center for Computational Biology,[3]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA
[4]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

Ellen Zhong

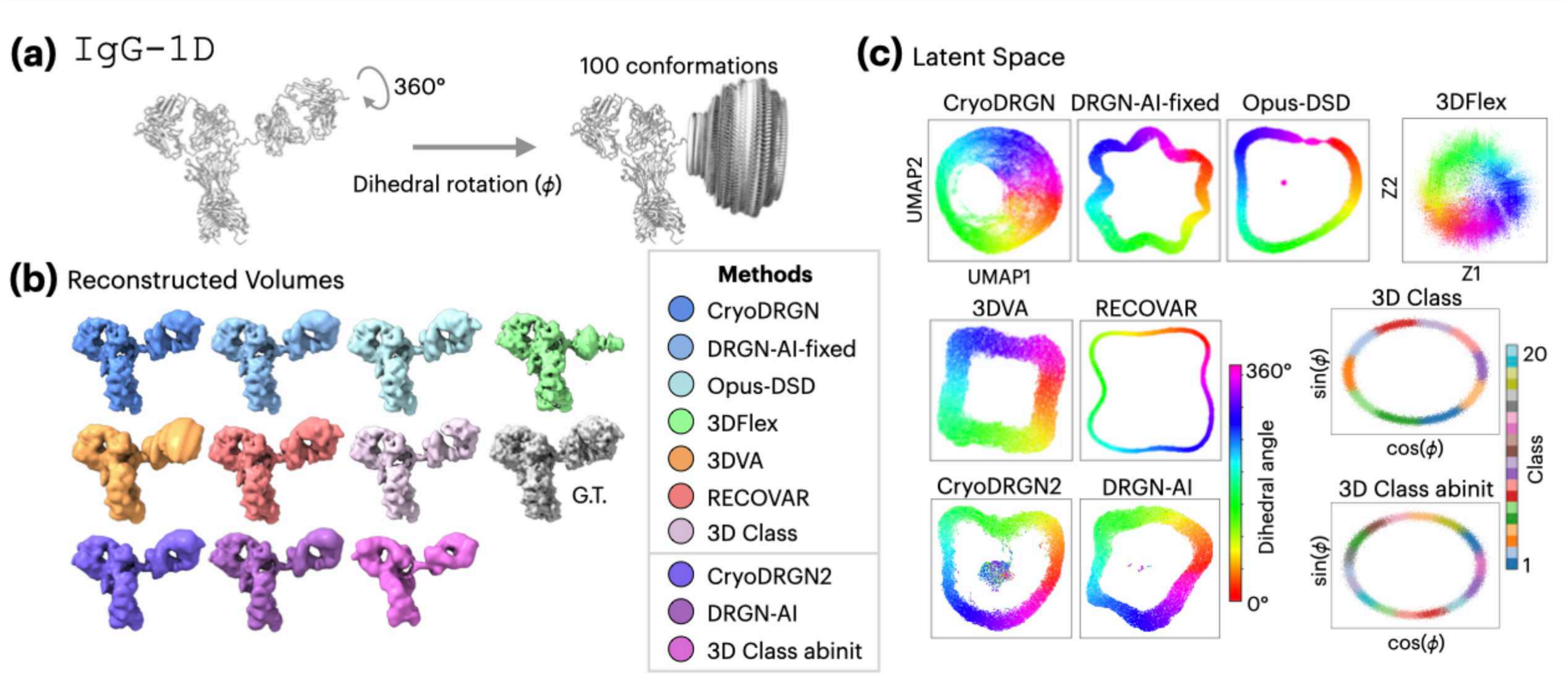
of cryoDRGN ❄️🐉
Nature Methods 2021



NeurIPS Benchmarks
2024 Spotlight

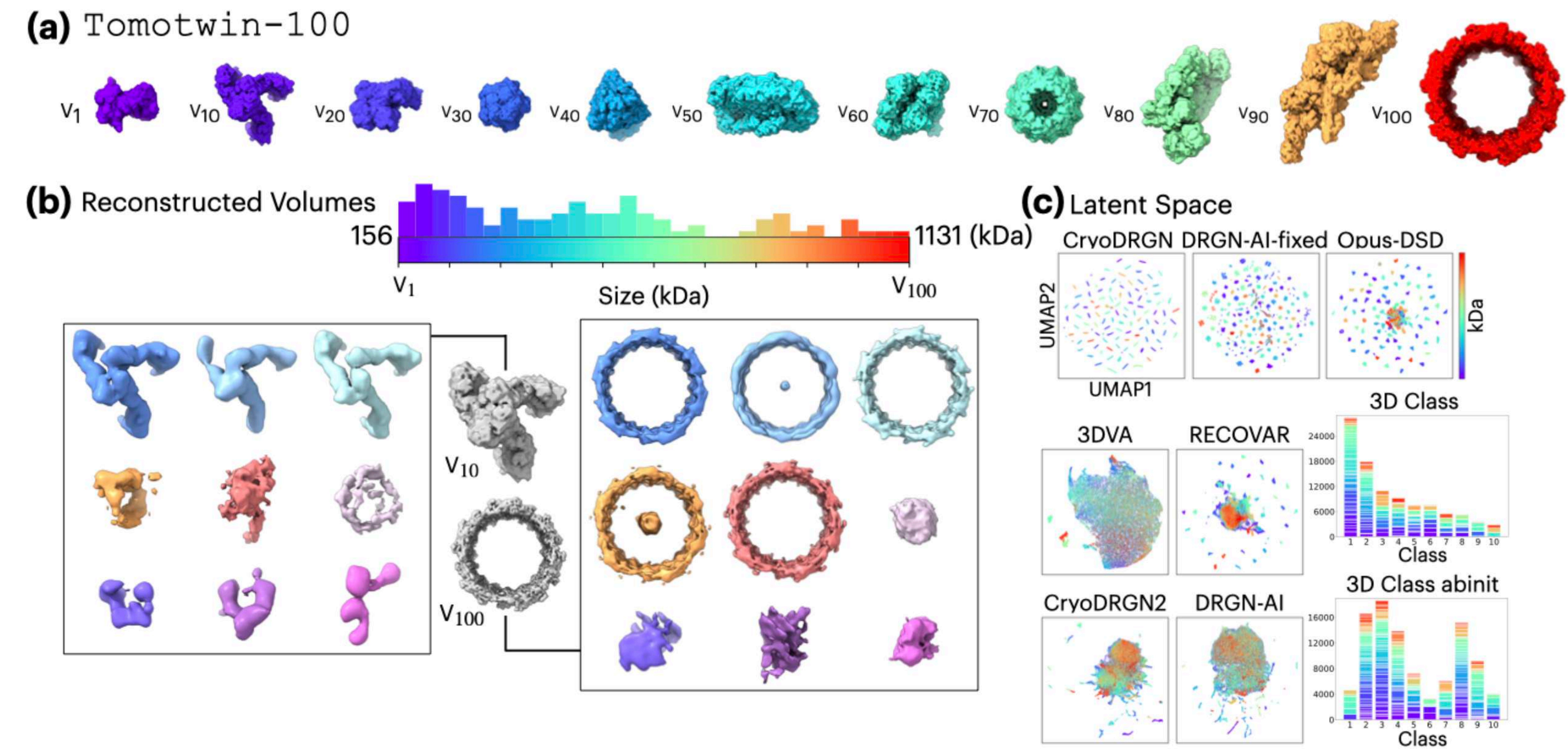
https://cryobench.cs.princeton.edu/

cryoBench ❄️🪑: From simple continuous motion.



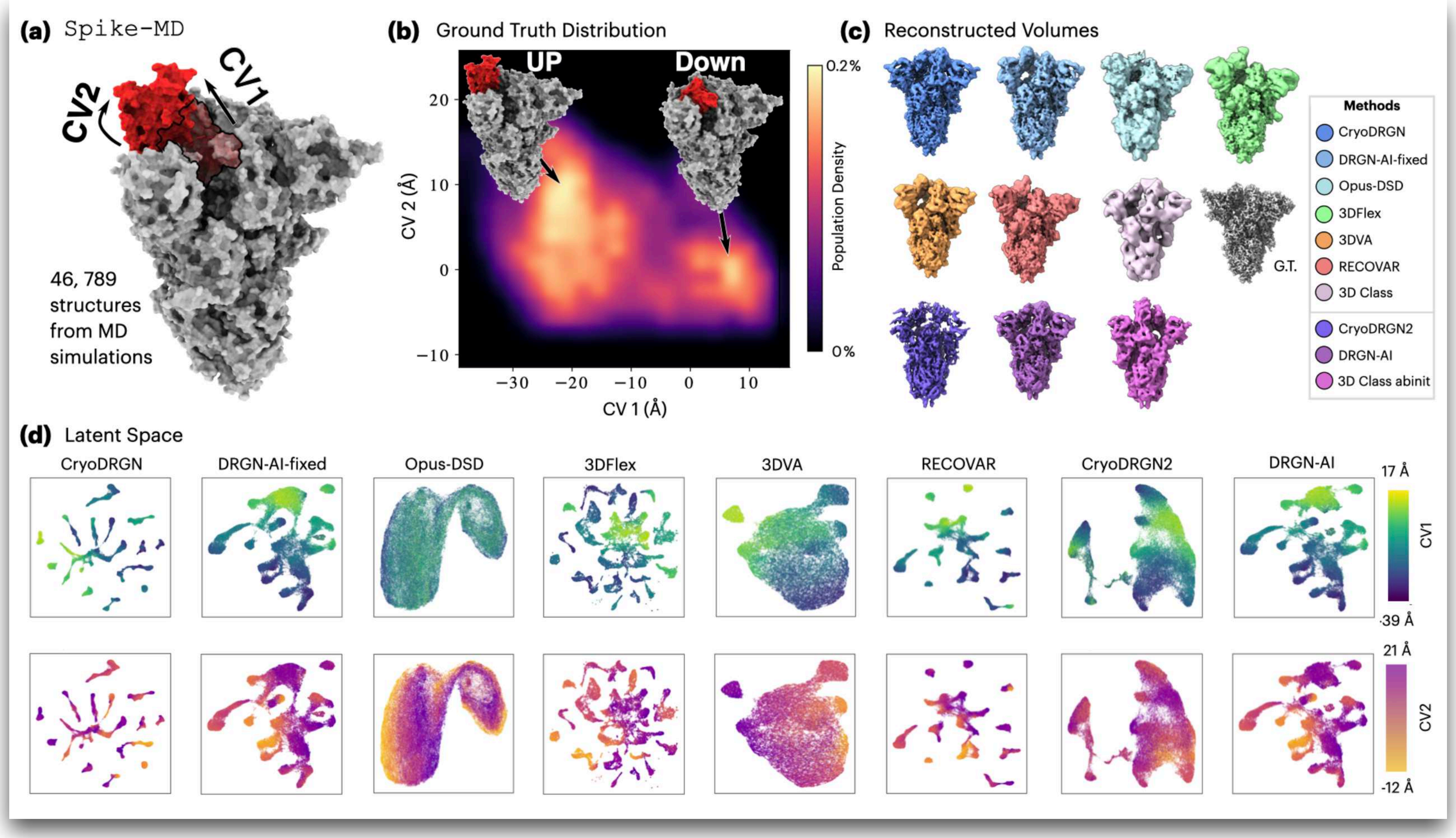
https://cryobench.cs.princeton.edu/

cryoBench ❄️🪑: To complex compositions.



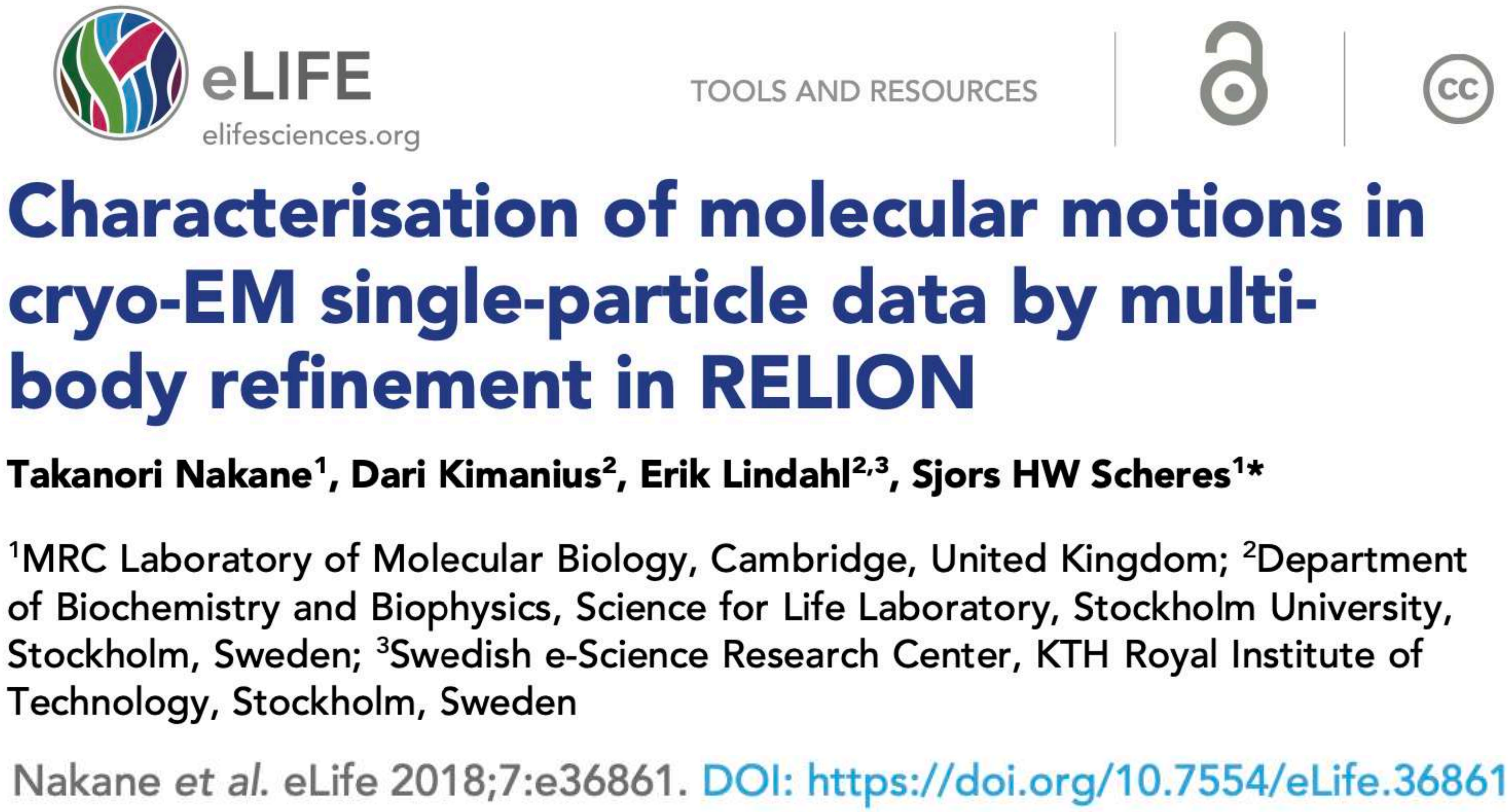
https://cryobench.cs.princeton.edu/

cryoBench ❄️🪑: To difficult full FE landscapes.



https://cryobench.cs.princeton.edu/

RELION: Multi-body Refinement



eLIFE
elifesciences.org

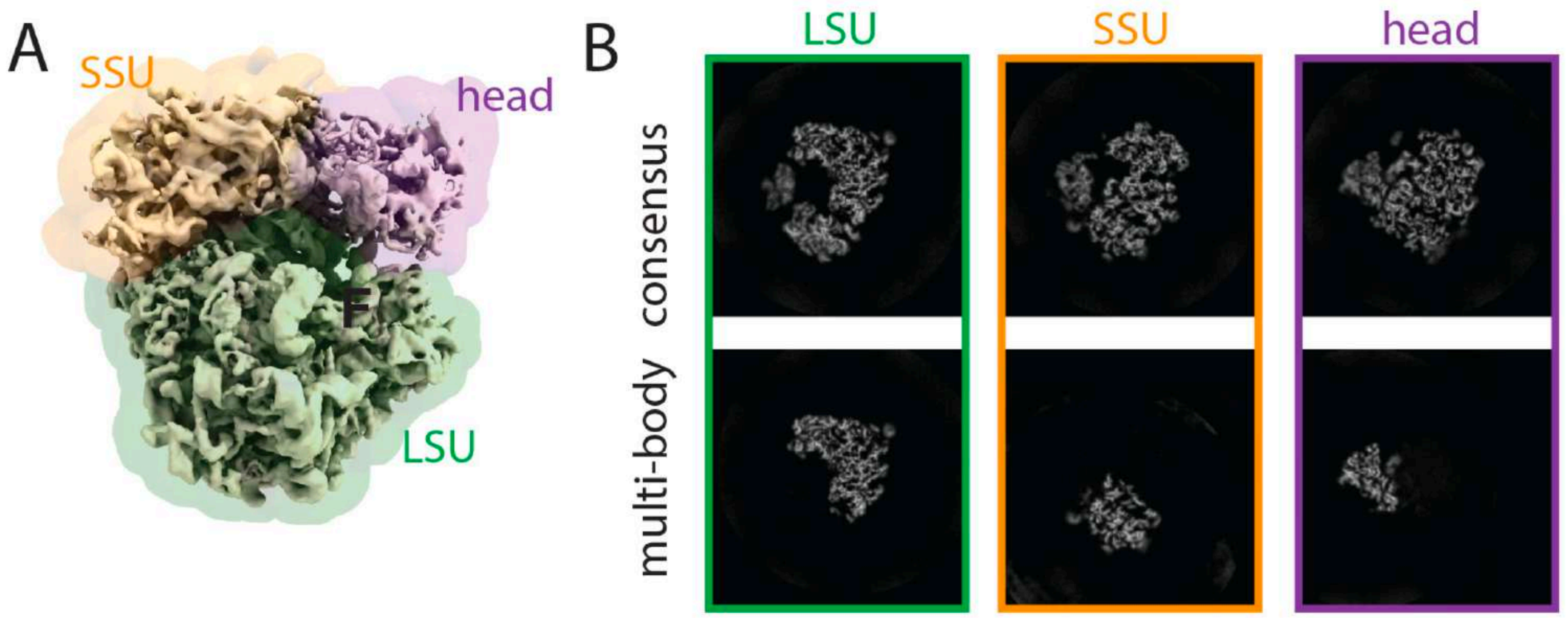
TOOLS AND RESOURCES

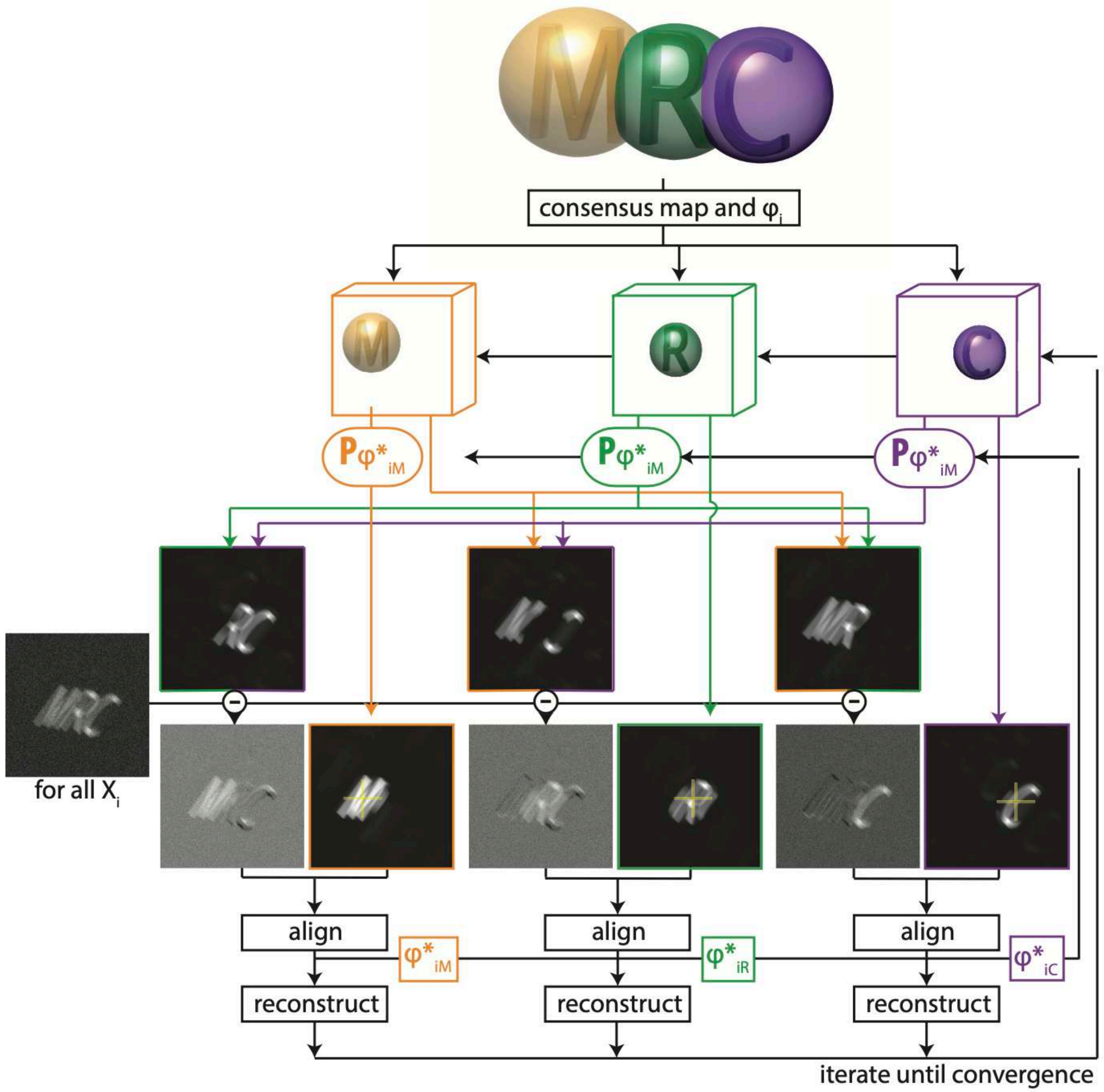
Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION

Takanori Nakane[1], Dari Kimanius[2], Erik Lindahl[2,3], Sjors HW Scheres[1]*

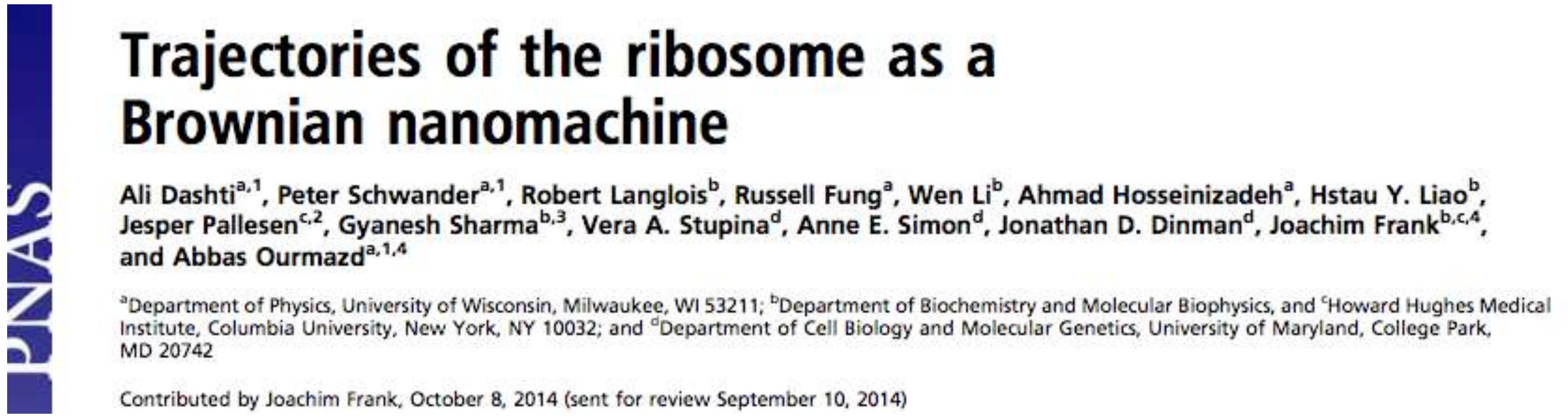
[1]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom; [2]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden; [3]Swedish e-Science Research Center, KTH Royal Institute of Technology, Stockholm, Sweden

Nakane *et al.* eLife 2018;7:e36861. DOI: https://doi.org/10.7554/eLife.36861





Initial TRPV1 analyses conducted in MATLAB version of ManifoldEM.



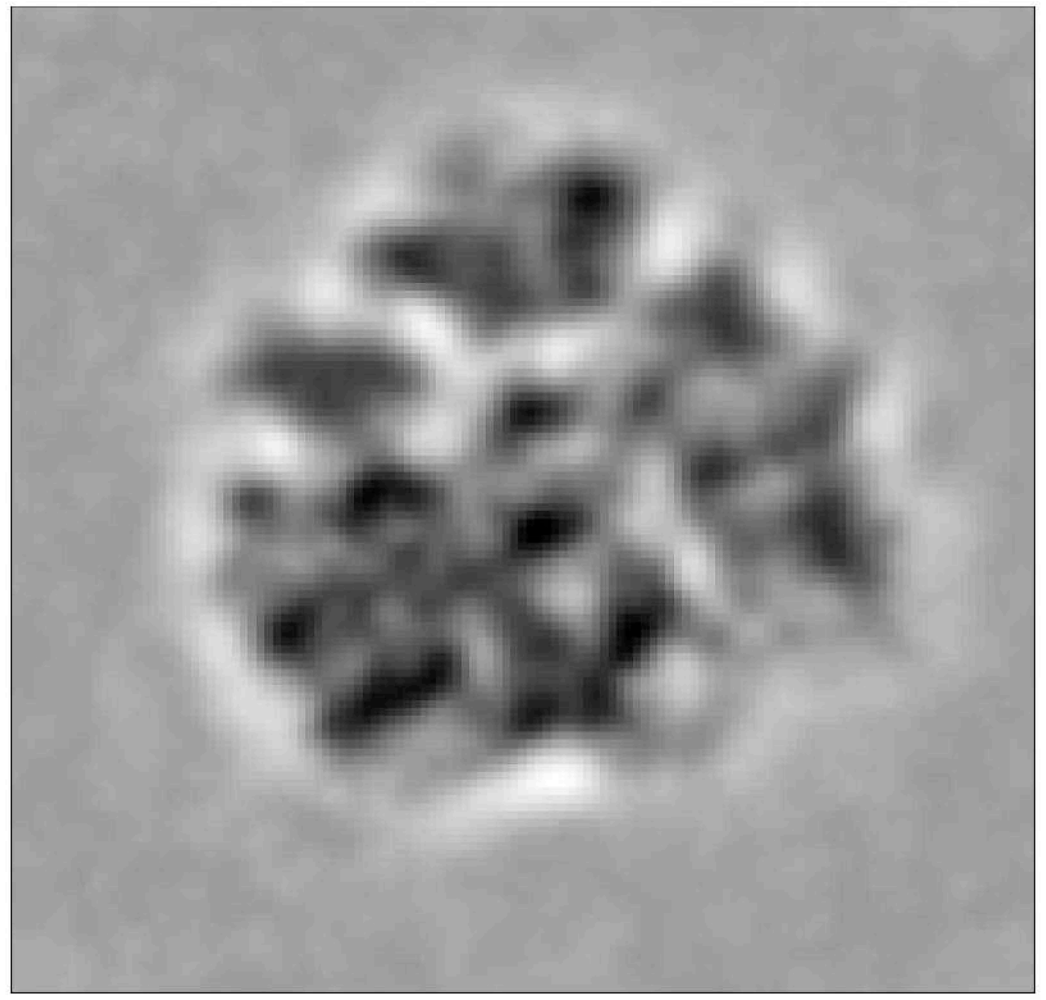
PNAS

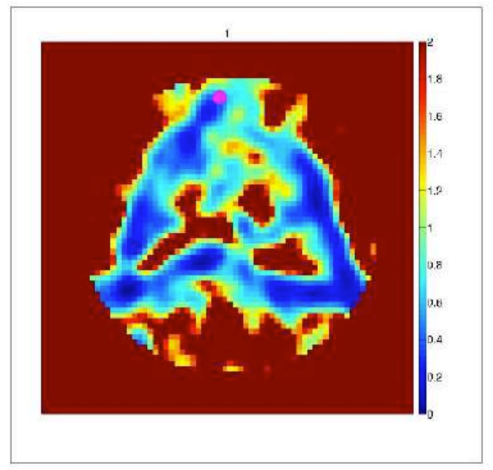
Trajectories of the ribosome as a Brownian nanomachine

Ali Dashti[a,1], Peter Schwander[a,1], Robert Langlois[b], Russell Fung[a], Wen Li[b], Ahmad Hosseinizadeh[a], Hstau Y. Liao[b], Jesper Pallesen[c,2], Gyanesh Sharma[b,3], Vera A. Stupina[d], Anne E. Simon[d], Jonathan D. Dinman[d], Joachim Frank[b,c,4], and Abbas Ourmazd[a,1,4]

[a]Department of Physics, University of Wisconsin, Milwaukee, WI 53211; [b]Department of Biochemistry and Molecular Biophysics, and [c]Howard Hughes Medical Institute, Columbia University, New York, NY 10032; and [d]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742

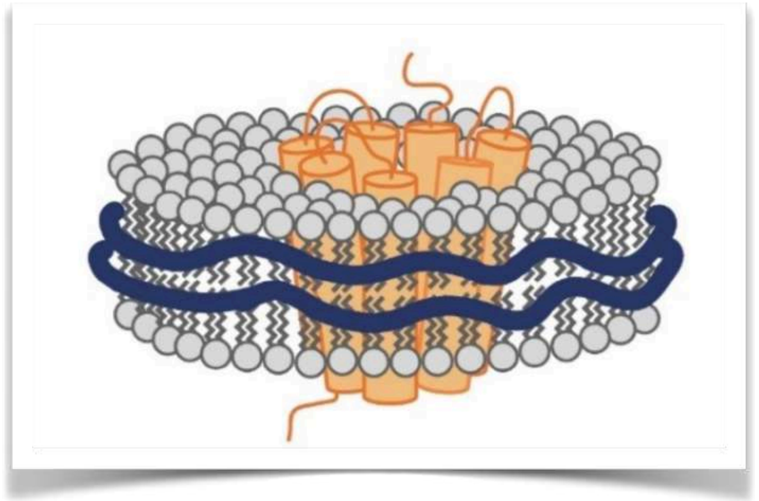
Contributed by Joachim Frank, October 8, 2014 (sent for review September 10, 2014)

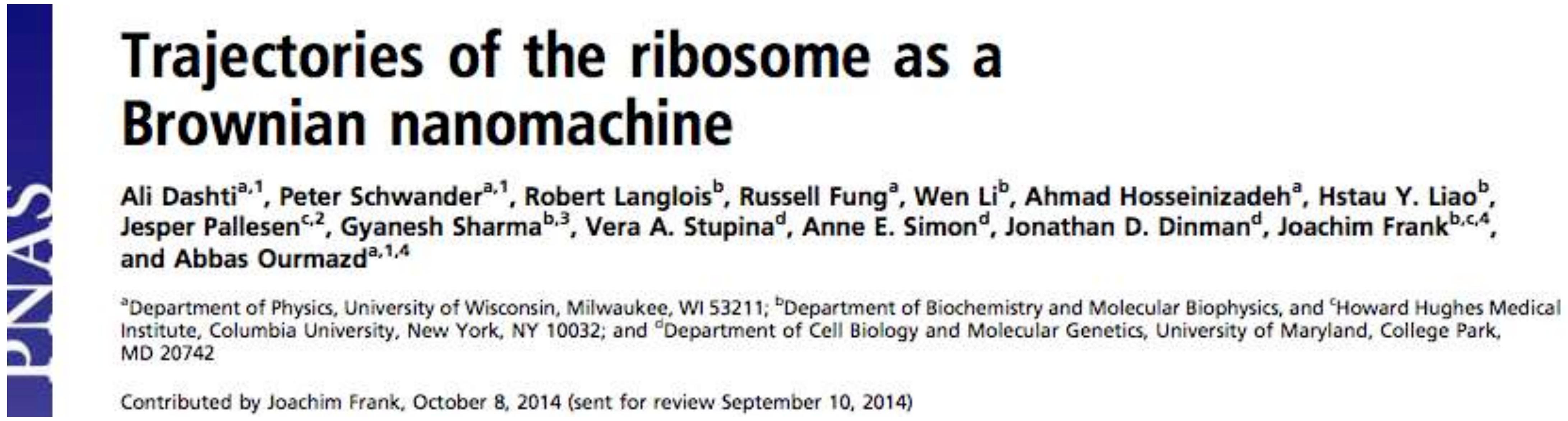




Dashti et al, *PNAS* (2014).

Initial TRPV1 analyses conducted in MATLAB version of ManifoldEM.





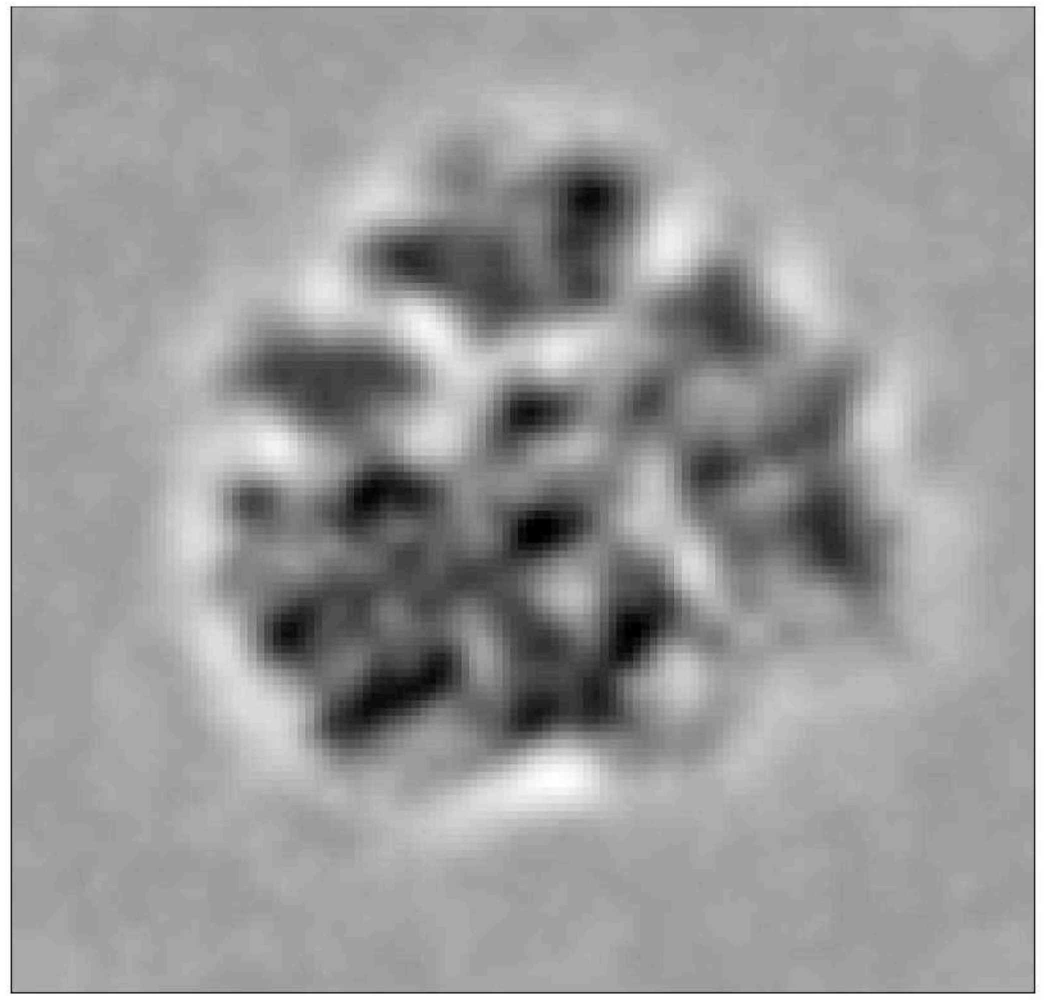
PNAS

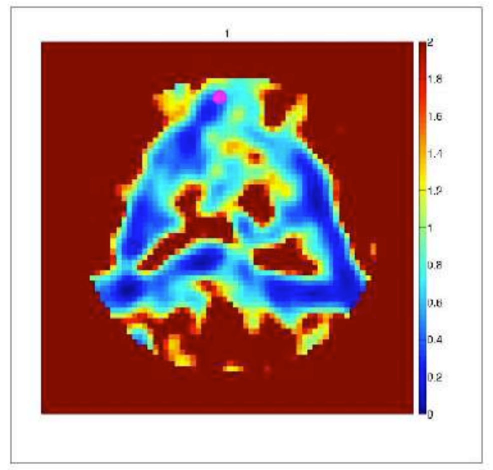
Trajectories of the ribosome as a Brownian nanomachine

Ali Dashti[a,1], Peter Schwander[a,1], Robert Langlois[b], Russell Fung[a], Wen Li[b], Ahmad Hosseinizadeh[a], Hstau Y. Liao[b], Jesper Pallesen[c,2], Gyanesh Sharma[b,3], Vera A. Stupina[d], Anne E. Simon[d], Jonathan D. Dinman[d], Joachim Frank[b,c,4], and Abbas Ourmazd[a,1,4]

[a]Department of Physics, University of Wisconsin, Milwaukee, WI 53211; [b]Department of Biochemistry and Molecular Biophysics, and [c]Howard Hughes Medical Institute, Columbia University, New York, NY 10032; and [d]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742

Contributed by Joachim Frank, October 8, 2014 (sent for review September 10, 2014)



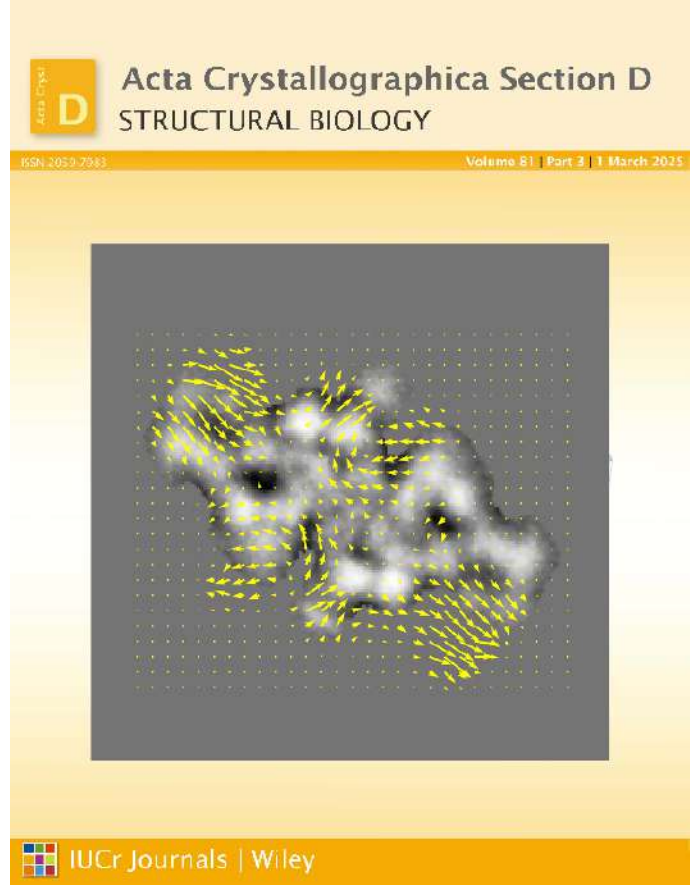


Dashti et al, *PNAS* (2014).

New Manifold Embedding Repository Maintained at Flatiron

- Pythonic
- Bug fixes
- New Features (CLI, ipynb's)
- Order of magnitude speedup

Robert Blackwell



Acta Crystallographica Section D
STRUCTURAL BIOLOGY

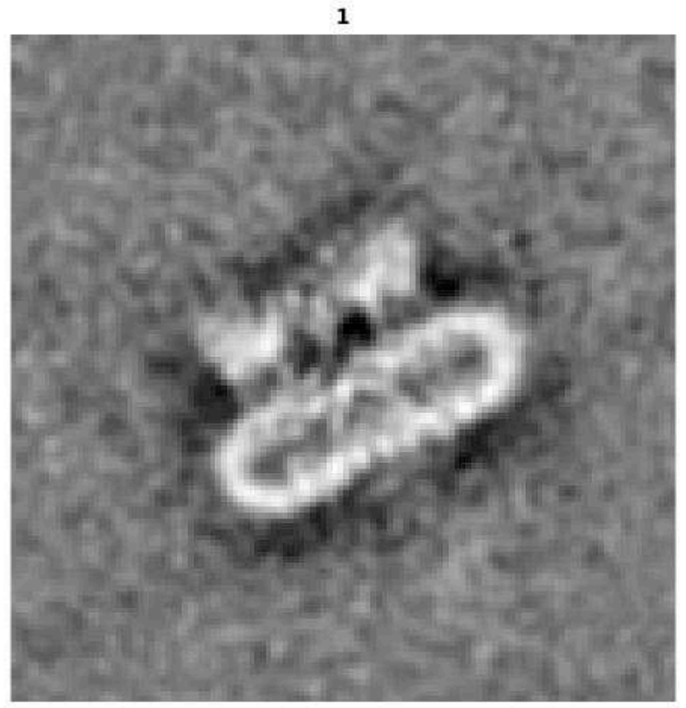
IUCr Journals | Wiley

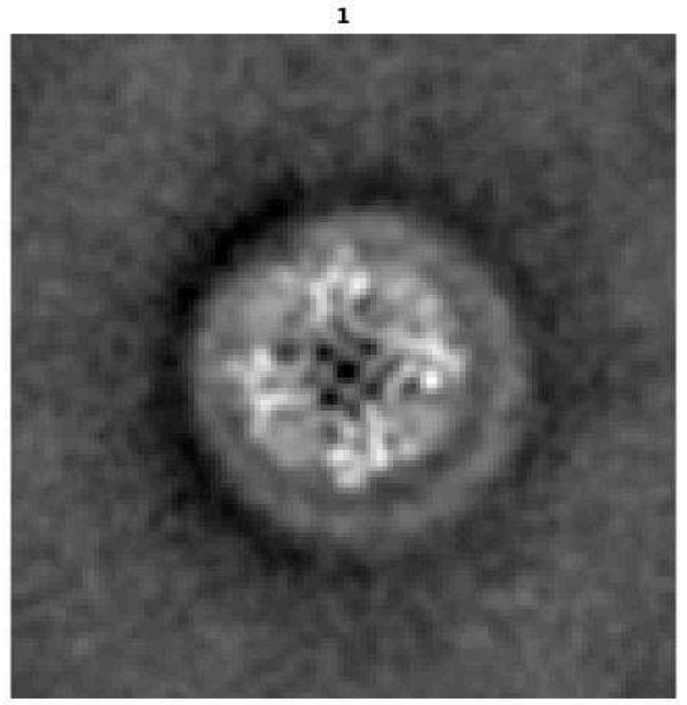
Ojha et al, *Acta Cryst D* (2025).

https://github.com/flatironinstitute/ManifoldEM

rTRPV1 + DkTx/RtX
in nanodiscs

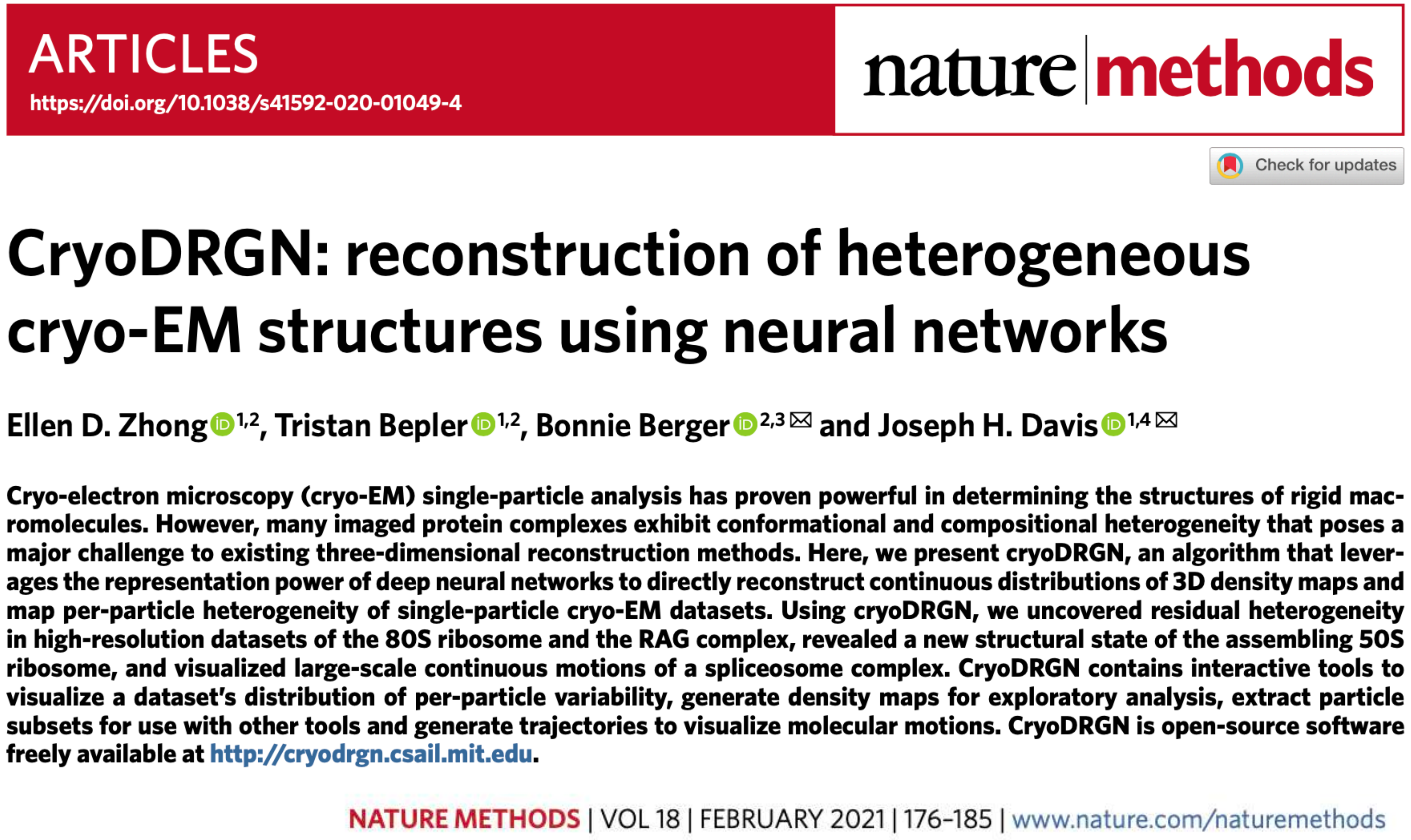
(EMPIAR 10059)





Preliminary data generated in Frank lab

CryoDRGN: The First Neural Network Approach


ARTICLES
https://doi.org/10.1038/s41592-020-01049-4

nature methods

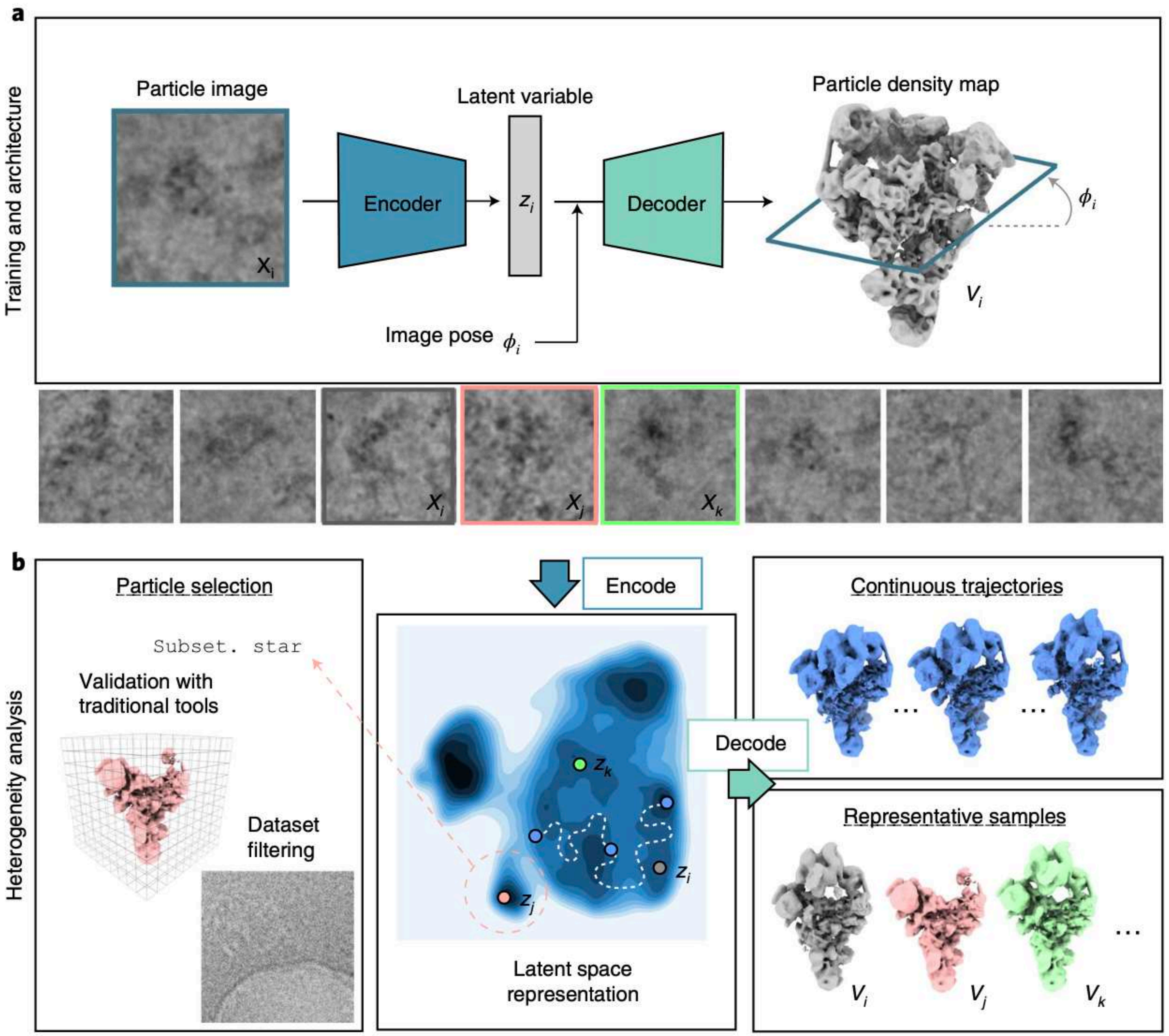
Check for updates

CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

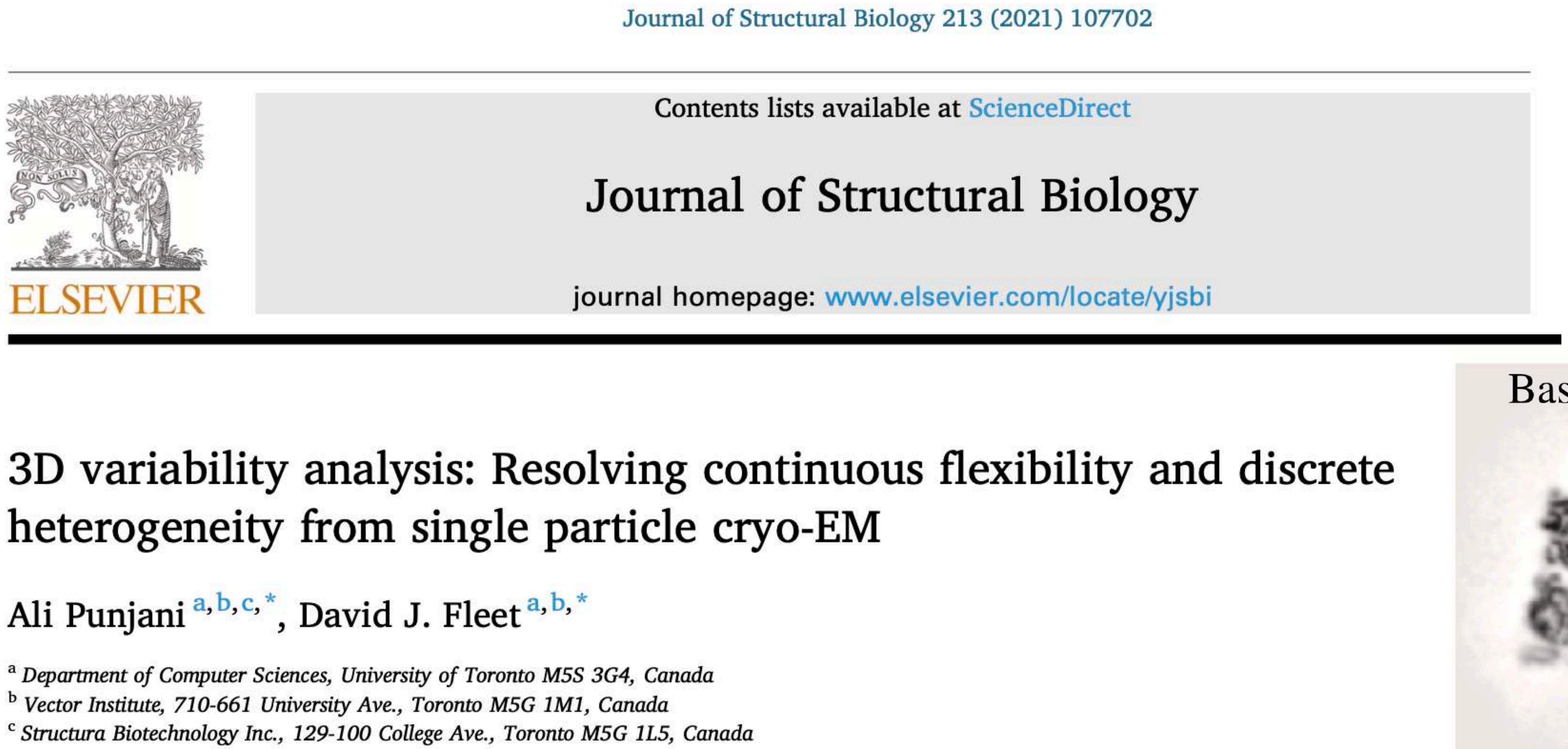
Ellen D. Zhong [1,2], Tristan Bepler [1,2], Bonnie Berger [2,3] ✉ and Joseph H. Davis [1,4] ✉

Cryo-electron microscopy (cryo-EM) single-particle analysis has proven powerful in determining the structures of rigid macromolecules. However, many imaged protein complexes exhibit conformational and compositional heterogeneity that poses a major challenge to existing three-dimensional reconstruction methods. Here, we present cryoDRGN, an algorithm that leverages the representation power of deep neural networks to directly reconstruct continuous distributions of 3D density maps and map per-particle heterogeneity of single-particle cryo-EM datasets. Using cryoDRGN, we uncovered residual heterogeneity in high-resolution datasets of the 80S ribosome and the RAG complex, revealed a new structural state of the assembling 50S ribosome, and visualized large-scale continuous motions of a spliceosome complex. CryoDRGN contains interactive tools to visualize a dataset's distribution of per-particle variability, generate density maps for exploratory analysis, extract particle subsets for use with other tools and generate trajectories to visualize molecular motions. CryoDRGN is open-source software freely available at http://cryodrgn.csail.mit.edu.

NATURE METHODS | VOL 18 | FEBRUARY 2021 | 176–185 | www.nature.com/naturemethods



3DVariability Analysis in cryoSPARC



Journal of Structural Biology 213 (2021) 107702

Contents lists available at ScienceDirect

Journal of Structural Biology

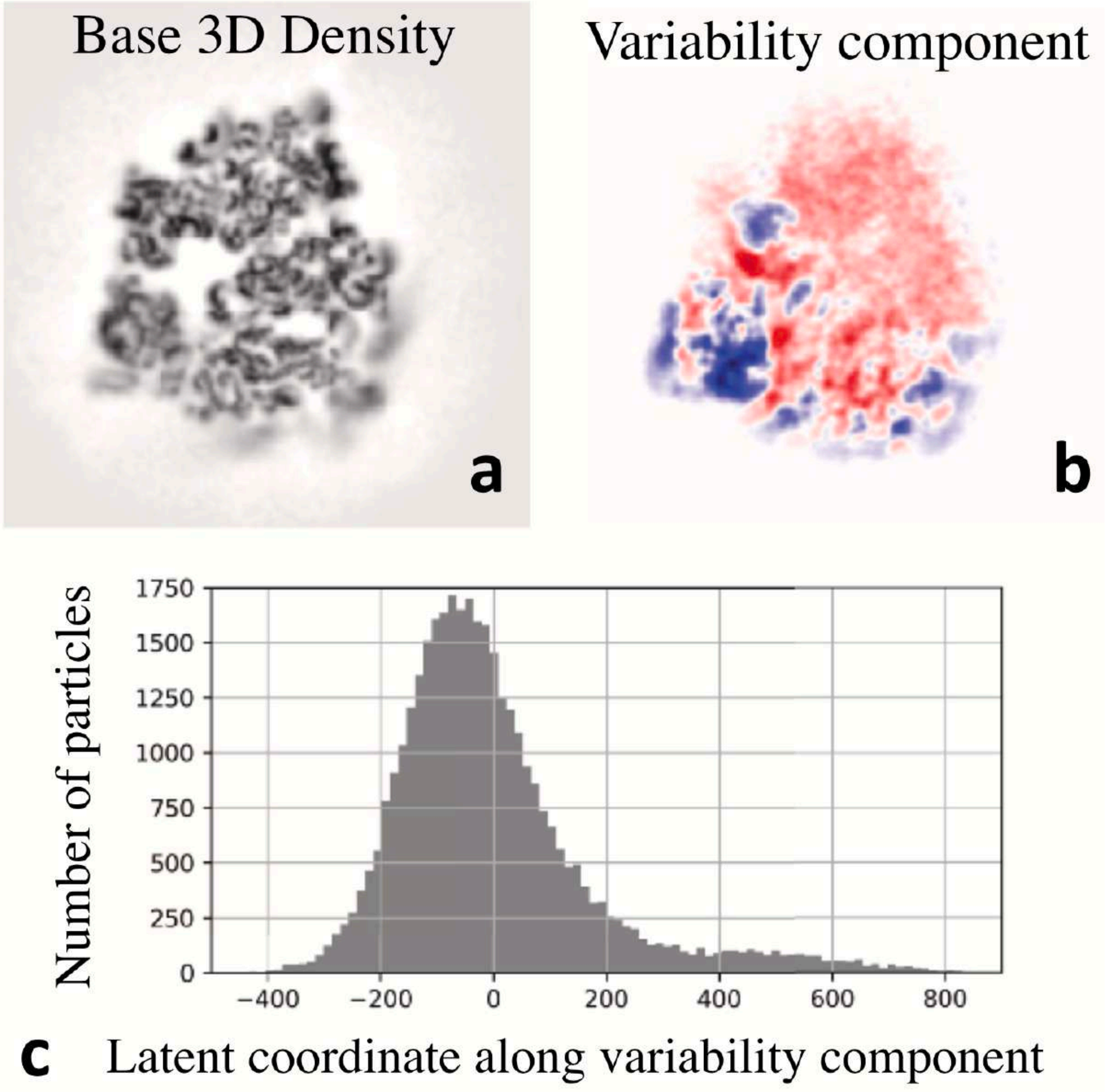
journal homepage: www.elsevier.com/locate/yjsbi

ELSEVIER

3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM

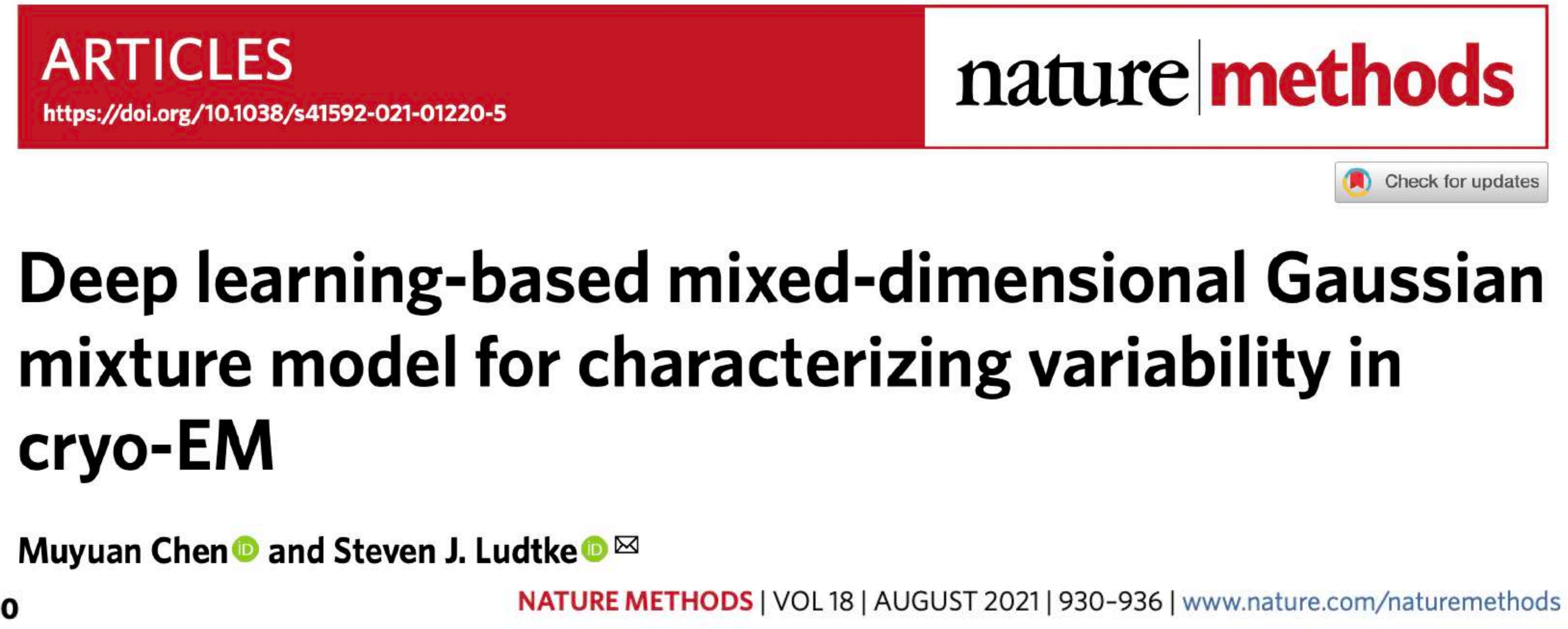
Ali Punjani [a,b,c,*], David J. Fleet [a,b,*]

[a] Department of Computer Sciences, University of Toronto M5S 3G4, Canada
[b] Vector Institute, 710-661 University Ave., Toronto M5G 1M1, Canada
[c] Structura Biotechnology Inc., 129-100 College Ave., Toronto M5G 1L5, Canada



- Very straightforward to use.
- Limited to linear motions.

Gaussian Mixture Model method in EMAN2



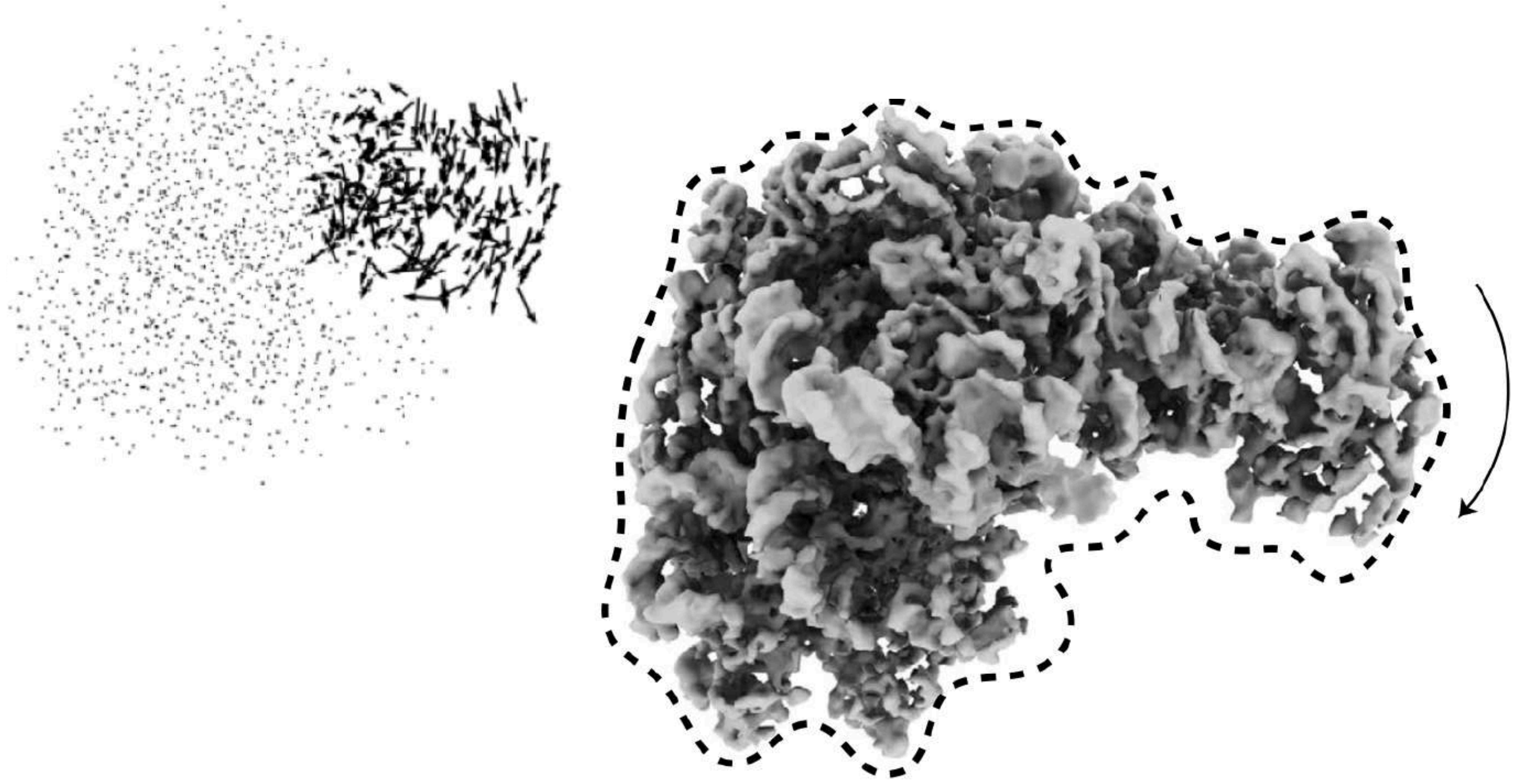
ARTICLES
https://doi.org/10.1038/s41592-021-01220-5
nature methods

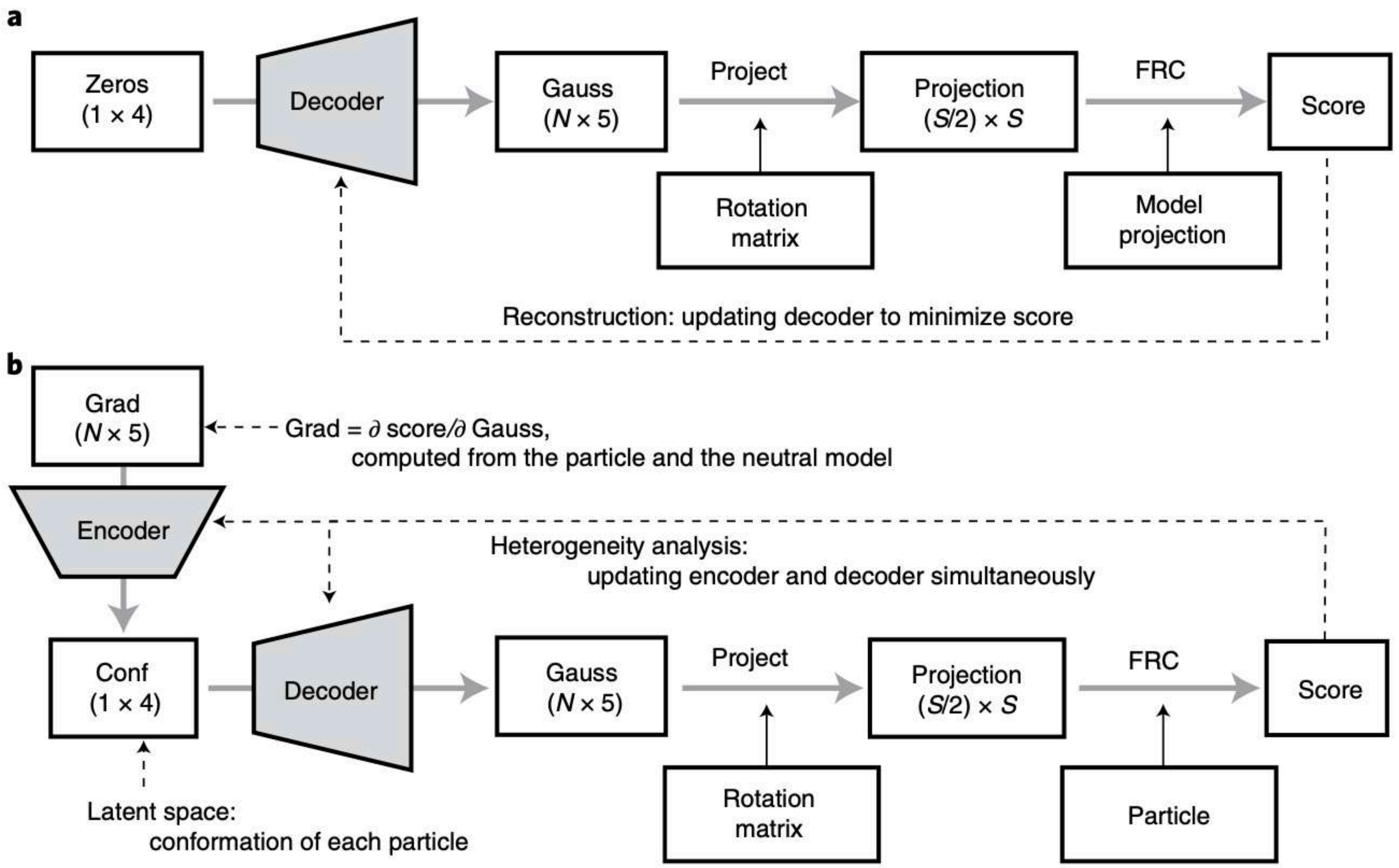
Check for updates

Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM

Muyuan Chen and Steven J. Ludtke

930 NATURE METHODS | VOL 18 | AUGUST 2021 | 930–936 | www.nature.com/naturemethods





- Incorporating the Gaussian Mixture Model (GMM) gets us toward atomic / molecular information of the dynamics.

cryoSPARC's more complex tool for heterogeneity analysis

nature methods

Article

https://doi.org/10.1038/s41592-023-01853-8

3DFlex: determining structure and motion of flexible proteins from cryo-EM

Received: 28 July 2022

Accepted: 16 March 2023

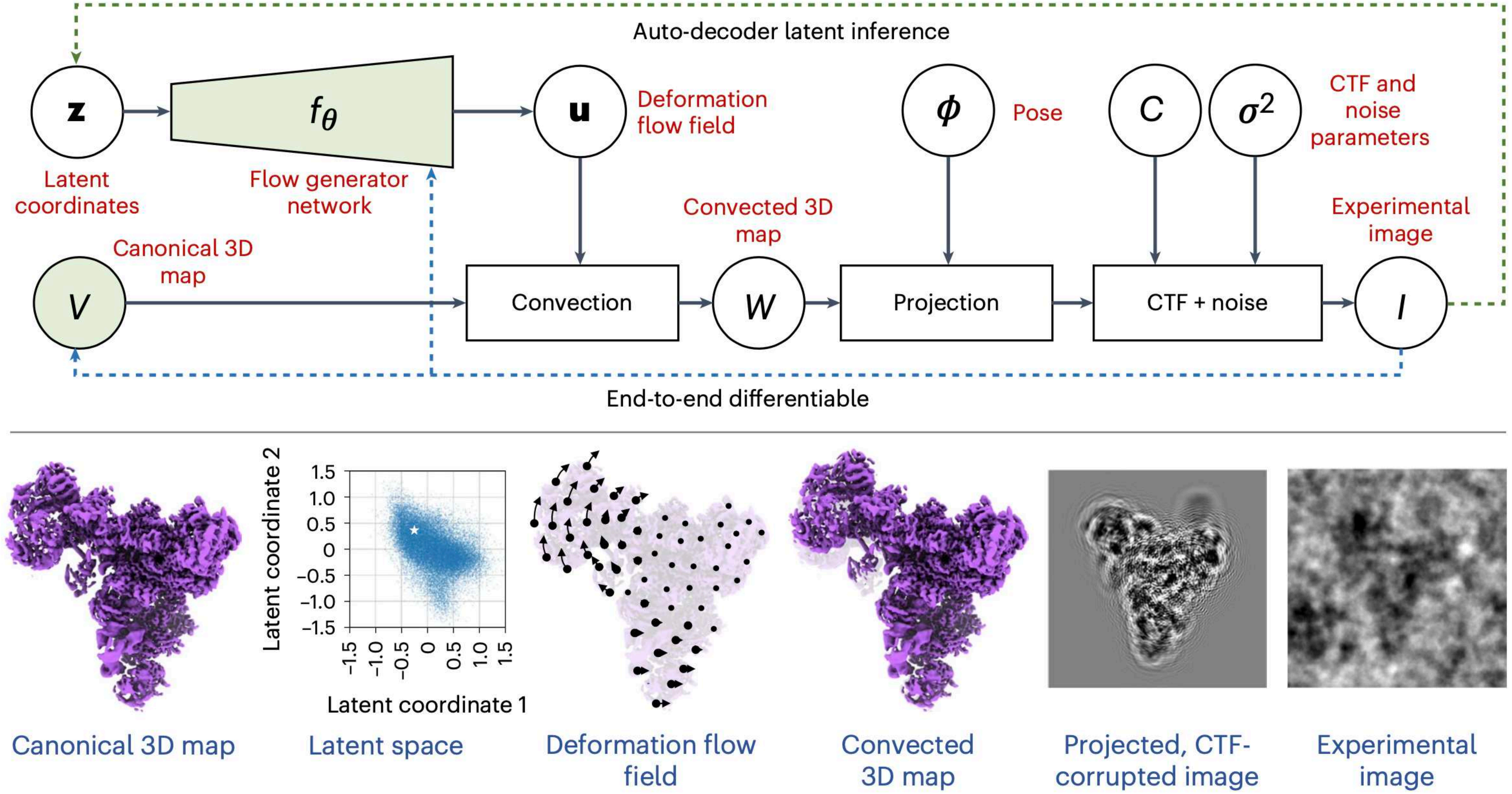
Published online: 11 May 2023

Check for updates

Ali Punjani [1,2,3] & **David J. Fleet** [1,2,4]

Modeling flexible macromolecules is one of the
single-particle cryogenic-electron microscopy (
to illuminate fundamental questions in structur
Three-Dimensional Flexible Refinement (3DFlex

Nature Methods | Volume 20 | June 2023 | 860–870



Canonical 3D map

Latent space

Deformation flow field

Convected 3D map

Projected, CTF-corrupted image

Experimental image

cryoSPARC's more complex tool for heterogeneity analysis

nature methods

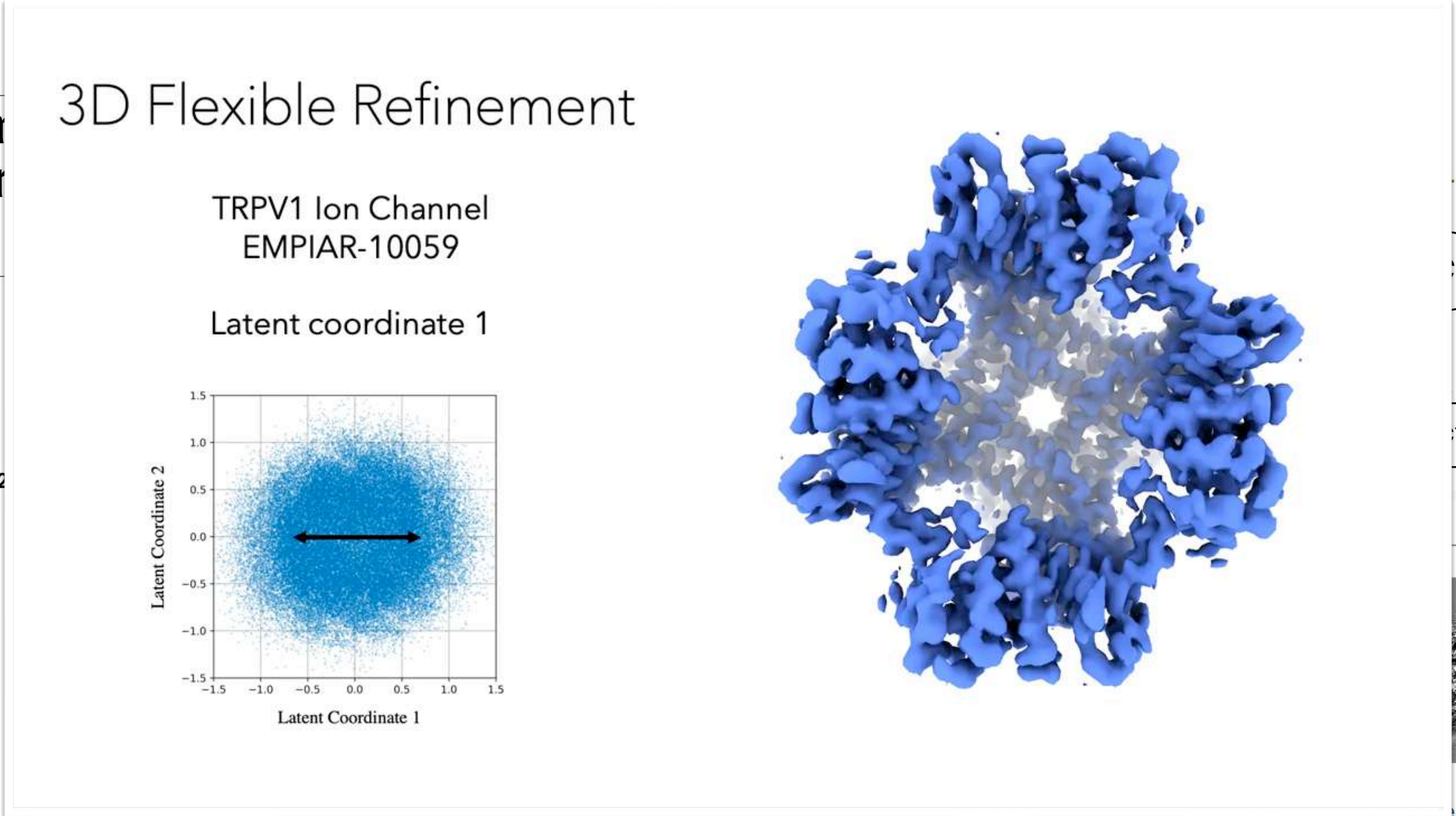
Article

3DFlex: determ
flexible protei

Received: 28 July 2022
Accepted: 16 March 2023

Published online: 11 May 2023
Check for updates

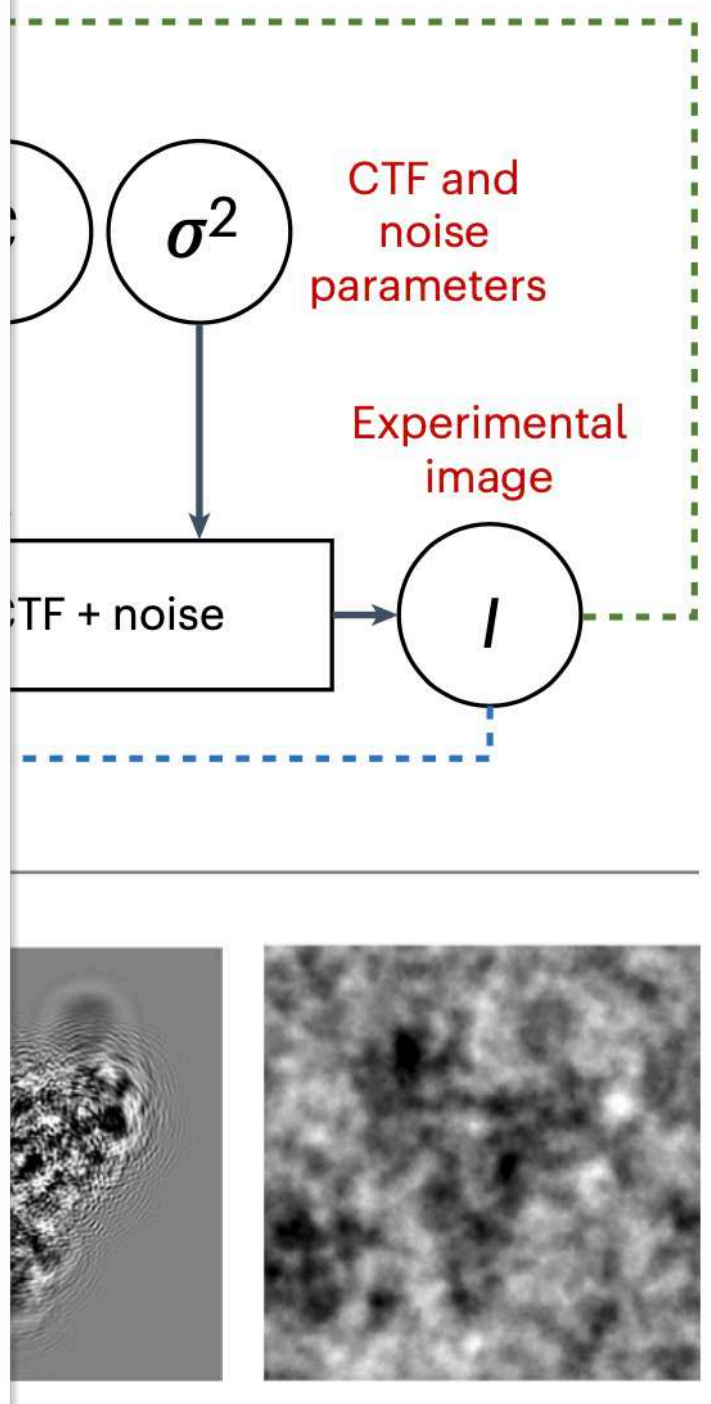
Nature Methods | Volume 20 | June 2



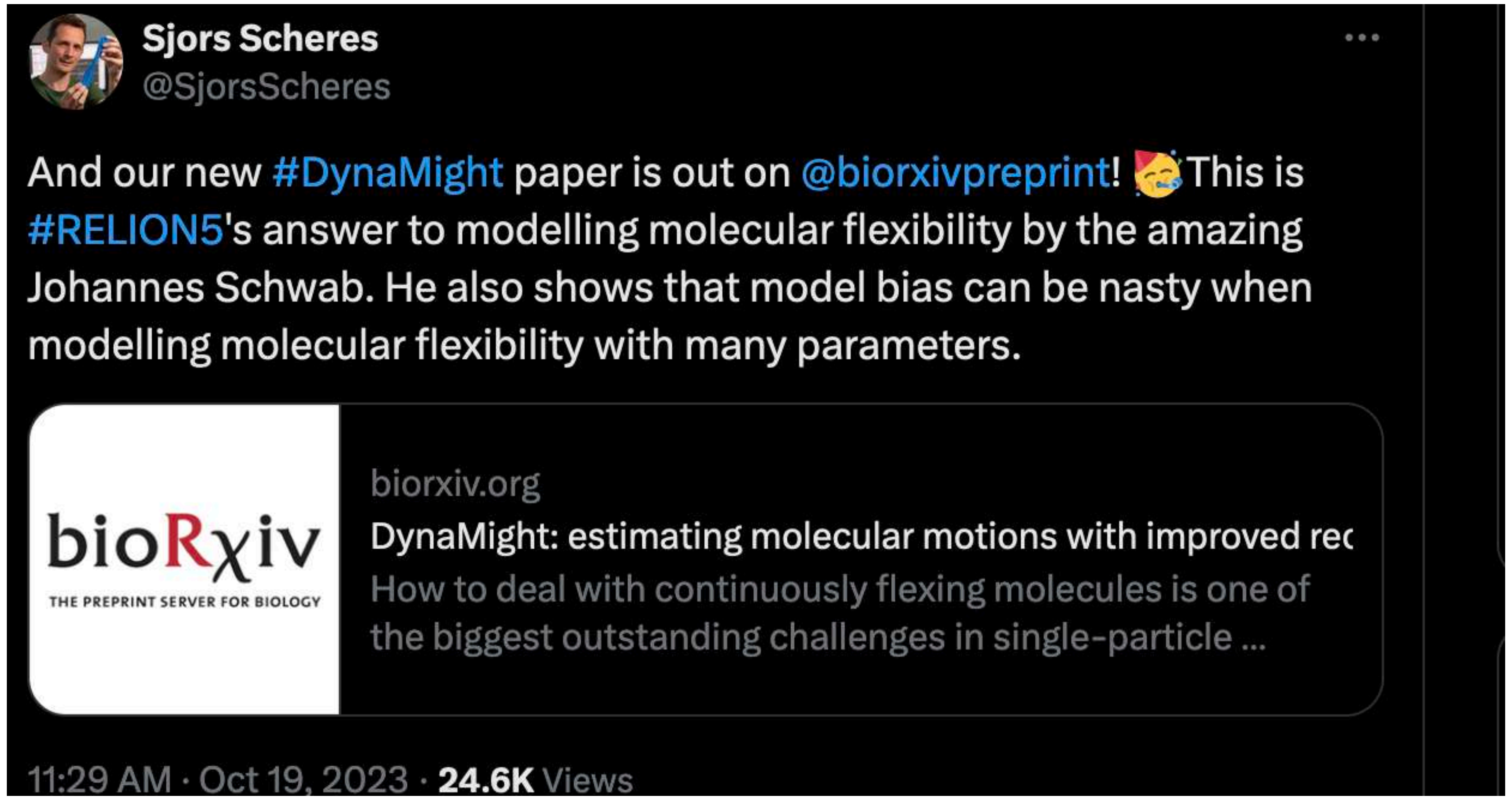
3D Flexible Refinement

TRPV1 Ion Channel
EMPIAR-10059

Latent coordinate 1



RELION also has a new ML-based solution for this problem



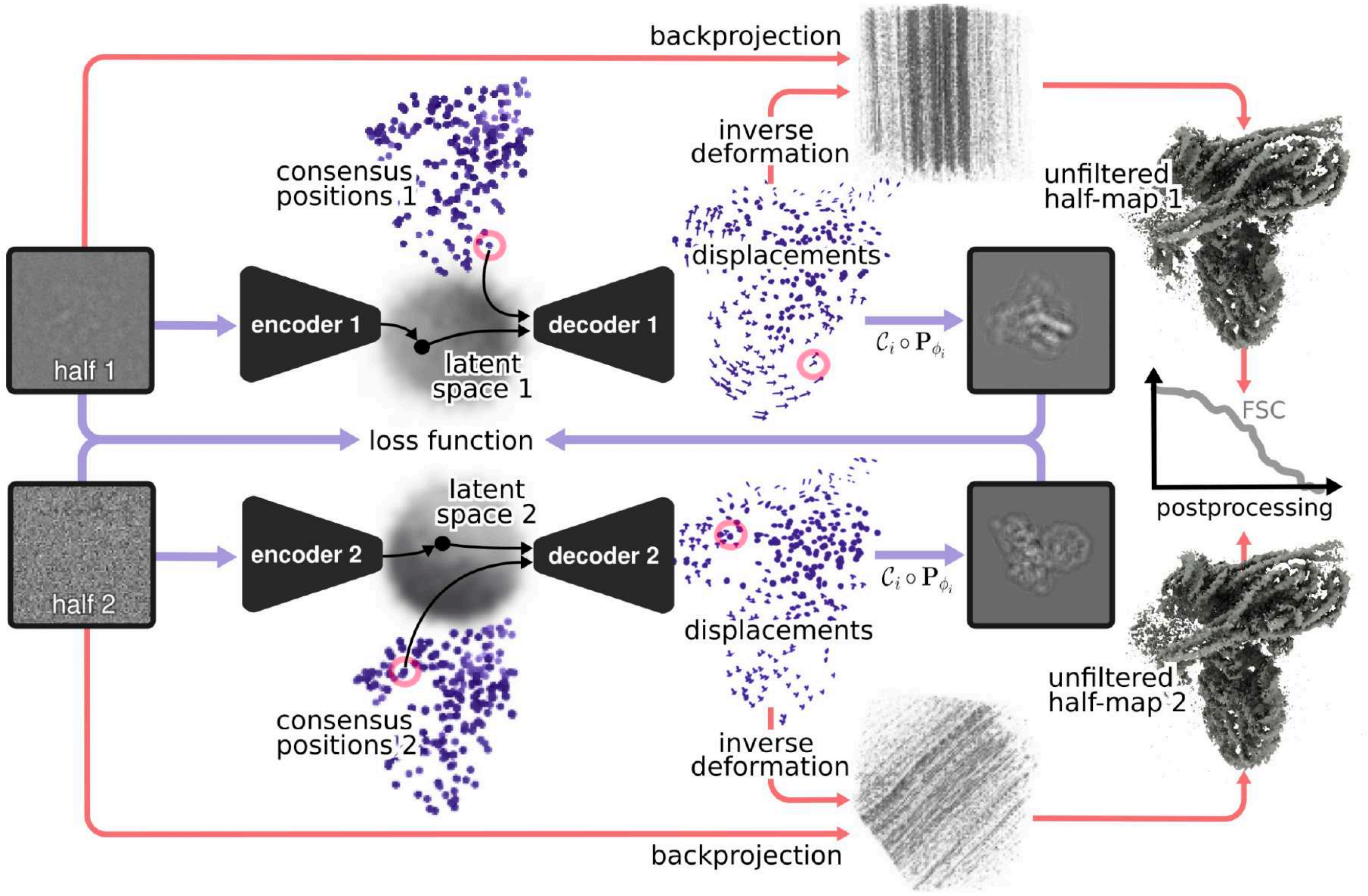
Sjors Scheres
@SjorsScheres

And our new #DynaMight paper is out on @biorxivpreprint! 🥳This is #RELION5's answer to modelling molecular flexibility by the amazing Johannes Schwab. He also shows that model bias can be nasty when modelling molecular flexibility with many parameters.

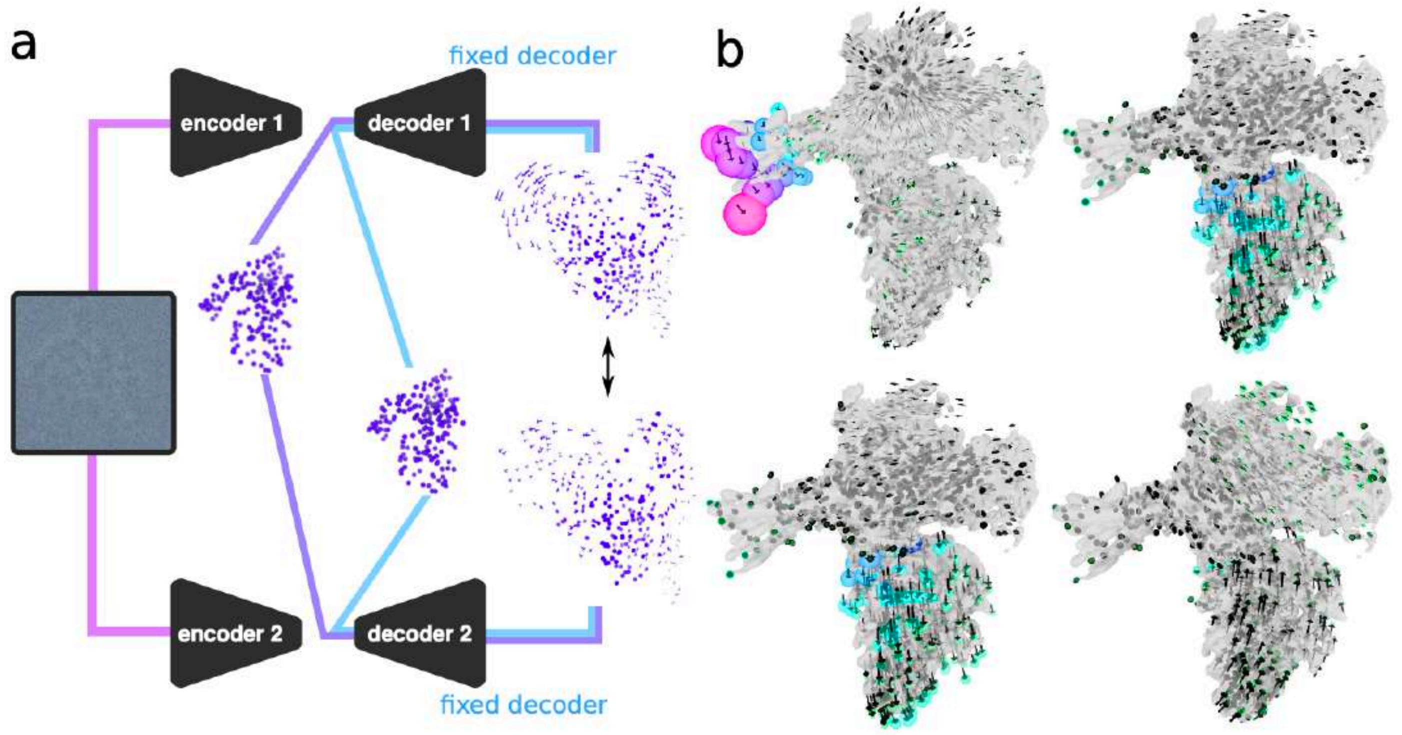
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

biorxiv.org
DynaMight: estimating molecular motions with improved rec
How to deal with continuously flexing molecules is one of the biggest outstanding challenges in single-particle ...

11:29 AM · Oct 19, 2023 · **24.6K** Views

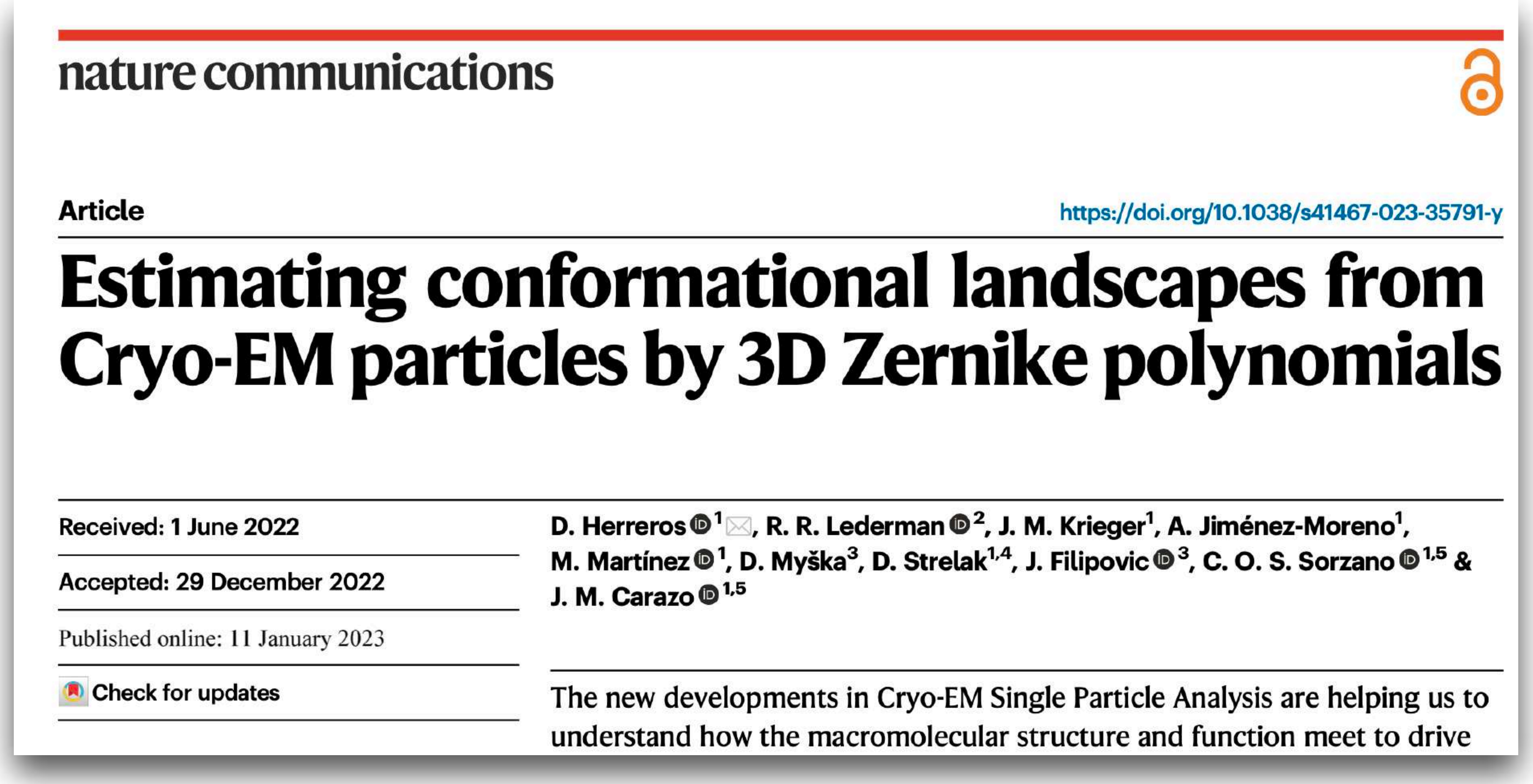


Has a method for
error estimation of the deformations.



Implemented in RELION-5

Multiple Methods in SciPion



nature communications

Article

https://doi.org/10.1038/s41467-023-35791-y

Estimating conformational landscapes from Cryo-EM particles by 3D Zernike polynomials

Received: 1 June 2022

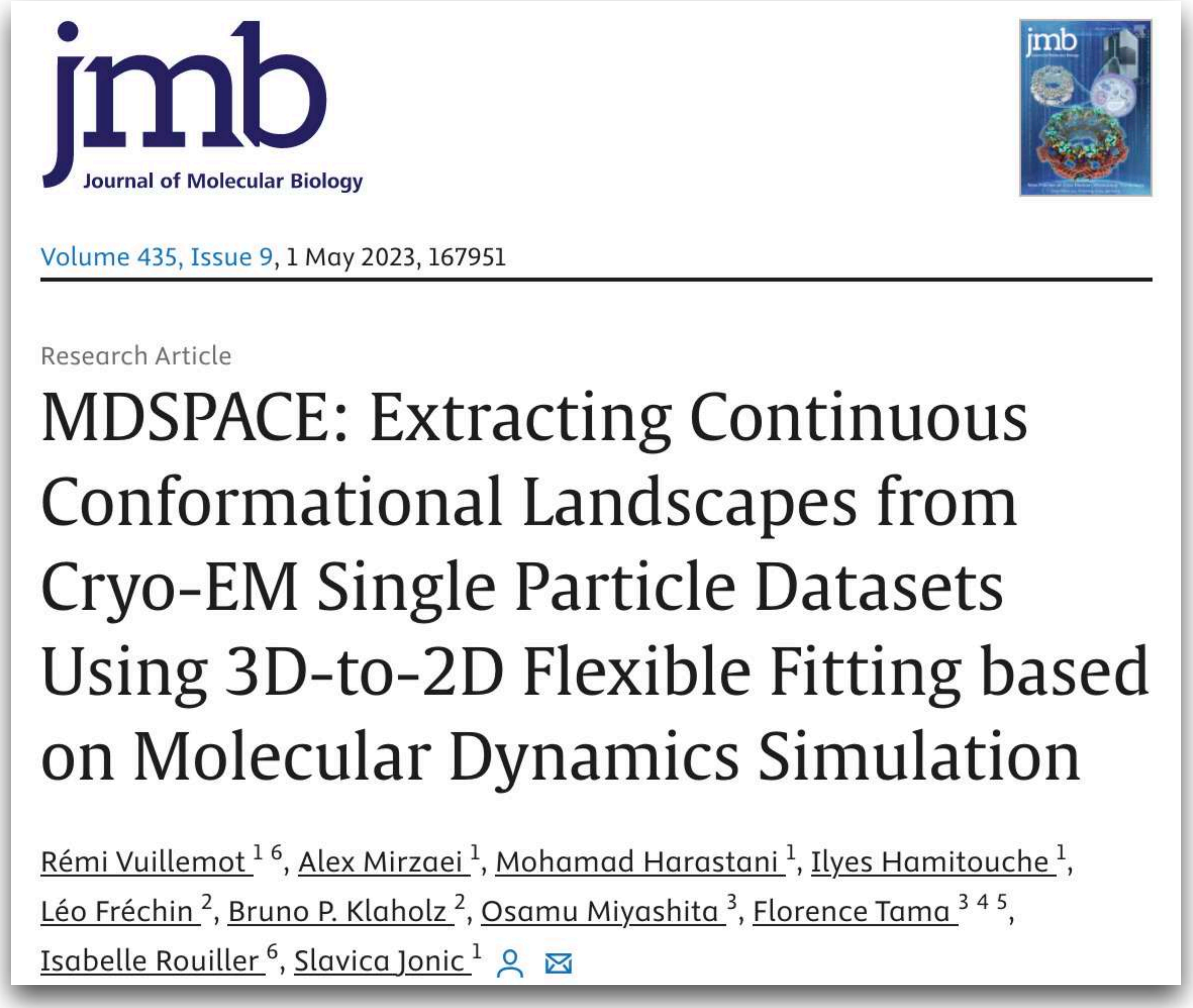
Accepted: 29 December 2022

Published online: 11 January 2023

Check for updates

D. Herreros[1], R. R. Lederman[2], J. M. Krieger[1], A. Jiménez-Moreno[1], M. Martínez[1], D. Myška[3], D. Strelak[1,4], J. Filipovic[3], C. O. S. Sorzano[1,5] & J. M. Carazo[1,5]

The new developments in Cryo-EM Single Particle Analysis are helping us to understand how the macromolecular structure and function meet to drive



jmb
Journal of Molecular Biology

Volume 435, Issue 9, 1 May 2023, 167951

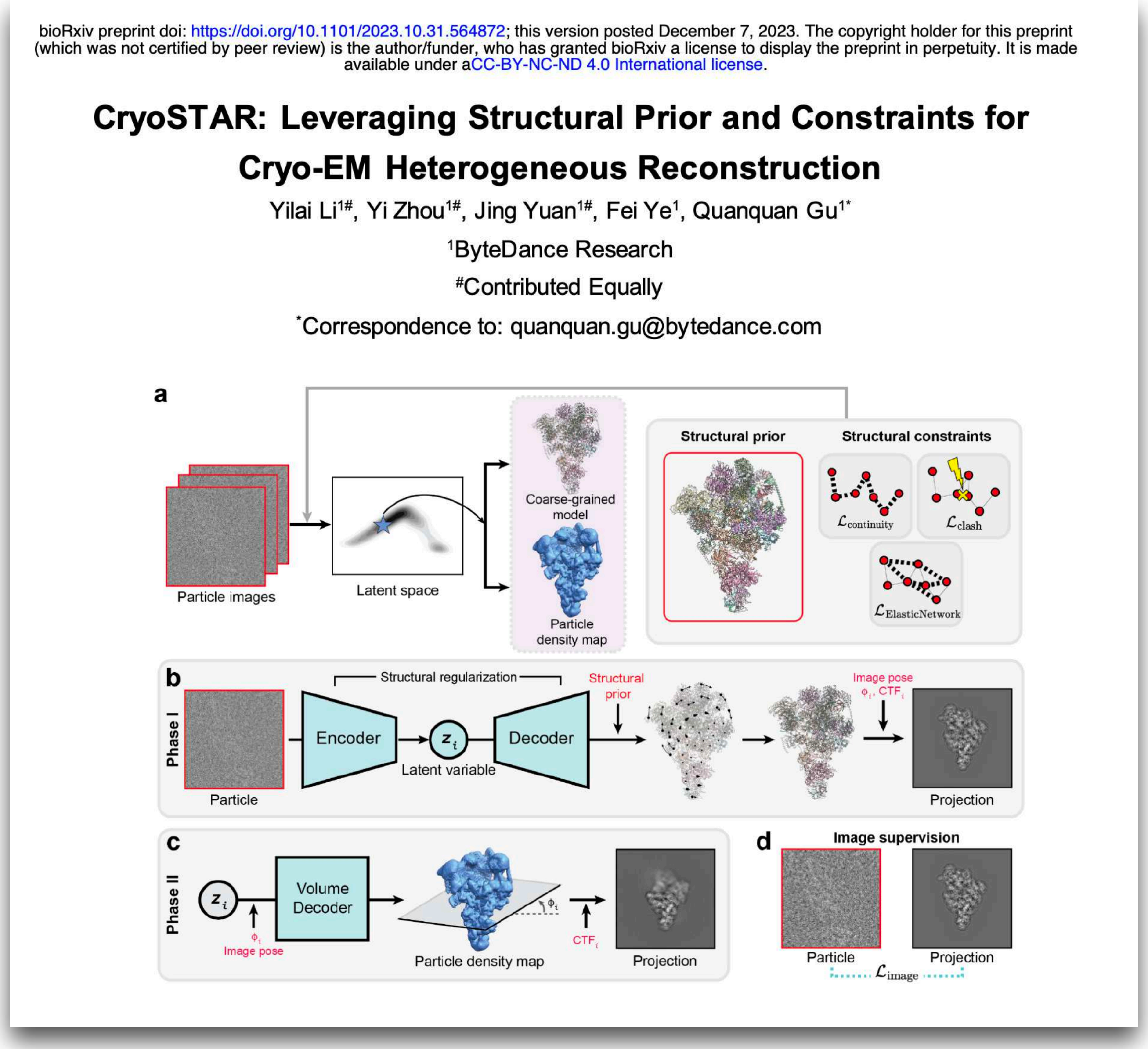
Research Article

MDSPACE: Extracting Continuous Conformational Landscapes from Cryo-EM Single Particle Datasets Using 3D-to-2D Flexible Fitting based on Molecular Dynamics Simulation

Rémi Vuillemot[1 6], Alex Mirzaei[1], Mohamad Harastani[1], Ilyes Hamitouche[1], Léo Fréchin[2], Bruno P. Klaholz[2], Osamu Miyashita[3], Florence Tama[3 4 5], Isabelle Rouiller[6], Slavica Jonic[1]

https://github.com/scipion-em/scipion-em-continuousflex

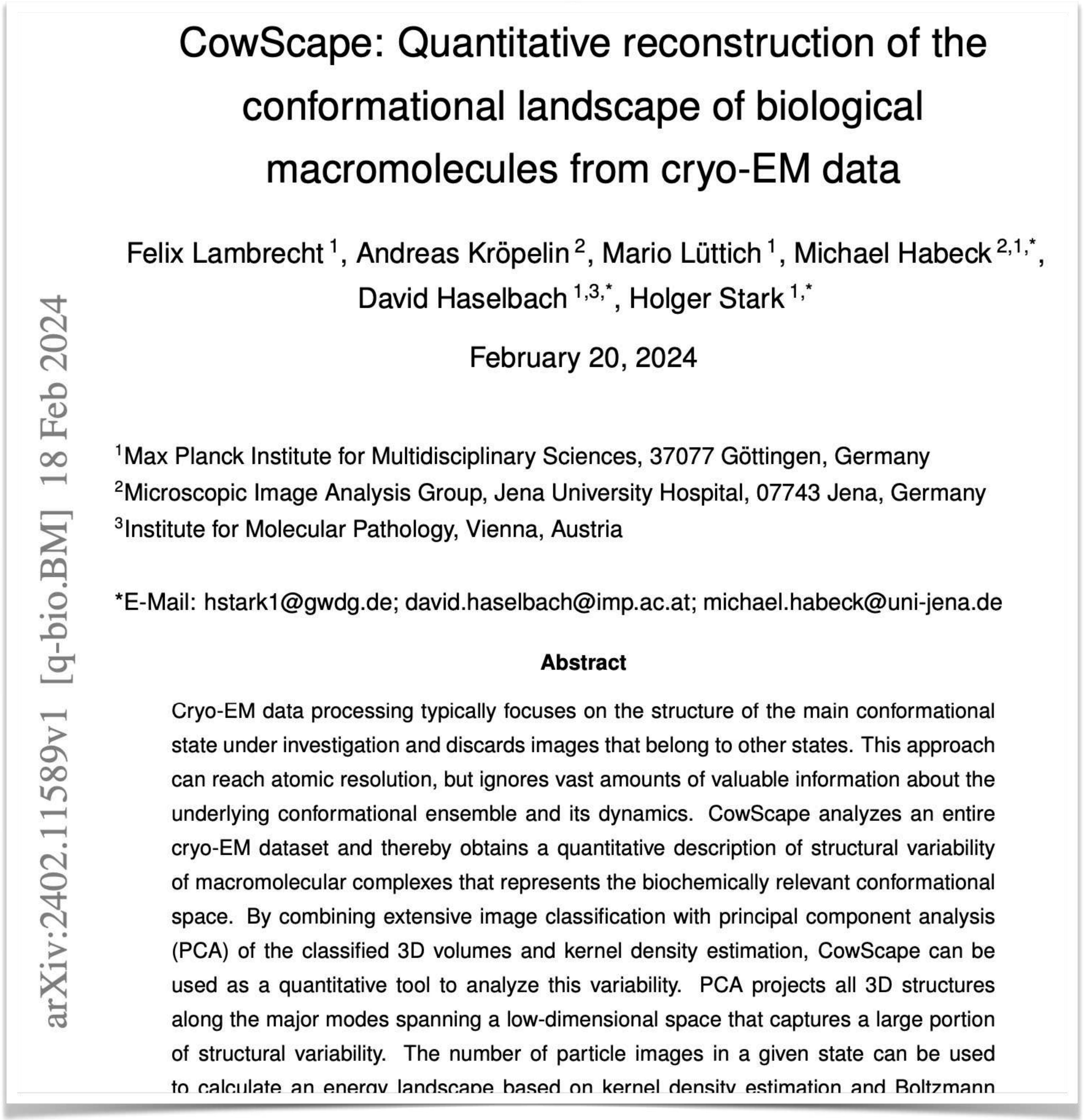
TikTok and Cows?



bioRxiv preprint doi: https://doi.org/10.1101/2023.10.31.564872; this version posted December 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CryoSTAR: Leveraging Structural Prior and Constraints for Cryo-EM Heterogeneous Reconstruction

Yilai Li[1#], Yi Zhou[1#], Jing Yuan[1#], Fei Ye[1], Quanquan Gu[1*]
[1]ByteDance Research
[#]Contributed Equally
[*]Correspondence to: quanquan.gu@bytedance.com



CowScape: Quantitative reconstruction of the conformational landscape of biological macromolecules from cryo-EM data

Felix Lambrecht[1], Andreas Kröpelin[2], Mario Lüttich[1], Michael Habeck[2,1,*], David Haselbach[1,3,*], Holger Stark[1,*]

February 20, 2024

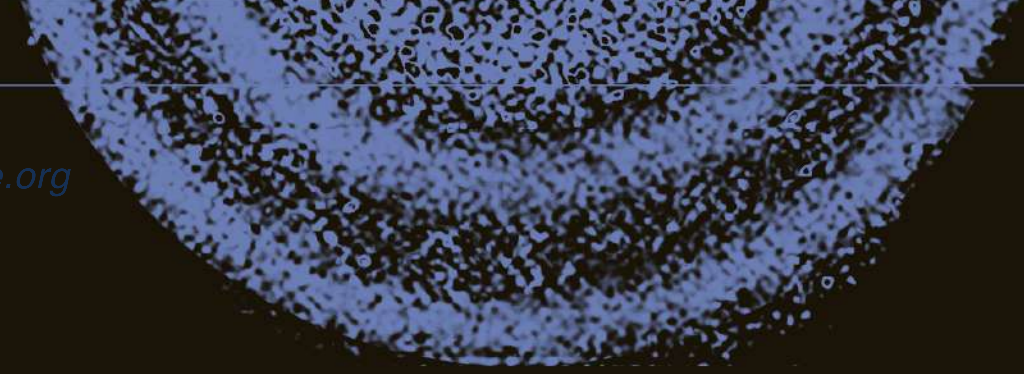
[1]Max Planck Institute for Multidisciplinary Sciences, 37077 Göttingen, Germany
[2]Microscopic Image Analysis Group, Jena University Hospital, 07743 Jena, Germany
[3]Institute for Molecular Pathology, Vienna, Austria

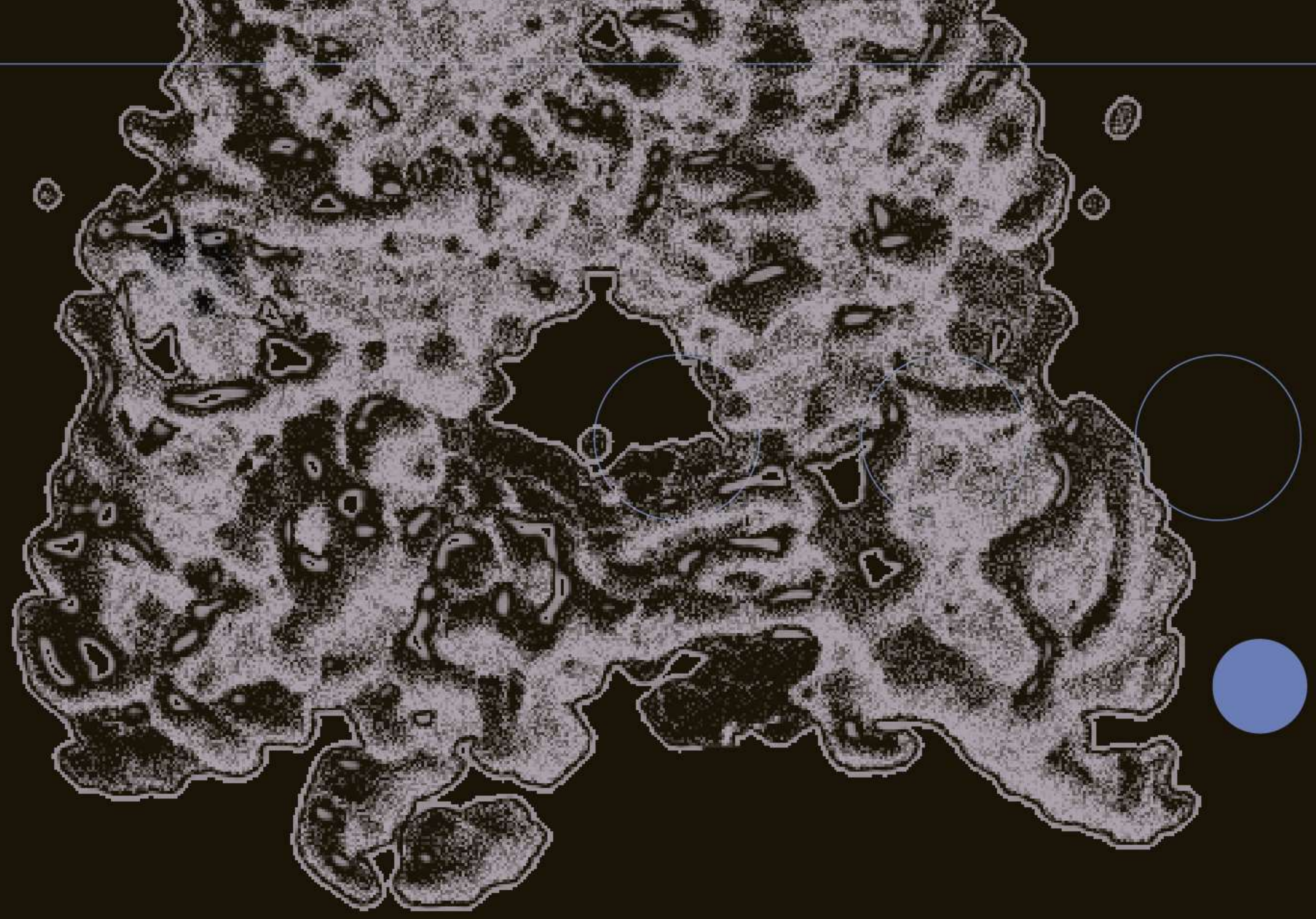
*E-Mail: hstark1@gwdg.de; david.haselbach@imp.ac.at; michael.habeck@uni-jena.de

arXiv:2402.11589v1 [q-bio.BM] 18 Feb 2024

Abstract

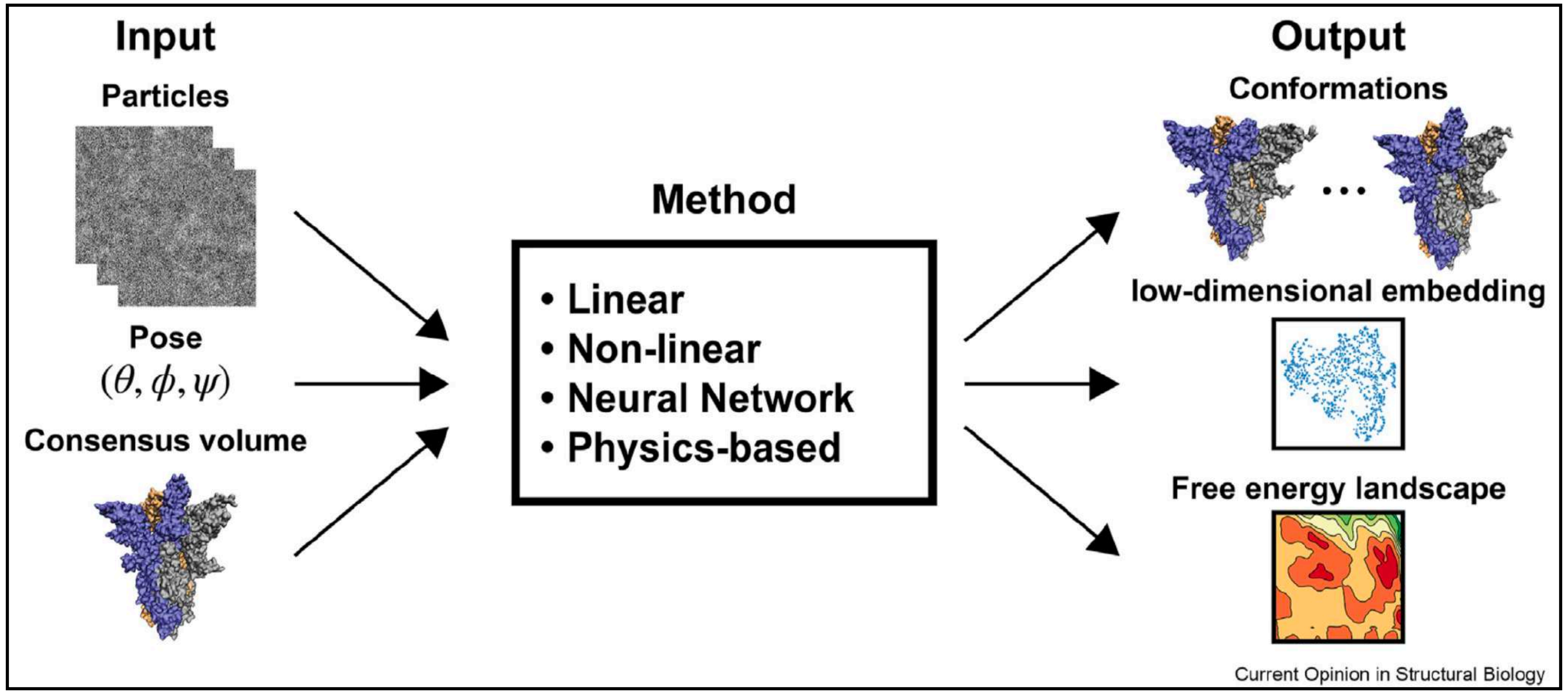
Cryo-EM data processing typically focuses on the structure of the main conformational state under investigation and discards images that belong to other states. This approach can reach atomic resolution, but ignores vast amounts of valuable information about the underlying conformational ensemble and its dynamics. CowScape analyzes an entire cryo-EM dataset and thereby obtains a quantitative description of structural variability of macromolecular complexes that represents the biochemically relevant conformational space. By combining extensive image classification with principal component analysis (PCA) of the classified 3D volumes and kernel density estimation, CowScape can be used as a quantitative tool to analyze this variability. PCA projects all 3D structures along the major modes spanning a low-dimensional space that captures a large portion of structural variability. The number of particle images in a given state can be used to calculate an energy landscape based on kernel density estimation and Boltzmann





Practical considerations when using continuous heterogeneity methods.

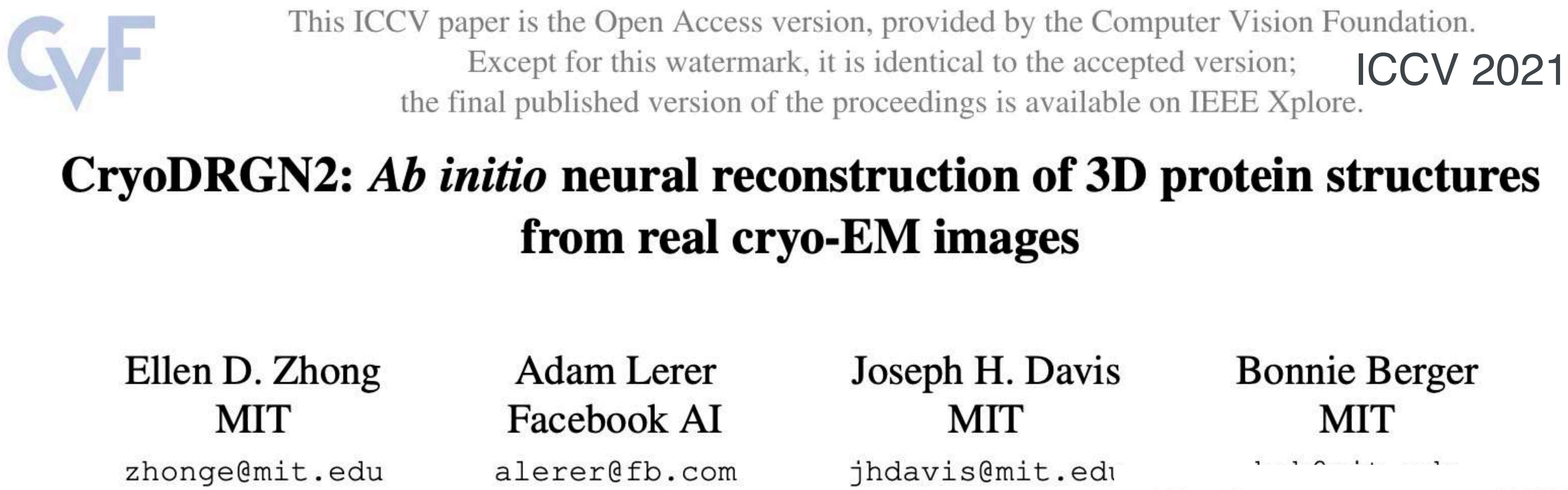
Most methods take in particle stacks after 3D Refinement from a previous tool.



Current Opinion in Structural Biology

- Many methods are already integrated into other tools that already do 3D Refinement (cryoSPARC/RELION/EMAN2) so this is straightforward
- cryoDRGN has a friendly easily installable Python pipeline and plays a little nicer with cryoSPARC inputs

Procedures to jointly optimize pose and conformation are still being improved.



This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ICCV 2021

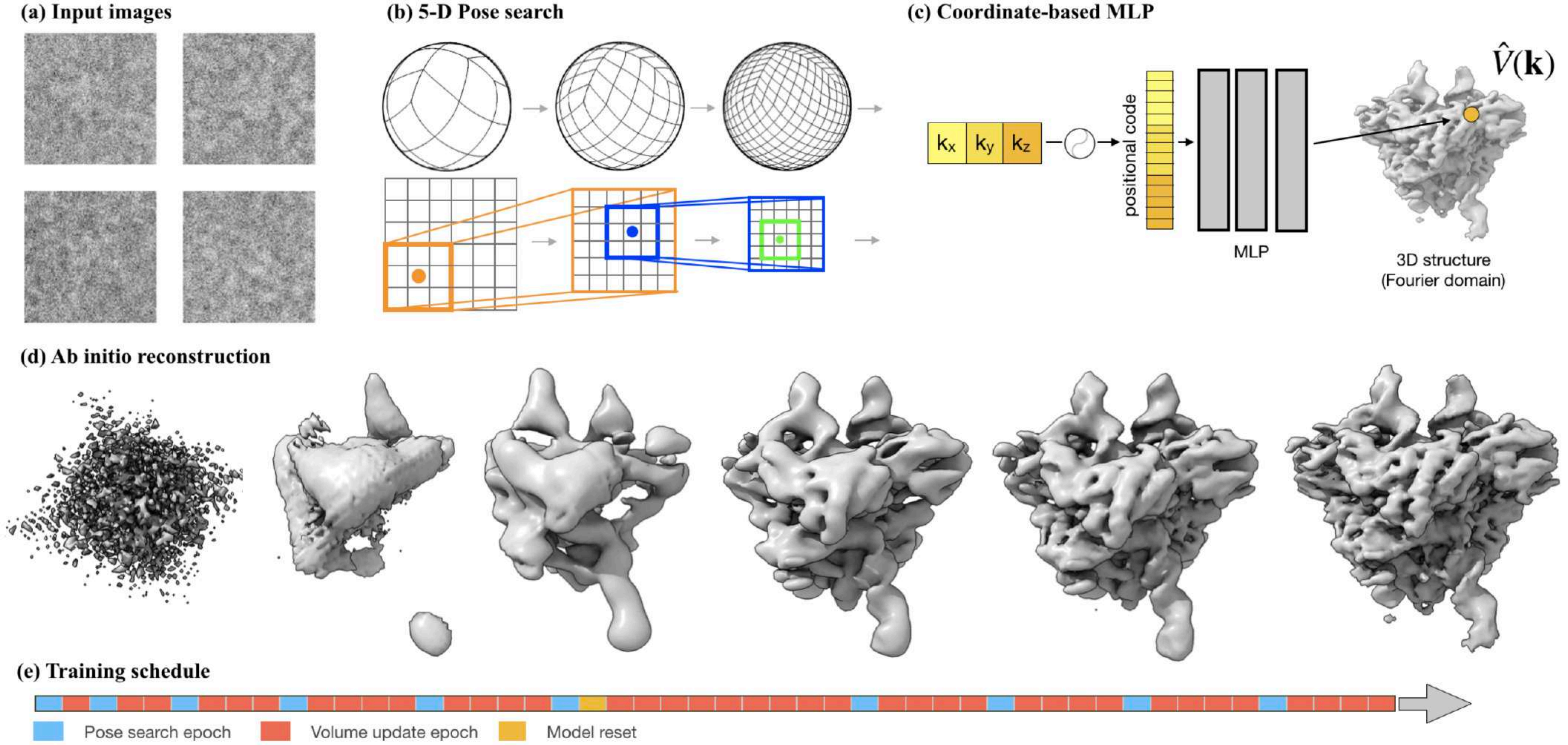
CryoDRGN2: *Ab initio* neural reconstruction of 3D protein structures from real cryo-EM images

Ellen D. Zhong
MIT
zhonge@mit.edu

Adam Lerer
Facebook AI
alerer@fb.com

Joseph H. Davis
MIT
jhdavis@mit.ed

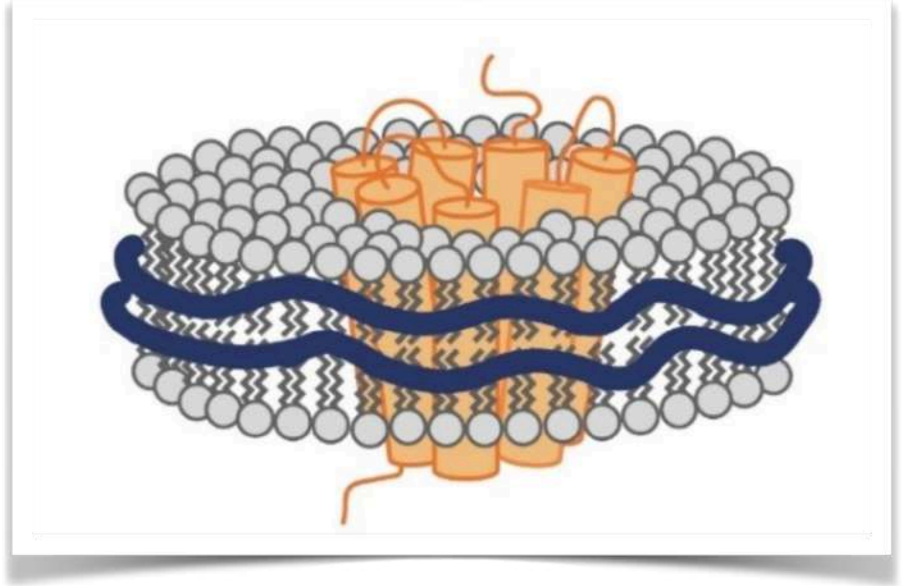
Bonnie Berger
MIT

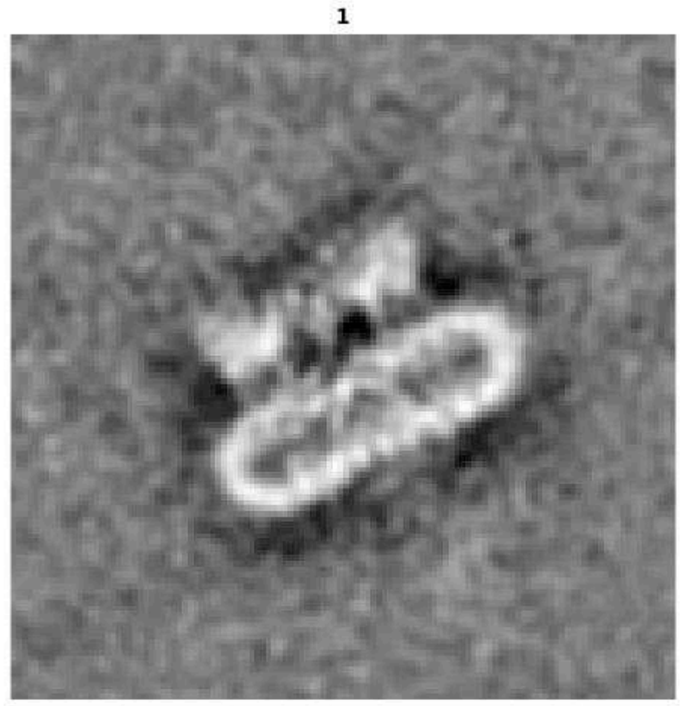


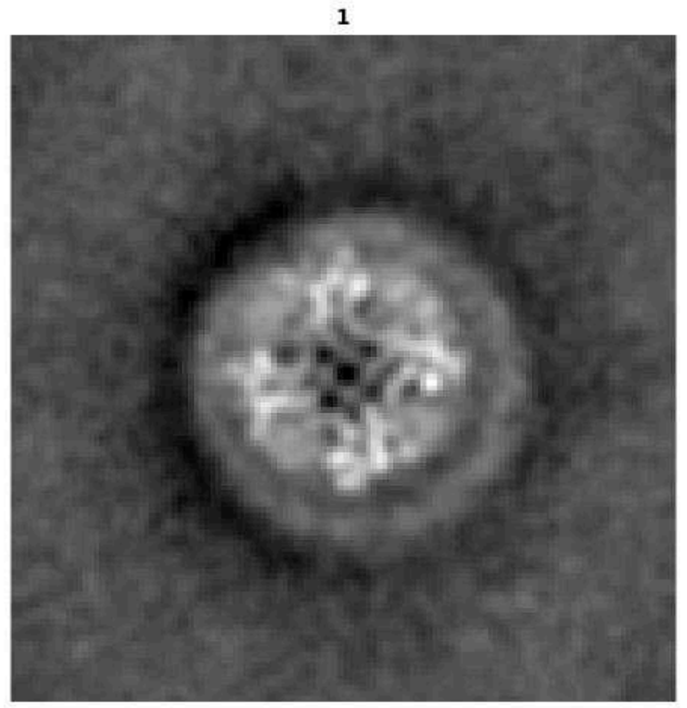
Nanodiscs can also be a challenge for continuous heterogeneity methods.

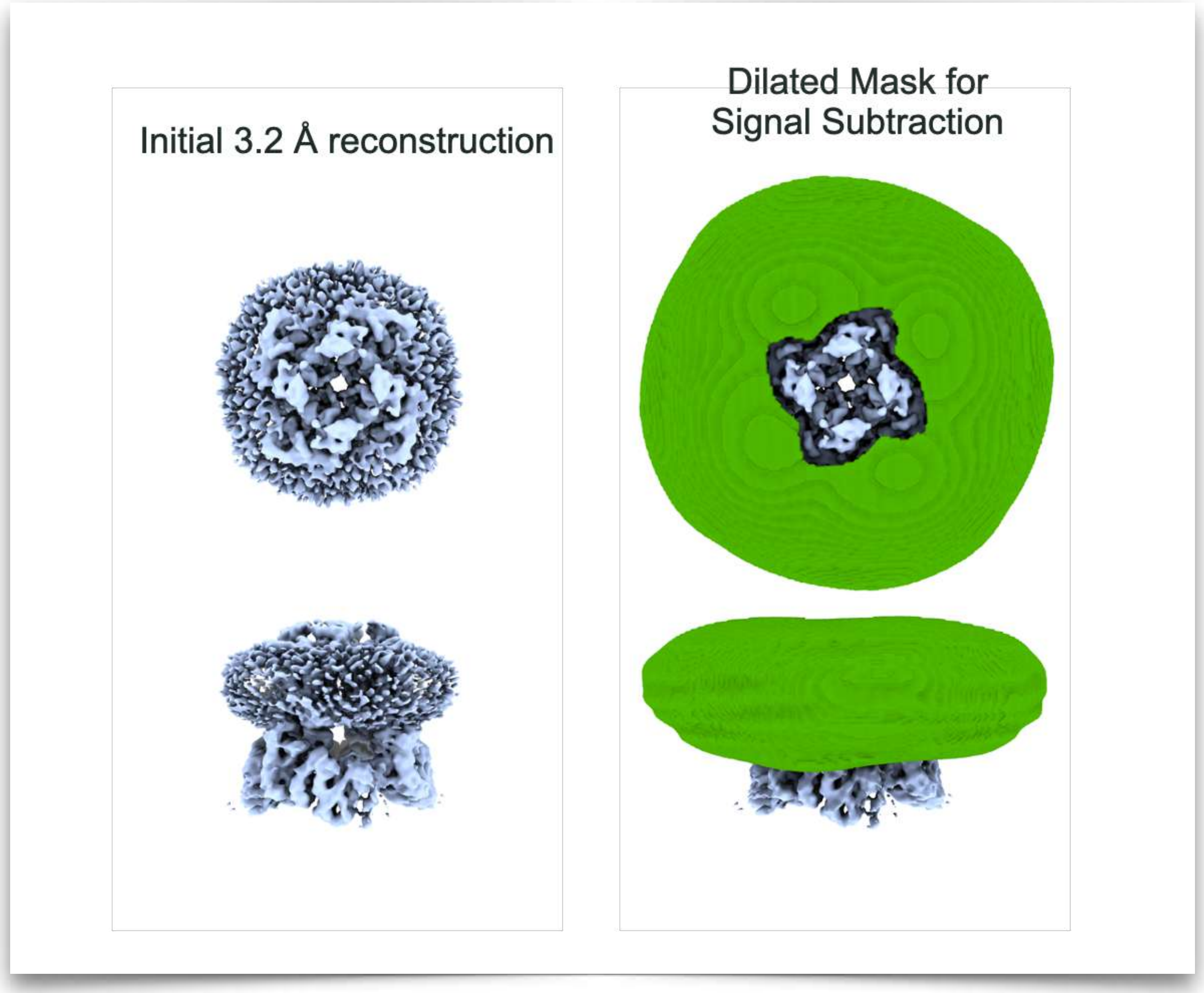
Analysis with ManifoldEM Matlab Code

rTRPV1 + DkTx/RtX
in nanodiscs
(EMPIAR 10059)



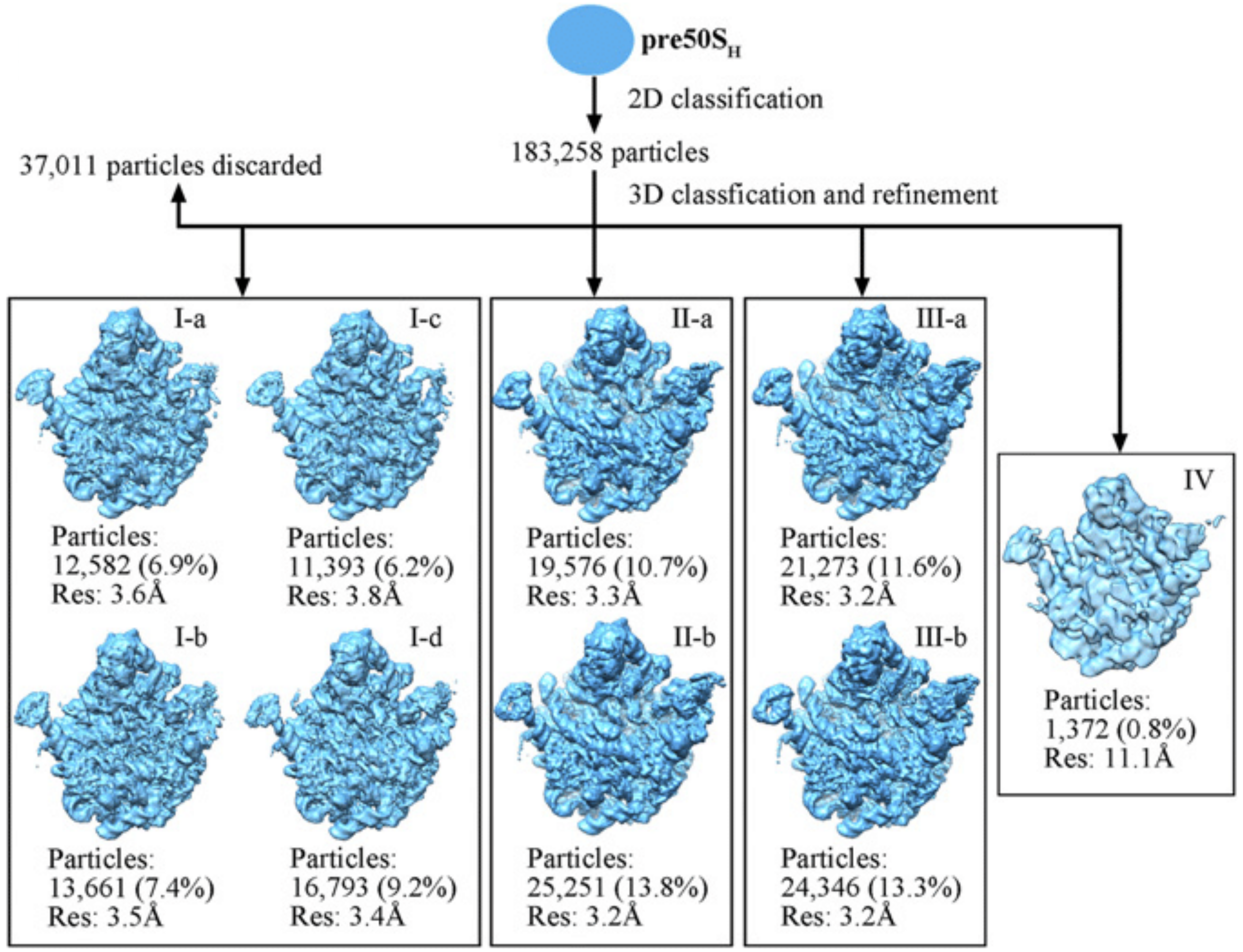


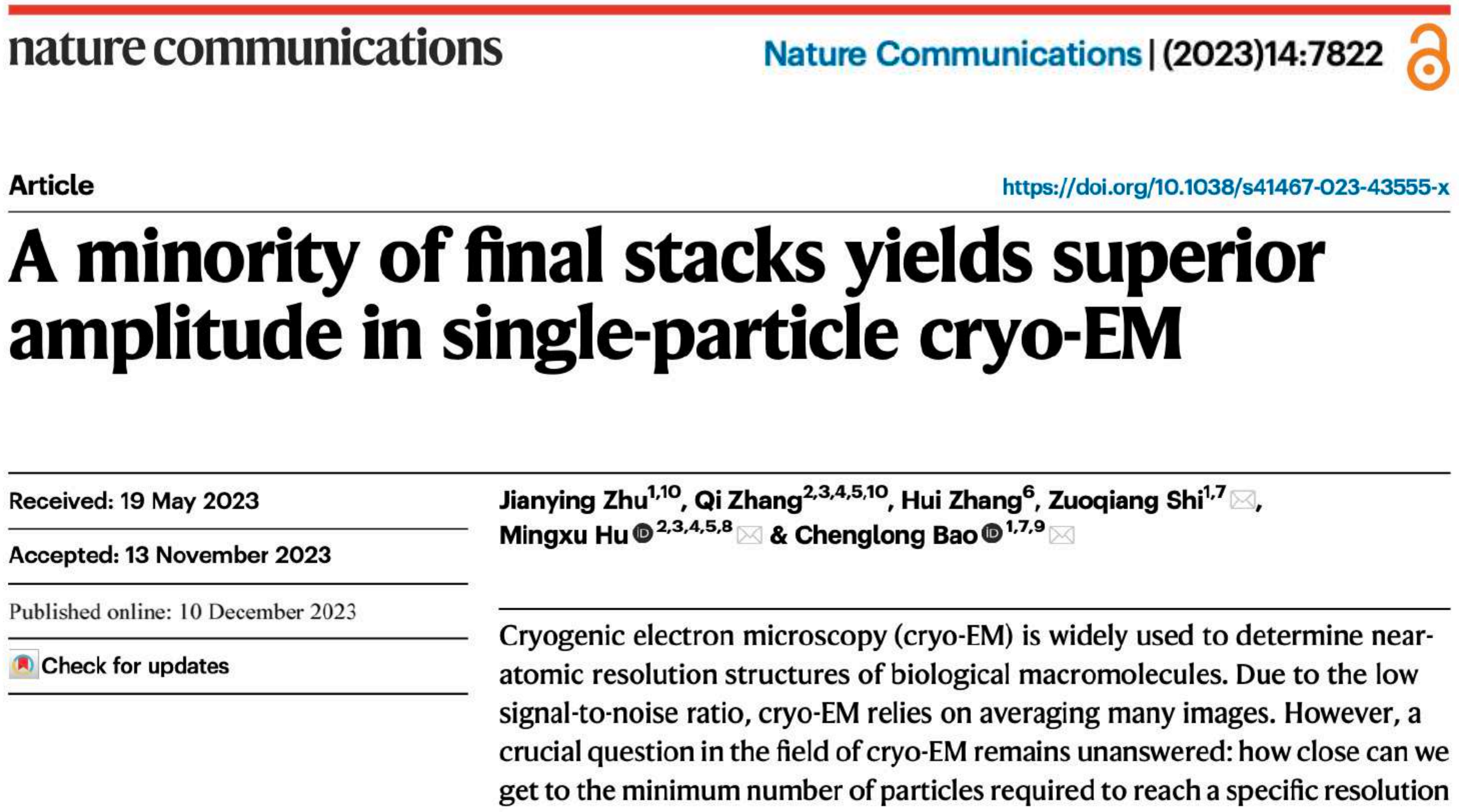




Using all vs. a subset of your data.

- Getting rid of junk is important.
- Making sure there isn't residual compositional heterogeneity is important
- Intuitively you wouldn't want particles too tightly classified in a single conformational state, but this is a little tricky to assess




nature communications

Nature Communications | (2023)14:7822

Article

https://doi.org/10.1038/s41467-023-43555-x

A minority of final stacks yields superior amplitude in single-particle cryo-EM

Received: 19 May 2023

Accepted: 13 November 2023

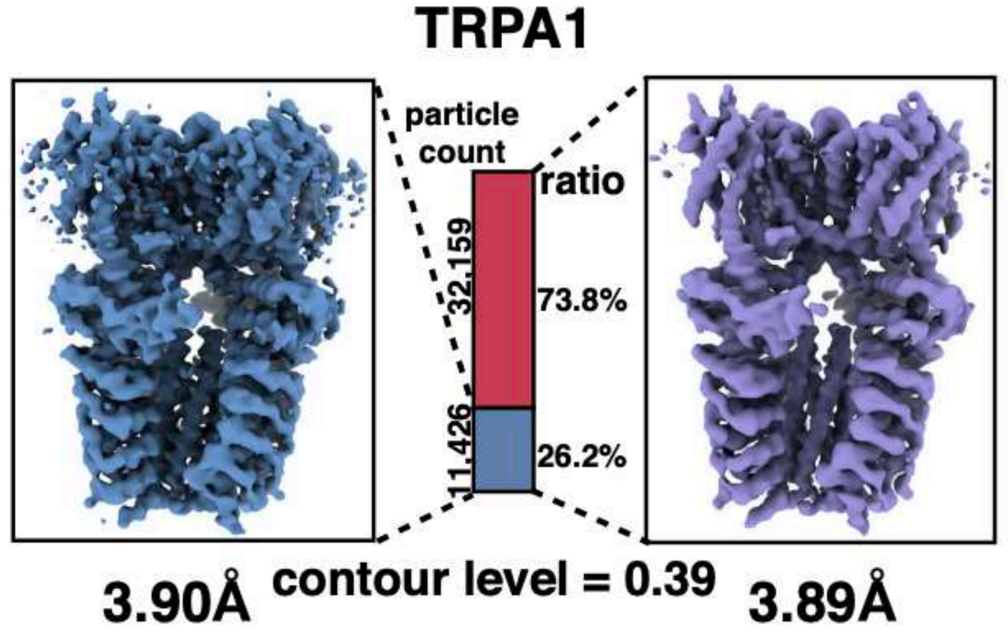
Published online: 10 December 2023

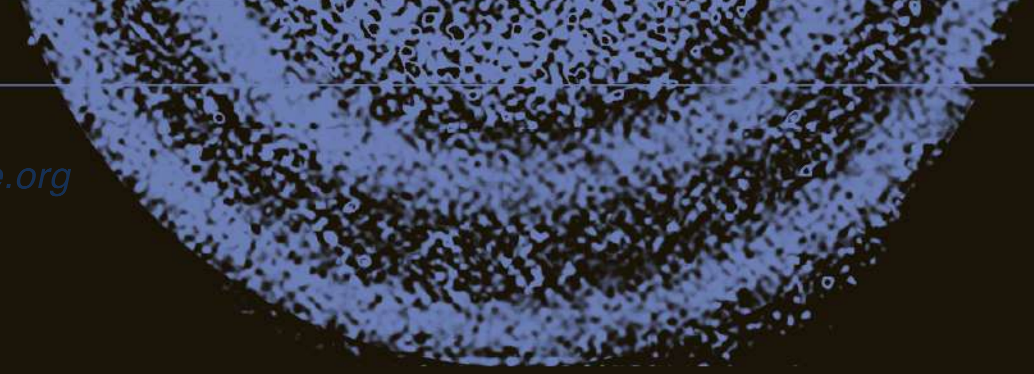
Check for updates

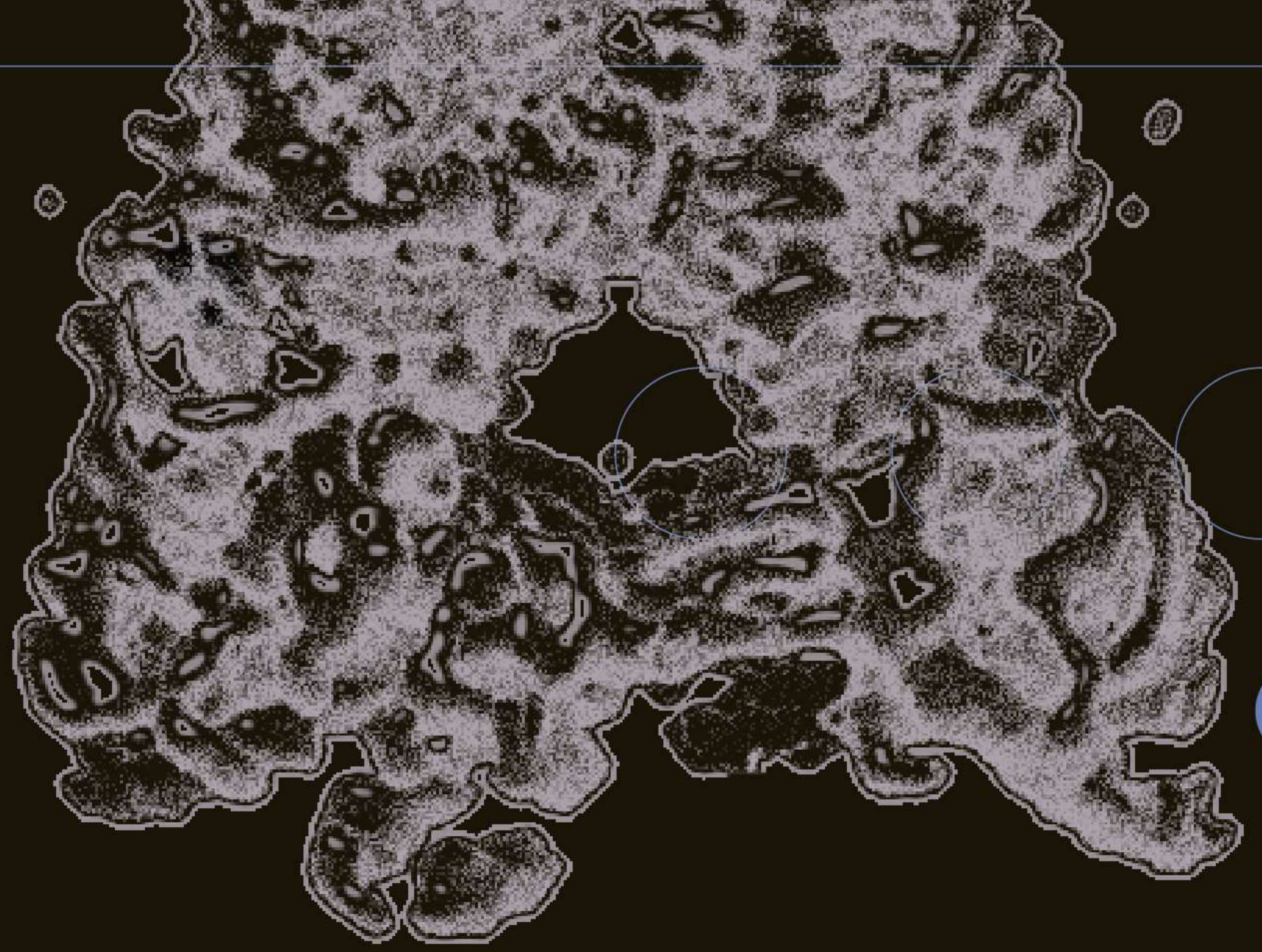
Jianying Zhu[1,10], Qi Zhang[2,3,4,5,10], Hui Zhang[6], Zuoqiang Shi[1,7], Mingxu Hu[2,3,4,5,8] & Chenglong Bao[1,7,9]

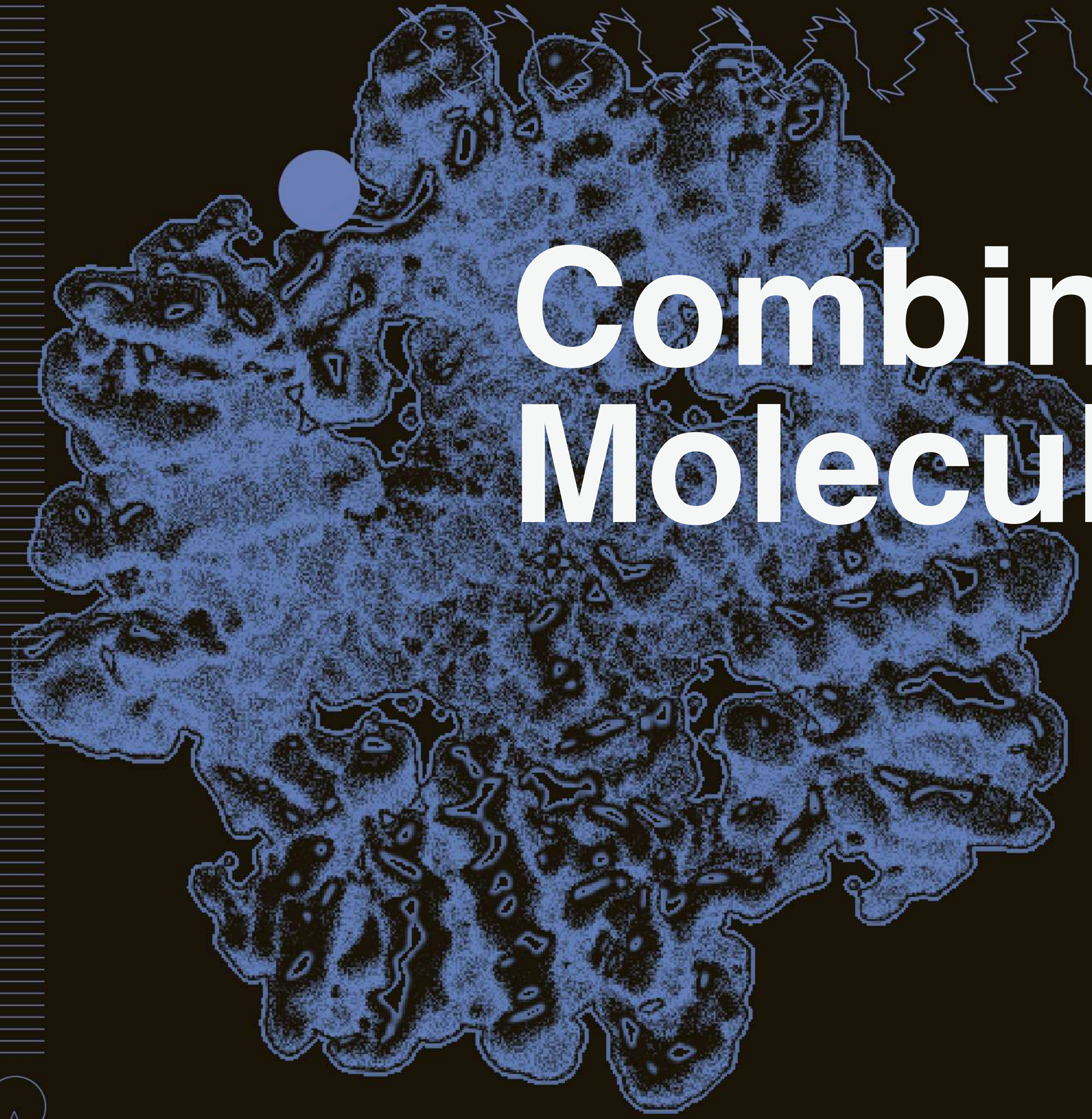
Cryogenic electron microscopy (cryo-EM) is widely used to determine near-atomic resolution structures of biological macromolecules. Due to the low signal-to-noise ratio, cryo-EM relies on averaging many images. However, a crucial question in the field of cryo-EM remains unanswered: how close can we get to the minimum number of particles required to reach a specific resolution

CryoSieve is a new tool that allows you to build a same resolution reconstruction with a fraction of the particles.



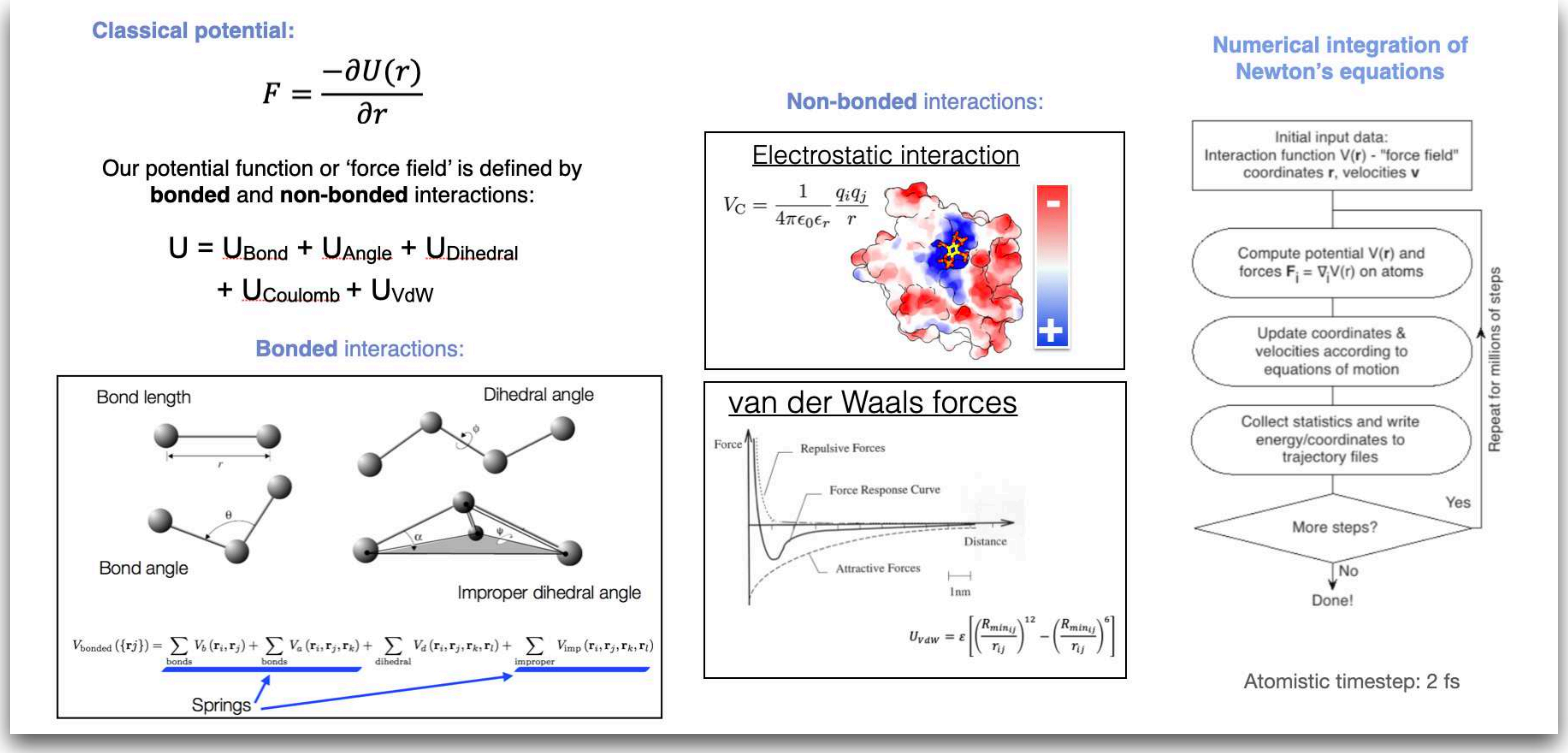






Combining cryo-EM and Molecular Dynamics Simulations

Molecular dynamics (MD) simulations as a key partner with cryo-EM for conformational heterogeneity analysis.



Classical potential:

$$F = \frac{-\partial U(r)}{\partial r}$$

Our potential function or 'force field' is defined by **bonded** and **non-bonded** interactions:

$$U = U_{Bond} + U_{Angle} + U_{Dihedral} + U_{Coulomb} + U_{VdW}$$

Bonded interactions:

$$V_{\text{bonded}}(\{\mathbf{r}j\}) = \underbrace{\sum_{\text{bonds}} V_b(\mathbf{r}_i, \mathbf{r}_j) + \sum_{\text{bonds}} V_a(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k)} + \sum_{\text{dihedral}} V_d(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) + \underbrace{\sum_{\text{improper}} V_{\text{imp}}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l)}$$

Springs

Non-bonded interactions:

Electrostatic interaction

$$V_C = \frac{1}{4\pi\epsilon_0\epsilon_r}\frac{q_i q_j}{r}$$

van der Waals forces

$$U_{vdw} = \varepsilon\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right]$$

Numerical integration of Newton's equations

Initial input data:
Interaction function V(**r**) - "force field"
coordinates **r**, velocities **v**

Compute potential V(**r**) and forces $\mathbf{F}_i = \nabla_i V(r)$ on atoms

Update coordinates & velocities according to equations of motion

Collect statistics and write energy/coordinates to trajectory files

More steps?

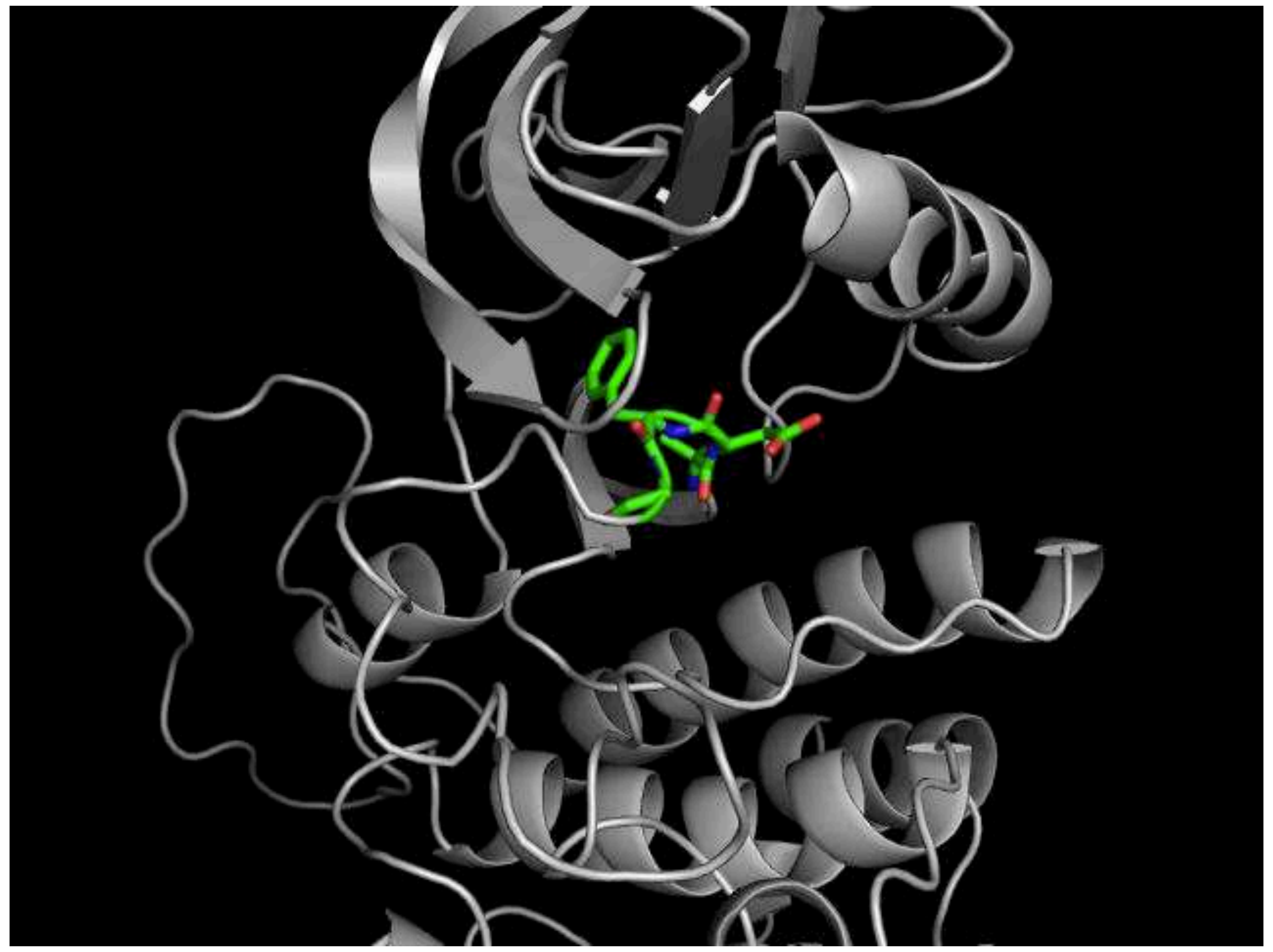
Yes

No

Done!

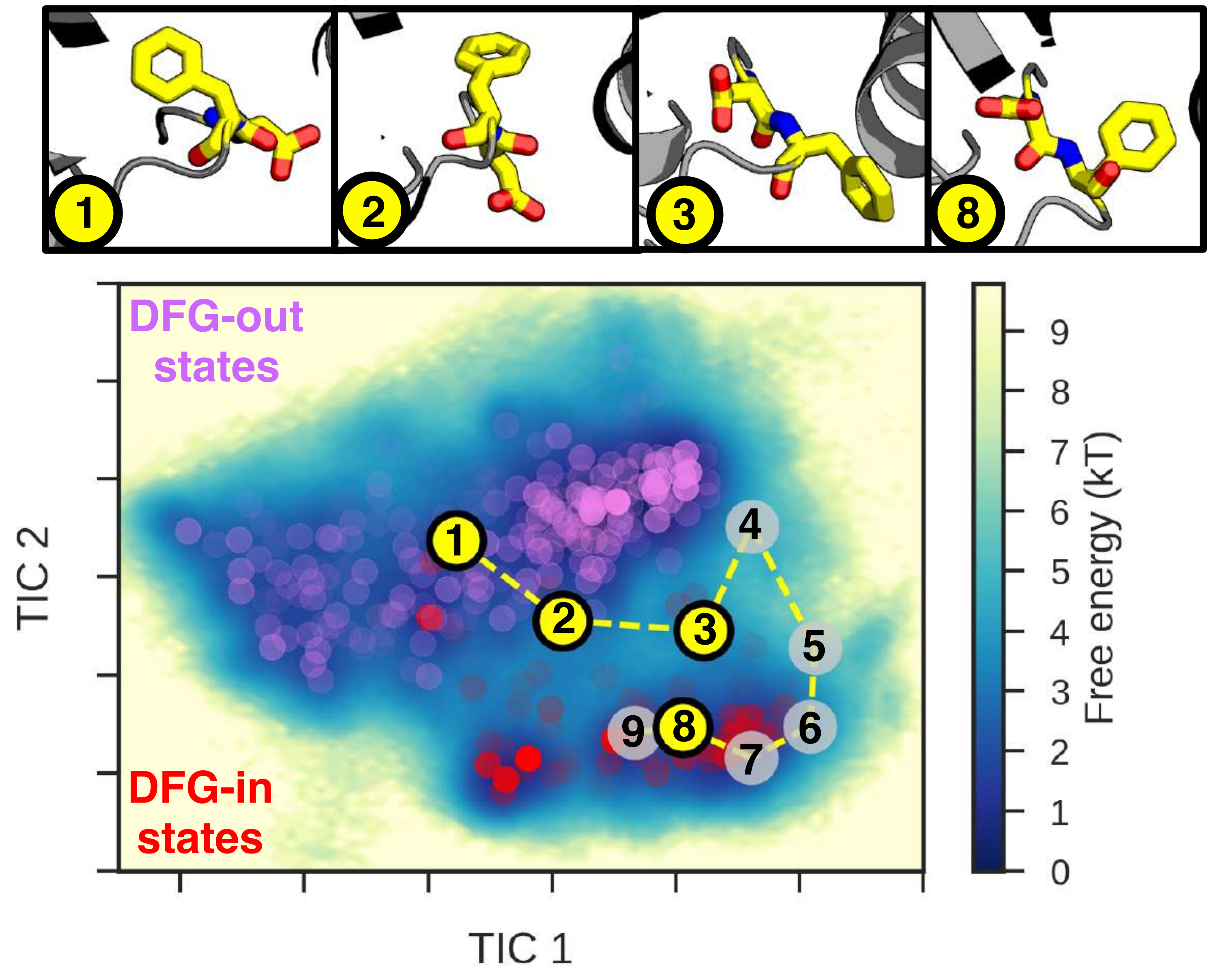
Repeat for millions of steps

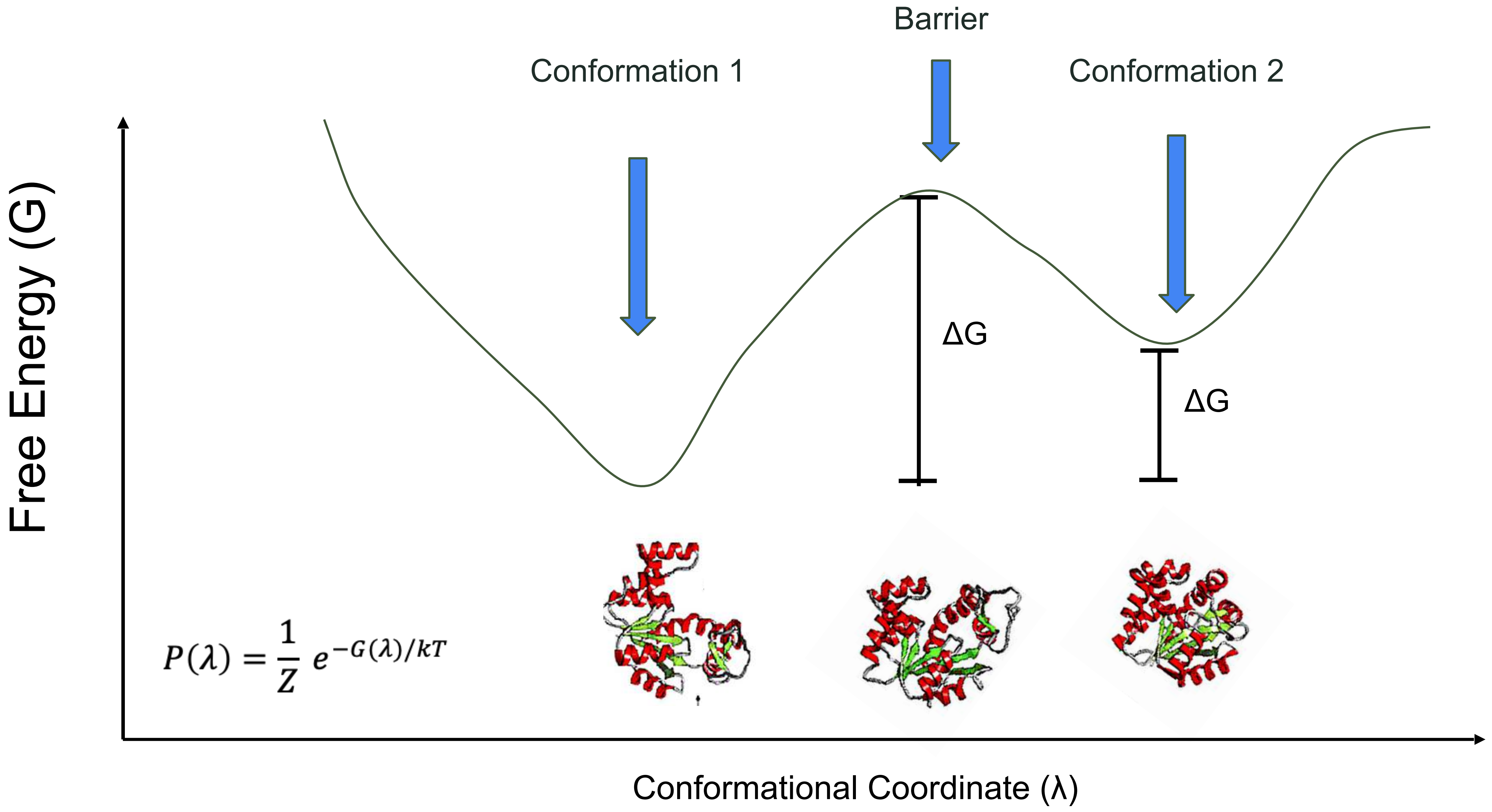
Atomistic timestep: 2 fs



Hanson*, Georghiou* et al, *Cell Chem Biol*, (2019).

In Molecular Dynamics we are used to think about Free Energies and populations of states.

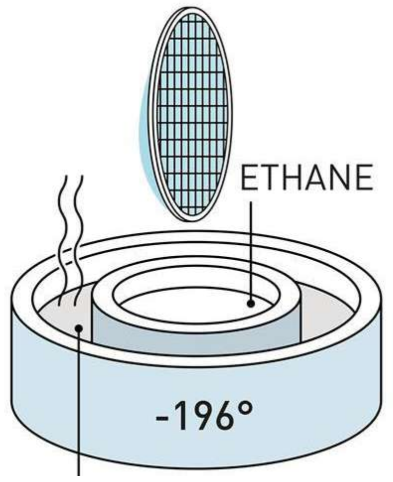




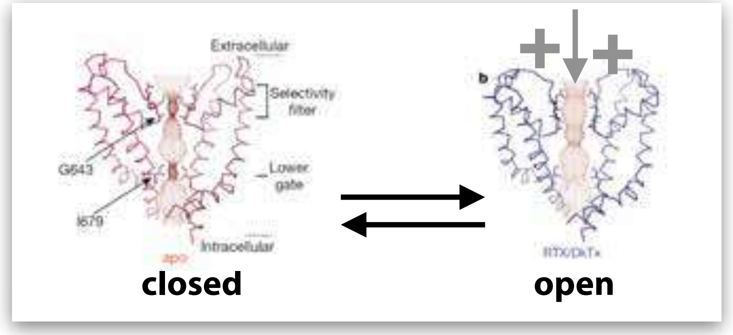
$$P(\lambda) = \frac{1}{Z}\, e^{-G(\lambda)/kT}$$

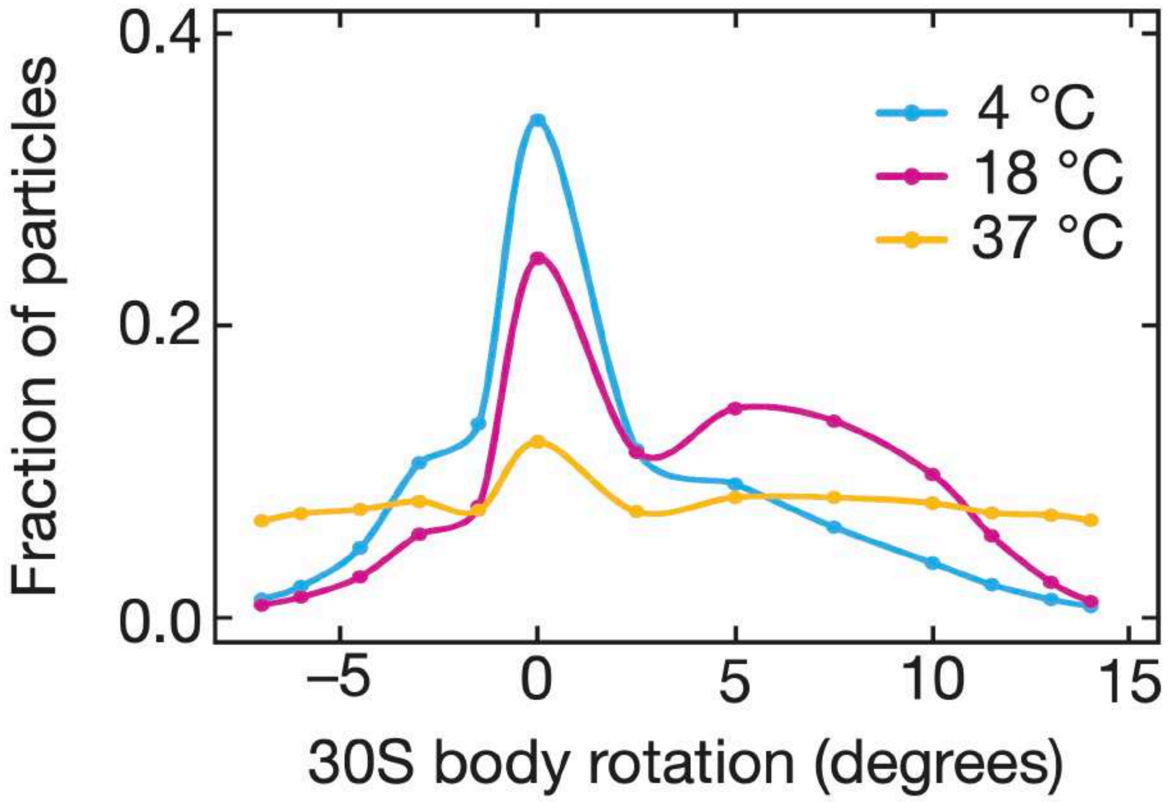
Hanson*, Georghiou* et al, *Cell Chem Biol*, (2019).

How much to worry about the ‘cryo’ of cryo-EM?



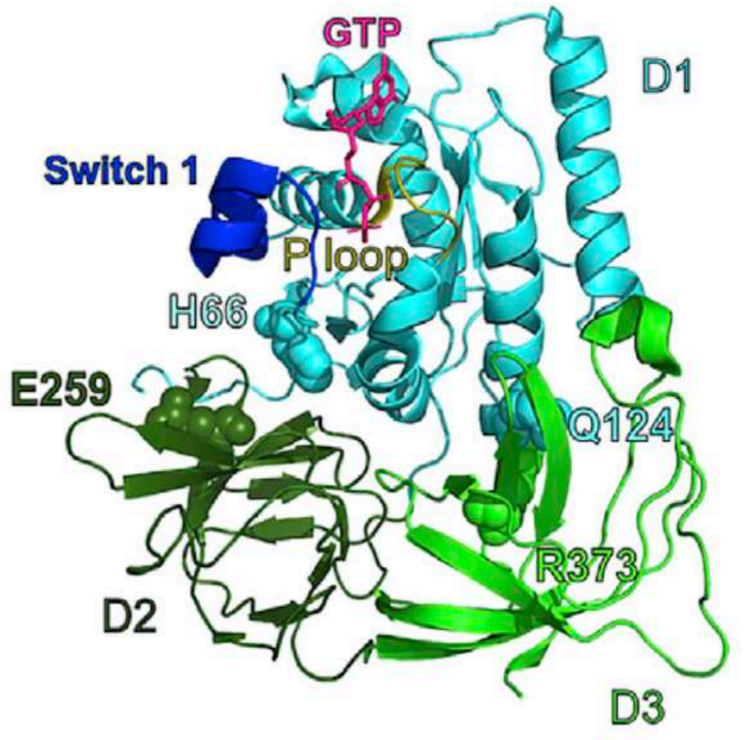
time to freeze during vitrification < 0.1 of ms*
current estimates of time scale of TRPV1 opening ~ 2-5 ms →



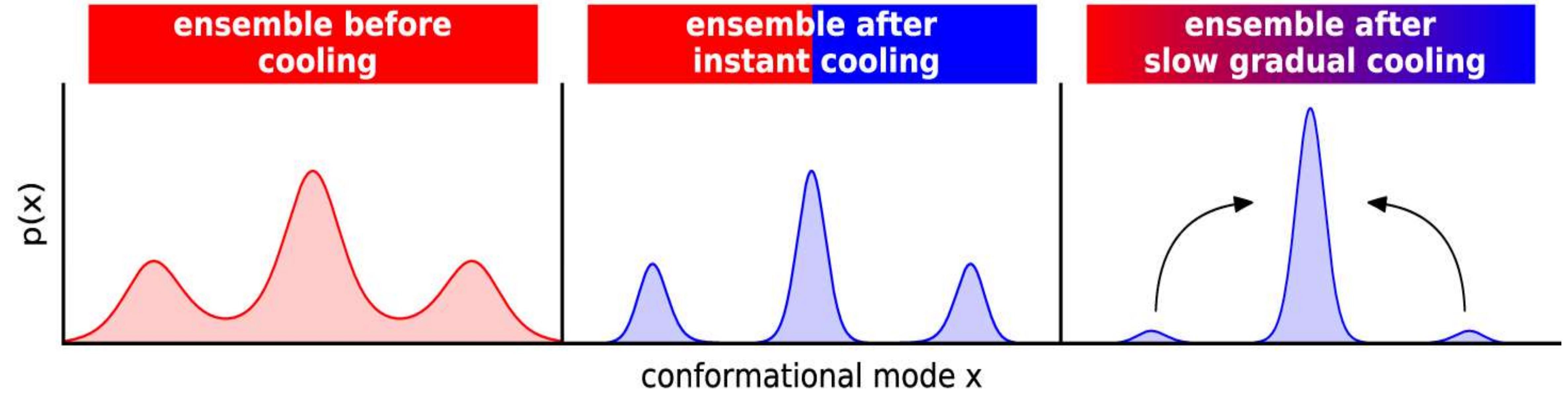


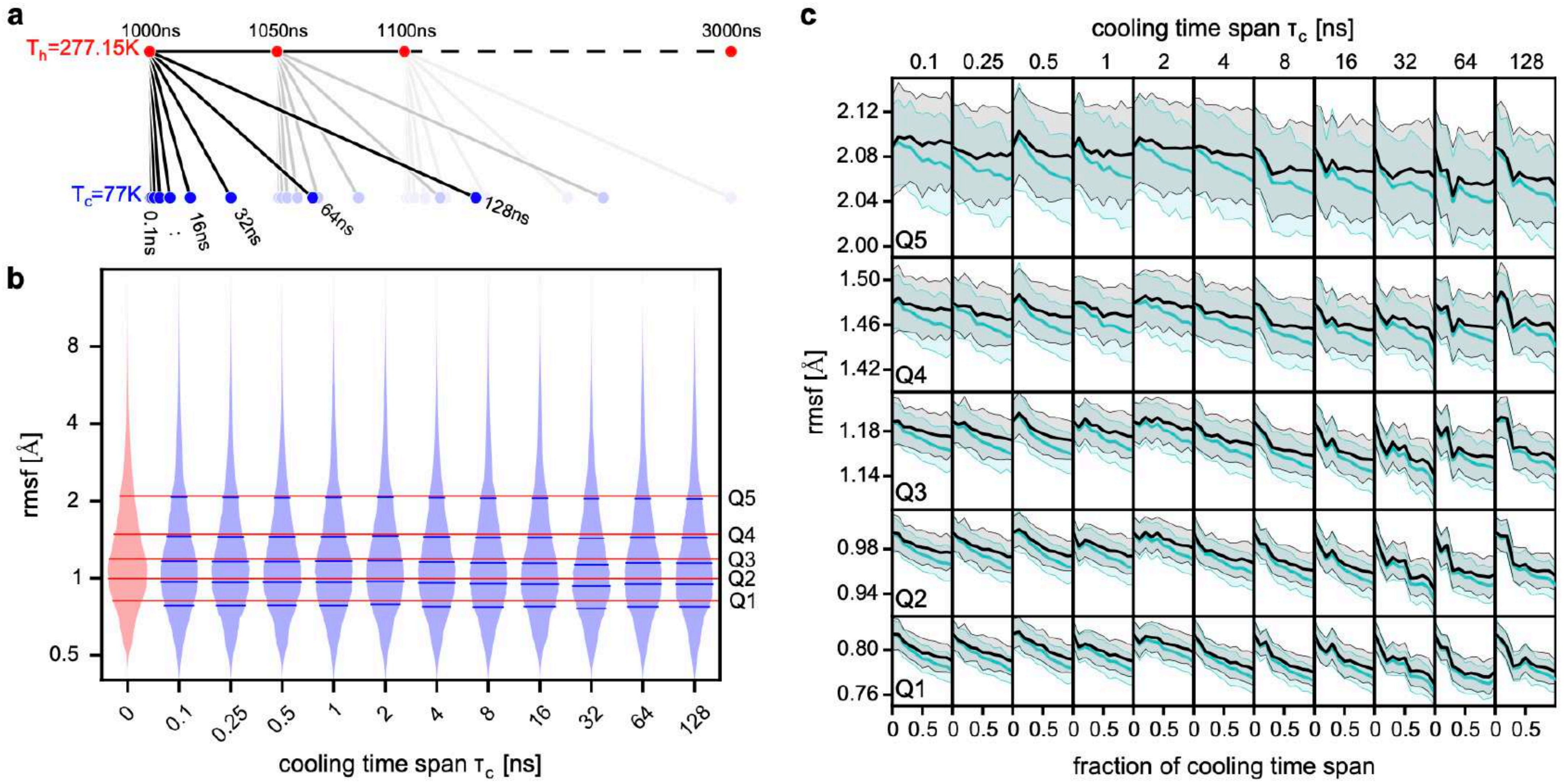
Fischer, Konevega, Wintermeyer, Rodnina,
& Stark, *Nature* (2010)

How much does vitrification perturb the equilibrium ensemble of our molecules?



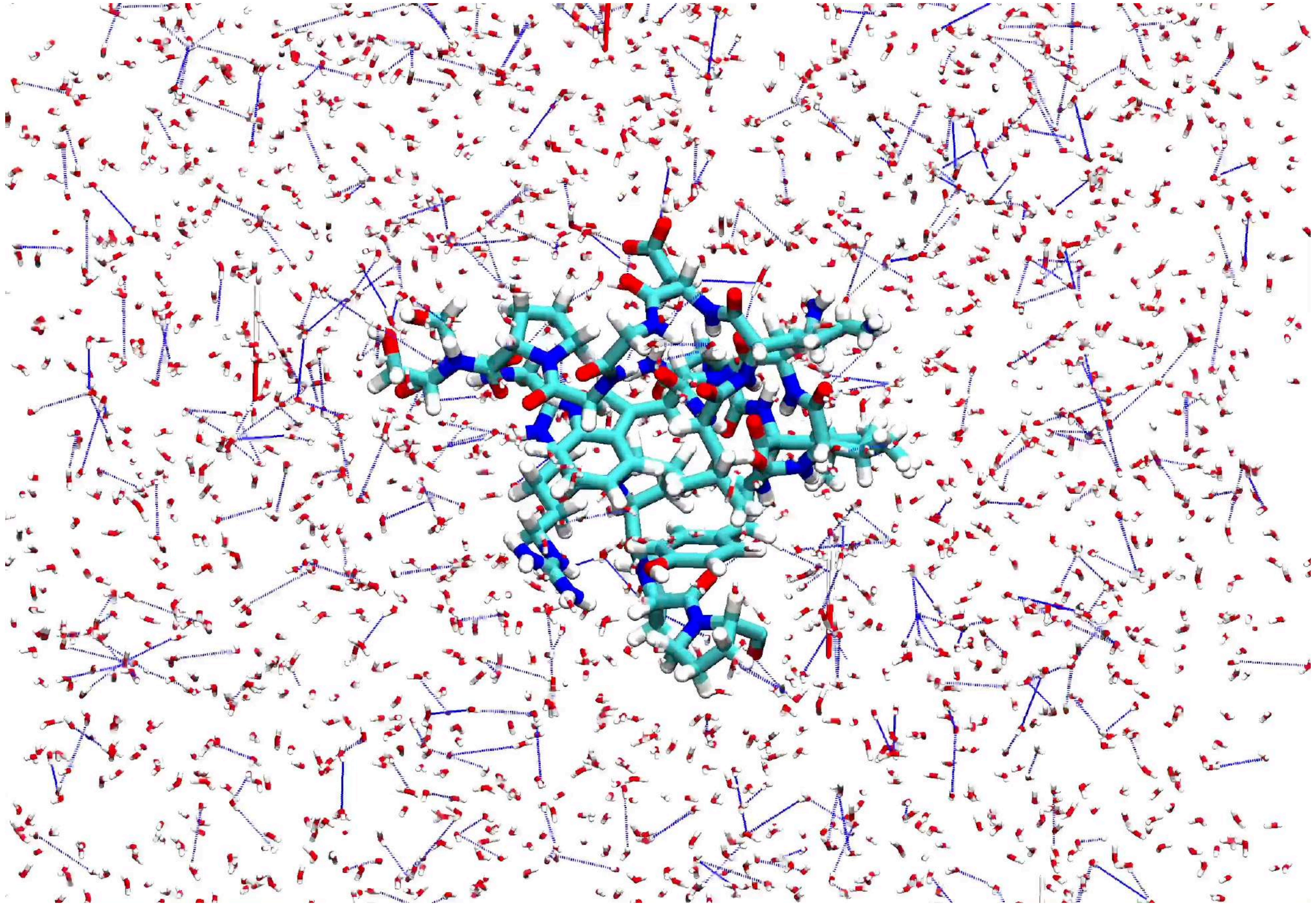
+ SPC/E water model



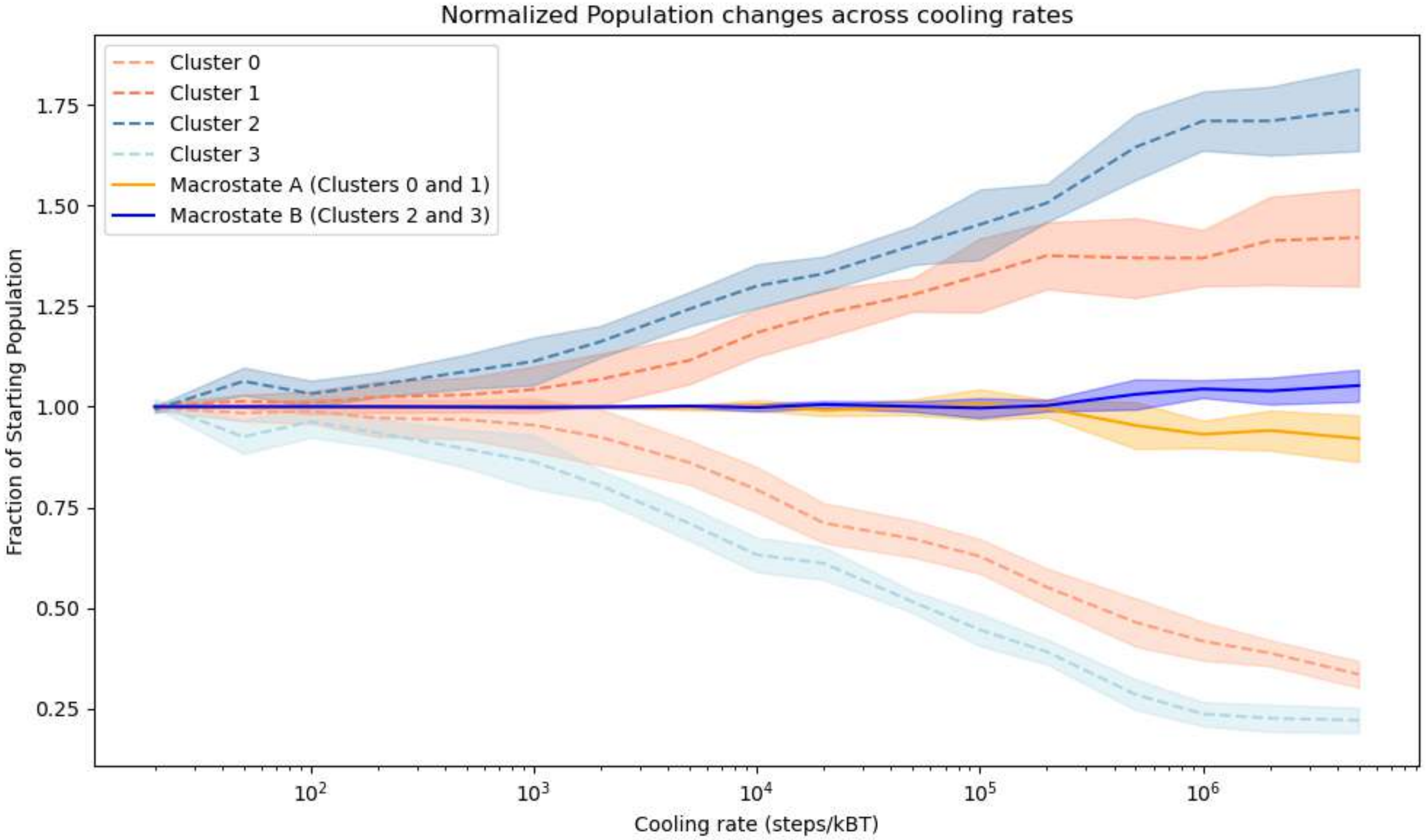


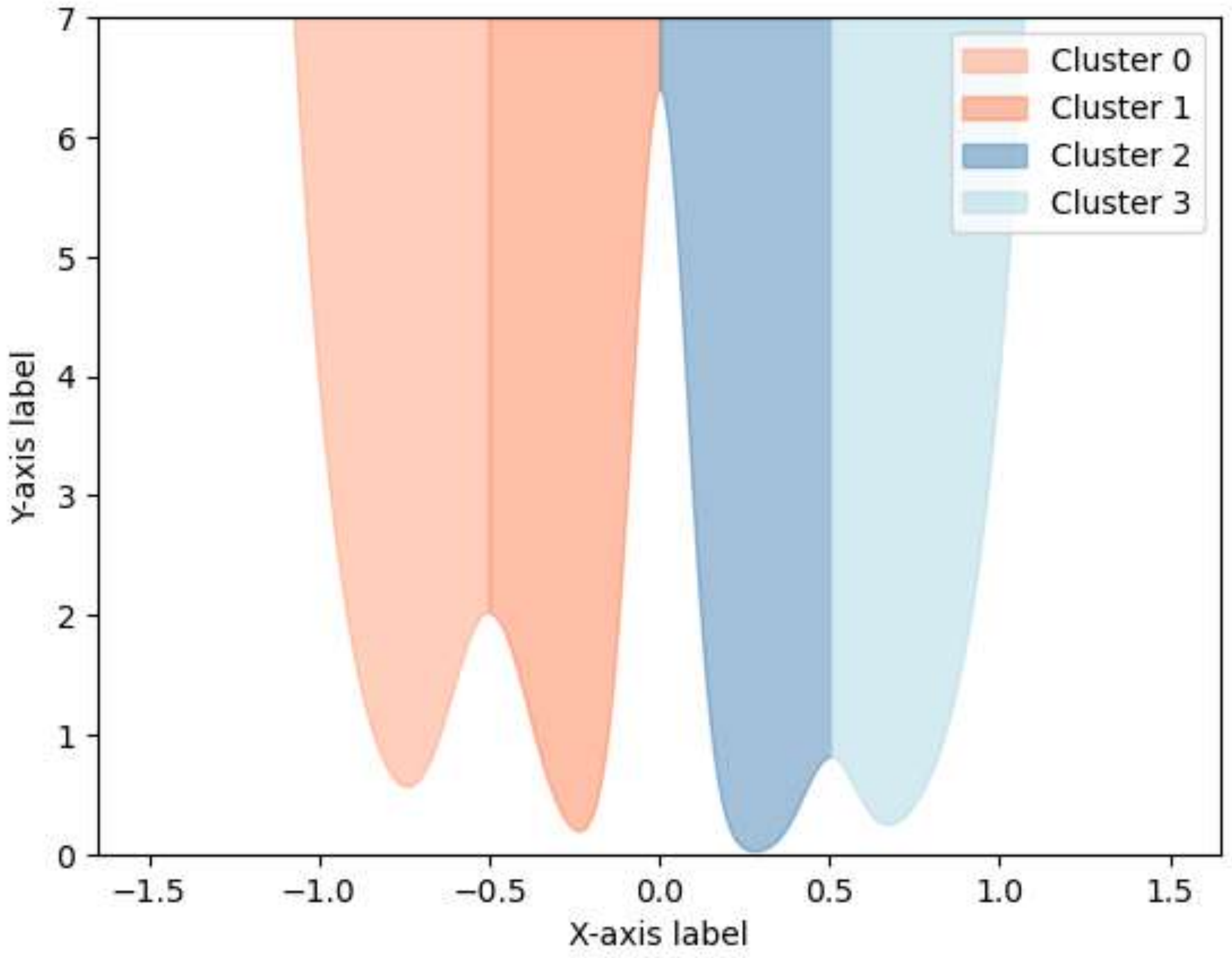
Bock & Grubmüller, *Nat. Comms.* (2022).

Approach and results so far with TRP-cage.



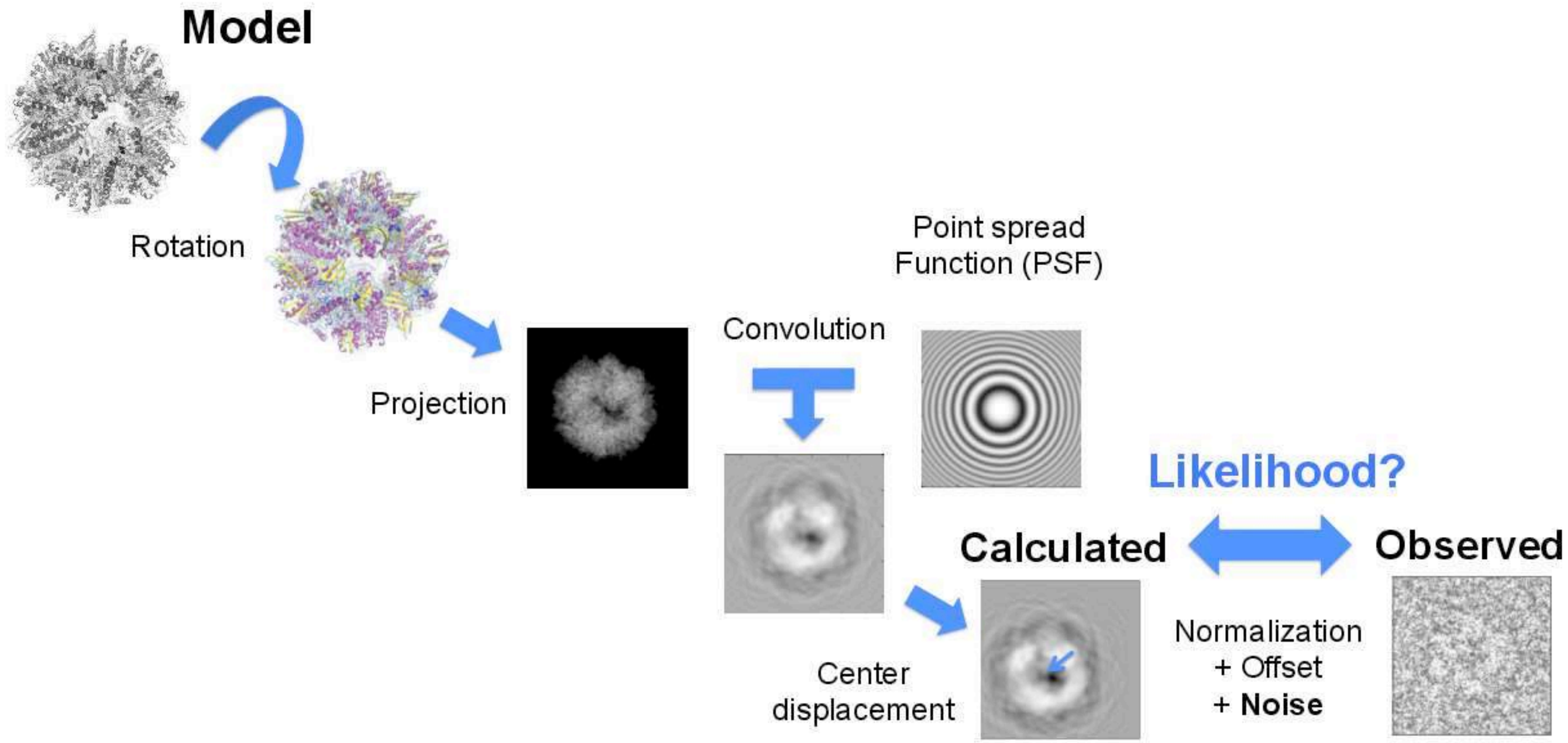
How much does vitrification perturb the equilibrium ensemble of our molecules?





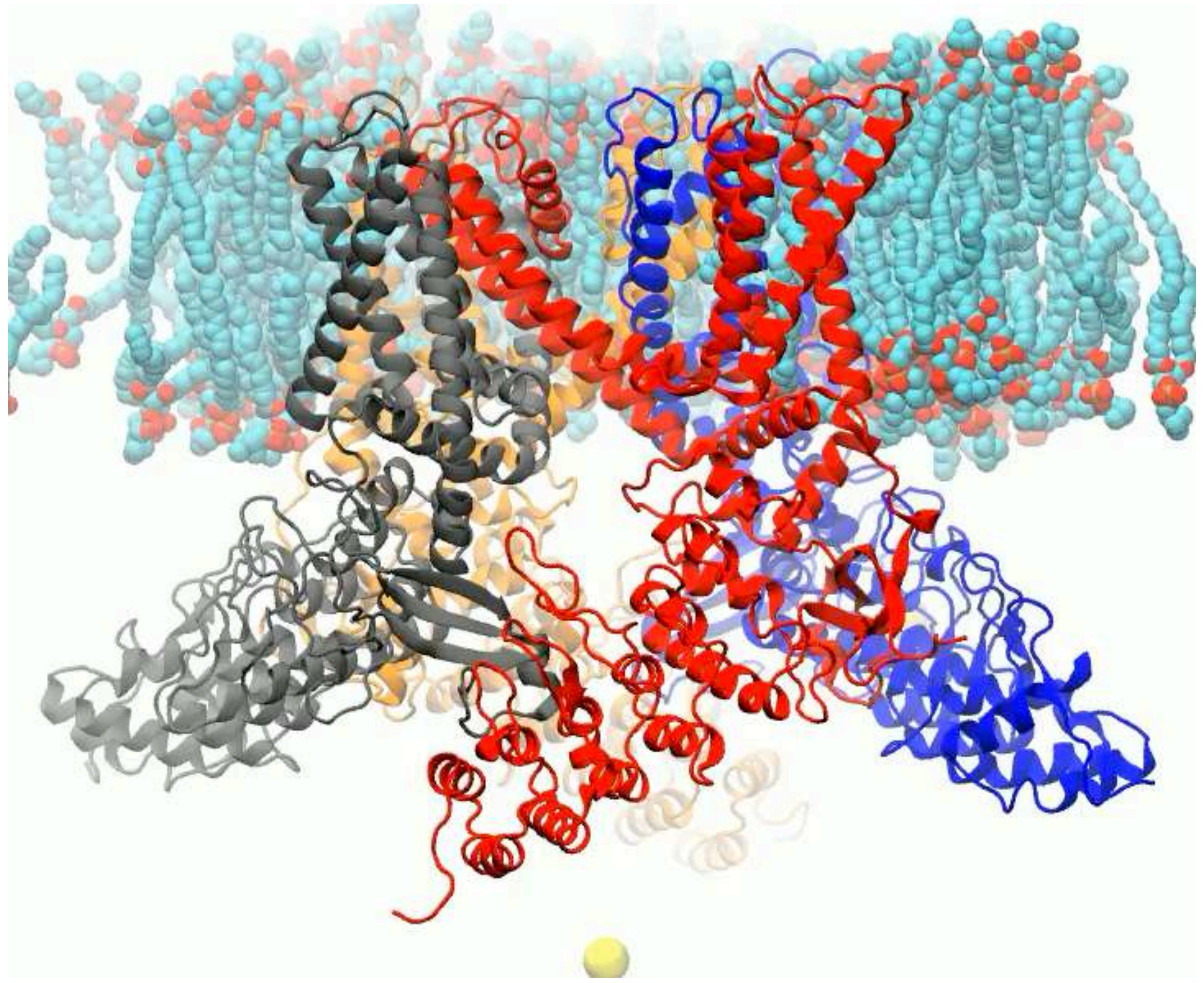
- Toy models to test if we can recapture the ‘hot ensemble’ from the ‘cooled ensemble’.

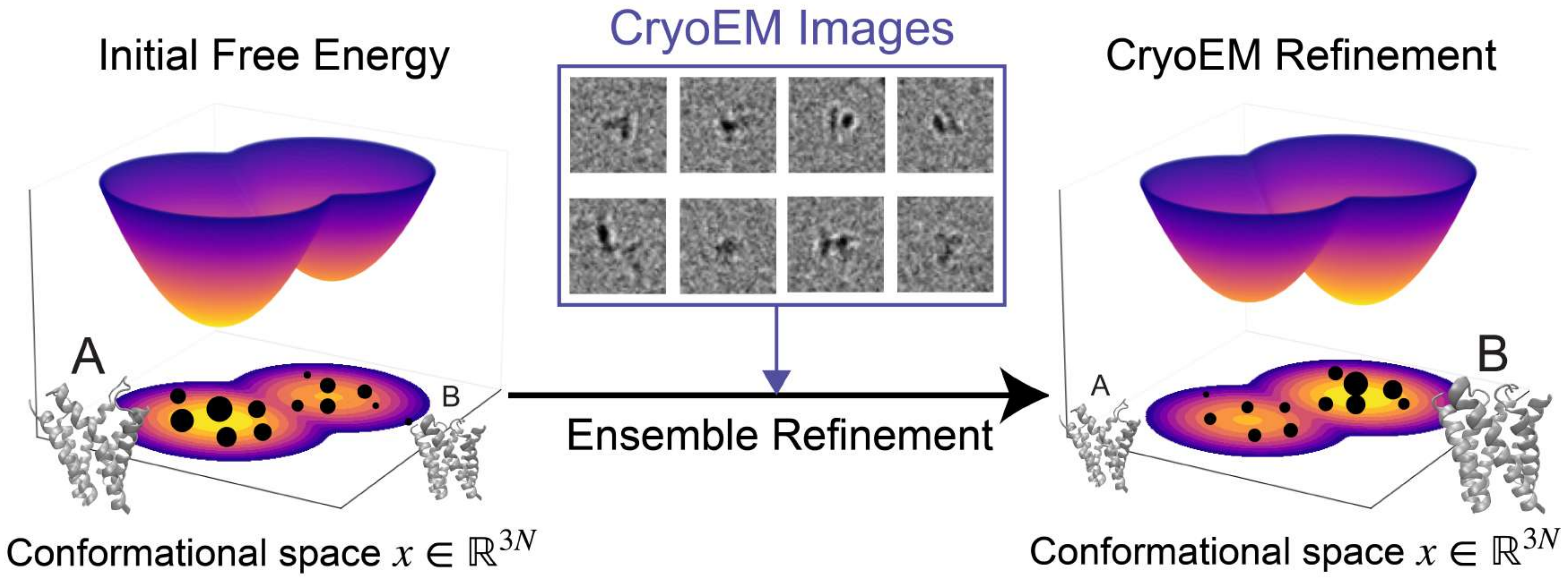
Bayesian Inference Methods For Comparing Single Particles to Atomic Models



Pilar Cossio et al. Computer Physics Communications (2017)

With this we can *bridge the gap* between cryo-EM data and long molecular dynamics simulations

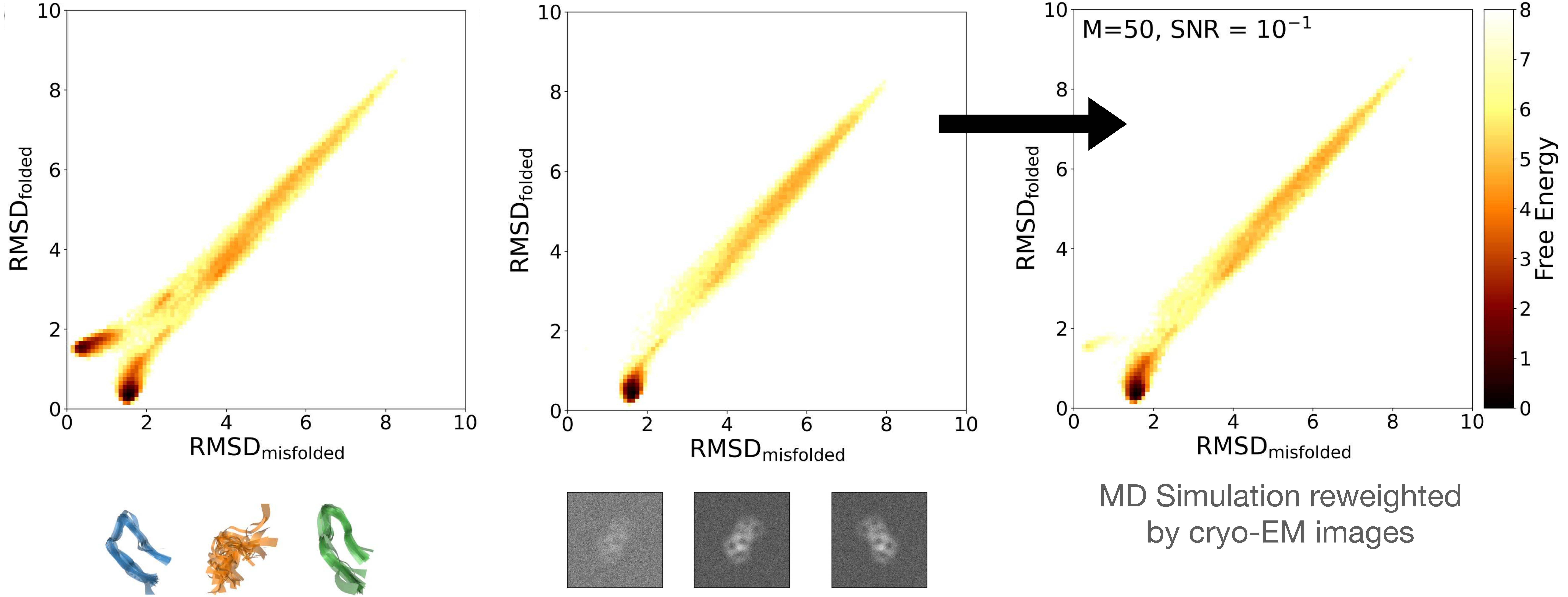




Cryo-EM ensemble reweighting

EH Thiede

WS Tang



MD Structural Ensemble

Synthetic Cryo-EM Images

Reweighted Ensemble

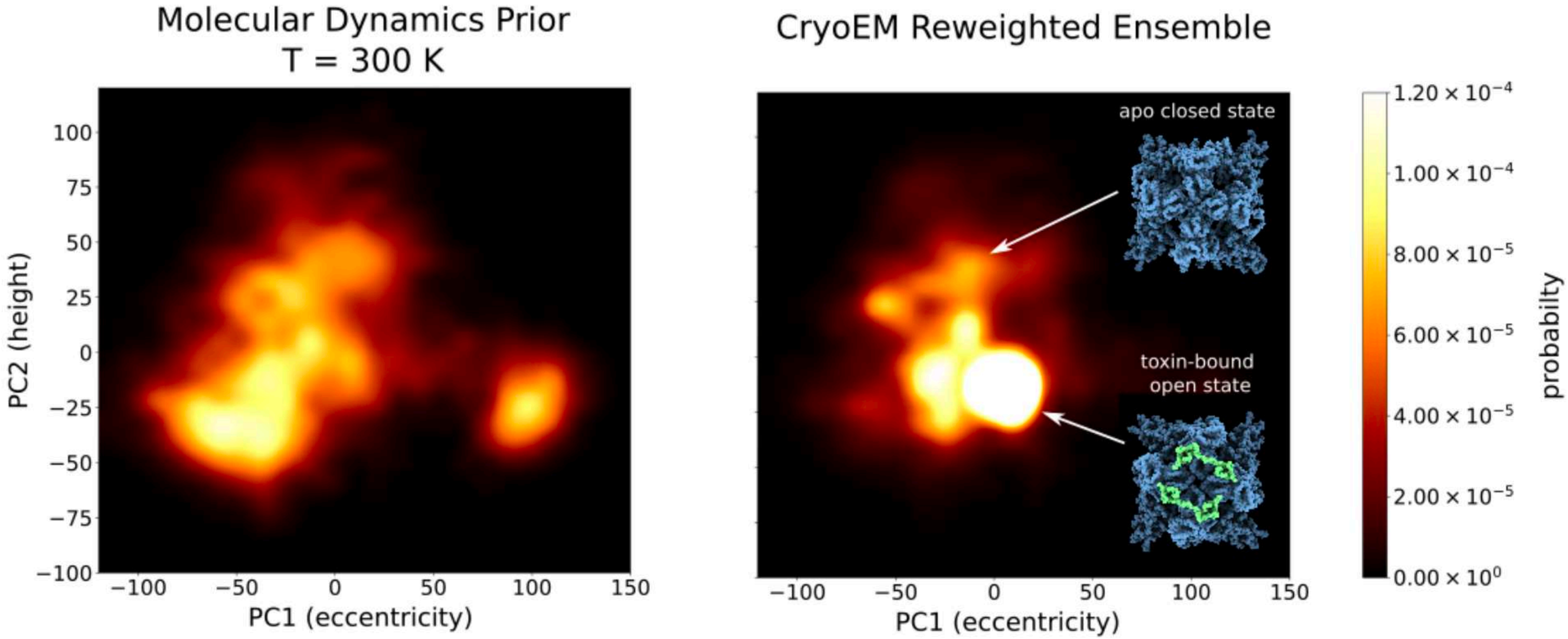
MD Simulation reweighted by cryo-EM images

(our simulation)

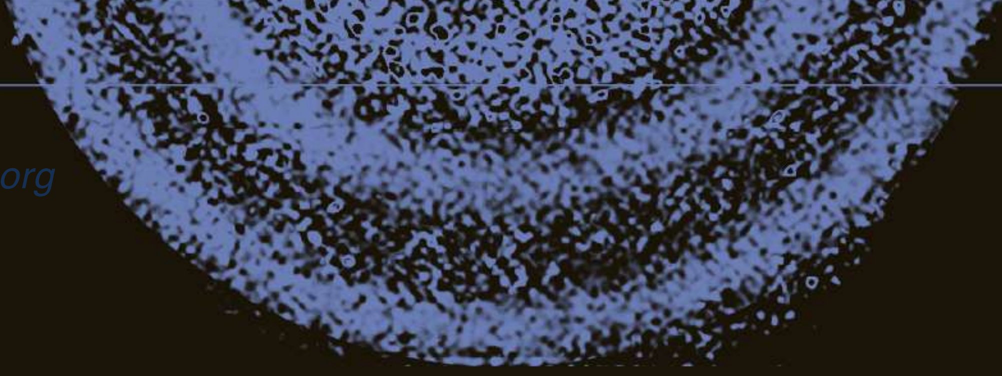
(DE Shaw data)

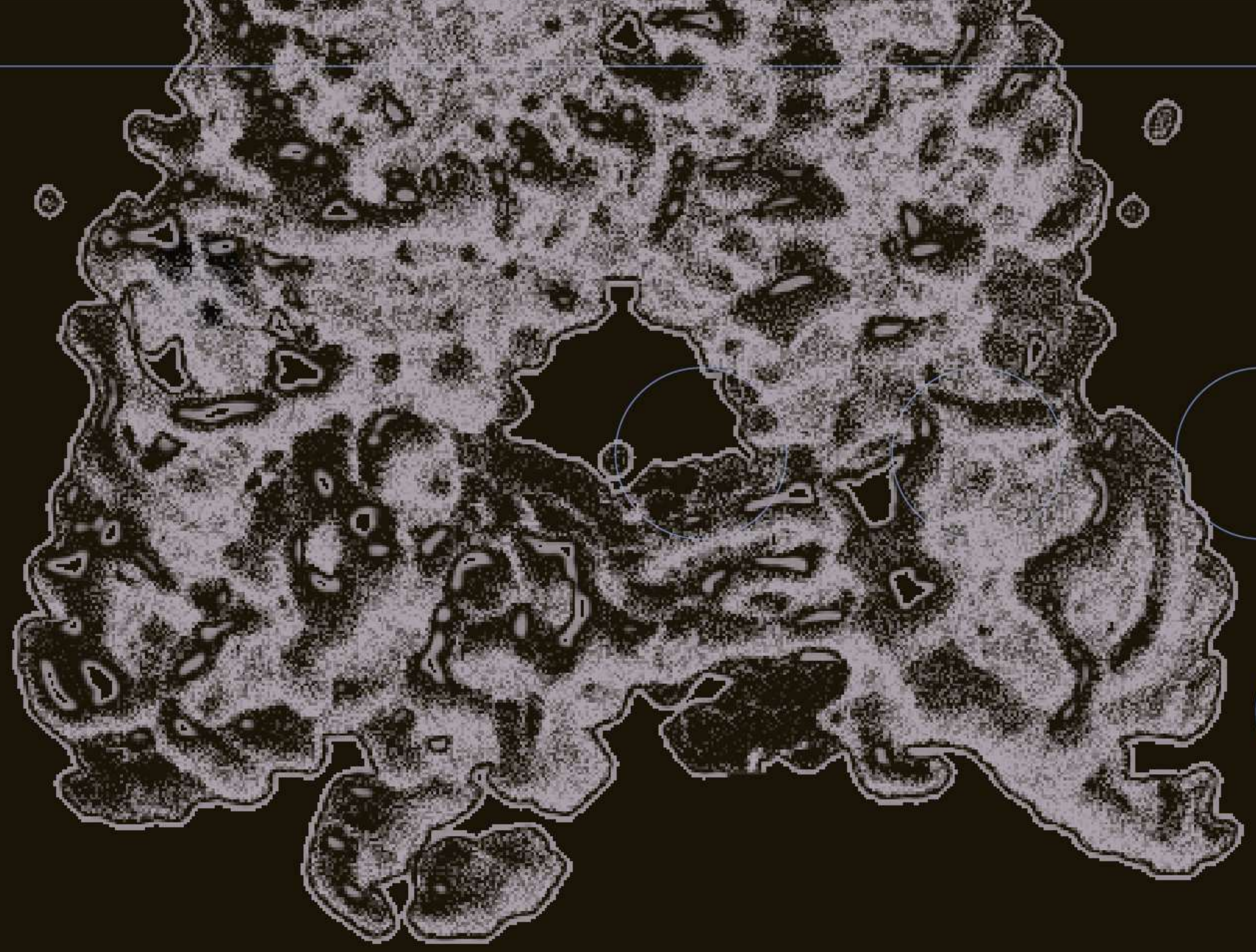
WS Tang et al. *J. Phys. Chem. B* **2023.**

And now a first illustration with real data: TRPV1.



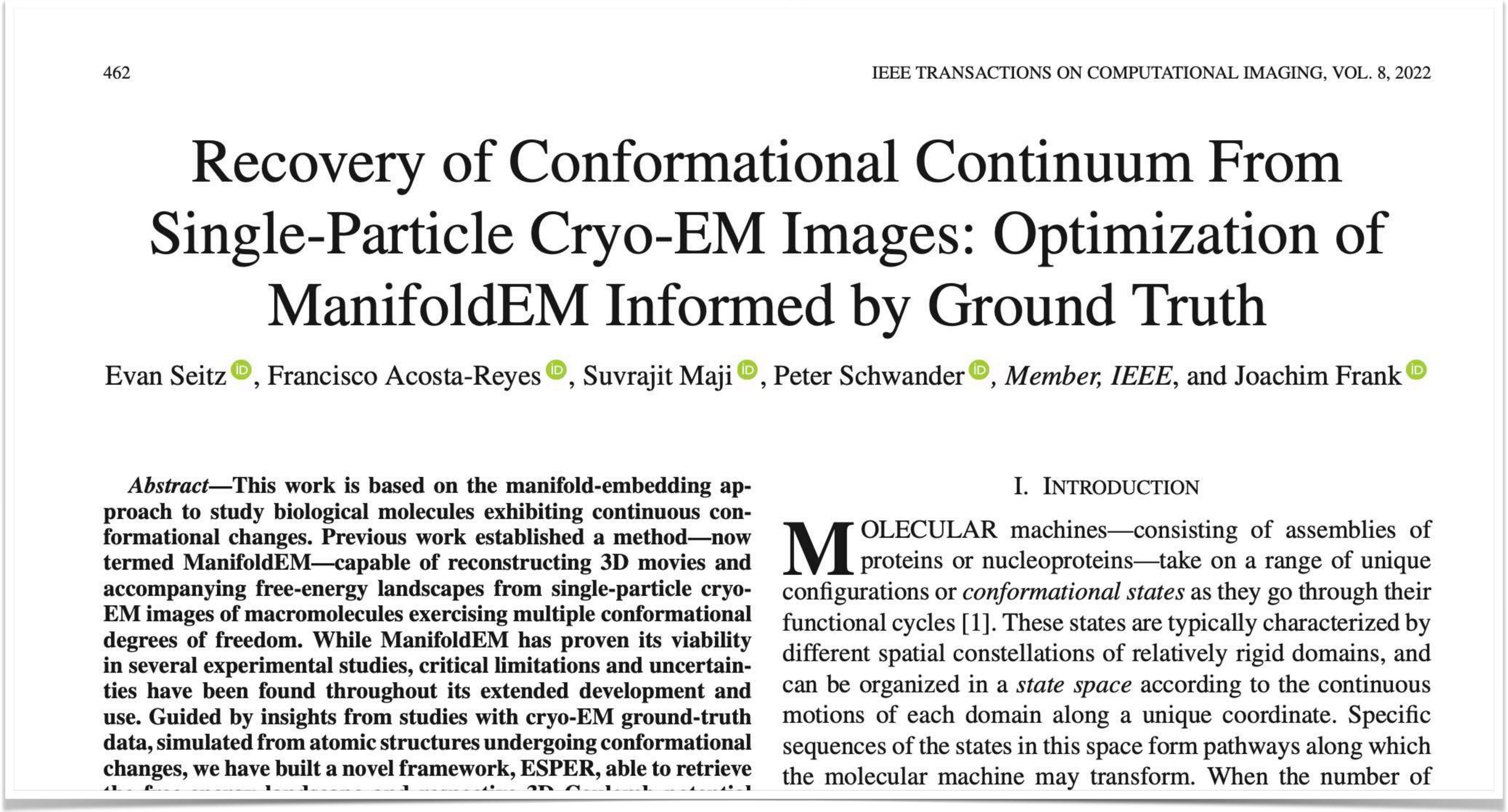
Astore et al, *bioRxiv* (2024).
https://doi.org/10.1101/2024.10.07.617120





Assessing the performance of continuous heterogeneity methods.

Are these motions real?



462 IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING, VOL. 8, 2022

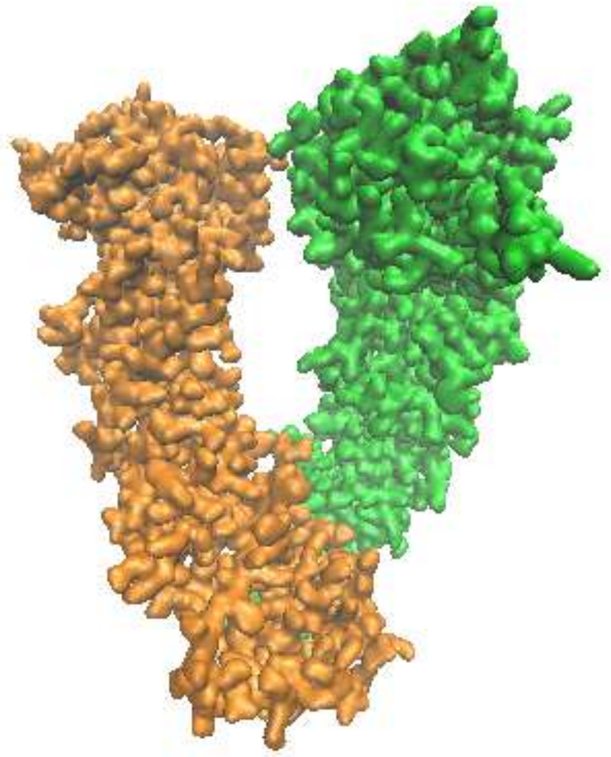
Recovery of Conformational Continuum From Single-Particle Cryo-EM Images: Optimization of ManifoldEM Informed by Ground Truth

Evan Seitz, Francisco Acosta-Reyes, Suvrajit Maji, Peter Schwander, *Member, IEEE*, and Joachim Frank

***Abstract*—This work is based on the manifold-embedding approach to study biological molecules exhibiting continuous conformational changes. Previous work established a method—now termed ManifoldEM—capable of reconstructing 3D movies and accompanying free-energy landscapes from single-particle cryo-EM images of macromolecules exercising multiple conformational degrees of freedom. While ManifoldEM has proven its viability in several experimental studies, critical limitations and uncertainties have been found throughout its extended development and use. Guided by insights from studies with cryo-EM ground-truth data, simulated from atomic structures undergoing conformational changes, we have built a novel framework, ESPER, able to retrieve**

I. INTRODUCTION

MOLECULAR machines—consisting of assemblies of proteins or nucleoproteins—take on a range of unique configurations or *conformational states* as they go through their functional cycles [1]. These states are typically characterized by different spatial constellations of relatively rigid domains, and can be organized in a *state space* according to the continuous motions of each domain along a unique coordinate. Specific sequences of the states in this space form pathways along which the molecular machine may transform. When the number of

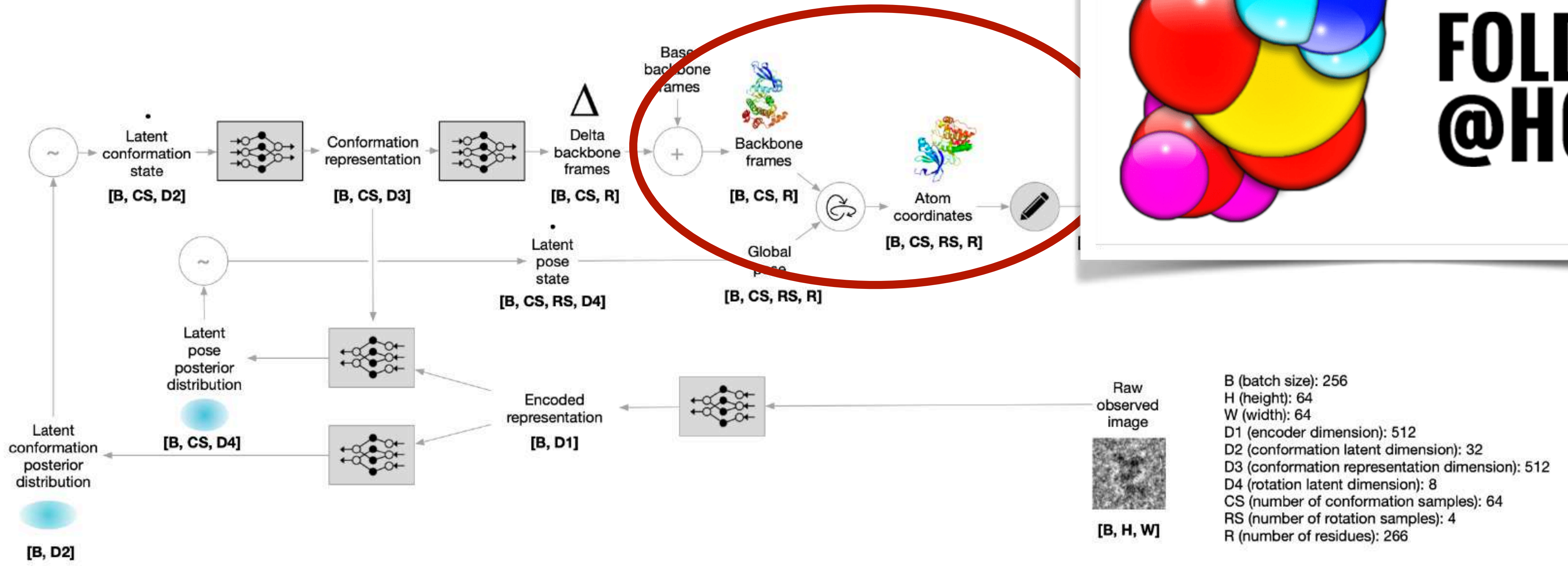


Seitz et al *bioRxiv* (2019)
https://doi.org/10.1101/864116

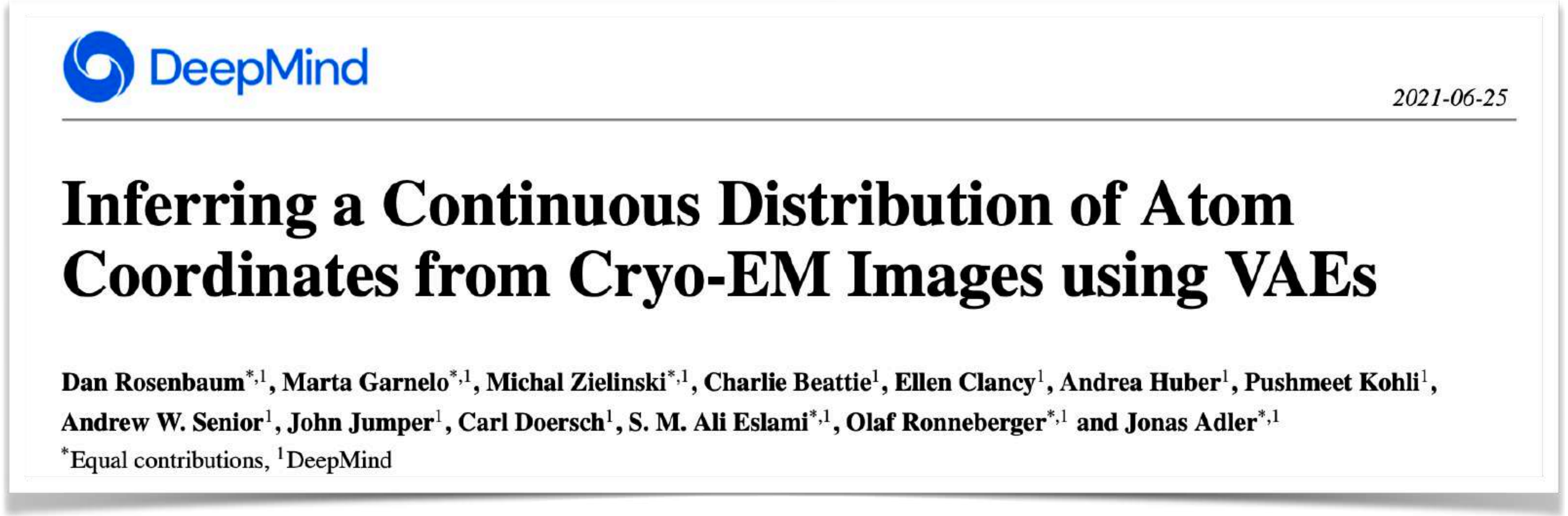
An apoferritin for heterogeneity?

FLATIRON INSTITUTE

- Simulated datasets.



FOLDING @HOME



DeepMind

2021-06-25

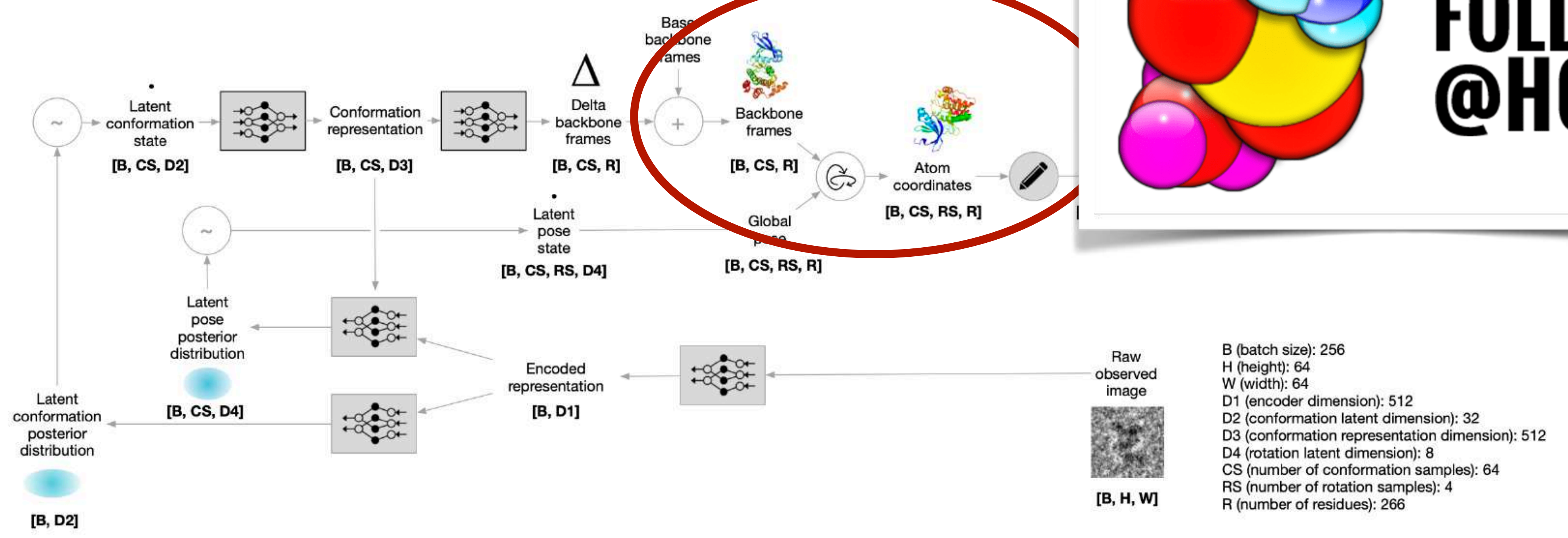
Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs

Dan Rosenbaum*,1, Marta Garnelo*,1, Michal Zielinski*,1, Charlie Beattie1, Ellen Clancy1, Andrea Huber1, Pushmeet Kohli1, Andrew W. Senior1, John Jumper1, Carl Doersch1, S. M. Ali Eslami*,1, Olaf Ronneberger*,1 and Jonas Adler*,1

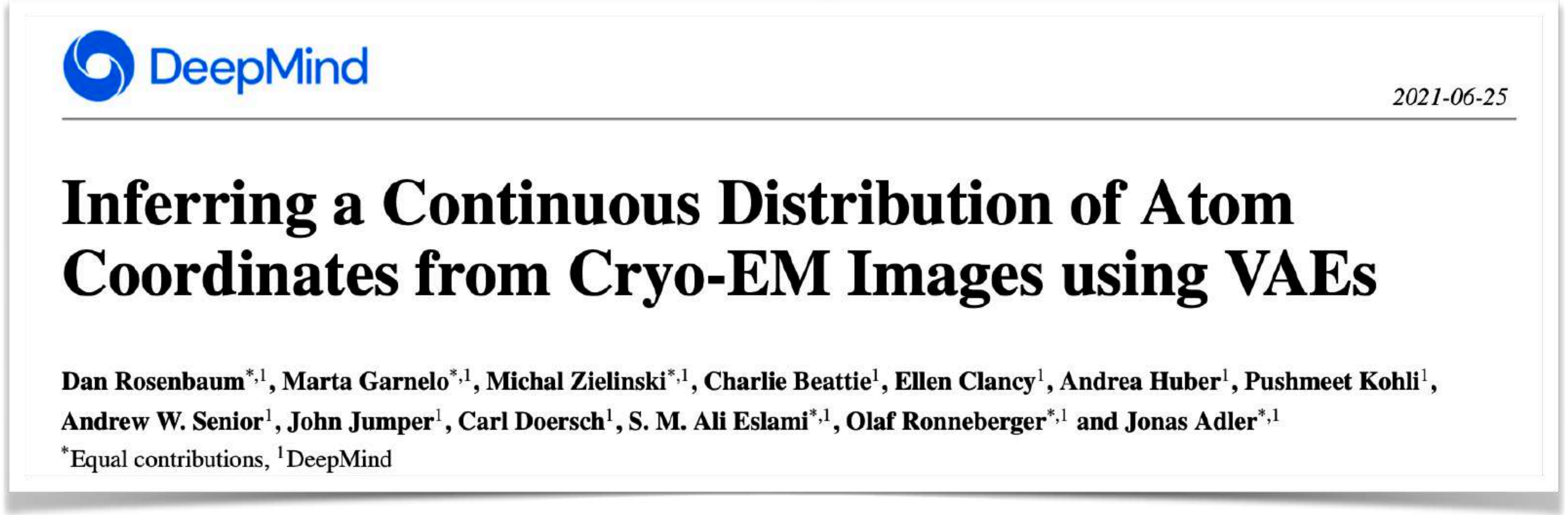
*Equal contributions, 1DeepMind

An apoferritin for heterogeneity?

- Simulated datasets.



FOLDING @HOME



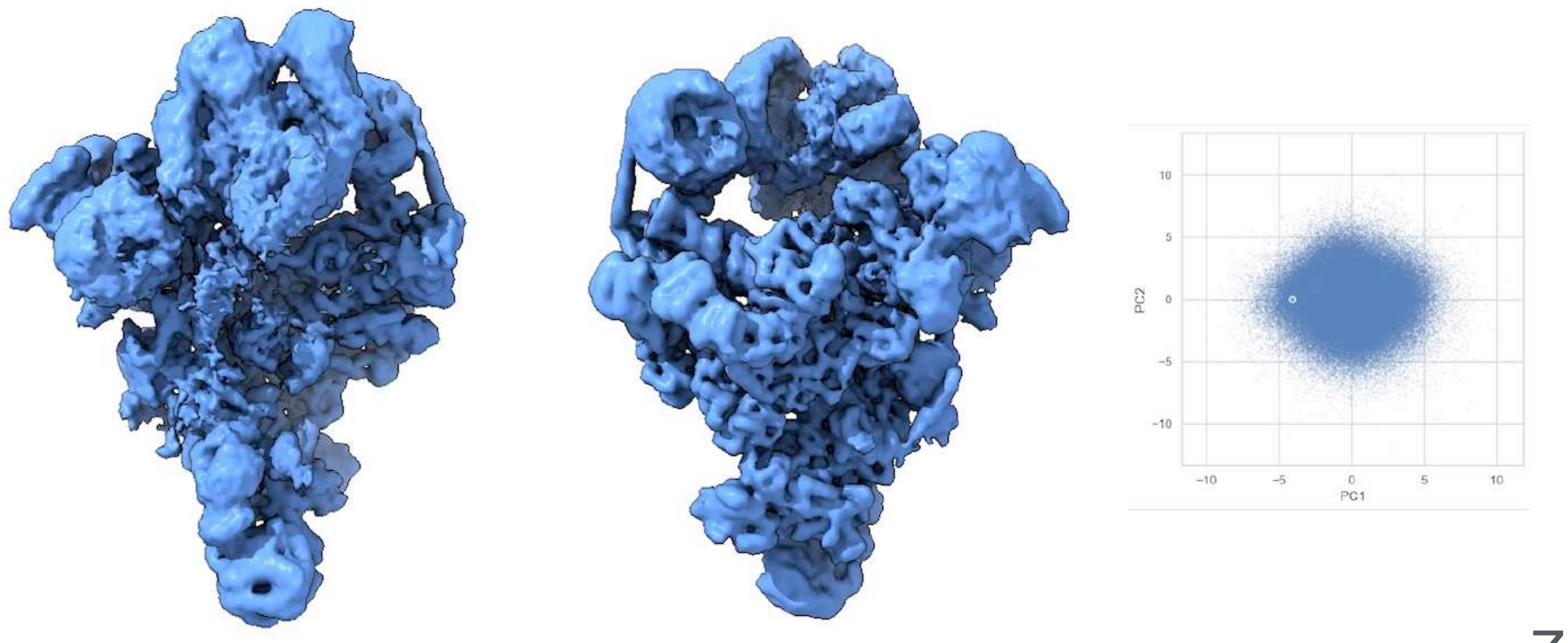
DeepMind 2021-06-25

Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs

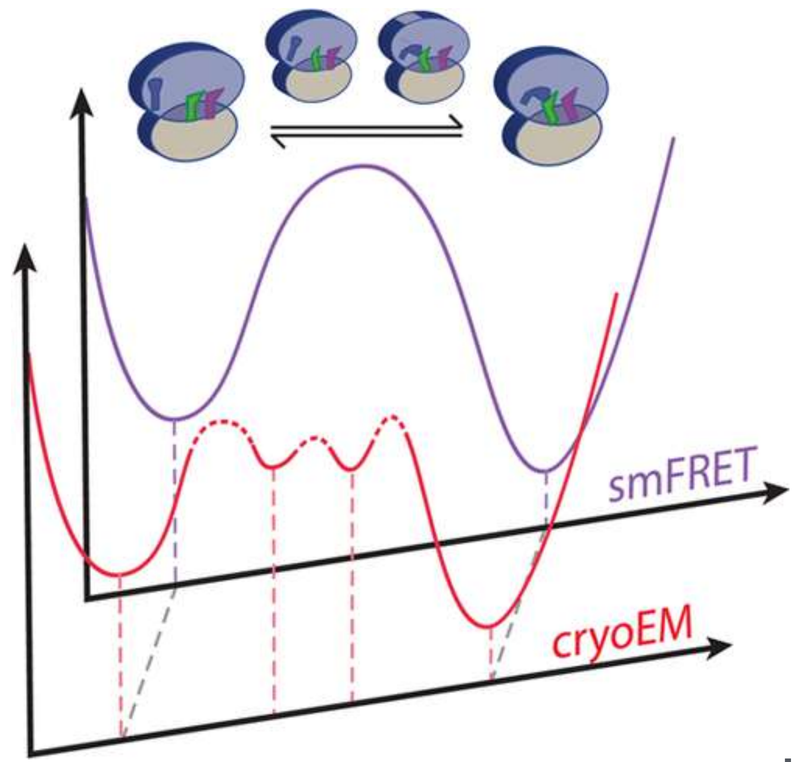
Dan Rosenbaum[*,1], Marta Garnelo[*,1], Michal Zielinski[*,1], Charlie Beattie[1], Ellen Clancy[1], Andrea Huber[1], Pushmeet Kohli[1], Andrew W. Senior[1], John Jumper[1], Carl Doersch[1], S. M. Ali Eslami[*,1], Olaf Ronneberger[*,1] and Jonas Adler[*,1]

[*]Equal contributions, [1]DeepMind

- Experimental datasets (w/ validation of the populations from another experimental method!)



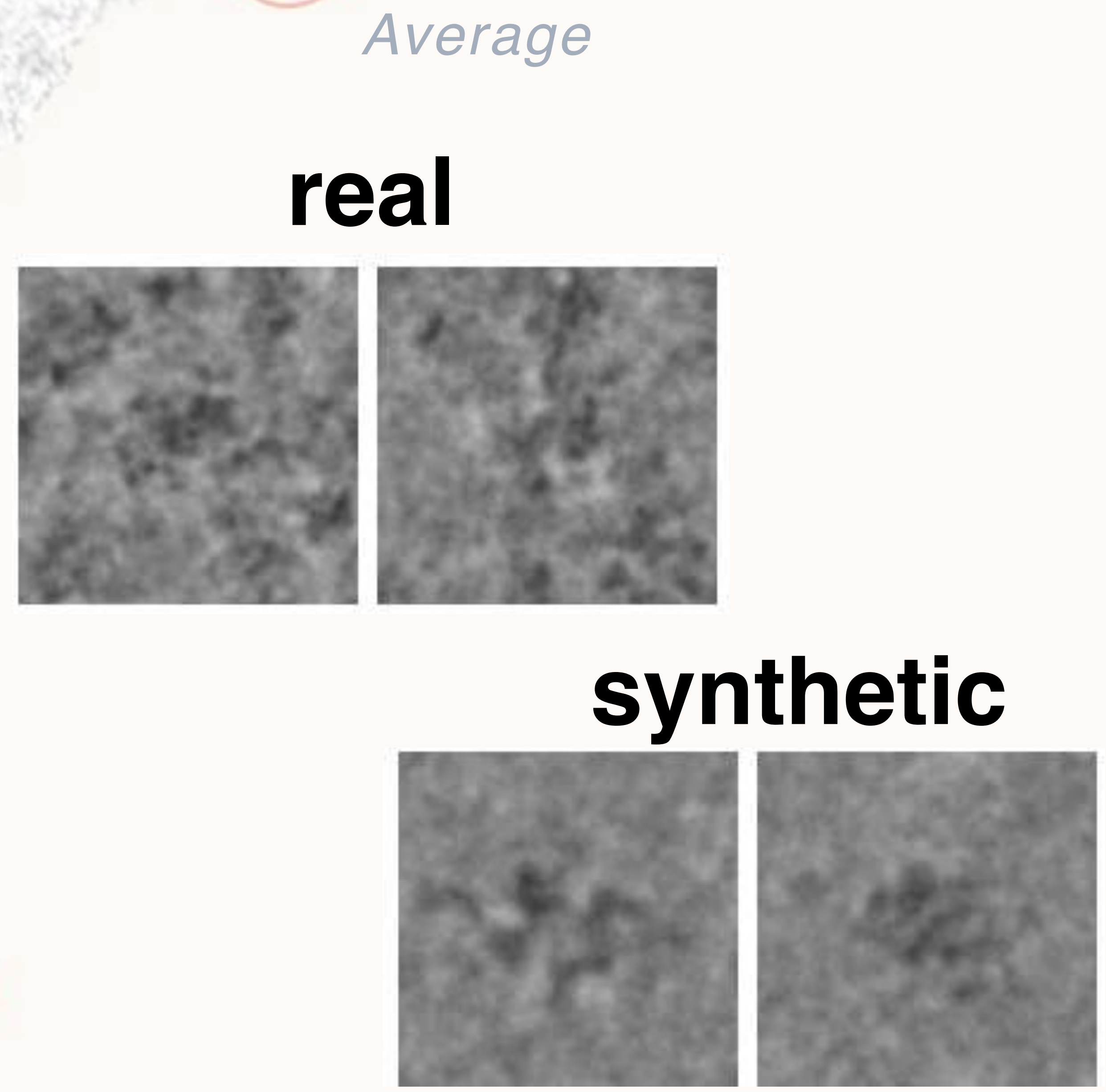
Zhong et al
Nature Methods (2021)

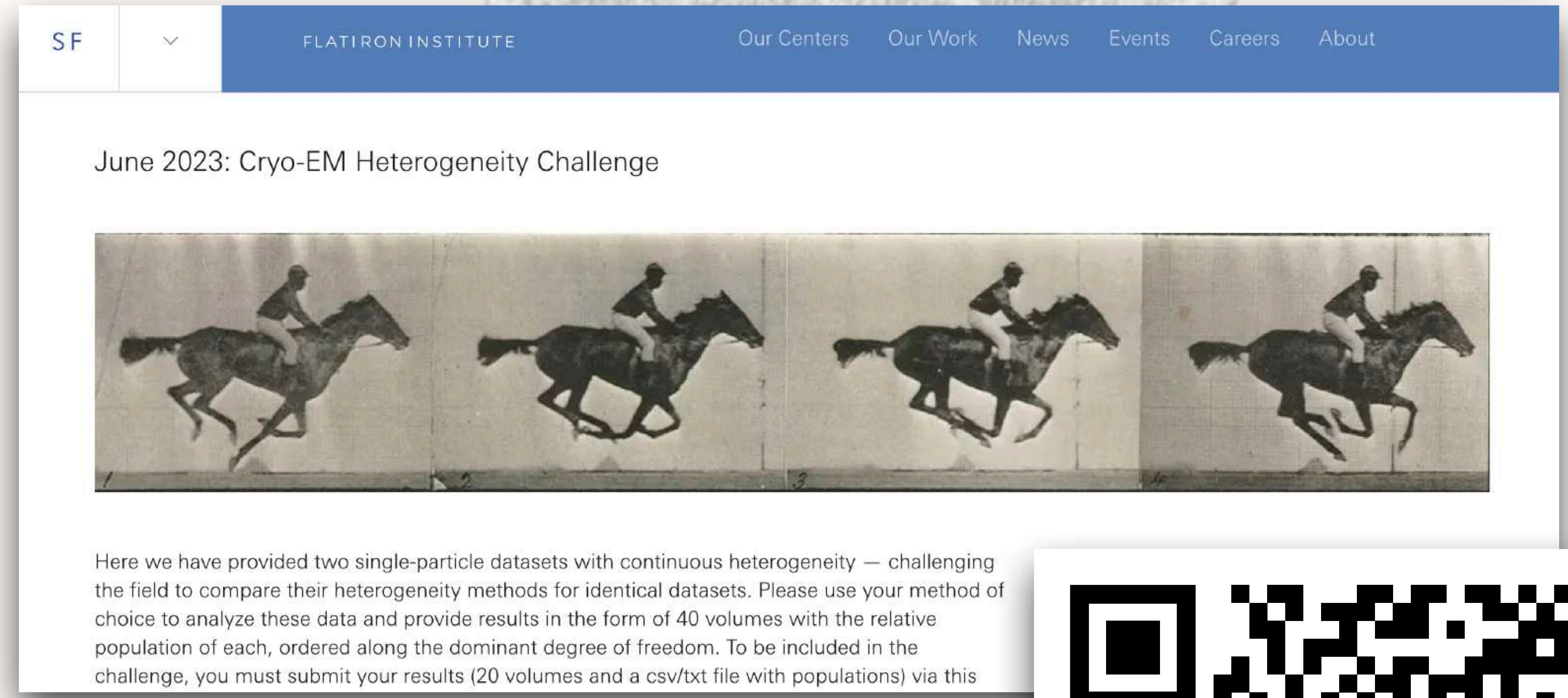


Kinz-Thompson et al
J Phys Chem B (2015)

The Inaugural Flatiron Institute Heterogeneity Challenge for cryo-EM!

The challenge was **announced in June 2023** at our summer cryo-EM workshop and submissions were due October 31 of the same year.





June 2023: Cryo-EM Heterogeneity Challenge

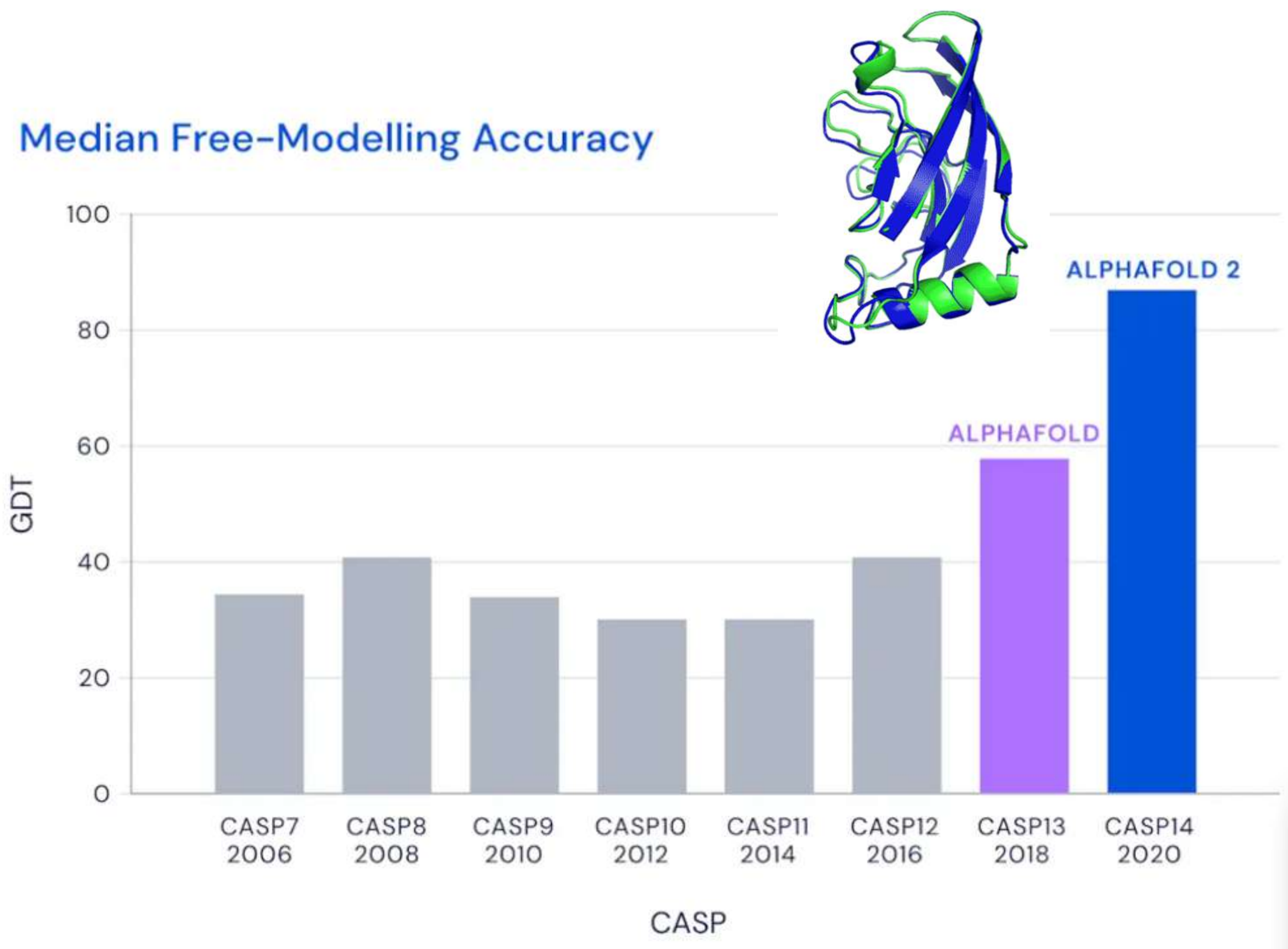
Here we have provided two single-particle datasets with continuous heterogeneity — challenging the field to compare their heterogeneity methods for identical datasets. Please use your method of choice to analyze these data and provide results in the form of 40 volumes with the relative population of each, ordered along the dominant degree of freedom. To be included in the challenge, you must submit your results (20 volumes and a csv/txt file with populations) via this

https://www.simonsfoundation.org/heterogeneity-in-cryo-electron-microscopy/

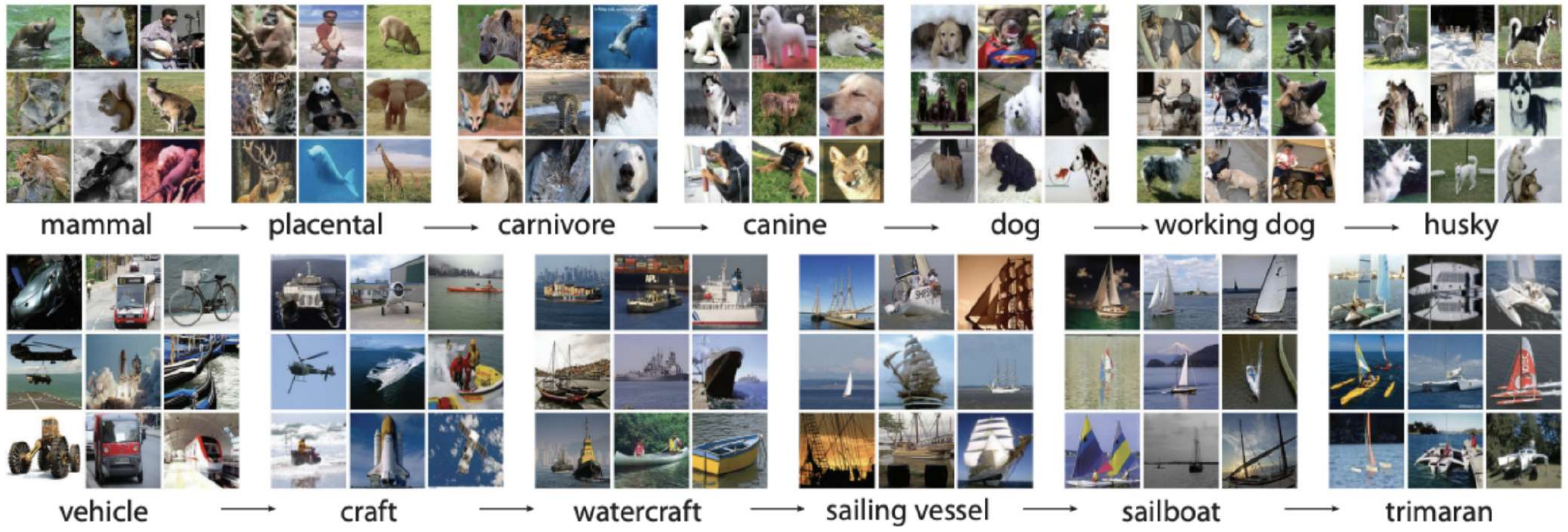
Benchmarks and solid metrics can drive fields forward.

MNIST

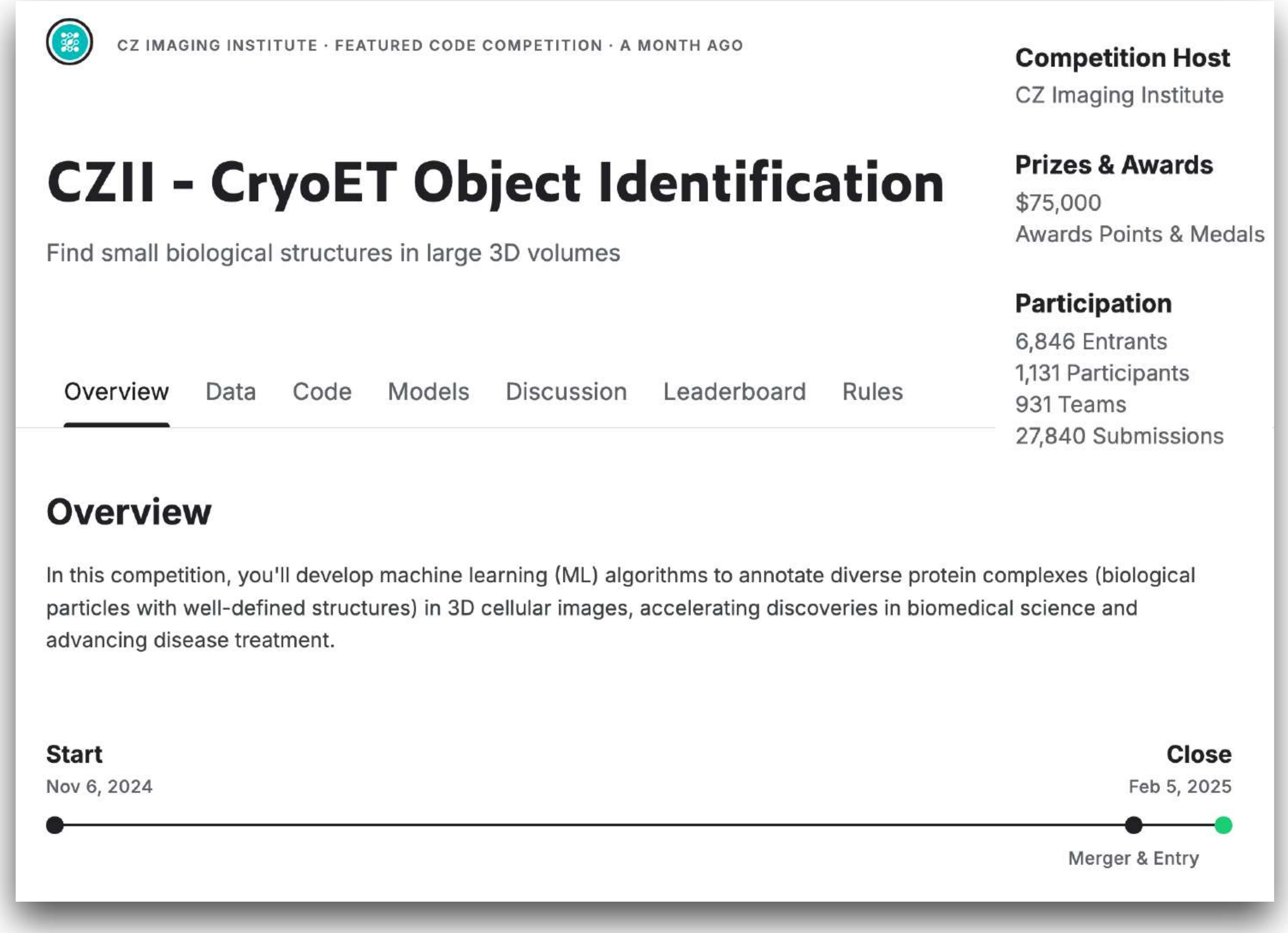
CASP



ImageNet



Kaggle



CZ IMAGING INSTITUTE · FEATURED CODE COMPETITION · A MONTH AGO

CZII - CryoET Object Identification

Find small biological structures in large 3D volumes

Overview Data Code Models Discussion Leaderboard Rules

Overview

In this competition, you'll develop machine learning (ML) algorithms to annotate diverse protein complexes (biological particles with well-defined structures) in 3D cellular images, accelerating discoveries in biomedical science and advancing disease treatment.

Start
Nov 6, 2024

Close
Feb 5, 2025

Merger & Entry

Competition Host
CZ Imaging Institute

Prizes & Awards
$75,000
Awards Points & Medals

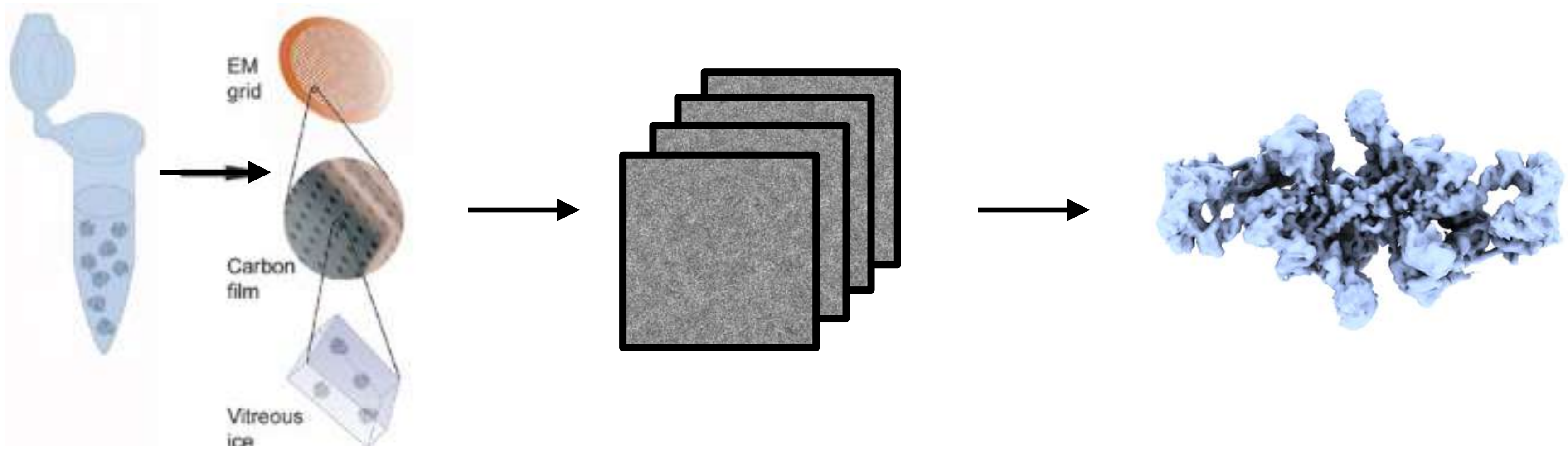
Participation
6,846 Entrants
1,131 Participants
931 Teams
27,840 Submissions

Inaugural Flatiron Institute Heterogeneity Community Challenge

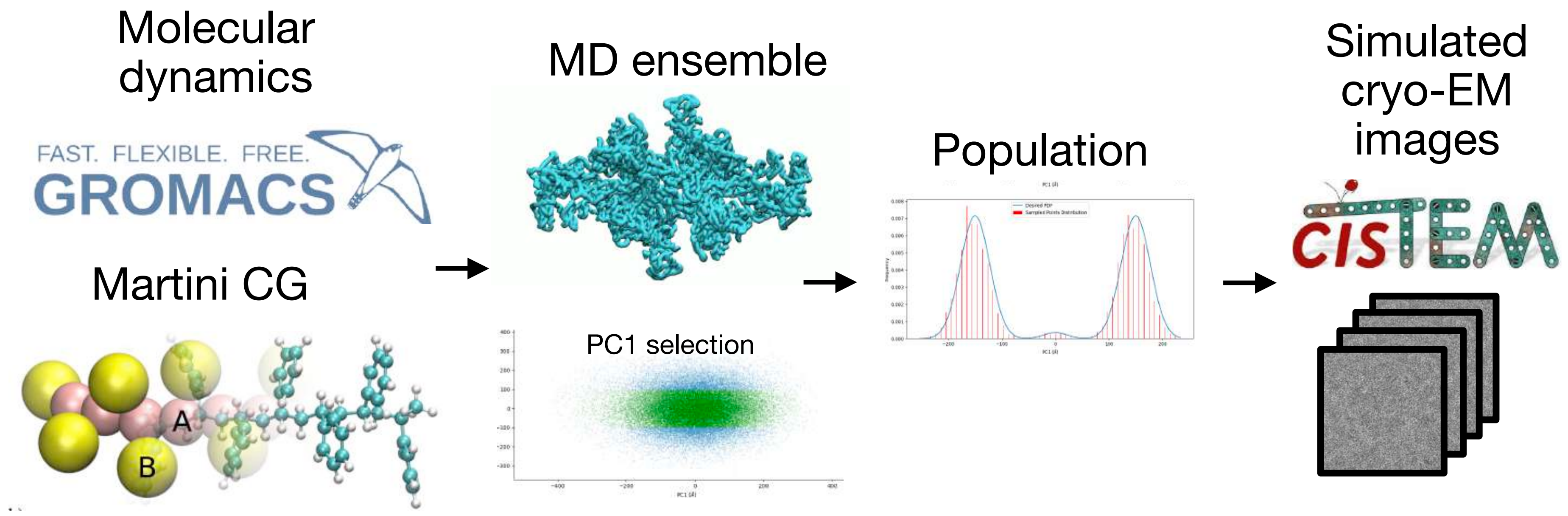
Launched in the summer 2023,

- **Two** datasets were prepared: one **real (NYSBC)** and one **synthetic** of thyroglobulin with ~33k particles from **MD simulations.**

1. Experimental



2. Synthetic



Molecular dynamics

FAST. FLEXIBLE. FREE. GROMACS

Martini CG

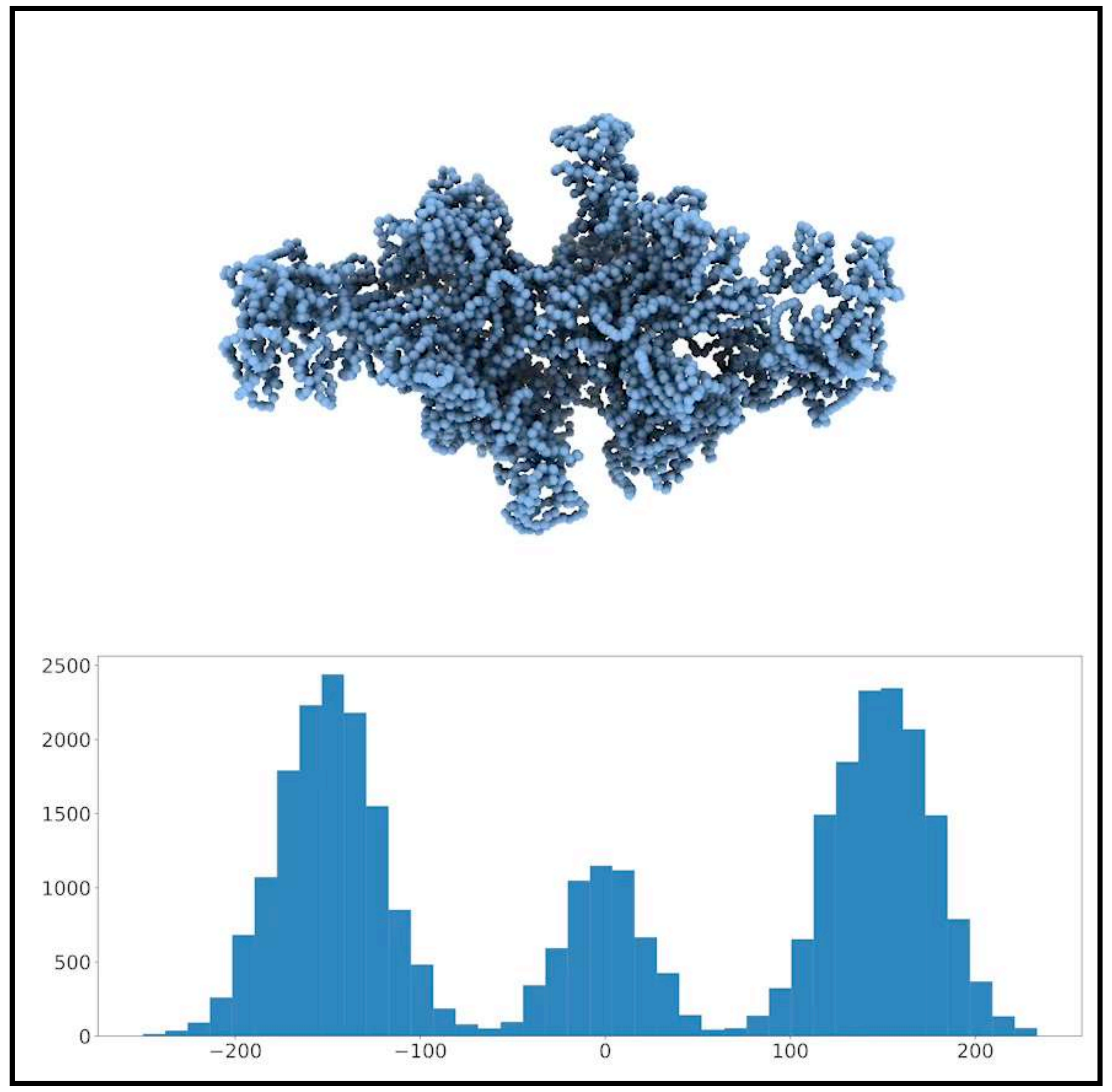
MD ensemble

PC1 selection

Population

Simulated cryo-EM images

cisTEM

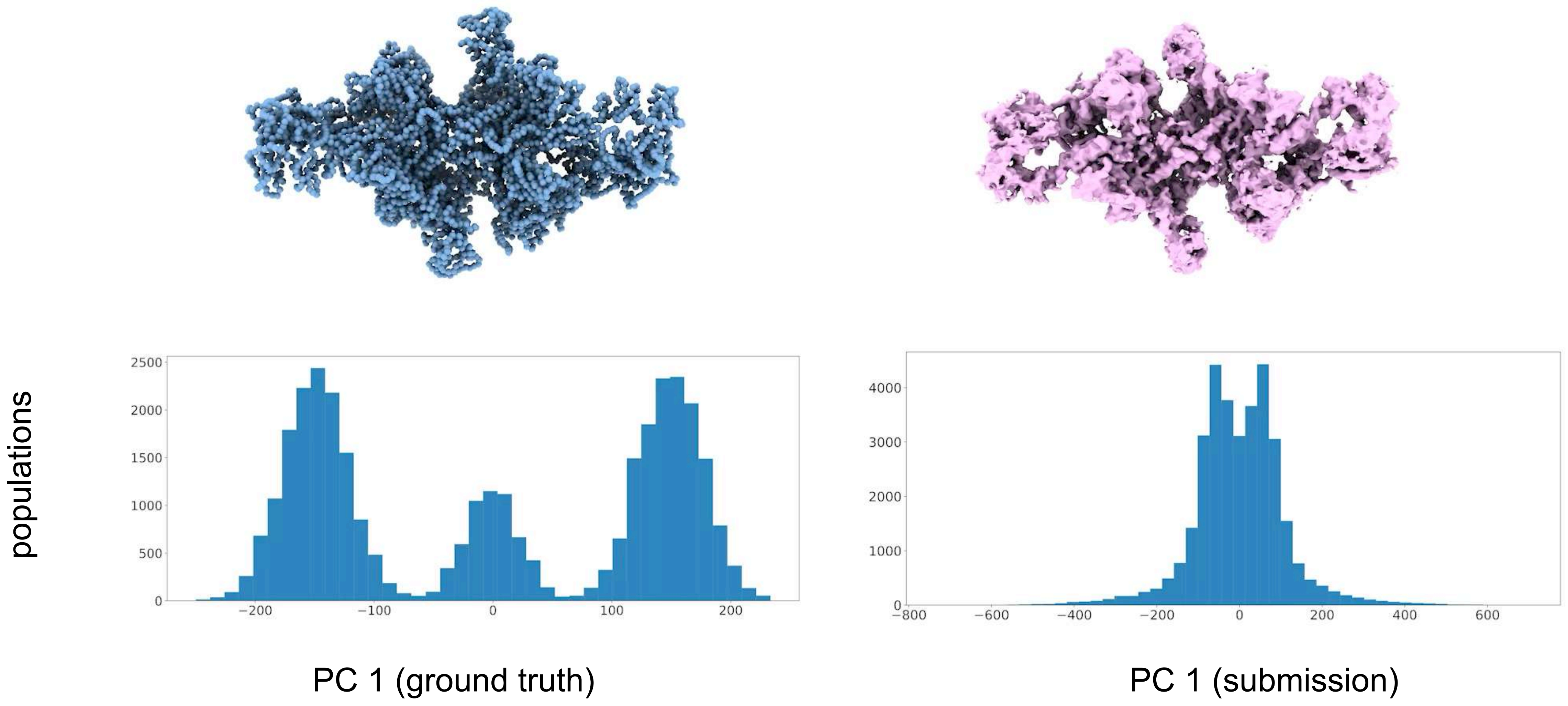


- We asked participants to submit **80 volumes along the dominant degree of freedom and the relative population of each volume.**

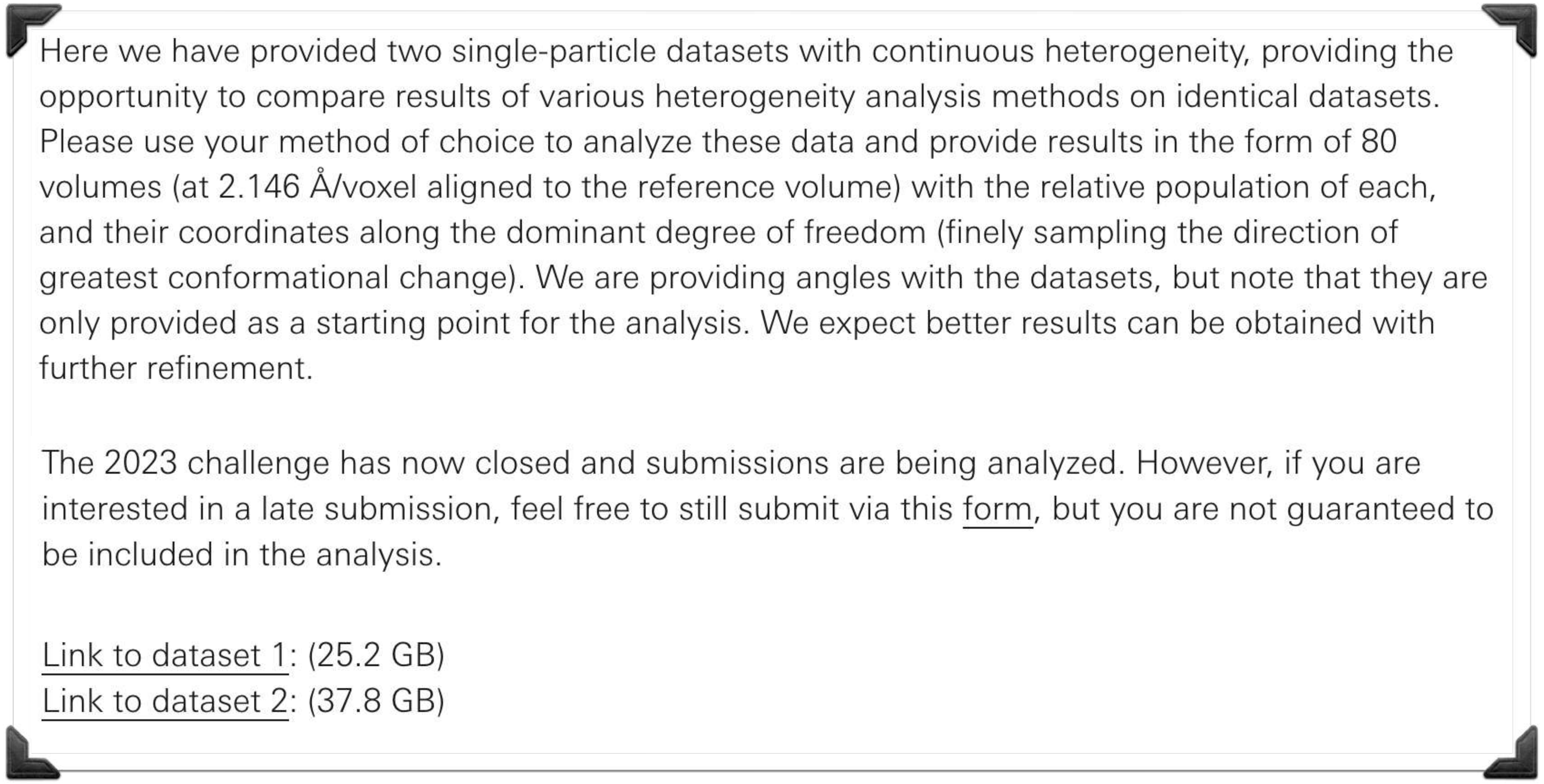
Miro Astore, Niko Grigorieff, (U Mass)
Misha Kopylov (NYSBC), David Silva,
Geoff Wollard, Participants....

What did challenge participants submit?

We asked participants to submit 80 volumes along the most dominant degree of freedom and the relative population of each volume for each of the two datasets (one real and one synthetic).



Synthetic Dataset



Here we have provided two single-particle datasets with continuous heterogeneity, providing the opportunity to compare results of various heterogeneity analysis methods on identical datasets. Please use your method of choice to analyze these data and provide results in the form of 80 volumes (at 2.146 Å/voxel aligned to the reference volume) with the relative population of each, and their coordinates along the dominant degree of freedom (finely sampling the direction of greatest conformational change). We are providing angles with the datasets, but note that they are only provided as a starting point for the analysis. We expect better results can be obtained with further refinement.

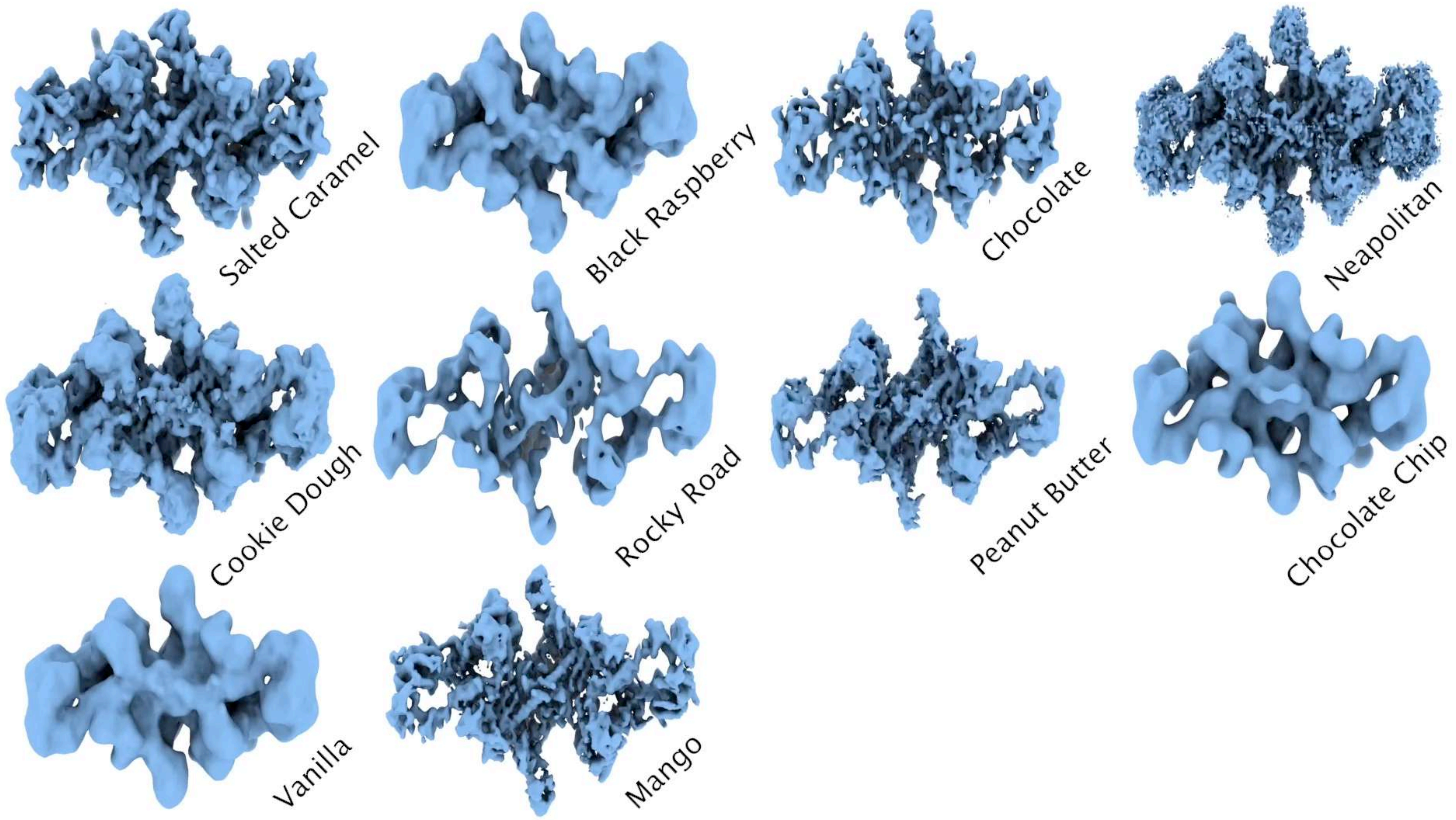
The 2023 challenge has now closed and submissions are being analyzed. However, if you are interested in a late submission, feel free to still submit via this form, but you are not guaranteed to be included in the analysis.

Link to dataset 1: (25.2 GB)
Link to dataset 2: (37.8 GB)

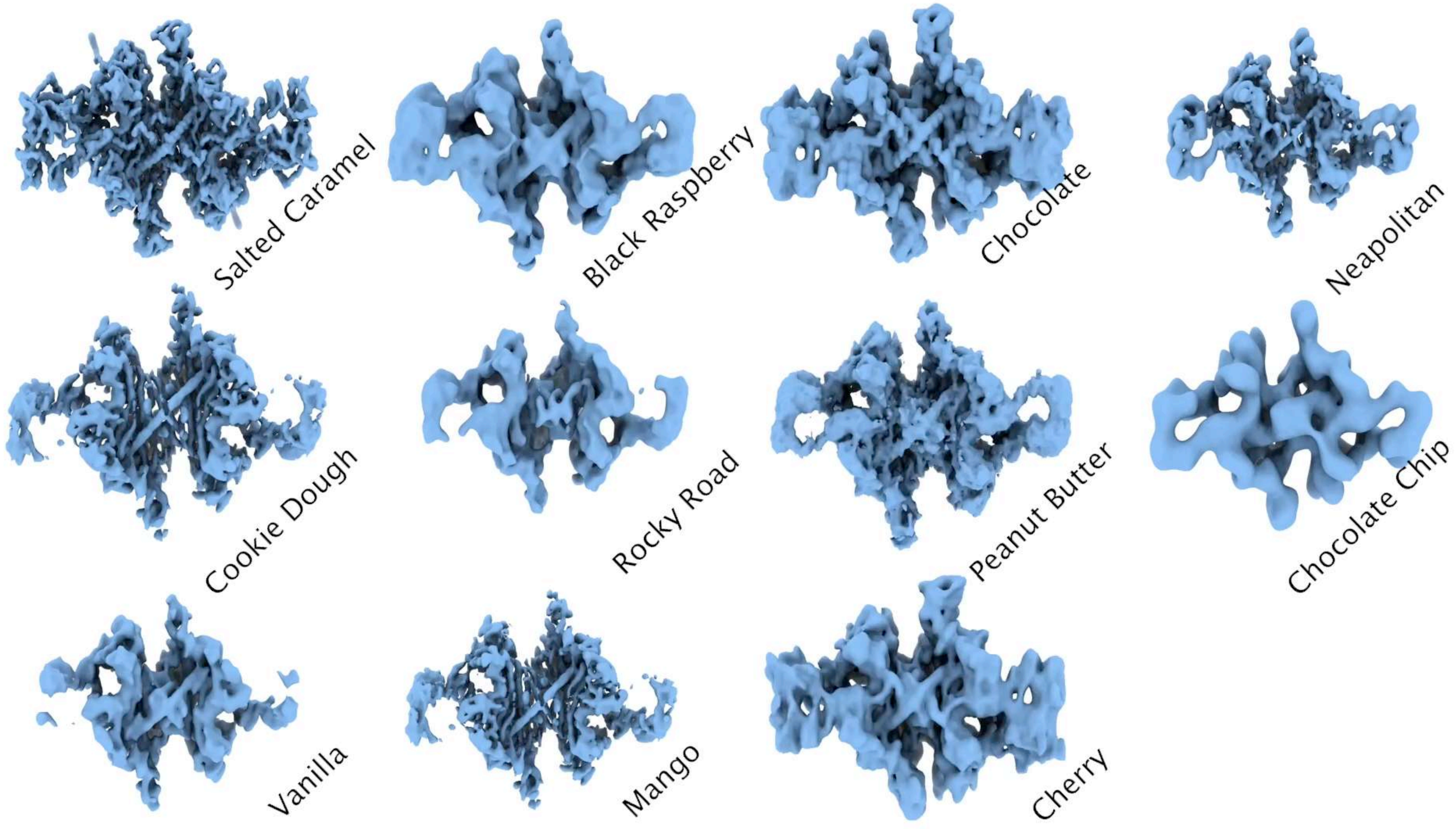
https://www.simonsfoundation.org/heterogeneity-in-cryo-electron-microscopy/

We had a great participation:
41 submissions from 9 groups.

Experimental Dataset



Synthetic Dataset



Results shown in this presentation have submissions anonymized and labeled by ice cream flavor.

First round submissions.

Finding a common subspace for the submissions.

Physics-informed 🟦🟦🟦

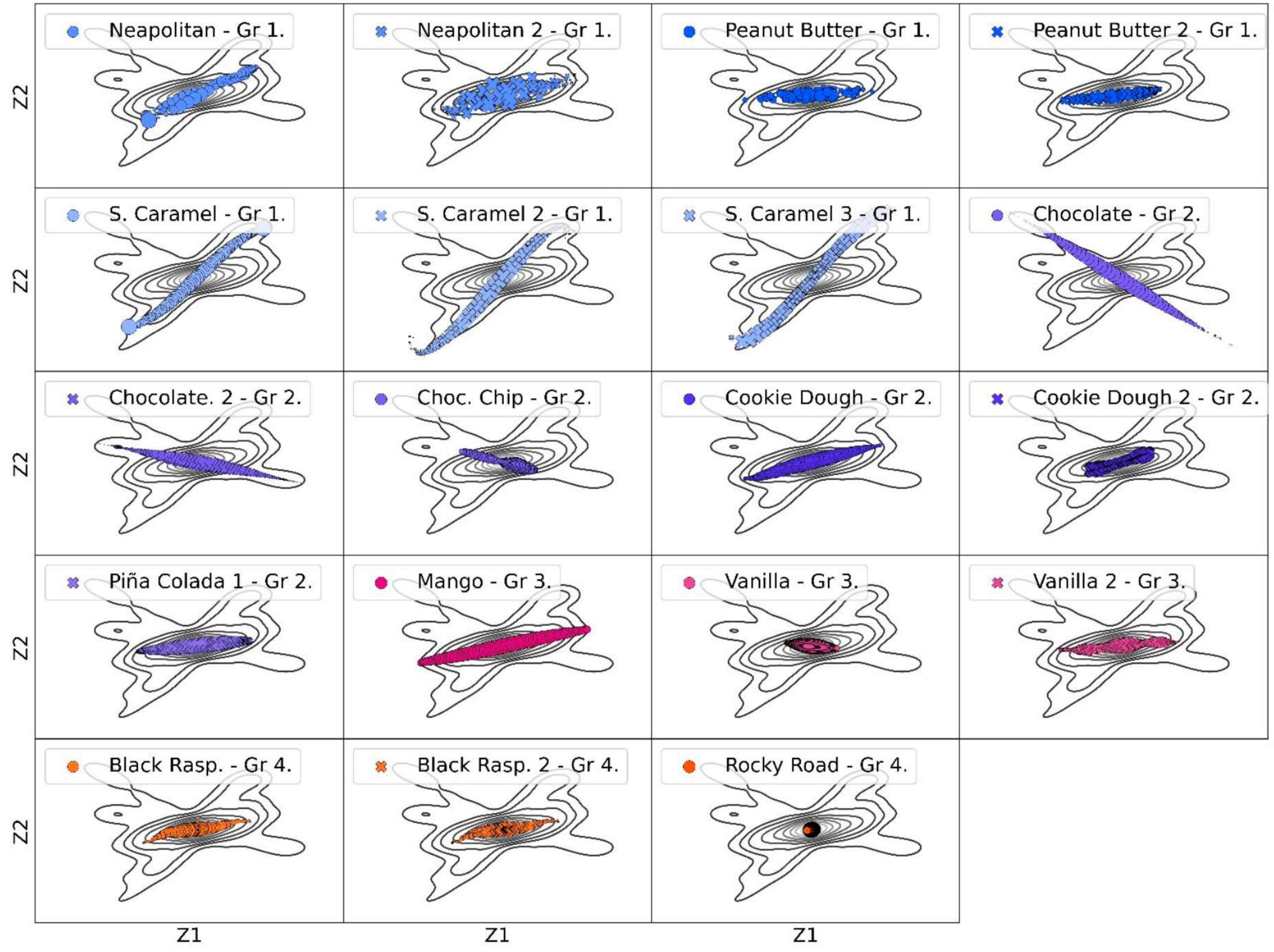
Neural network 🟪🟪🟪

Non-linear 🟧🟧

Linear 🟥🟥

Ground Truth 🟨

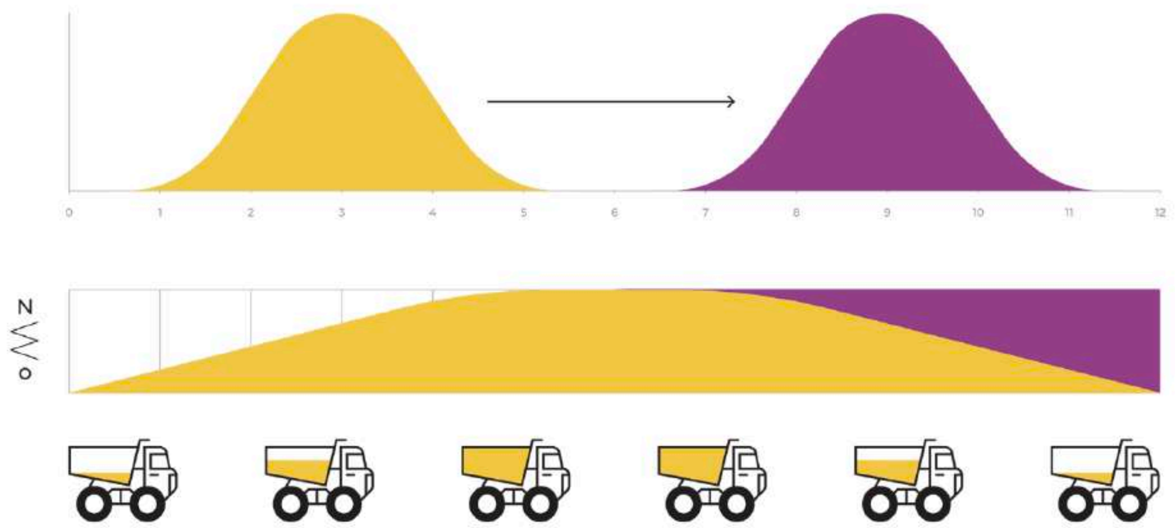
Experimental Dataset

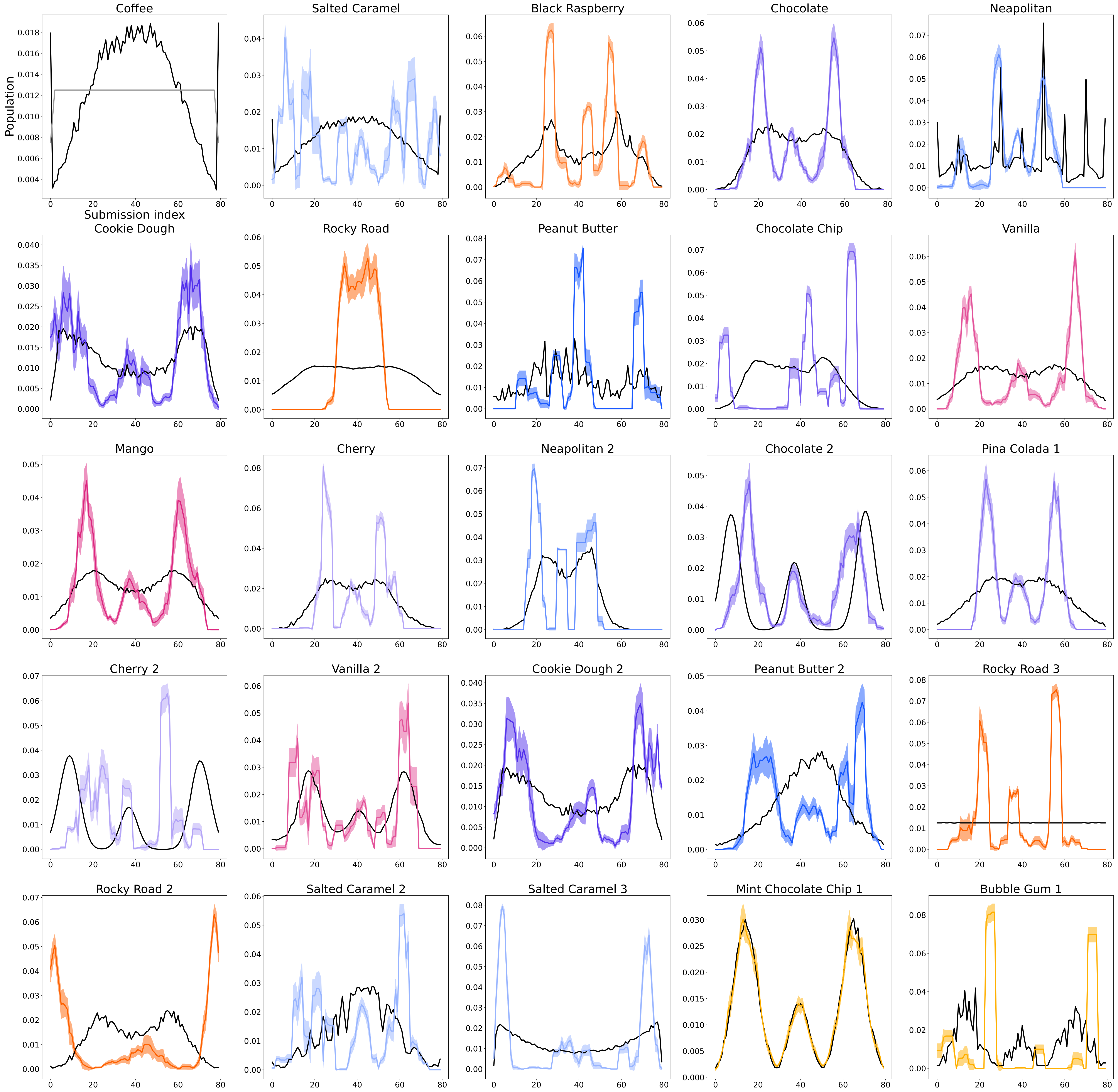


Obtain optimal distribution in the submitted volume space.

Black line
= submitted distributions
Color line
= optimal distributions

$$\mathrm{EMD} = \min_{T\in\mathbb{R}^{|\mathrm{GT}|\times|\mathrm{sub}|}} \sum_{ij} T_{ij} d_{ij}^{(\mathrm{FSC})}$$



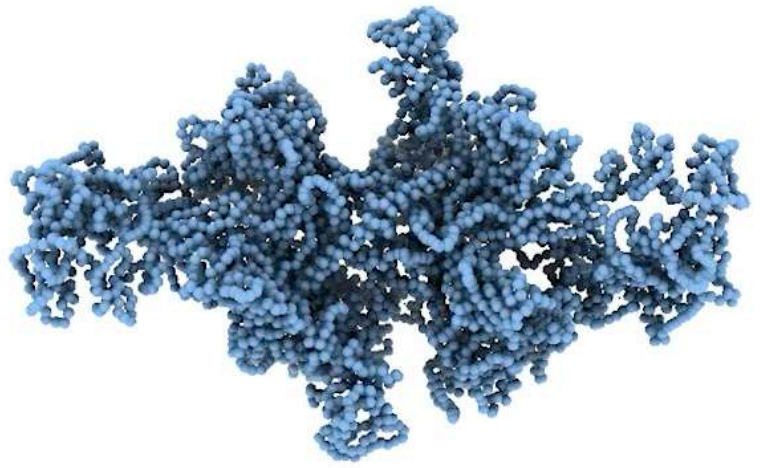


Geoffrey Woollard

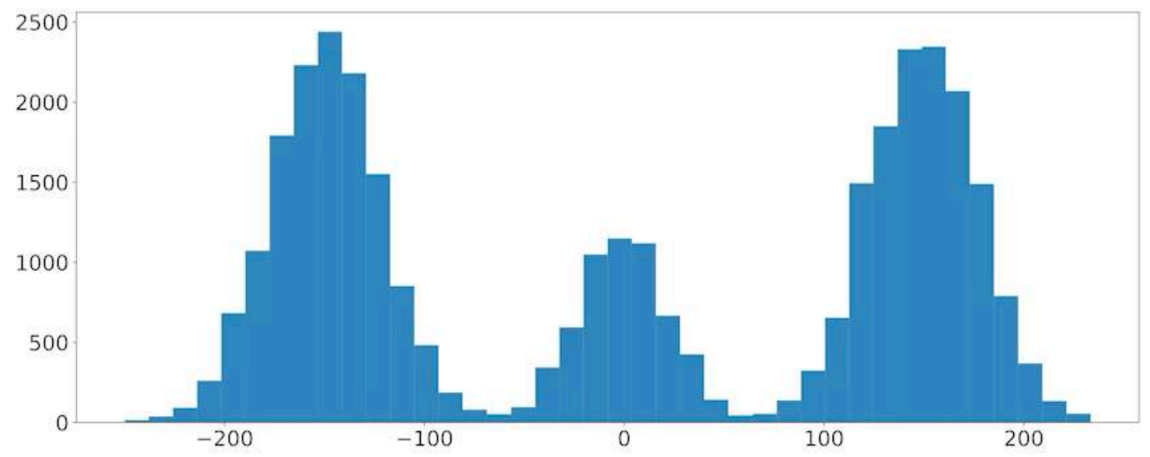
Synthetic Data

Are we restricted to heterogeneity metric where a ground truth is known?

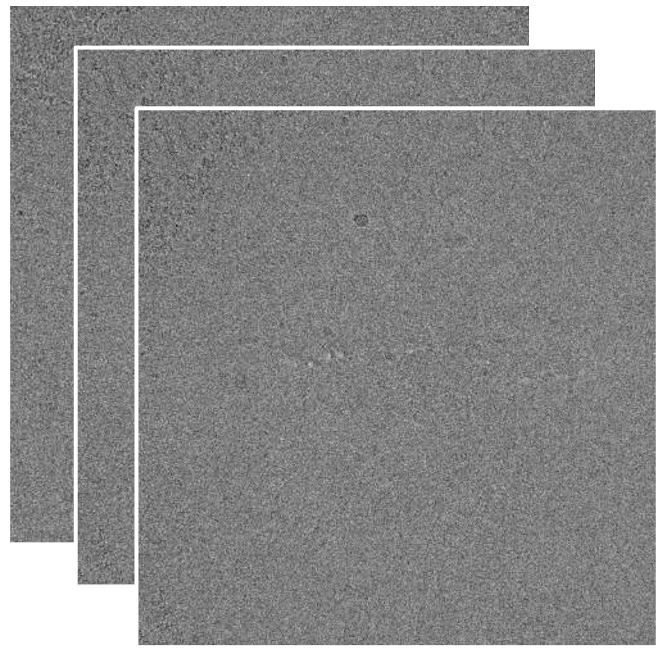
Known Volumes



Known Distribution



What we provided:



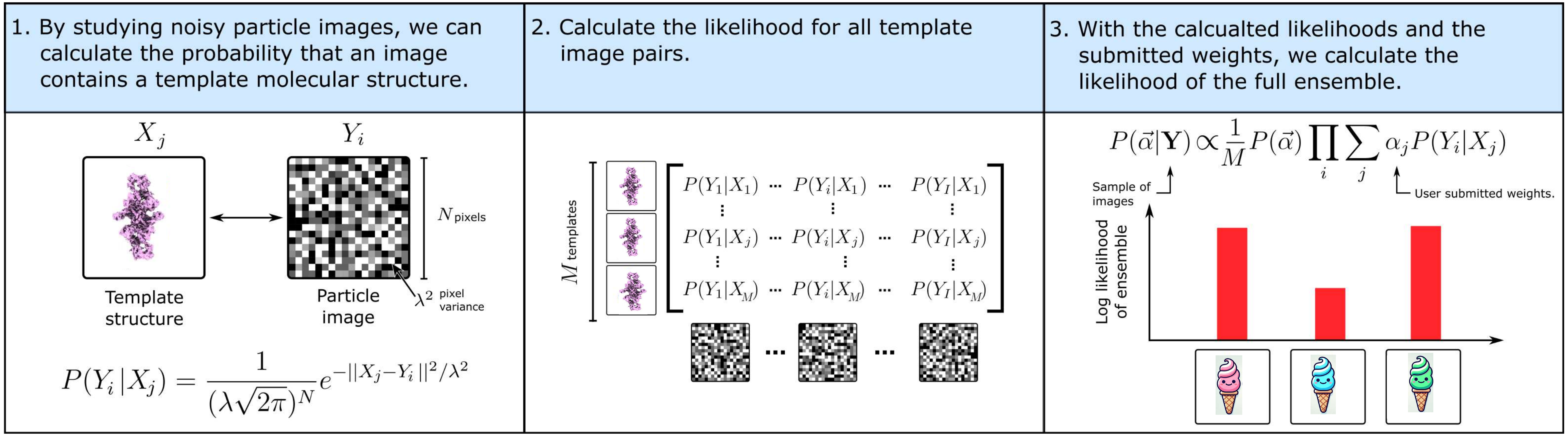
33 742 particles

Are we restricted to heterogeneity metric where a ground truth is known?

CryoLike: A python package for cryo-electron microscopy image-to-structure likelihood calculations

Wai Shing Tang, Jeff Soules, Aaditya Rangan, Pilar Cossio

doi: https://doi.org/10.1101/2024.10.18.619077



1. By studying noisy particle images, we can calculate the probability that an image contains a template molecular structure.
2. Calculate the likelihood for all template image pairs.
3. With the calcualted likelihoods and the submitted weights, we calculate the likelihood of the full ensemble.

$$P(Y_i|X_j) = \frac{1}{(\lambda\sqrt{2\pi})^N} e^{-||X_j - Y_i||^2/\lambda^2}$$

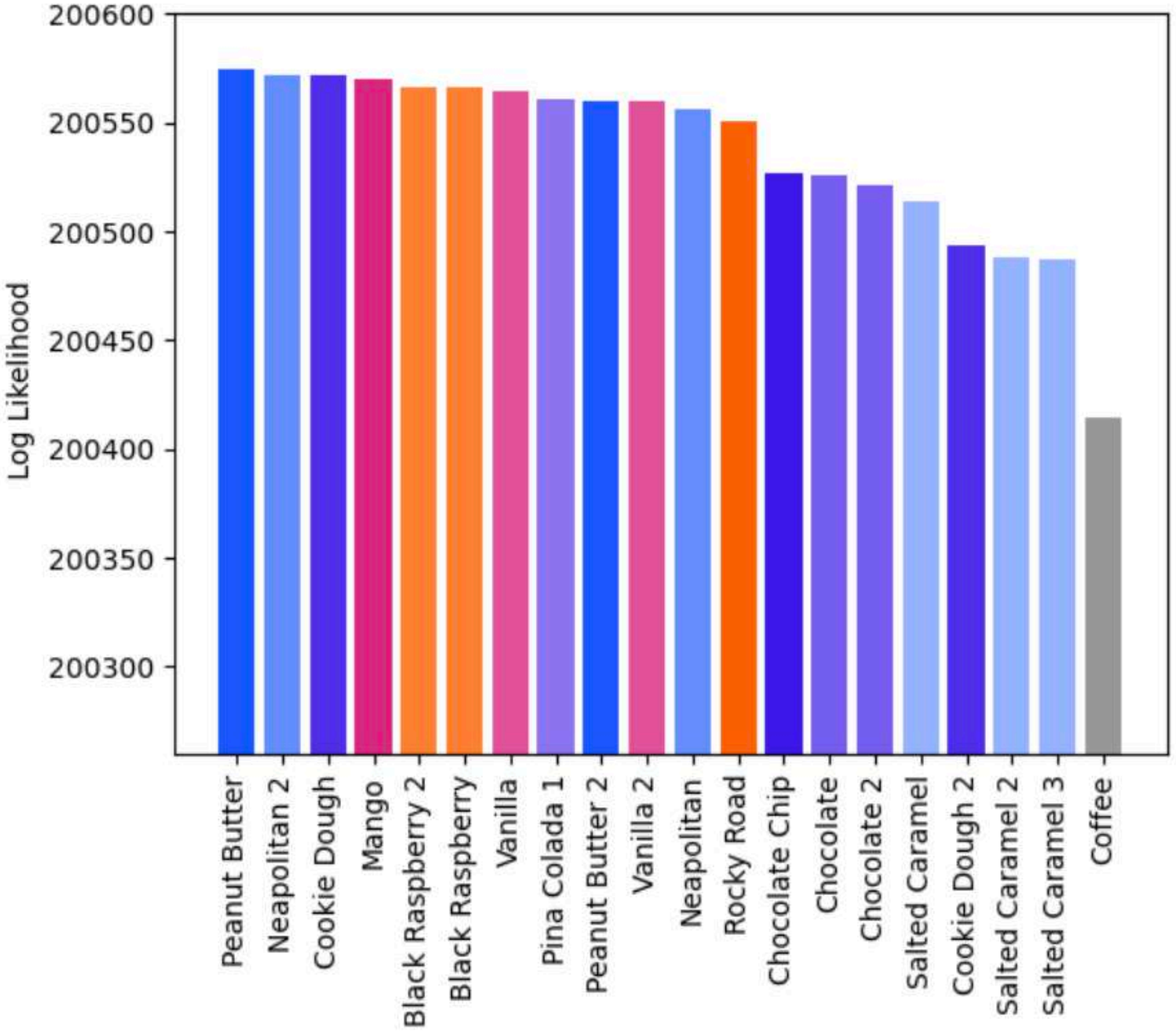
$$\begin{bmatrix} P(Y_1|X_1) & \cdots & P(Y_i|X_1) & \cdots & P(Y_I|X_1) \\ \vdots & & \vdots & & \vdots \\ P(Y_1|X_j) & \cdots & P(Y_i|X_j) & \cdots & P(Y_I|X_j) \\ \vdots & & \vdots & & \vdots \\ P(Y_1|X_M) & \cdots & P(Y_i|X_M) & \cdots & P(Y_I|X_M) \end{bmatrix}$$

$$P(\vec{\alpha}|\mathbf{Y}) \propto \frac{1}{M} P(\vec{\alpha}) \prod_i \sum_j \alpha_j P(Y_i|X_j)$$

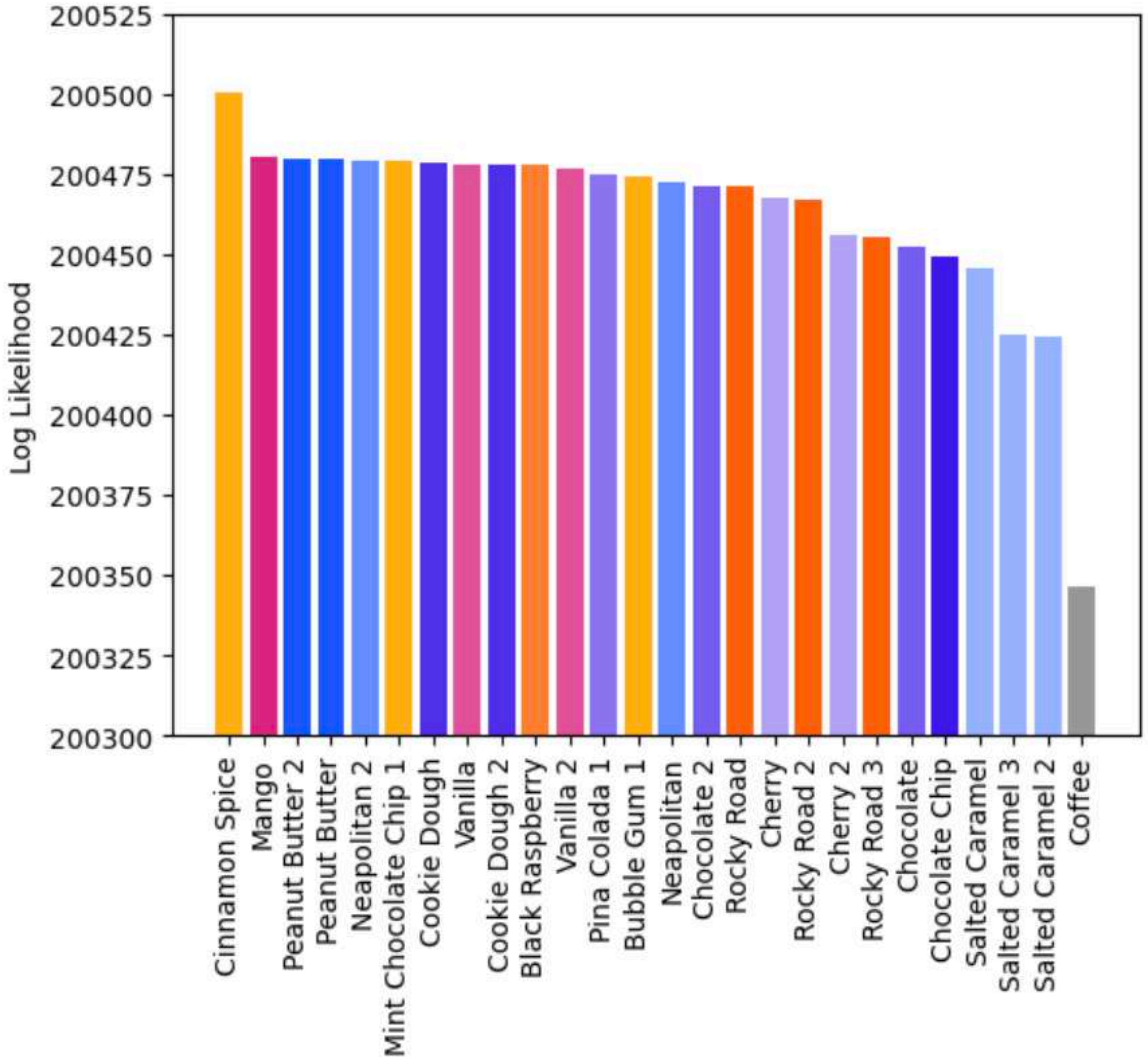
https://github.com/flatironinstitute/CryoLike

A new metric: Directly calculate the likelihood of an image having been created from a given volume.

FLATIRON INSTITUTE



Experimental Dataset



Synthetic Dataset

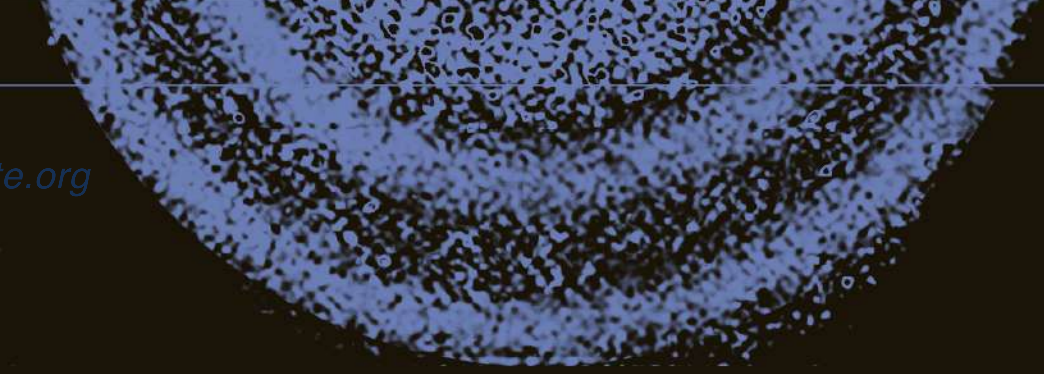
DISCUSSION TOPICS

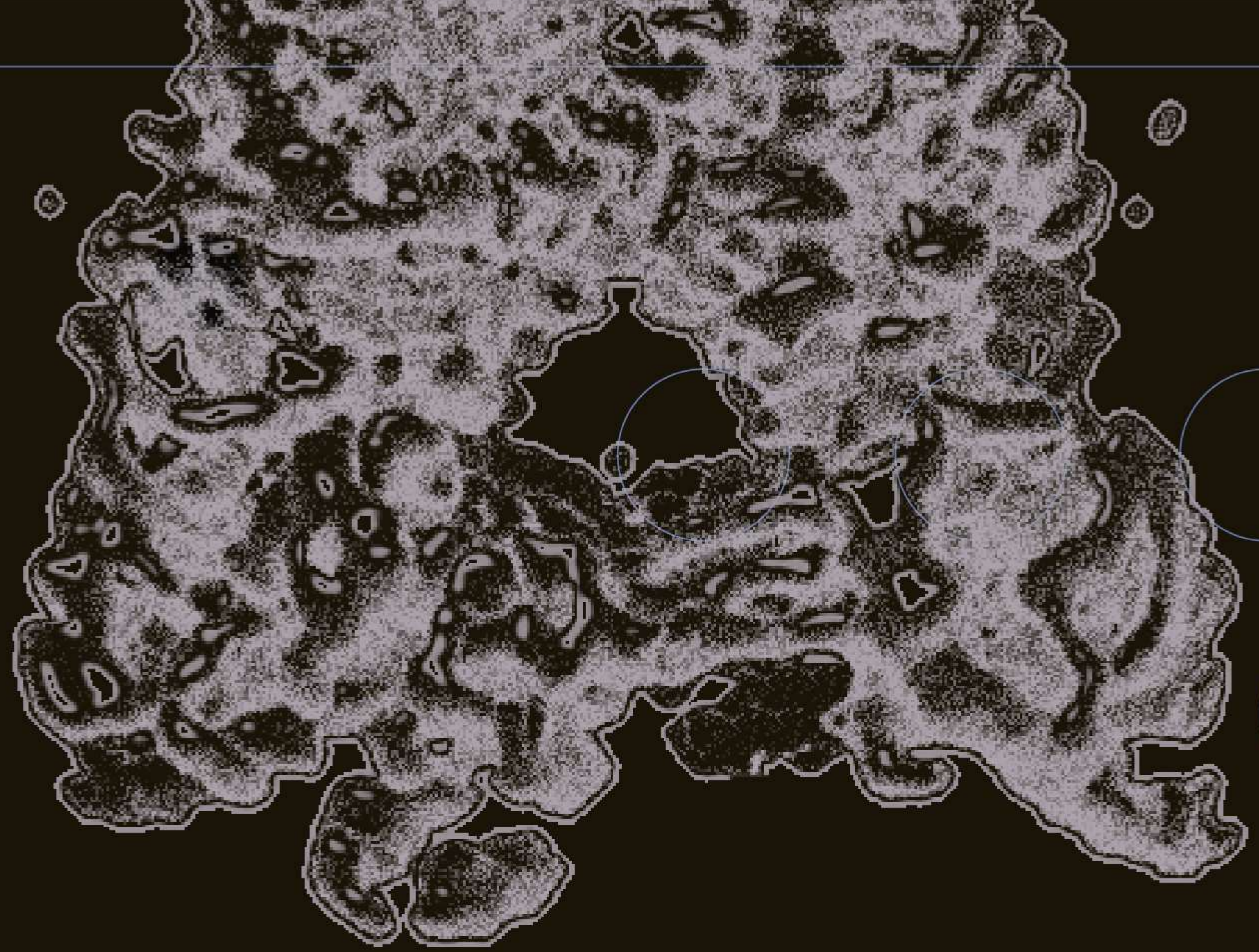
How should we benchmark methods in continuous heterogeneity for cryo-EM?

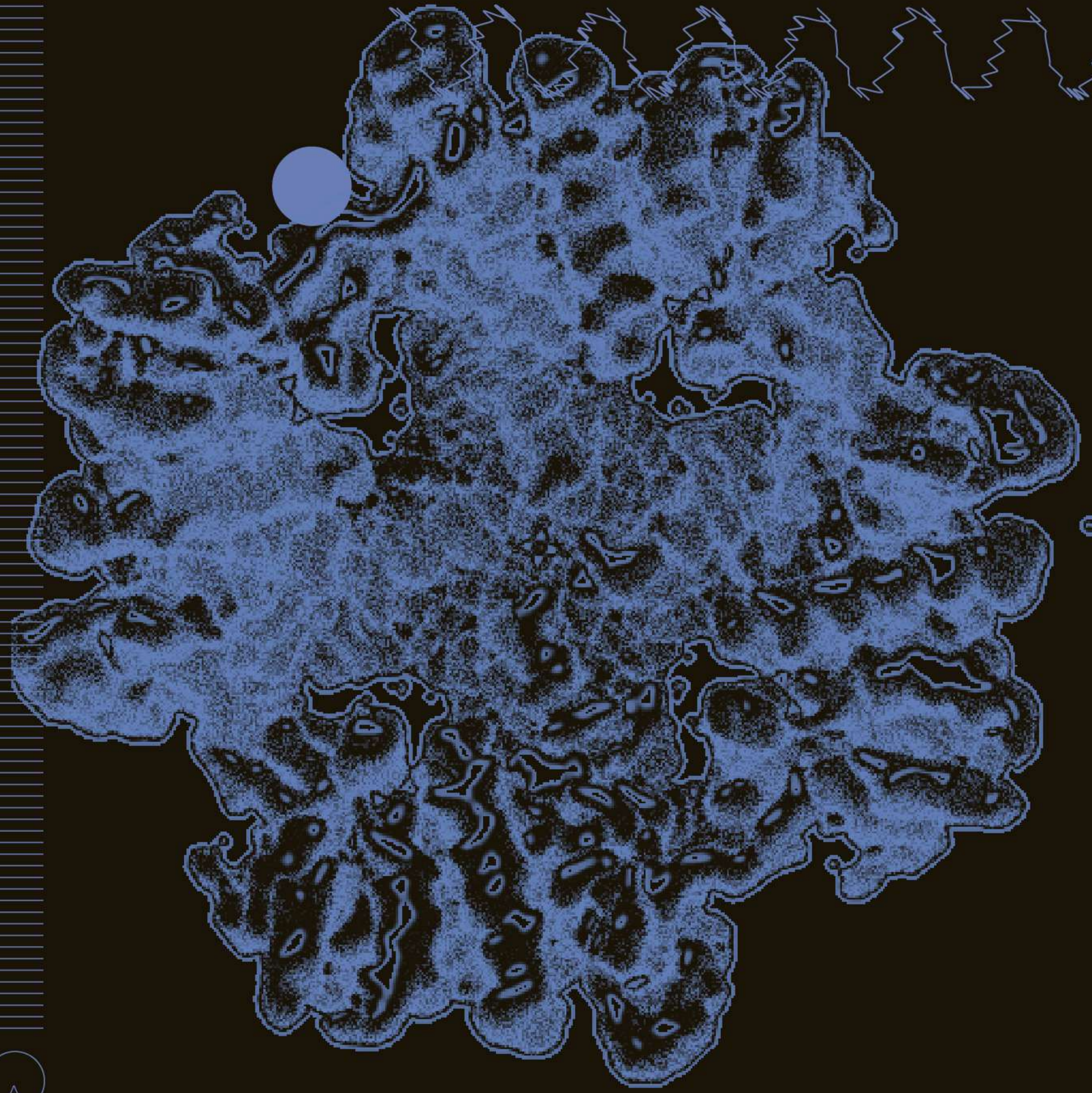
How much should we prioritize getting populations?

How much do we think the vitrification process affects our ensembles?

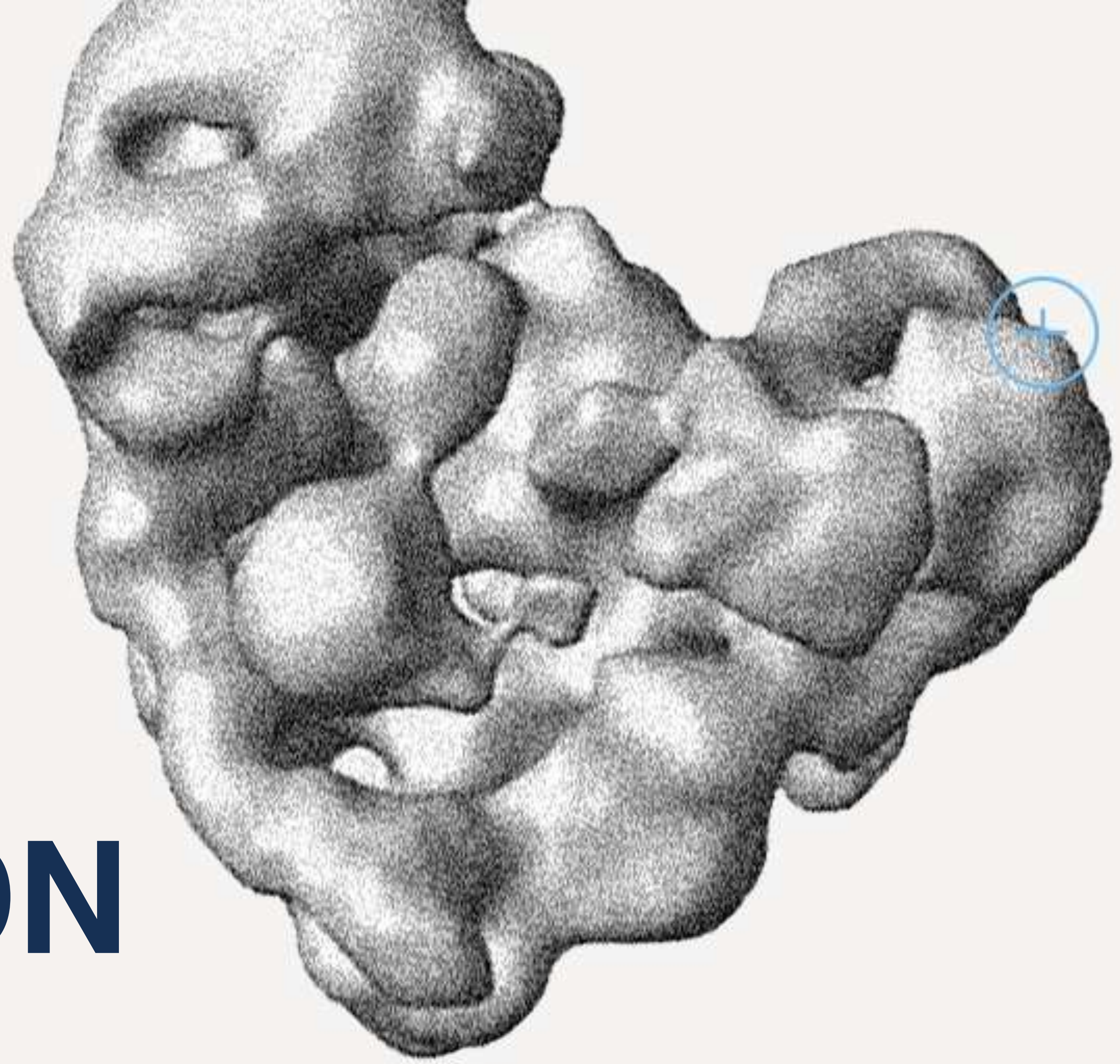
et cetera







Toward continuous heterogeneity in *in situ* cryo-EM and cryo-ET.



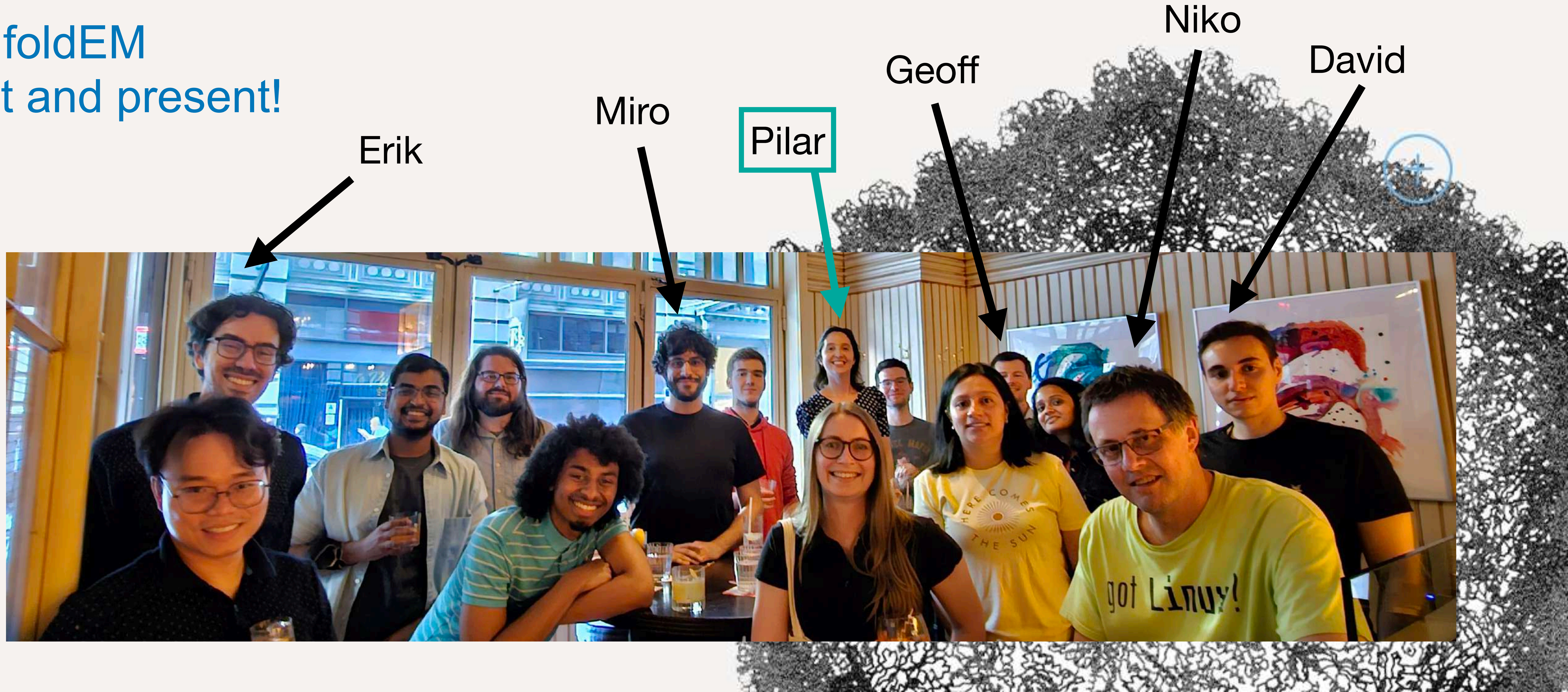
THANKS FOR YOUR ATTENTION

Pablo Debenedetti
Syma Khalid
Louis Smith

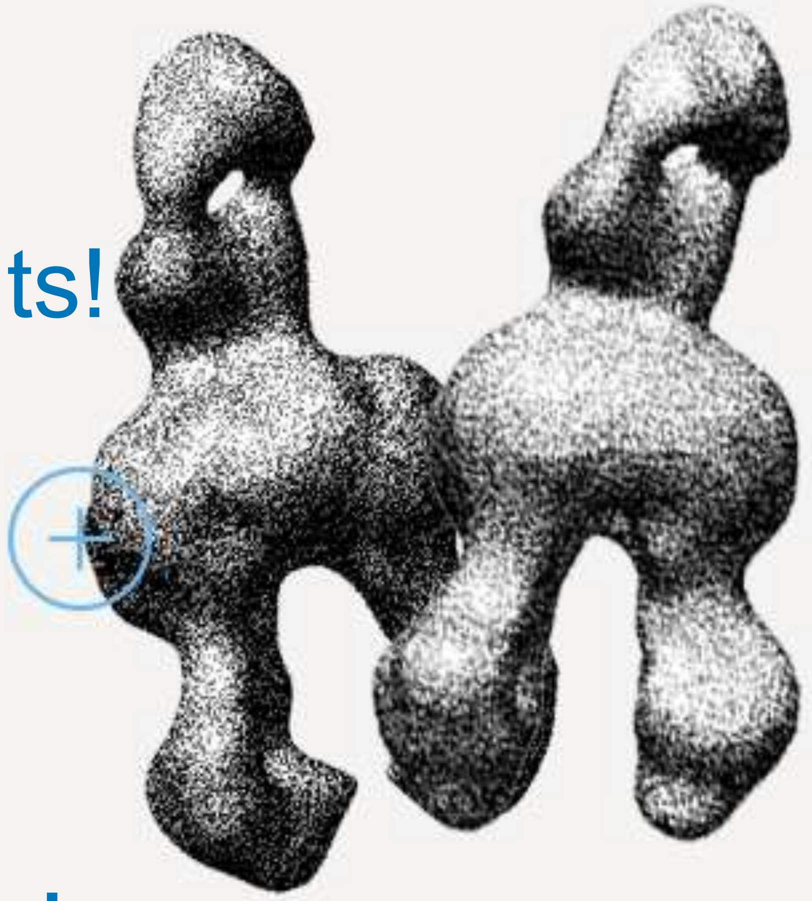
Wai Shing Tang

Robbie Clark

Thanks to ManifoldEM developers past and present!



Thanks to all challenge participants!



^^^ Thanks to SMBp!