

# Consistency Criterion for Particle Sorting in Single-Particle Cryo-EM

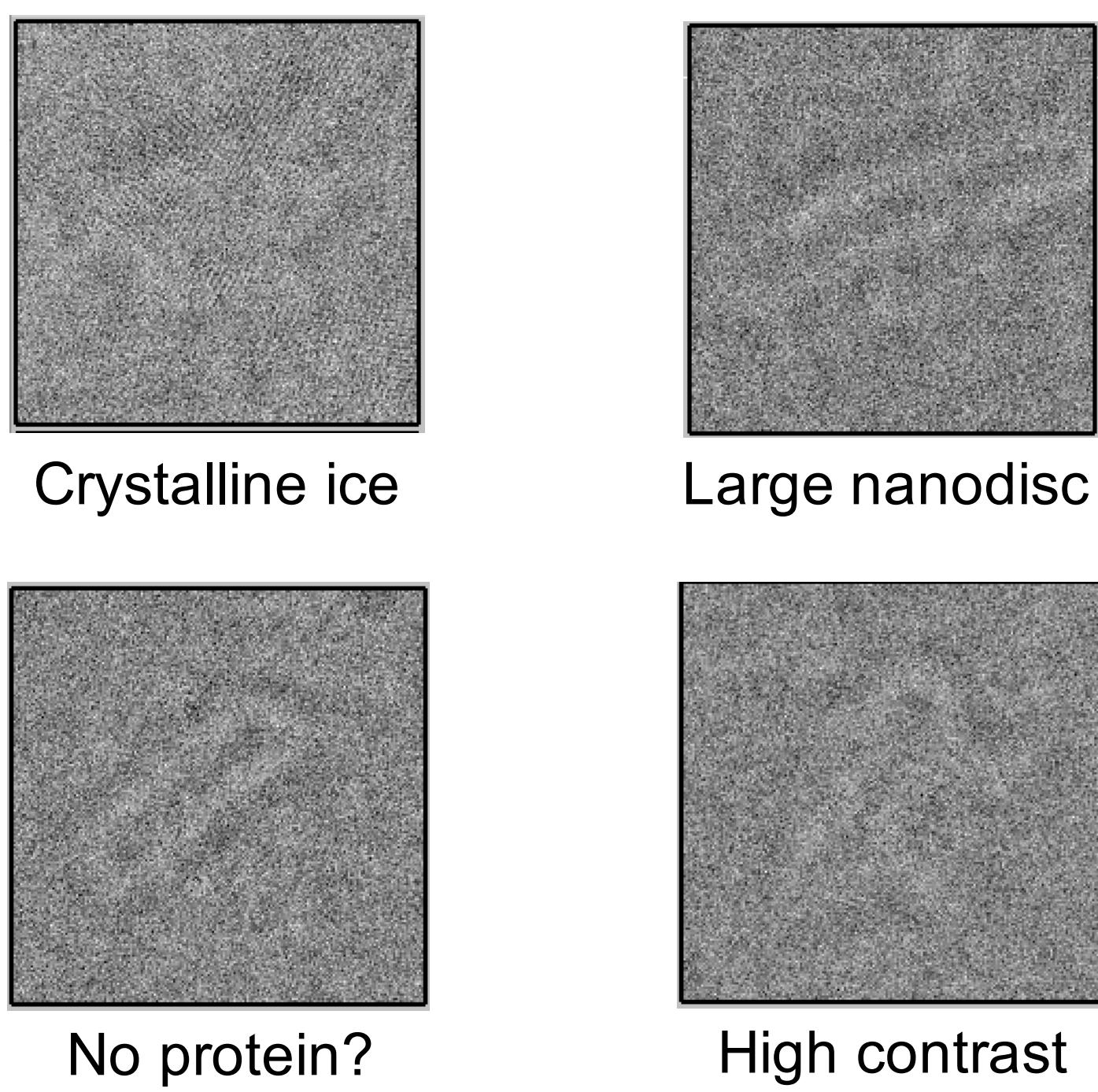


Daniel Asarnow<sup>1</sup>, and Yifan Cheng<sup>1,2</sup>

<sup>1</sup>Department of Biochemistry and Biophysics, University of California, San Francisco <sup>2</sup>Howard Hughes Medical Institute, University of California, San Francisco

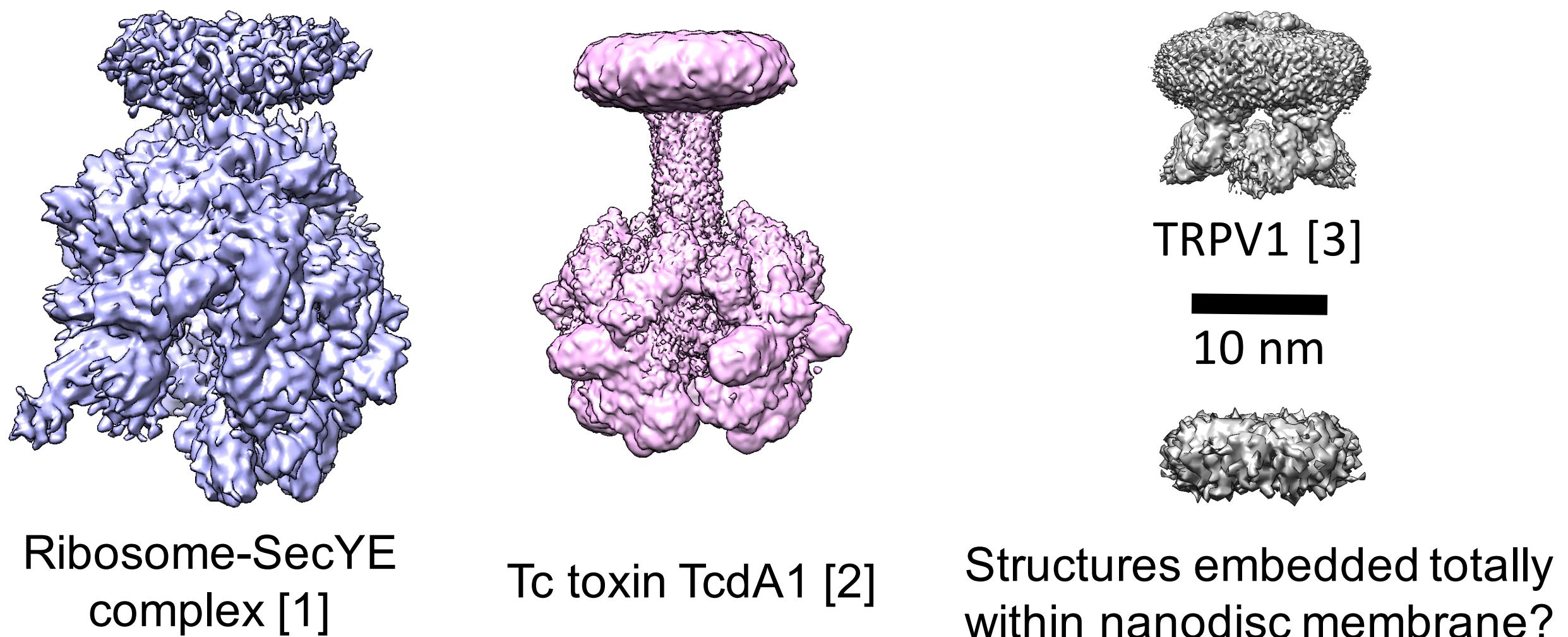
## Introduction

- Current cryo-EM datasets contain tens of thousands to millions of high quality single-particle images.
- They also contain many “bad” particles: damaged or unfolded protein or other contaminants.
- High-resolution reconstructions hinge on:
  - Identifying and eliminating “bad” particles
  - Separating remaining particles into distinct classes
  - Obtaining accurate orientation estimates within each class
- “Particle sorting” attempts to differentiate “good” and “bad” particles.
- Typically does not address misalignment of “good” particles.



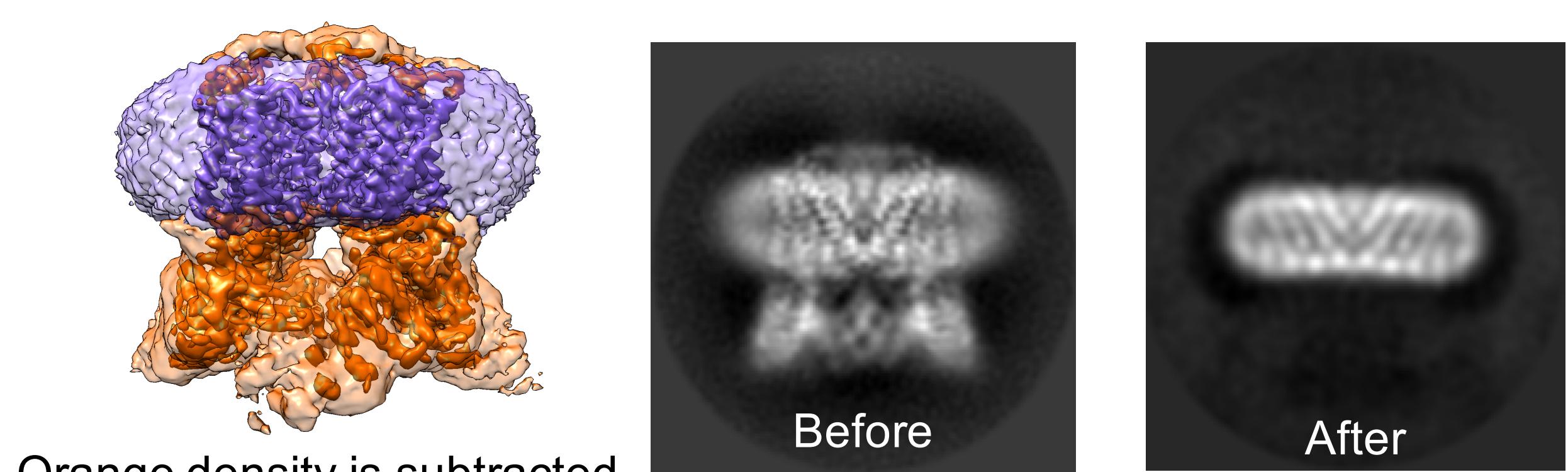
## Cryo-EM of small membrane proteins

- Membrane proteins must be solubilized in some kind of membrane mimic before imaging.
- Native-like membrane mimics, such as lipid nanodiscs, contain significant amounts of unstructured mass.
- All bilayer-bound cryo-EM structures to date feature large soluble domains, often dwarfing the membrane component.
- In contrast, membrane proteins without soluble domains lack clear structural features to cue alignment of particle images.



## Synthetic “disc-only” TRPV1 dataset

- In order to study the impact of the membrane environment, we used partial signal subtraction [4] to simulate cryo-EM data for a hypothetical membrane protein lacking any soluble domain and reconstituted in lipid nanodiscs.
- Starting from TRPV1 in lipid nanodiscs [3], we subtracted soluble parts of the structure, leaving the TM region and disc untouched.
- The approach allows direct comparison with the original data, which contains significant structural features protruding from the disc.
- EMPIAR (acc. 10059), EMDB (acc. 8117), PDB (acc. 5irx).
- Subtraction code available at <http://github.com/asarnow/pyem>.

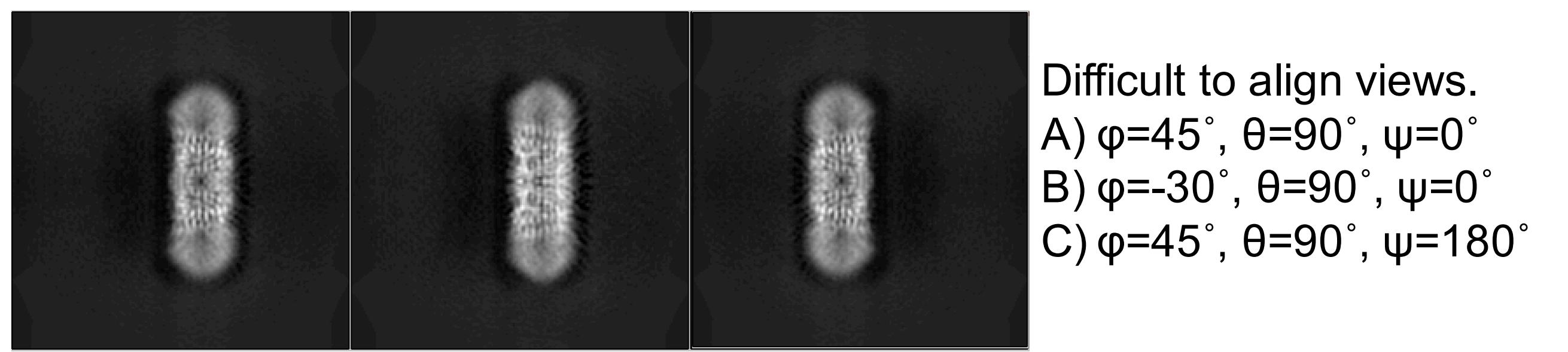


## The membrane hinders alignment

- Relion [5] auto-refine yields at 3.2 Å raw map (C4 symmetry).
- This is >0.2 Å worse than the original, whole TRPV1 map.
- Refinements of subsampled datasets in Cryosparc [6] demonstrate more disc-only particles are needed to obtain the same resolution.

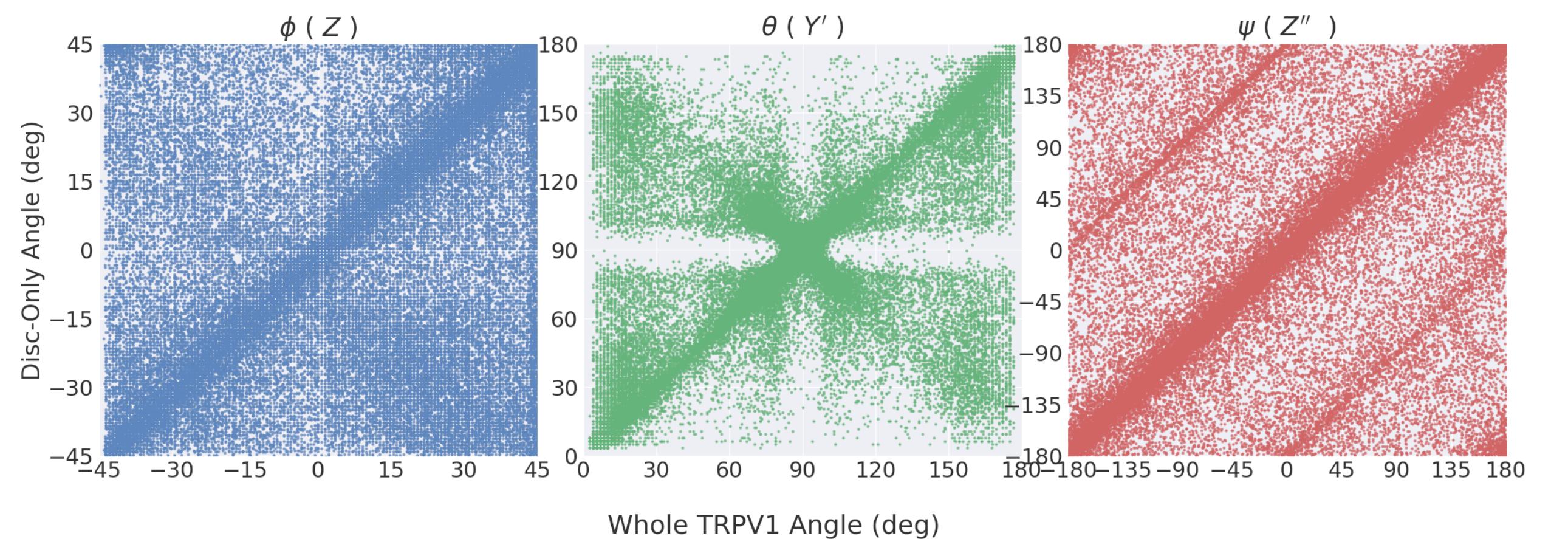
TRPV1 in lipid nanodiscs					Soluble elements subtracted				
FSC resolution (Å) vs. data size (# particles / fraction)									
73,928	55,446	36,964	16,303	7,361	73,928	55,446	36,964	16,303	7,361
1.0	0.75	0.5	0.25	0.1	1.0	0.75	0.5	0.25	0.1
2.97	3.01	3.12	3.34	3.58 A	3.48	4.31	8.27	9.28	-

- Example projections suggest that the subtraction introduces new D4 and C<sub>∞</sub> pseudo-symmetries (TRPV1 is actually in C4).
- These pseudosymmetries can degrade the accuracy of orientation estimates or induce wholesale misalignments.



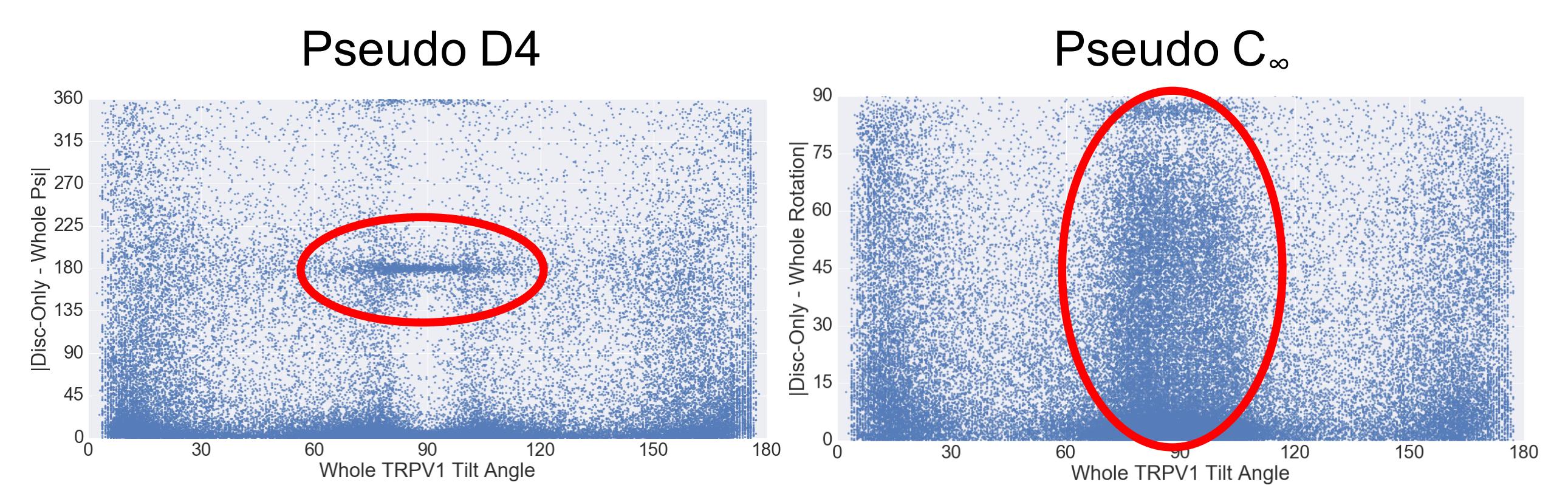
## Misalignment is frequent and stereotypical

- Euler angle assignments for the same particles in the whole and synthetic disc-only datasets show widespread disagreement.
- By convention, Euler angles,  $\phi$  (rotation),  $\theta$  (tilt), and  $\psi$ , take an object from standard orientation to final pose by rotating around Z, the new Y, and the new Z axis, respectively.



- The distribution of differences between angle assignment show certain views are stereotypically misaligned.
- This observations supports the pseudosymmetry hypothesis.
- Plots of the absolute angular difference in either  $\phi$  or  $\psi$  with respect to  $\theta$  show this effect clearly.

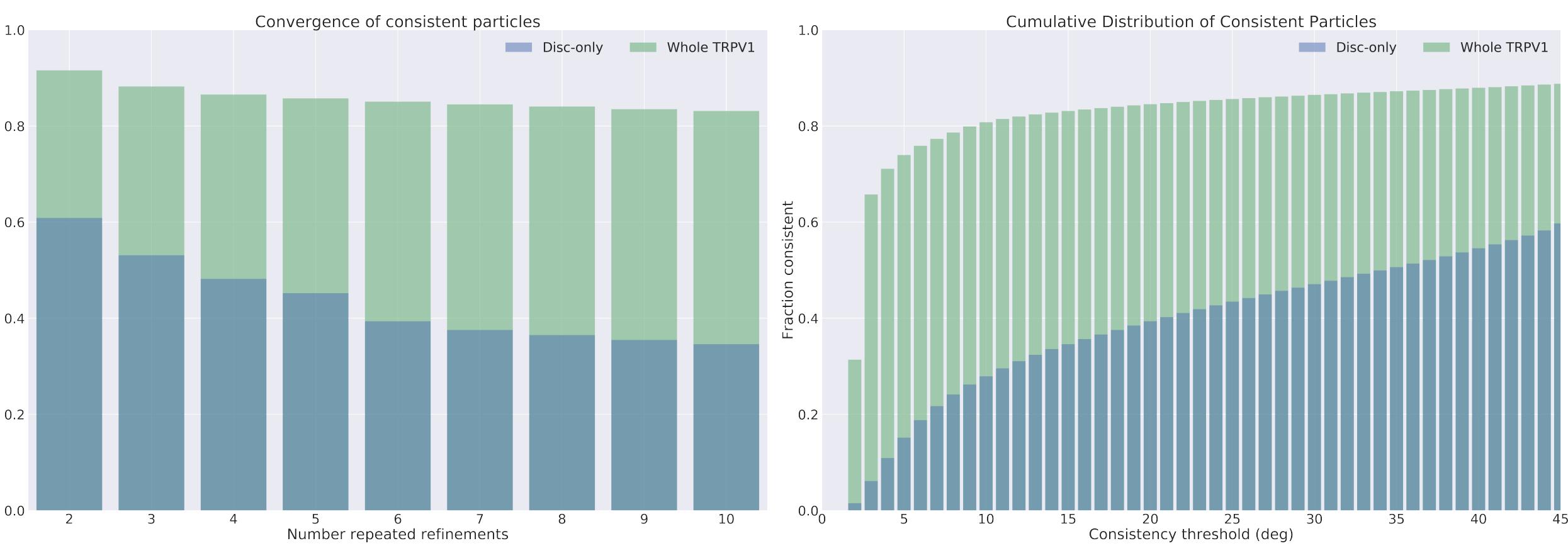
Side views rotated around:



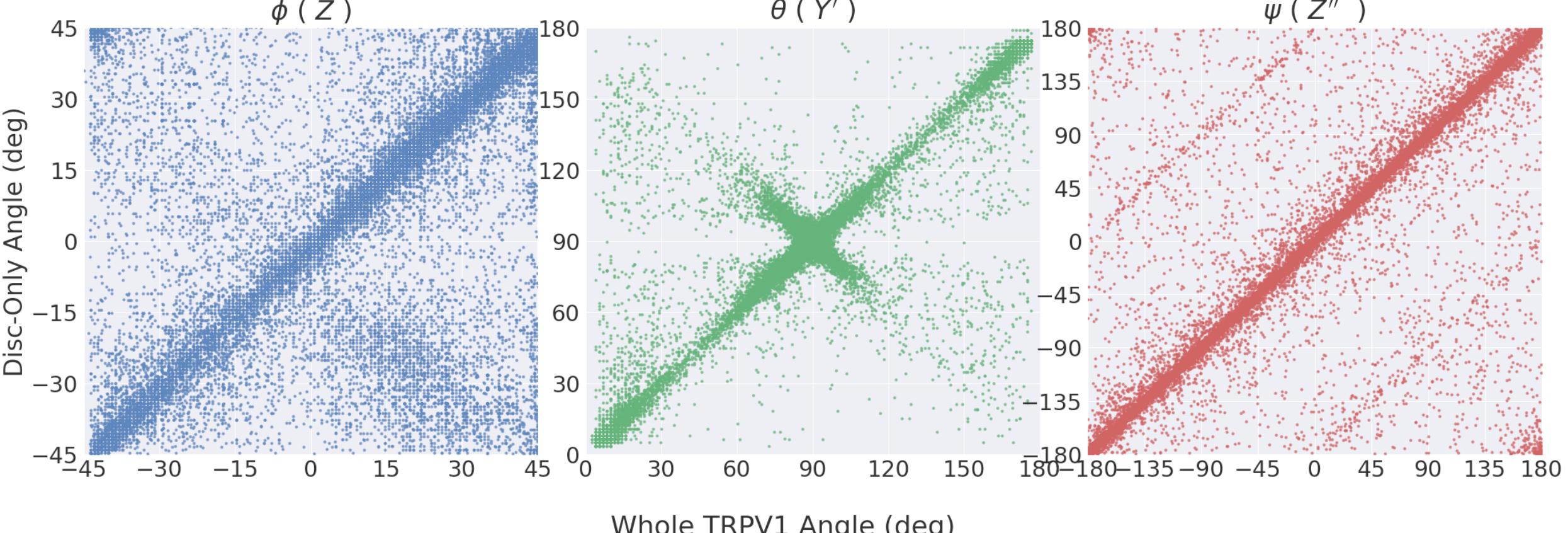
- An approach is needed to identify misaligned particles without reference to known values (supplied above by very high resolution refinements of the intact TRPV1 particle).

## Measuring orientation consistency

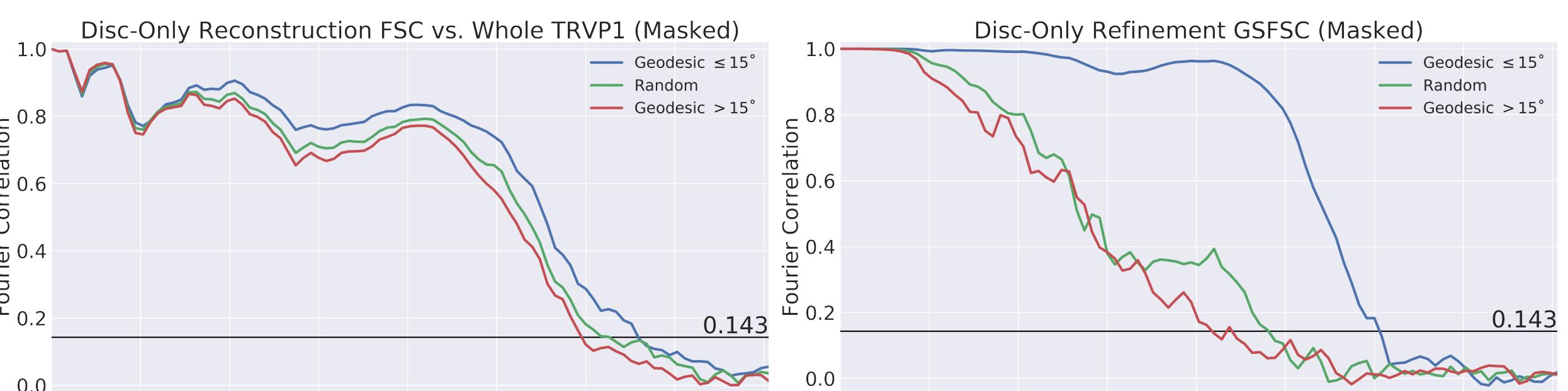
- We observed that misaligned particles also obtained divergent orientation assignments upon repetition of *de novo* 3D refinement.
- Thus, we attempt to differentiate “misaligned” and “well-aligned” particles based on consistency of particle alignment across multiple *de novo* 3D refinements with different random initializations.
- Particles are considered “consistent” if the geodesic distance in SO(3) [7] between angle assignments is never greater than a threshold value, across all pairs of repeated refinements (e.g. 15°).
- Particles are sampled randomly to yield datasets of constant size. The TRPV1 data contains 73,928 particles; 25,580 particles were consistent to 15° and this number was used in analysis.
- Refinement was repeated N=10 times (45 distances/particle), at which point the set of consistent particles was essentially converged.



- Angles for consistent, disc-only particles are more congruent with those for the intact TRPV1 particle.



- Consistent particles yield subjectively higher quality maps.
- This is recapitulated by FSC curves against the whole TRPV1 map and by GSFSC for *de novo* refinement of these particles alone.



## Discussion

- High-resolution cryo-EM of relatively small proteins completely embedded in discoidal lipid bilayers appears feasible.
- However, the presence of a contrastive medium around the protein hinders alignment via introduction of pseudo-symmetries.
- Misaligned particles can be identified by their inconsistent angle assignments during multiple *de novo* refinements (different seeds).
- Excluding such particles enhances reconstruction quality, and well-aligning particles are required for successful 3D refinement.
- Back projection of inconsistent particles with angles refined in the presence of consistent particles, yields relatively good maps.
- This observation suggests that such particles are likely misaligned “good” particles, rather than residual “bad” particles persisting through 3D classification.
- Misalignments are likely commonplace in single-particle analysis, and may explain why mostly non-overlapping sets of particles from the same micrographs can often be refined to similar final resolutions using different image processing suites.

## References and acknowledgement

- Funding by Howard Hughes Medical Institute and US National Institutes of Health 5R01GM098672-07 and 5T32GM008284-27.
- [1] J. Fraunfelder et al., *Nat. Struct. Mol. Biol.*, vol. 18, no. 5, pp. 614–621, May 2011.
  - [2] C. Gatsogiannis et al., *Nat. Struct. Mol. Biol.*, vol. 23, no. 10, pp. 884–890, Oct. 2016.
  - [3] Y. Gao, E. Cao, D. Julius, and Y. Cheng, *Nature*, vol. 534, no. 7607, pp. 347–351, Jun. 2016.
  - [4] S. H. W. Scheres, *B.-M. in Enzymology*, Ed. Academic Press.
  - [5] S. H. W. Scheres, *J. Struct. Biol.*, vol. 180, no. 3, pp. 519–530, Dec. 2012.
  - [6] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, *Nat. Methods*, Feb. 2017.
  - [7] D. Q. Huynh, “Metrics for 3D Rotations: Comparison and Analysis,” *J. Math. Imaging and Vision*, vol. 35, no. 2, pp. 155–164, Oct. 2009.