Non-Euclidean Metrics for Cryo-EM Analysis

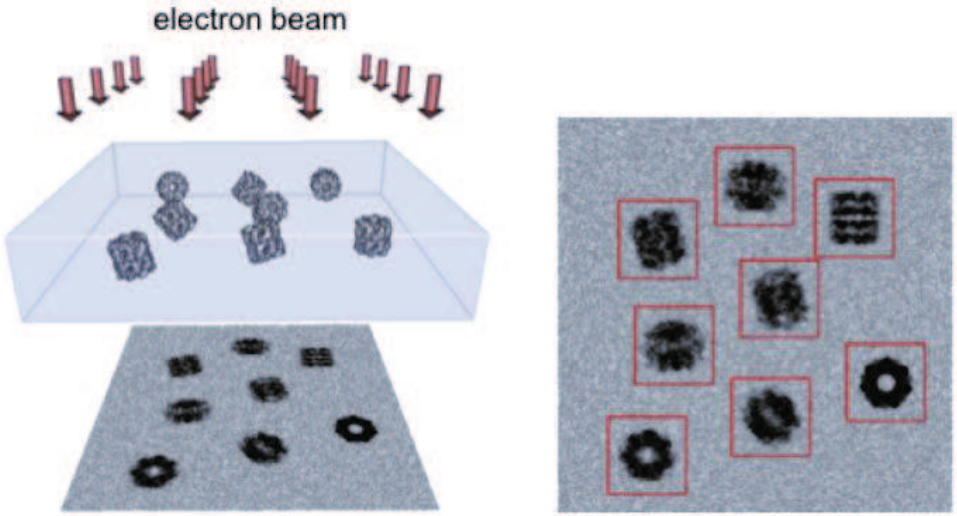
Amit Singer

Department of Mathematics and Program in Applied and Computational Mathematics

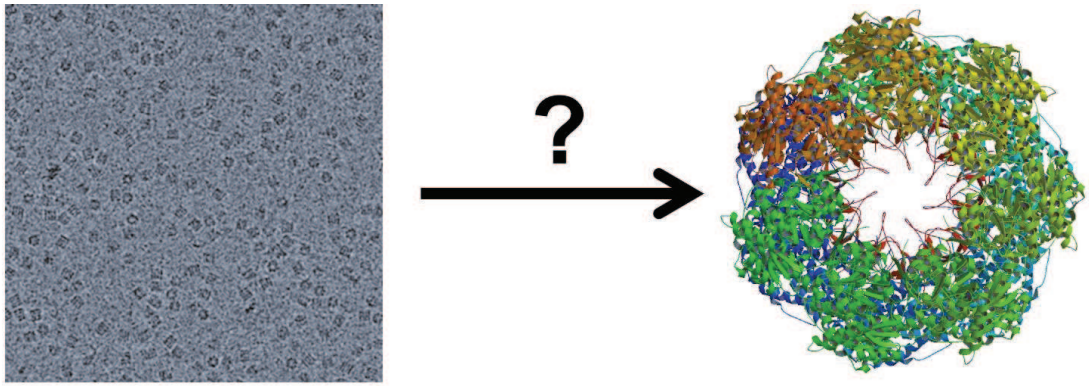
September 6, 2023

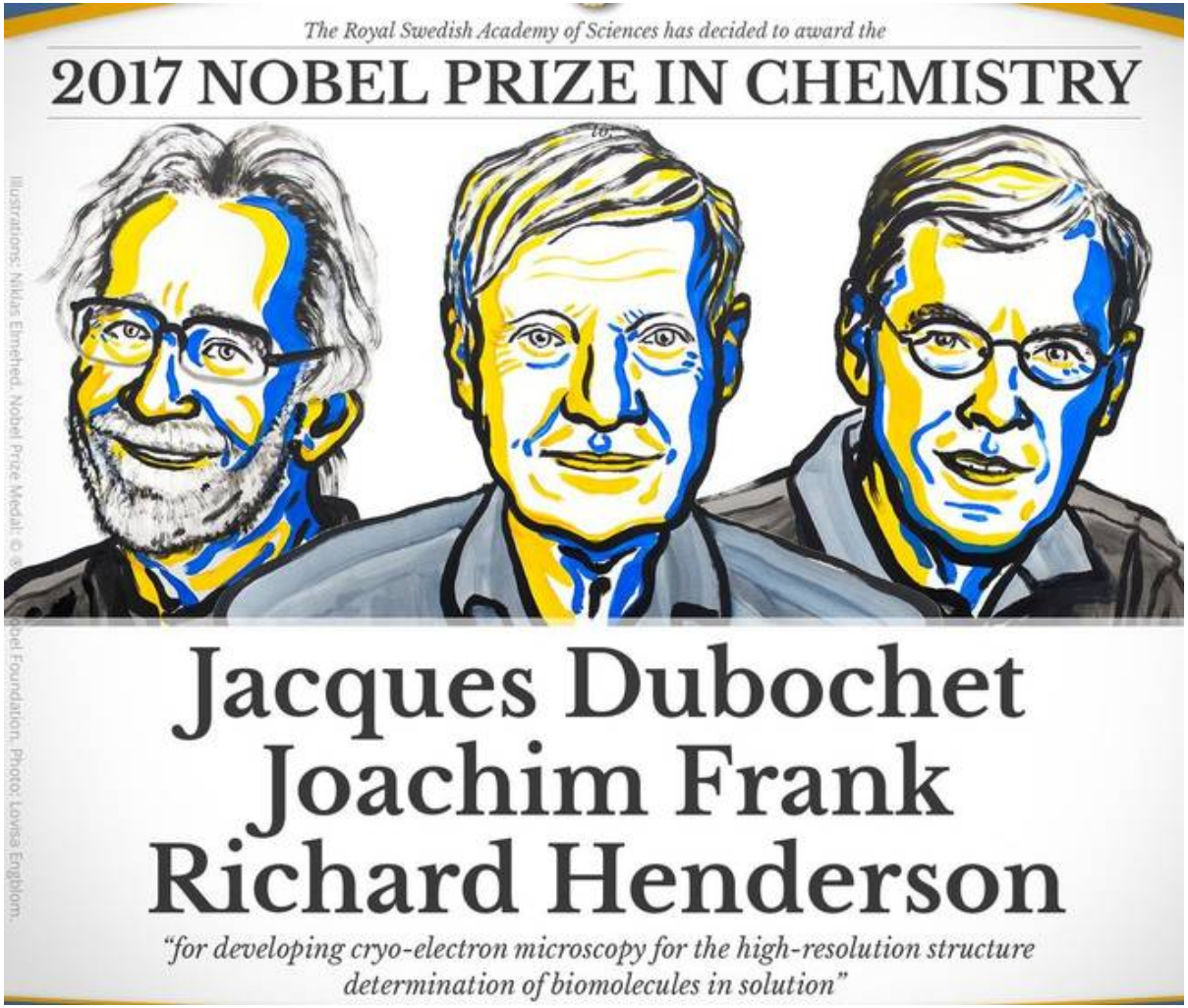
What is single particle cryo-EM?

Schematic drawing of the imaging process:



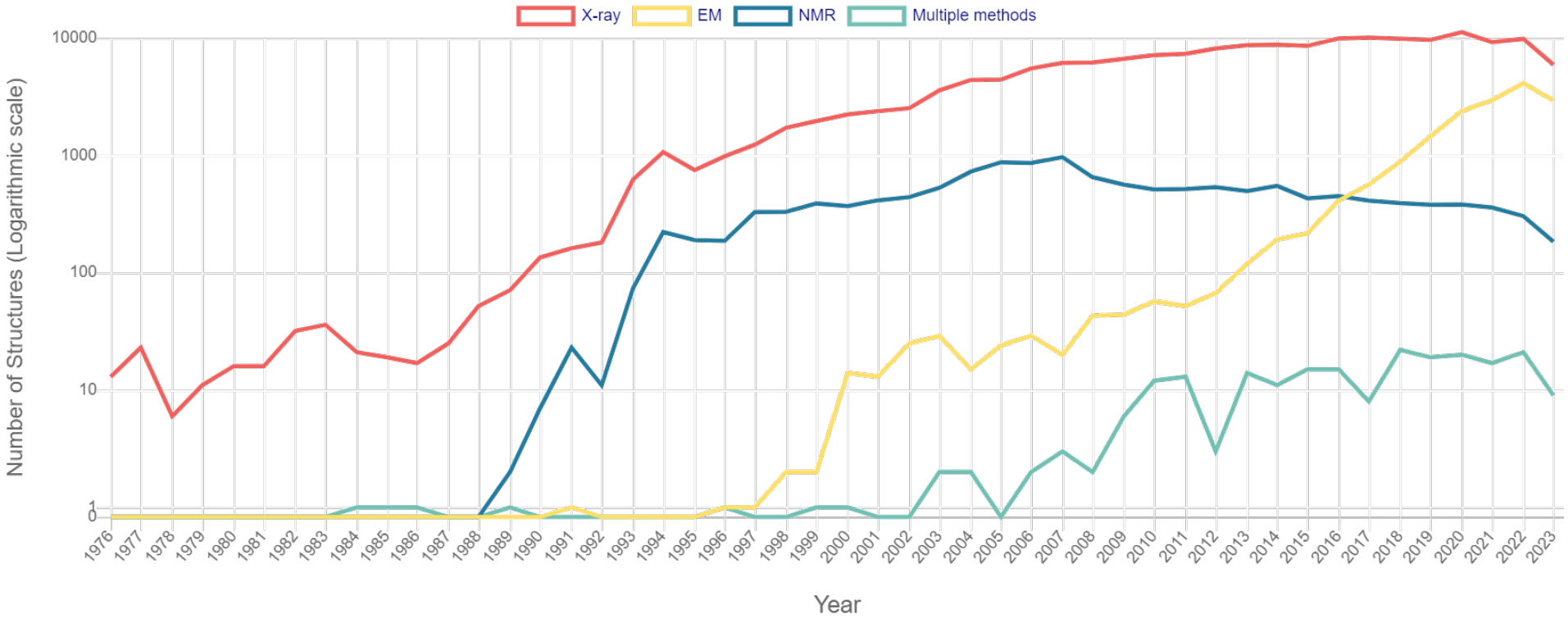
The standard cryo-EM reconstruction problem:



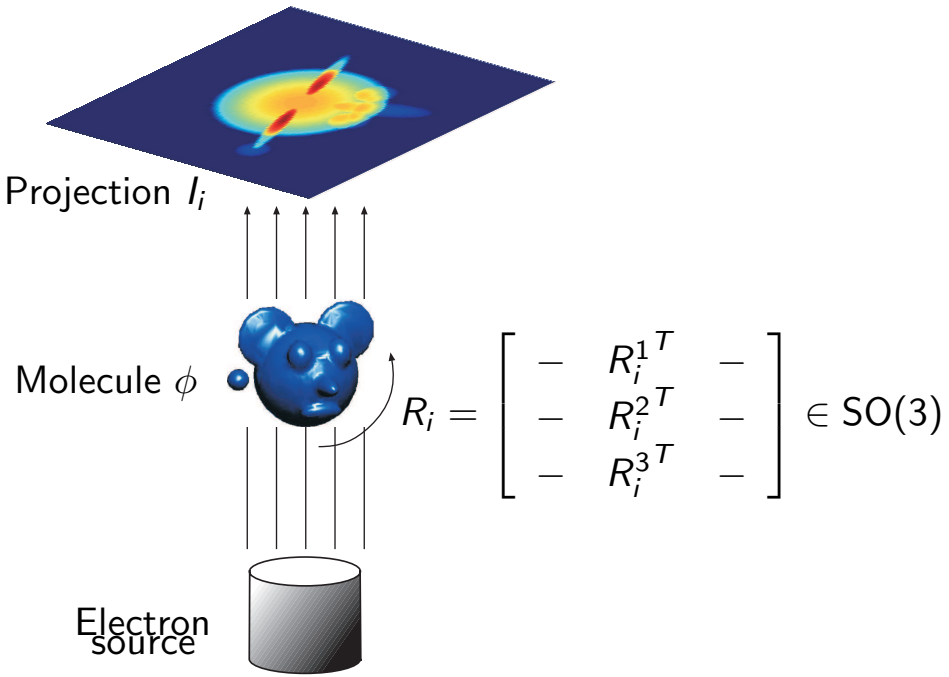


The Nobel Prize in Chemistry 2017 is awarded to Jacques Dubochet, Joachim Frank and Richard Henderson for the development of cryo-electron microscopy, which both simplifies and improves the imaging of biomolecules. This method has moved biochemistry into a new era. (The Royal Swedish Academy of Sciences)

Number of Released Protein Data Bank (PDB) Structures per Year



Cryo-EM reconstruction from picked particles



- Forward model (assuming perfectly centered picked particles and known CTF):

$$I_i(x,y) = h_i * \int_{-\infty}^{\infty} \phi(xR_i^1 + yR_i^2 + zR_i^3)\, dz + \text{“noise"}$$

- n images ($i = 1, \ldots, n$) of size $L \times L$ pixels
- $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is the electrostatic potential created by the molecule.
- The basic “reconstruction” problem: Estimate ϕ given $I_1, \ldots, I_n$.
- The “heterogeneity” problem: Estimate the distribution of ϕ given $I_1, \ldots, I_n$.

Experimental noisy images

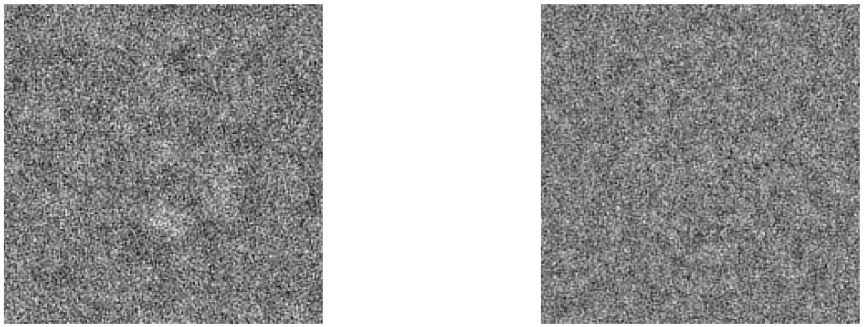


Figure: Kv1.2 Ion Channel. $16,911$ images, 192×192 pixels. Data courtesy of Dr. Fred Sigworth (Yale).

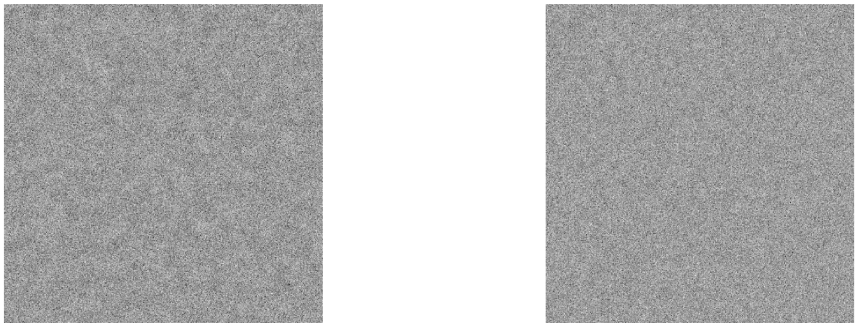


Figure: eIF2B. $99,526$ images, 458×458 pixels. Data courtesy of Dr. Adam Frost (UCSF) (Tsai et al., Science 2018)

Euclidean distances in cryo-EM analysis

- Noise is modeled as additive Gaussian (not necessarily white).
- Likelihood based methods lead to (weighted) Euclidean distances between images in 2-D classification, iterative 3-D model refinement (projection matching), and more.
- Error in reconstructed 3-D maps also assumed to be Gaussian.
- Alignment of 3-D "half maps" and 3-D heterogeneity analysis are based on Euclidean distances between 3-D maps.
- Does it make sense to use non-Euclidean metrics for cryo-EM analysis?

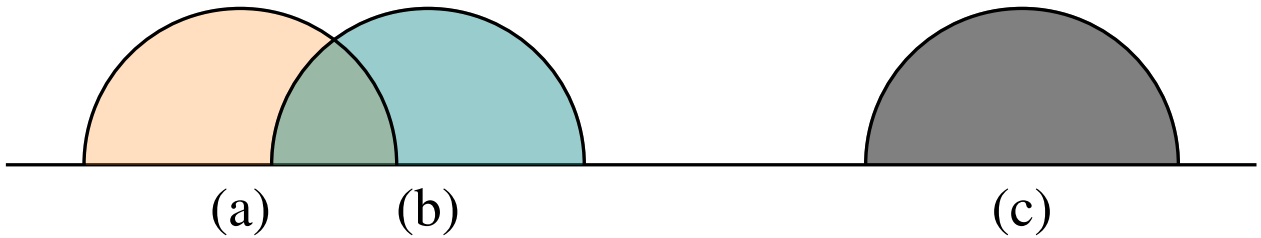
Non-Euclidean distances in cryo-EM analysis: Three Examples

- 3-D heterogeneity analysis
- 2-D classification
- Alignment of 3-D maps

Non-Euclidean distances for 3-D heterogeneity analysis

Zelesko, Moscovich, Kileel, S; ISBI 2020

- The goal is to learn the manifold of molecular conformations (represented as 3-D density maps).
- Euclidean distances between conformations are sensitive to deformations and movements (rigid and non-rigid).
- A large number of samples is therefore required for manifold learning techniques such as diffusion maps.
- The Earthmover's distance (EMD) changes more gradually and is meaningful for larger deformations and movements.



EMD vs. Euclidean distance for translational motion: Euclidean distance is only meaningful for measuring small displacements. The distance between half-disks (c) and (a) is the same as between (c) and (b). By contrast, for any translational motion, the EMD is its magnitude.

Non-Euclidean distances for 3-D heterogeneity

Zelesko, Moscovich, Kileel, S; ISBI 2020

- The Earthmover's distance (EMD) changes more gradually and is meaningful for larger deformations and movements.
- Fewer samples are therefore required by EMD for manifold learning.
- However, computation of EMD between all pairs of 3-D maps is costly.
- Remedy: replace EMD with another metric that can be efficiently computed and like EMD changes gradually with movements and deformations:

$$d_{\mathrm{EMD}}(x_i, x_j) := \min_{\pi \in \Pi(x_i, x_j)} \sum_{u \in [L]^3} \sum_{v \in [L]^3} \pi(u, v) \|u - v\|_2,$$

where $\Pi(x_i, x_j)$ is the set of joint probability measures on $[L]^3 \times [L]^3$ with marginals x_i and x_j, respectively.

$$d_{\mathrm{WEMD}}(x_i, x_j) := \sum_{\lambda} 2^{-5s/2} \, |\mathcal{W}x_i(\lambda) - \mathcal{W}x_j(\lambda)|,$$

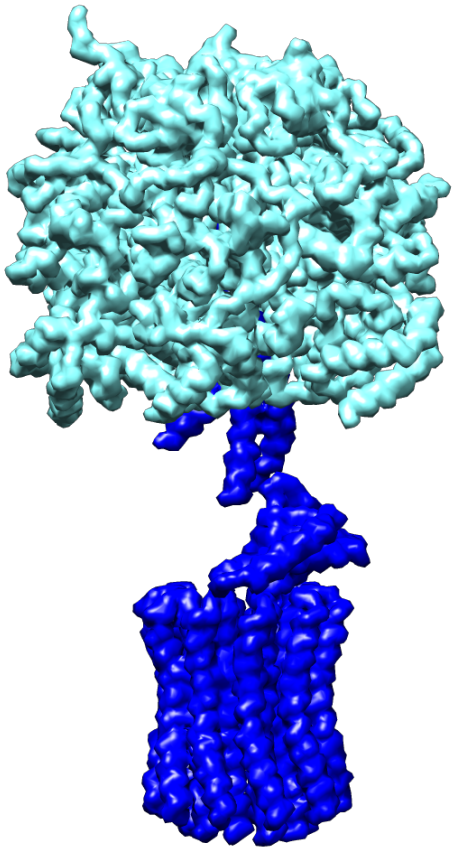
where, $\mathcal{W}x$ denotes a 3D wavelet transform of x.

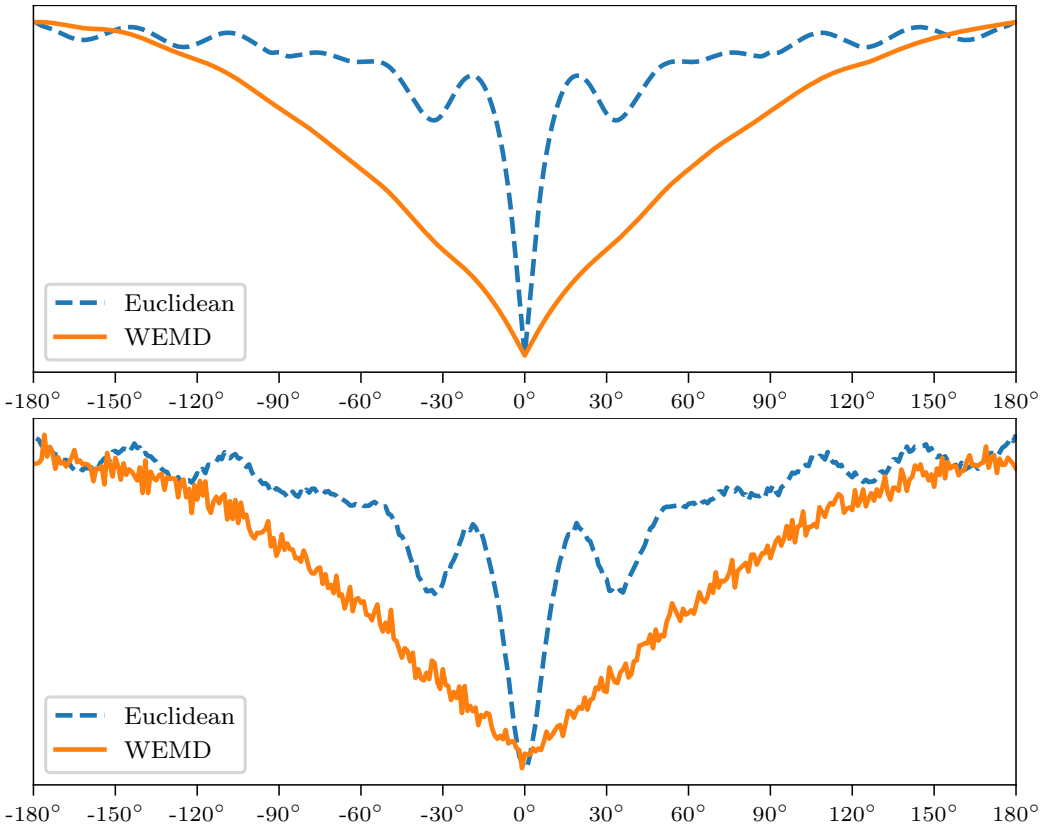
- The wavelet transform is computed in linear time $O(L^3)$.
- d_{WEMD} is simply a weighted ℓ_1 distance between wavelet coefficients.

Euclidean vs WEMD for 3-D shape analysis

Zelesko, Moscovich, Kileel, S; ISBI 2020

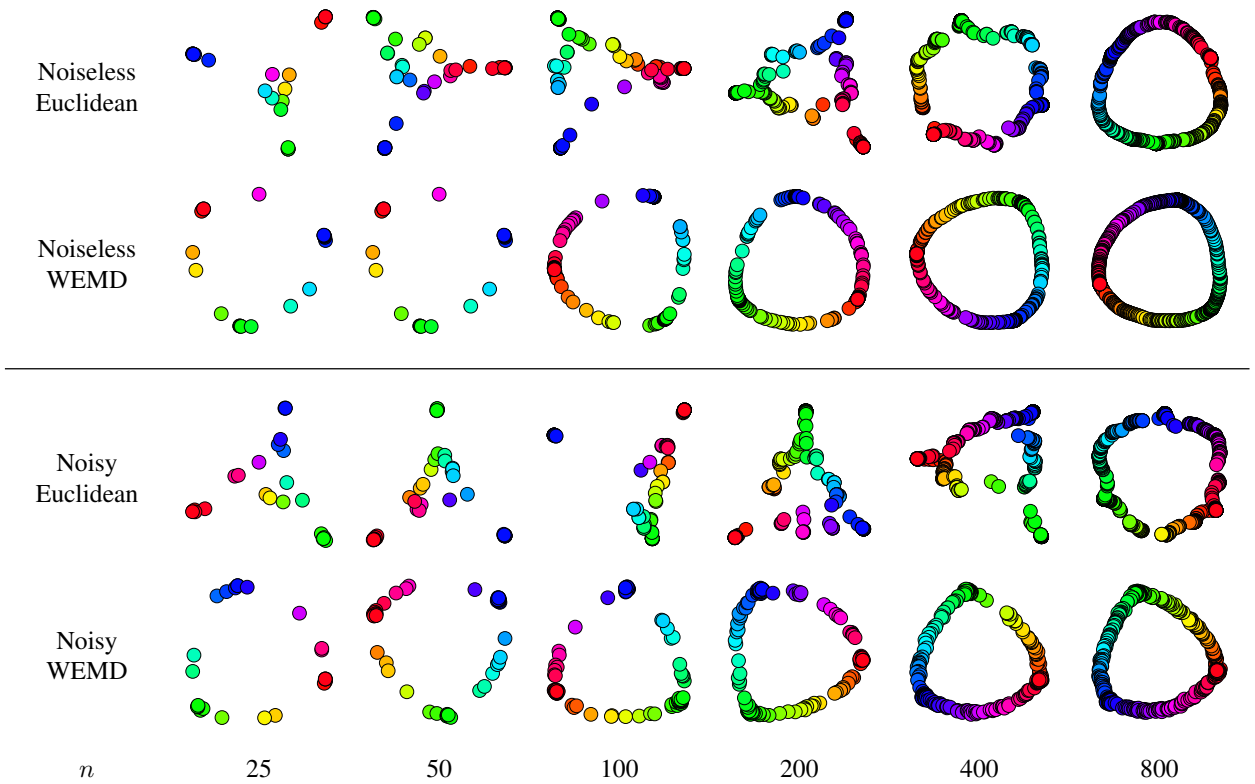
- Simulated data: rotating blue shaft of the ATP synthase





Euclidean vs WEMD for 3-D shape analysis

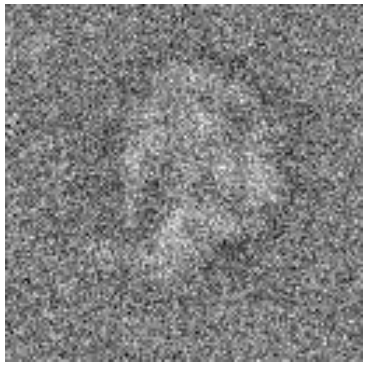
Zelesko, Moscovich, Kileel, S; ISBI 2020

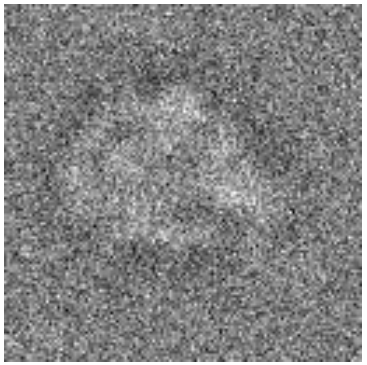


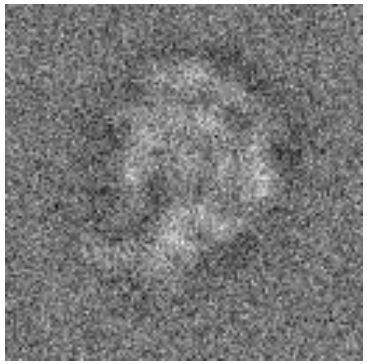
Euclidean vs. WEMD-based diffusion mappings on the clean and noisy ATP synthase datasets for sample sizes $n = 25, 50, 100, 200, 400, 800$. The Euclidean diffusion maps need more than 400 samples to capture the intrinsic geometry whereas WEMD manages to do so with merely $n = 25$ samples. The colors encode the (ground truth) angle.

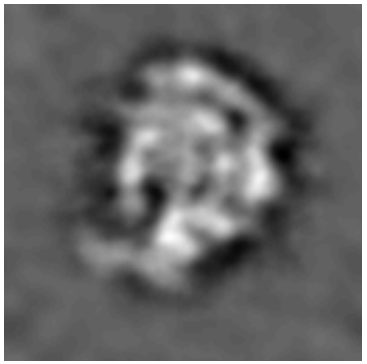
Optimal transport for 2D class averaging

Rao, Moscovich, S; NeurIPS 2020









Why 2-D Class Averaging?

- Image denoising: boost the SNR for 3-D ab-initio modeling.
- Quick assessment of sample preparation and data collection quality and first glance on how 2-D projections look like, before extensive usage of microscope time.
- Revealing possible non-trivial symmetry of the molecule.
- Particle picking procedures use 2-D class averaging as a step in their pipeline.

Basic principle of class averaging

- Find images believed to have similar viewing directions, perform in-plane rotational and translational alignment of neighboring images, and average to suppress noise.
- Main problem: How to find images with similar viewing directions?
- Challenges:
  1. Low SNR: difficult to detect images with similar viewing directions, signal is buried in noise.
  2. How to compare images? Which metric?
  3. Computational time: Pairwise comparison of all images together with in-plane alignment is computationally costly. It would be preferable to have an algorithm that scales linearly with the number of images.

Wasserstein K-Means for Clustering Tomographic Projections

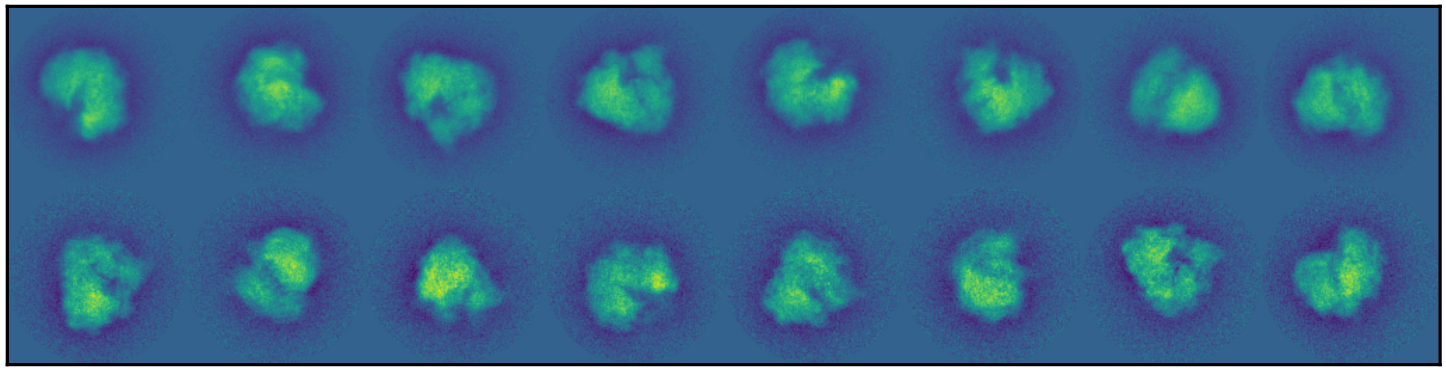
Rao, Moscovich, S; NeurIPS 2020

- Input: n noisy images $I_1, \ldots, I_n$.
- Initialize K centers $C_1, \ldots, C_K$, e.g., by randomly choosing K of the input images.
- For each image assign its closest center up to in-plane rotation in terms of d_{WEMD} (a total of Kn pairwise comparisons).
- Form new centers by aligning and averaging the assigned images.
- Repeat as long as loss function decreases.

Wasserstein K-Means for Clustering Tomographic Projections

Rao, Moscovich, S; NeurIPS 2020

- Synthetic dataset of $n = 10,000$ tomographic projections (no CTF, no shifts) of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine (EMD-2660)
- $K = 150$ clusters
- A visual comparison of the centroids based on rotation-aligned Euclidean (top) vs. WEMD (bottom) (SNR=1/16)

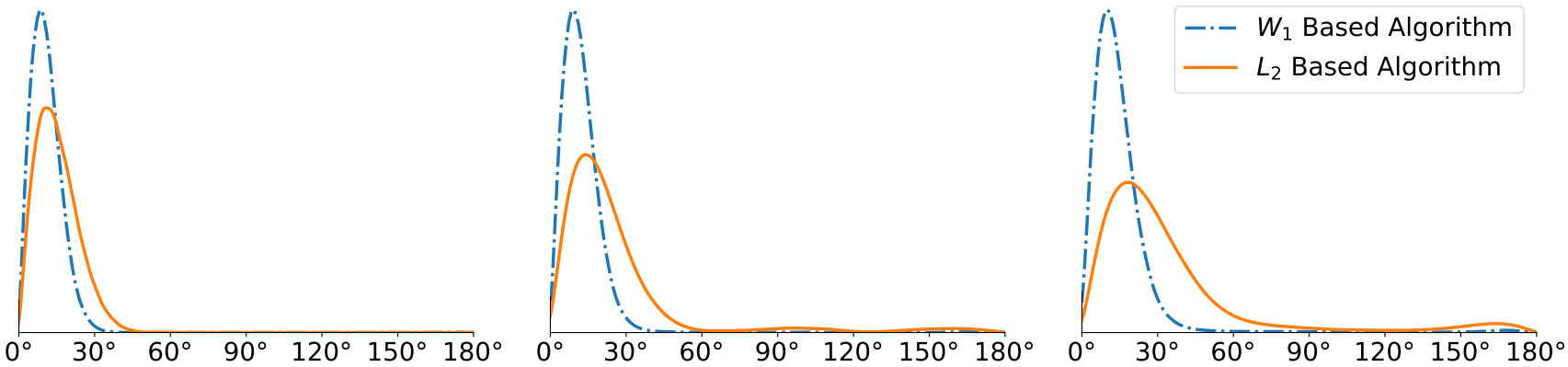


- The WEMD based clusters seem to preserve more details than those using Euclidean distances.

Wasserstein K-Means for Clustering Tomographic Projections

Rao, Moscovich, S; NeurIPS 2020

- Within-cluster angular differences (left to right: SNR $= 1/8, 1/12, 1/16$)



- The WEMD clusters have better angular coherency

EMD between tomographic projections

Rao, Moscovich, S; NeurIPS 2020

- Let $\rho : \mathbb{R}^3 \to \mathbb{R}_{\geq 0}$ be a probability distribution supported on the 3D unit ball and let I_1 and I_2 be its tomographic projections along the vectors u and v respectively. Denote by $\measuredangle(u, v) \in [0, \pi]$ the angle between the vectors, then

$$W_p^R(I_1, I_2)^p \leq [2 \sin(\measuredangle(u, v)/2)]^p \leq \measuredangle(u, v)^p$$

  where W_p^R is the rotationally-invariant Wasserstein metric.
- A similar upper-bound for the rotationally-invariant L_2 distance cannot hold for all densities ρ. To see why, consider an off-center point mass. Any two projections at slightly different viewing angles will have a large L_2^R distance no matter how small their angular difference is.
- However, for densities with bounded gradients it is possible to produce upper bounds. Let $B = \sup_{\mathbf{x}} |\nabla \rho(\mathbf{x})|$ be an upper bound on the absolute gradient of the density. Then,

$$L_2^R(I_1, I_2) \leq 2\sqrt{\pi} B \measuredangle(u, v).$$

- This bound suggests that L_2^R is a reasonable metric to use for very smooth signals. For non-smooth signals, or signals with very large B, this means that there is no guarantee that the L_2^R distance will assign a small distance between projections with a small viewing angle.

Non-Euclidean distances for 3-D volume alignment

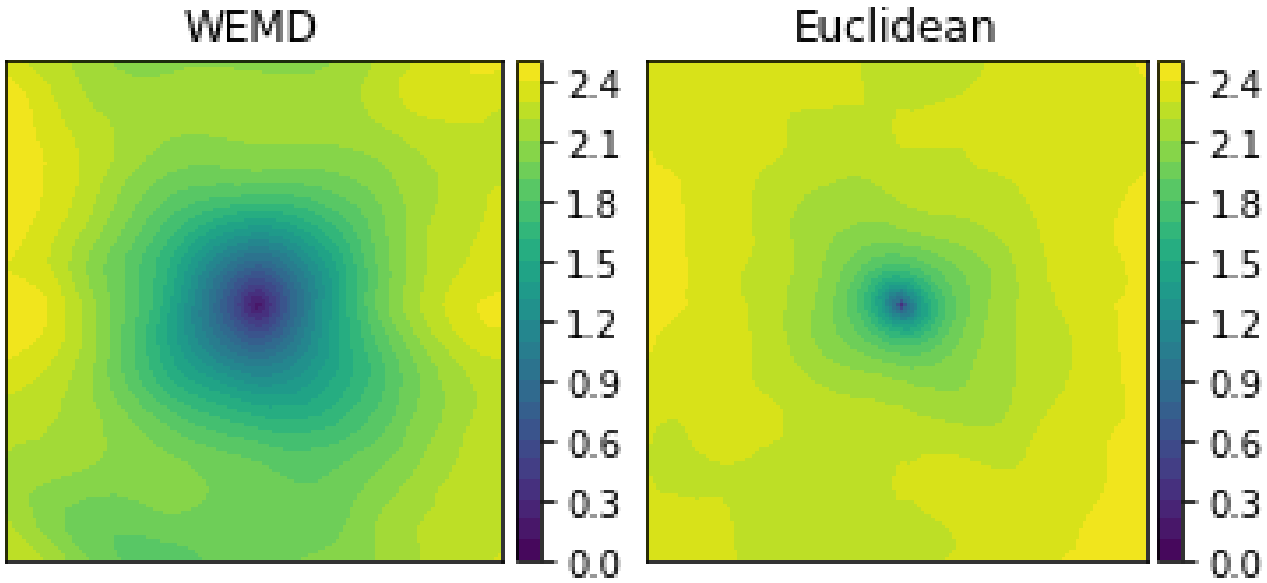
S, Yang; arXiv 2023

- The goal is to recover the relative rotation that best aligns two given volumes ϕ_1, ϕ_2 (represented as 3-D density maps):

$$R^* = \underset{R \in SO(3)}{\arg\min}\ d\big(\phi_1(\cdot), \phi_2(R\cdot)\big) =: \underset{R \in SO(3)}{\arg\min}\ F_d(R),$$

  where d is a distance function.

- Setting d as WEMD creates a better landscape for F_d:



Local landscapes of $F_d(R)$ when d is WEMD and Euclidean (L^2) for $R = R_z(\gamma) \cdot R_y(\beta)$, $\gamma, \beta \in [-\pi/2, \pi/2]$.

Non-Euclidean distances for 3-D volume alignment

S, Yang; arXiv 2023

- Employ Bayesian optimization for solving

$$R^* = \underset{R\in SO(3)}{\arg\min}\ d_{\mathrm{WEMD}}\left(\phi_1(\cdot), \phi_2(R\cdot)\right) =: \underset{R\in SO(3)}{\arg\min}\ F_{\mathrm{WEMD}}(R).$$

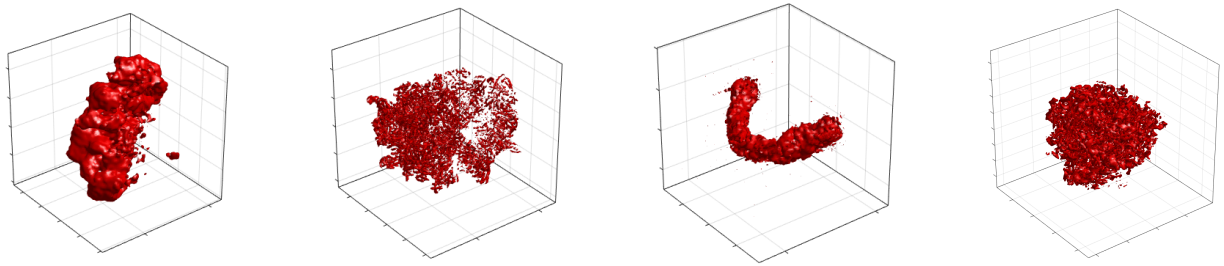
- Bayesian optimization is a global optimization method, hence less prone to get stuck at local optima than gradient based methods, improving accuracy.

- Bayesian optimization explores only "high probability regions", therefore requiring fewer evaluations of F_{WEMD} than exhaustive search based methods, improving efficiency.

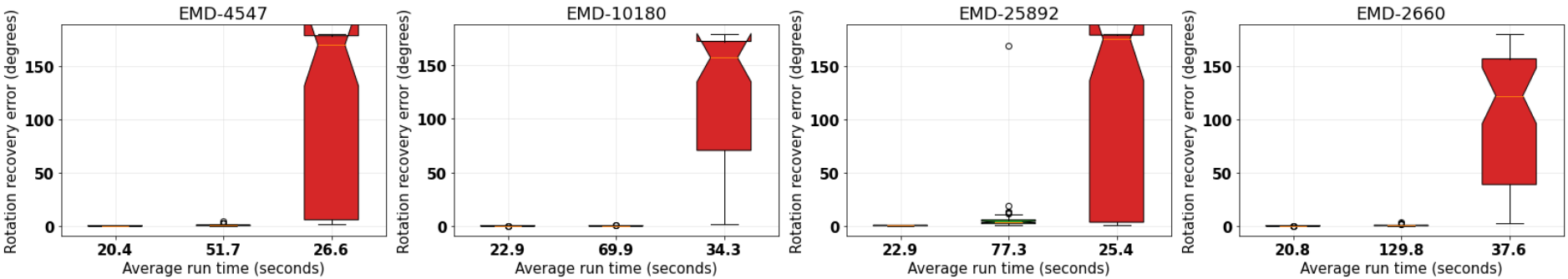
Non-Euclidean distances for 3-D volume alignment

Comparison with existing methods for the following test volumes: *S, Yang*; arXiv 2023

EMD-4547 EMD-10180 EMD-25892 EMD-2660



$L = 280$ $L = 320$ $L = 320$ $L = 360$



The three boxplots in each subfigure correspond to (from left to right) BOTalign (our method), EMalign (Harpaz and Shkolnisky, 2023), and AlignOT (Riahi et al, 2022). The vertical axis represents rotation recovery error in degrees. The tick labels record the average run time in seconds.

ASPIRE

Algorithms for Single Particle Reconstruction

ASPIRE Python Pip CI passing | codecov 89% | DOI 10.5281/zenodo.5657281 | downloads/month 525

ASPIRE - Algorithms for Single Particle Reconstruction - v0.12.0

The ASPIRE-Python project supersedes Matlab ASPIRE.

ASPIRE is an open-source software package for processing single-particle cryo-EM data to determine three-dimensional structures of biological macromolecules. The package includes advanced algorithms based on rigorous mathematics and recent developments in statistics and machine learning. It provides unique and improved solutions to important computational challenges of the cryo-EM processing pipeline, including 3-D *ab-initio* modeling, 2-D class averaging, automatic particle picking, and 3-D heterogeneity analysis.

For more information about the project, algorithms, and related publications please refer to the ASPIRE Project website.

For full documentation and tutorials see the docs.

Please cite using the following DOI. This DOI represents all versions, and will always resolve to the latest one.

```
ComputationalCryoEM/ASPIRE-Python: v0.12.0 https://doi.org/10.5281/zenodo.5657281
```

https://github.com/ComputationalCryoEM/ASPIRE-Python

http://spr.math.princeton.edu/

Summary

- Three examples (heterogeneity analysis, 2-D classification, 3-D alignment) from cryo-EM analysis where non-Euclidean metrics (Wasserstein and related distances) outperform Euclidean distances.

.

- Noise statistics suggests optimality of Euclidean distances, but the underlying signals (projection images, density maps) are better compared using non-Euclidean distances.

- More applications and other metrics (work in progress)

Thank You!


NSF


AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE


NIH
NIGMS


GORDON AND BETTY
MOORE
FOUNDATION


SIMONS
FOUNDATION