CryoEM with Spider Kernel Graphs

Amit Singer

Department of Mathematics
Program in Applied Mathematics
Yale University

Joint work with...

- *Ronald R. Coifman* (Yale University, Applied Mathematics)
- *Yoel Shkolnisky* (Yale University, Applied Mathematics)
- *Fred Sigworth* (Yale School of Medicine, Cellular & Molecular Physiology)

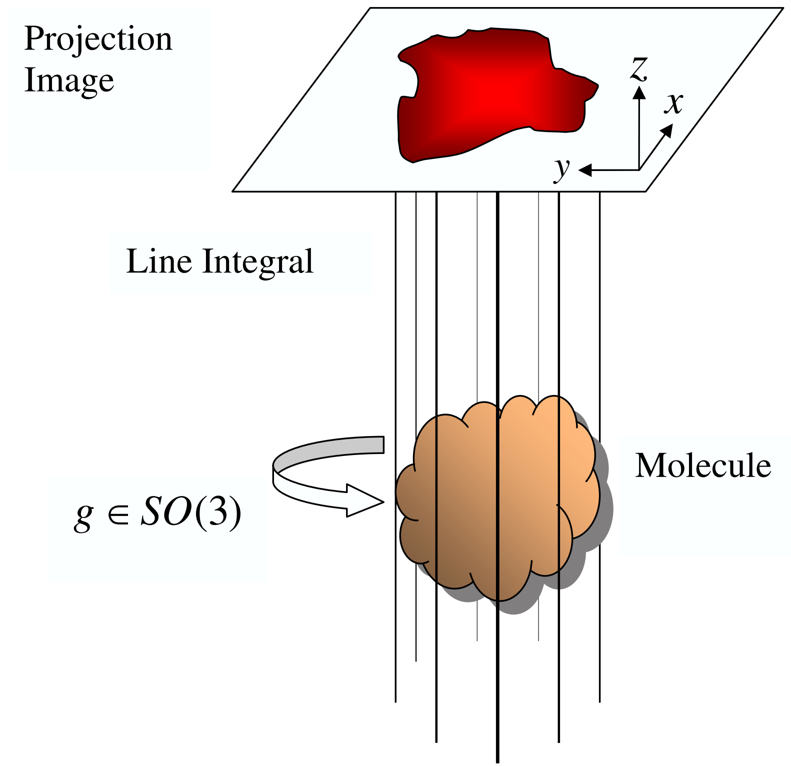
Structuring of Protein Channels

- Rod MacKinnon was co-awarded the 2003 Nobel Prize in Chemistry for structuring the Potassium channel in 1998.
- Proteins were crystallized (all share the same space orientation).
- Classical X-ray Computational Tomography (CT).
- A few other proteins had been structured since.
- However, most channels cannot be crystallized.
- Can a protein be structured without being crystallized?

Cryo Electron Microscopy

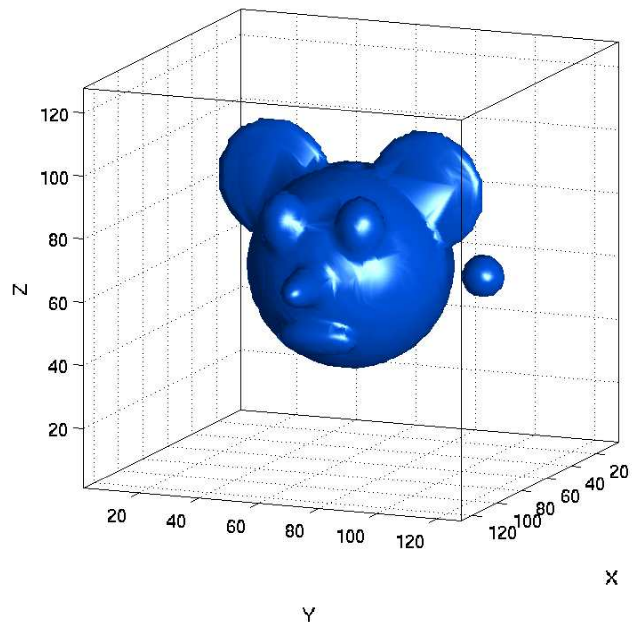
- CryoEM: Electron Microscope imaging of proteins ”freezed” in liquid nitrogen.
- Thousands of images: every image corresponds to a different protein frozen in a different space orientation.
- Orientations are random and unknown.
- Highly intense electron beam destroys protein while being imaged: a single protein can be imaged only once.
- Images are very noisy (low SNR)
- Images are 100×100 pixels.

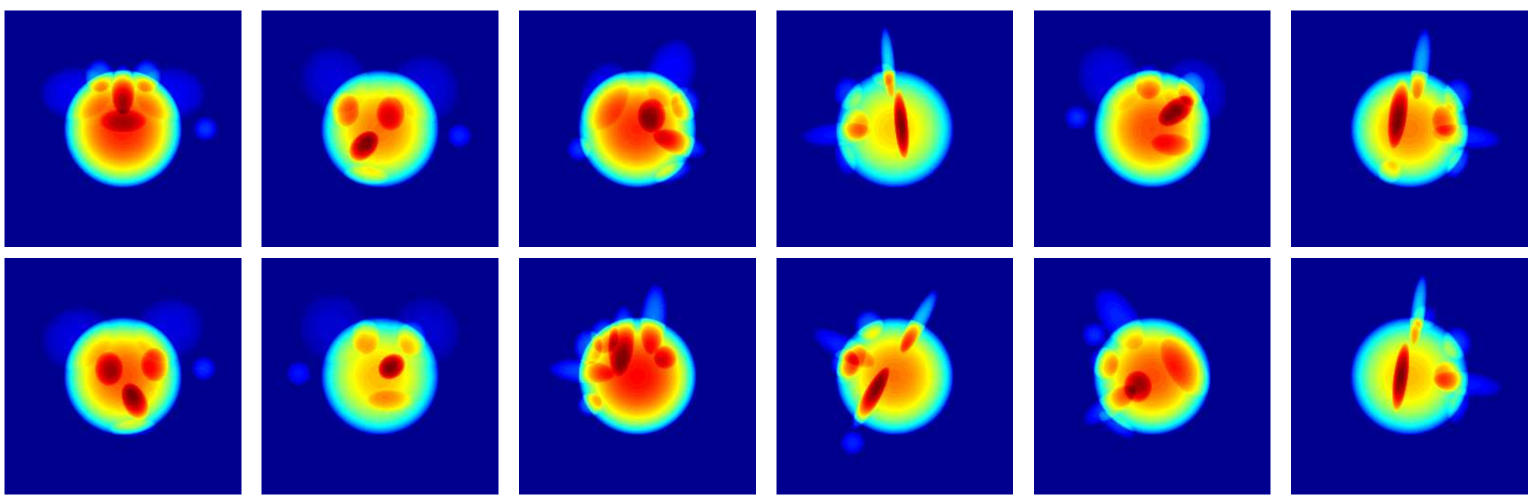
Projection Images



- The projection image is $P_g(x,y) = \int_{-\infty}^{\infty} \phi_g(x,y,z)\,dz$.
- $\phi(r)$ is the electric potential of the molecule, $\phi_g(r) = \phi(g^{-1}r)$.

Projection Images: Toy Example





The Fourier projection-slice theorem

- $\theta \in S^2$ beaming direction, $\theta^\perp$ orthogonal plane.
- The 2D FT of the projection image is the double integral

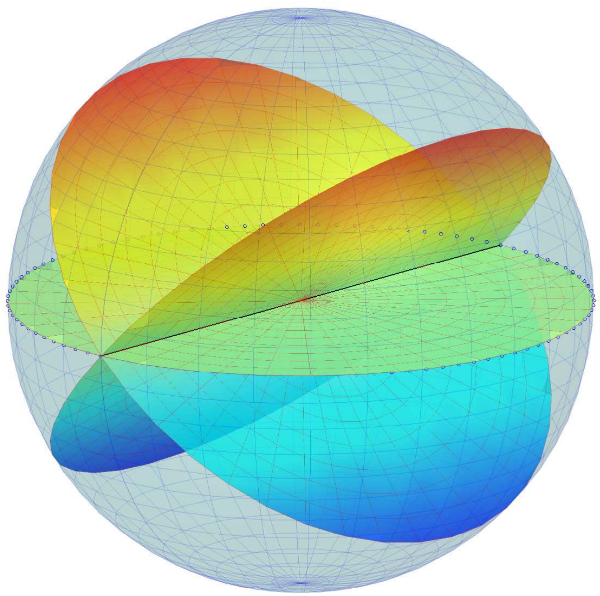
$$\hat{P}_\theta(\xi) = \int_{\theta^\perp} e^{-ir\cdot\xi} P_\theta(r)\, dr.$$

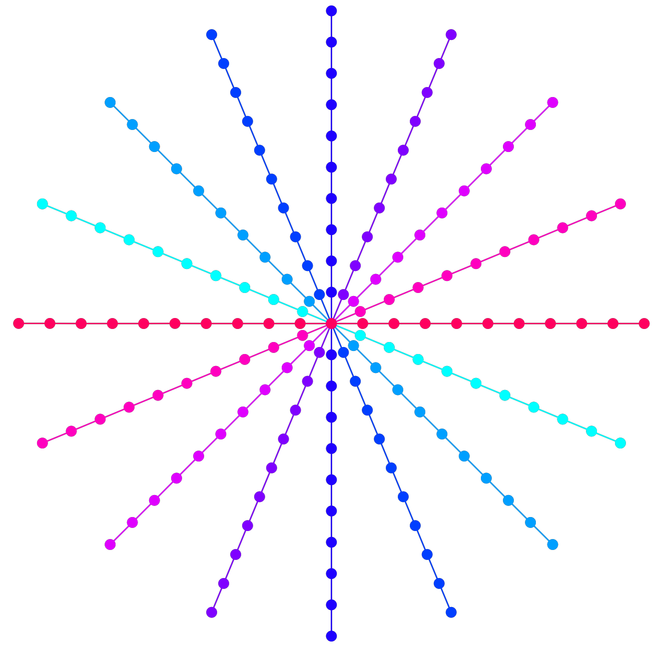
- The 3D FT of the molecule is the triple integral

$$\hat{\phi}(\xi) = \int_{\mathbb{R}^3} e^{-ir\cdot\xi} \phi(r)\, dr.$$

- Slice Theorem: $\hat{P}_\theta(\eta) = \hat{\phi}(\eta), \quad \eta \in \theta^\perp.$

The Geometry of the slice theorem





- Every image is a great circle over S^2.
- Any pair of images have a common line, or
- Any pair of great circles meet at two antipodal points.

Three Dimensional Puzzle

- The radial lines are the puzzle pieces.
- Every image is a circular chain of pieces.
- Common line: meeting point

The Spider Kernel: It's the Network

- K projection images
- L radial lines
- We build a weighted directed graph $G = (V, E, W)$.
- The vertices are the radial lines ($|V| = KL$)

$$V = \{(k,l) : 1 \le k \le K,\ 0 \le l \le L-1\}.$$

- The heart of the algorithm is the definition of arrows and weights

$$E = \{((k_1,l_1),(k_2,l_2)) : (k_1,l_1) \text{ points to } (k_2,l_2)\}.$$

- W is a sparse weight matrix of size $KL \times KL$

$$((k_1,l_1),(k_2,l_2)) \notin E \Longrightarrow W_{(k_1,l_1),(k_2,l_2)} = 0.$$

Weights

- All weights are taken from a single (sparse) symmetric circular weight vector of length L

$$w = (w_0, w_1, \ldots, w_{L-1}),$$

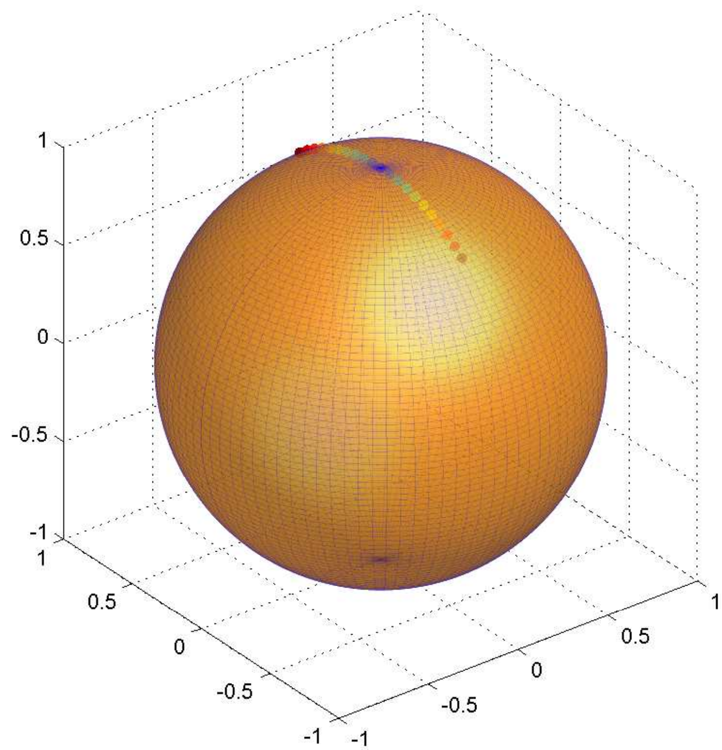
$$w_l = w_{-l}.$$

- Example:

$$w = (1, 1, \ldots, 1) = \mathbf{1}$$

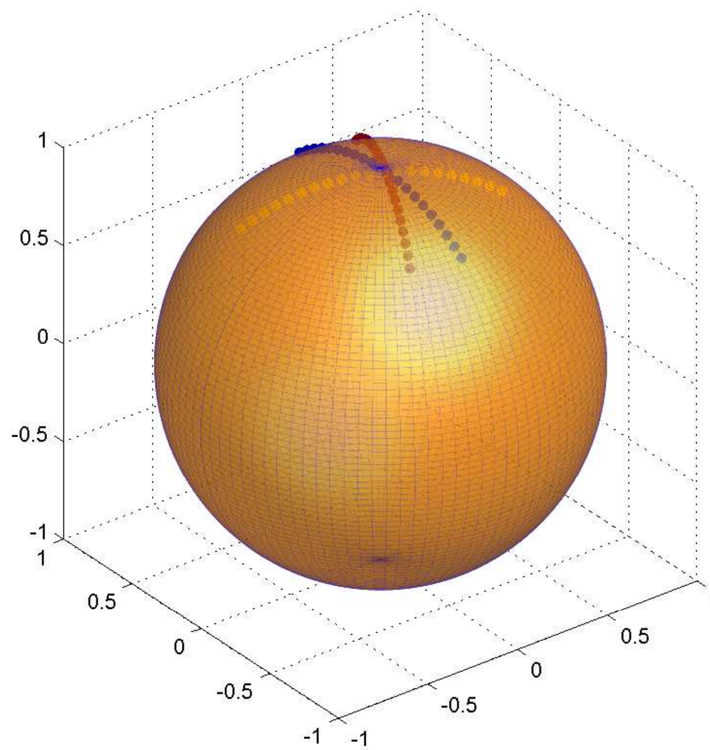
renders W the adjacency matrix of the graph.

Spider first pair of legs



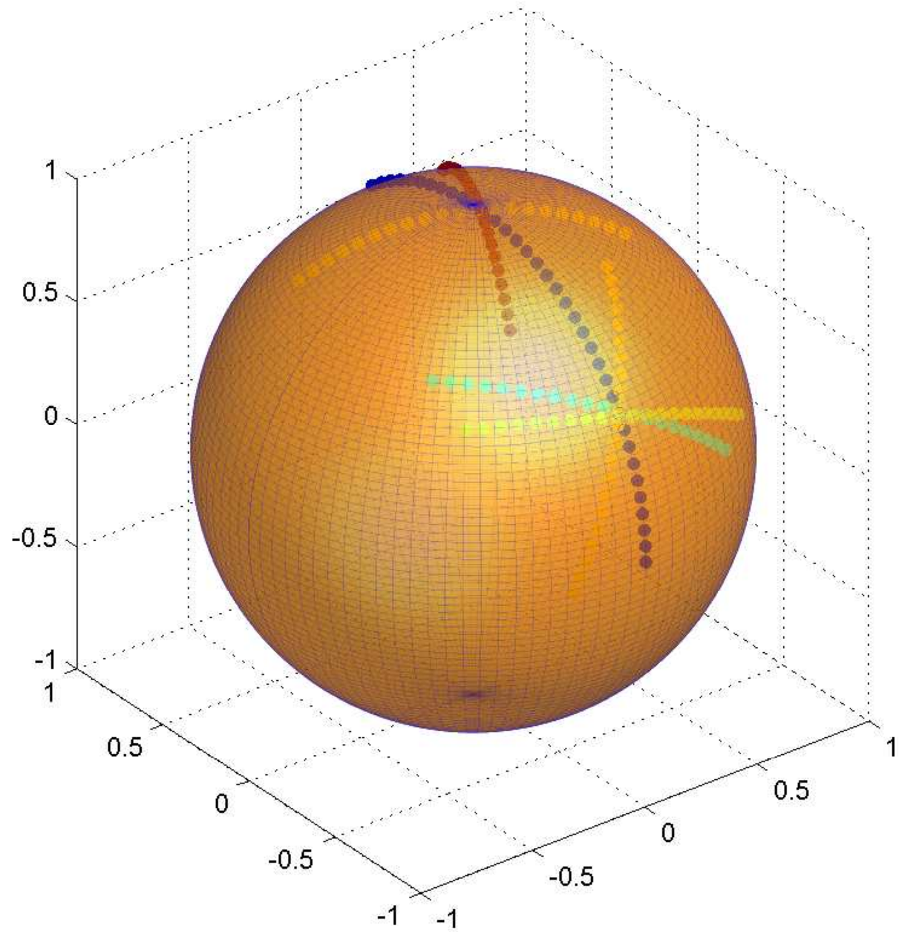
- Blue vertex (k_1, l_1) is the head of the spider
- Linked vertices: $(k_1, l_1 + l)$, $-d \leq l \leq d$ (same image radial lines)
- Weights: $W_{(k_1,l_1),(k_1,l_1+l)} = w_l$.

Spider: remaining legs



- (k_1, l_1) and (k_2, l_2) are common radial lines of different images.
- Links: $((k_1, l_1), (k_2, l_2 + l)) \in E \quad \text{for } -d \leq l \leq d.$
- Weights: $W_{(k_1,l_1),(k_2,l_2+l)} = w_l.$

Communicating Spiders



Sparse weight matrix

- W is sparse: its number of nonzero entries is only

$$|E| = (2d+1)\left[KL + 2K(K-1)\right].$$

- There are KL spiders with first pair legs of size $2d+1$.
- There are $2\binom{K}{2} = K(K-1)$ intersection points (with antipodals). Every meeting point belongs to two different circles so it appears in two different spiders. In every spider it contributes two legs of total length $2d+1$.
- Algorithm is linear in number of lines and intersection points for $d = O(1)$ (small spiders).

Averaging operator

- Row sums of W depend on the number of pair of legs $M_{k,l}$

$$\sum_{(k',l')\in V} W_{(k,l),(k',l')} = M_{k,l} \sum_{l=-d}^{d} w_l.$$

- The outdegree $d_{k,l}$ of the (k,l)'th vertex is

$$d_{k,l} = |\{(k',l') : ((k,l),(k',l')) \in E\}| = M_{k,l}(2d+1).$$

- We normalize the weight matrix W to have constant row sums by dividing each row by its outdegree:
- $A = D^{-1}W$, with D diagonal $D_{(k,l),(k,l)} = d_{k,l}$.
- The row sums of A are identical and equal

$$\sum_{(k',l')\in V} A_{(k,l),(k',l')} = \frac{1}{2d+1} \sum_{l=-d}^{d} w_l, \quad \forall (k,l) \in V.$$

Averaging operator

- A is a spider weighted averaging operator

$$(Af)(k_1, l_1) = \sum_{((k_1,l_1),(k_2,l_2)) \in E} A_{(k_1,l_1),(k_2,l_2)} f(k_2, l_2).$$

- Example: $w = (1, 1, \ldots, 1) = \mathbf{1}$
  A is row stochastic, weighted average = non-weighted average

$$(Af)(k_1, l_1) = \frac{1}{d_{k,l}} \sum_{((k_1,l_1),(k_2,l_2)) \in E} f(k_2, l_2).$$

- We call A the spider kernel.

The spectrum of the spider kernel

- A and W are not symmetric, their spectrum may be complex.
- A has constant row sums: $\psi_0 = 1$ is a trivial eigenvector

$$(A\psi_0)(k,l) = \left(\frac{1}{2d+1} \sum_{l=-d}^{d} w_l \right) \psi_0(k,l), \quad \forall (k,l) \in V.$$

- Example: $w = (1, 1, \ldots, 1) = \mathbf{1}$
  A is row stochastic, $\lambda_0 = 1$, remaining spectrum $|\lambda| < 1$.
- Much more can be said on the spectrum!

Spherical Harmonics

- The spherical harmonics Y_l^m are the eigenfunctions of the Laplacian on the sphere

$$\Delta_{S^2} Y_l^m = -l(l+1) Y_l^m, \quad l = 0, 1, 2, \ldots, \quad m = -l, \ldots, l.$$

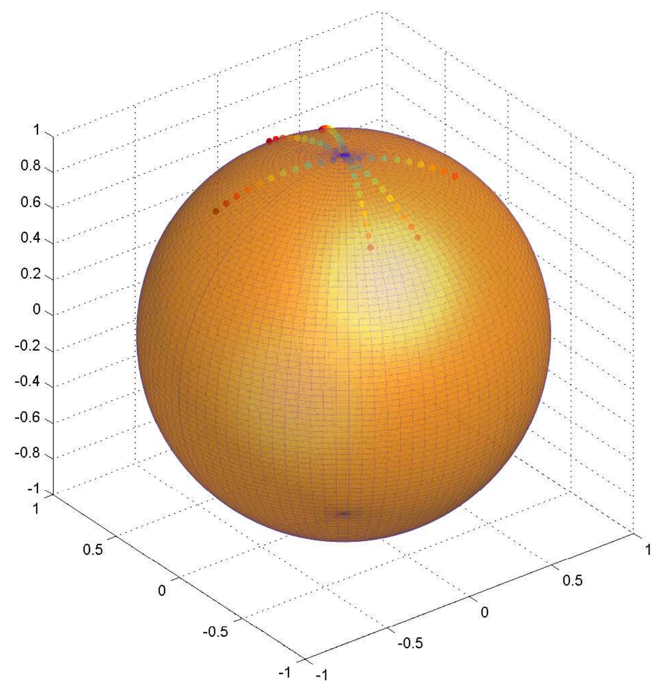
- Funk-Hecke: The spherical harmonics are the eigenfunctions of any integral operator that commutes with rotations:

$$\begin{aligned} (\mathcal{K} f)(\beta) &= \int_{S^2} k(\langle \beta, \beta' \rangle) f(\beta')\, dS_{\beta'}, \\ \mathcal{K} Y_l^m &= \lambda_l Y_l^m. \end{aligned}$$

- The spider kernel commutes with rotations only on average, so spherical harmonics are not guaranteed.
- The three linear spherical harmonics are exact eigenfunctions of the spider kernel.

Linear Eigenfunctions

- Linear functions $f(x, y, z) = a_1x + a_2y + a_3z$ are eigenfunctions
- The center of mass of every spider is beneath the spider's head: any pair of opposite legs balance each other – w is symmetric.



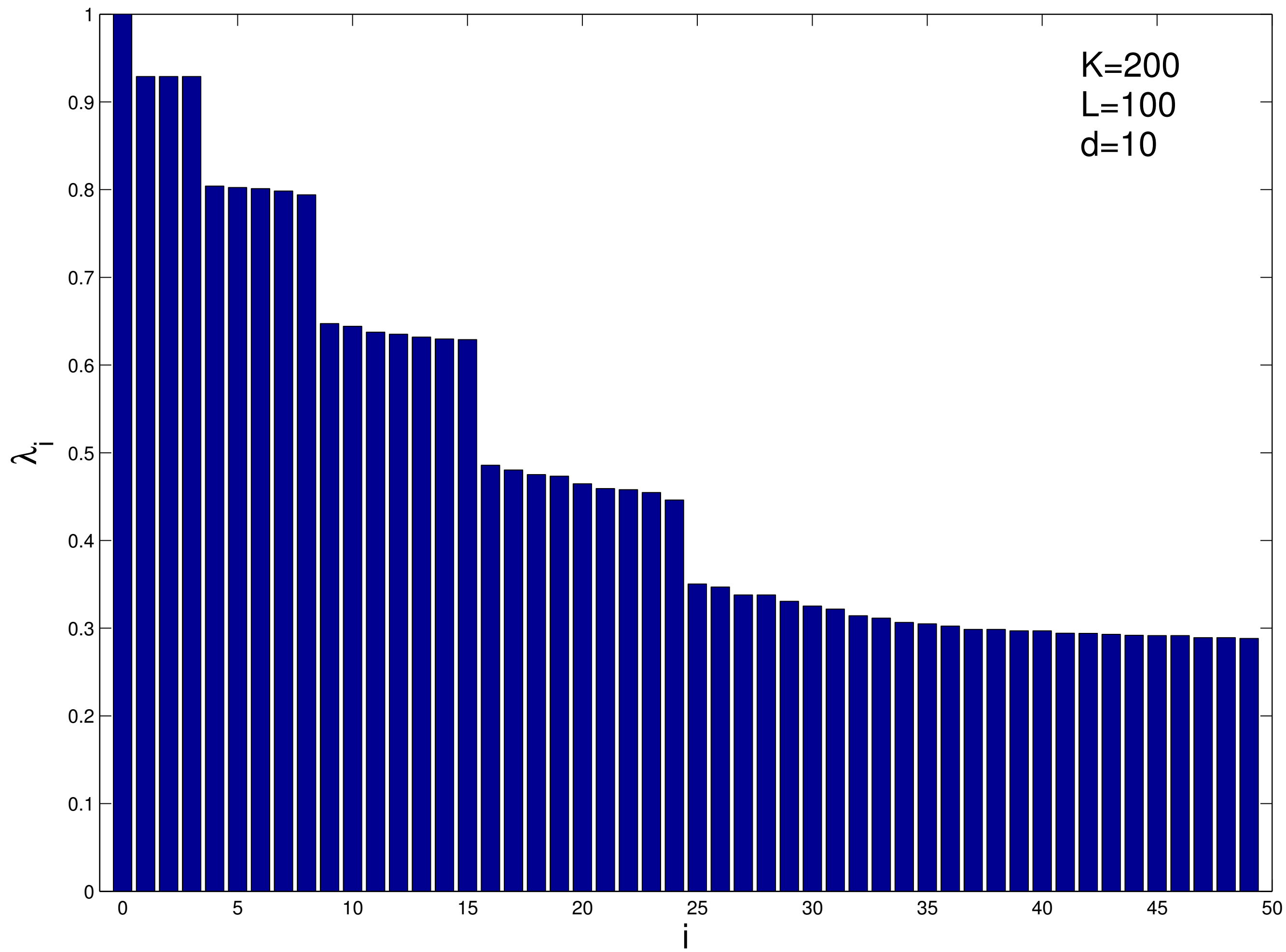
Spider kernel embedding and algorithm

- Find the common lines for all pairs of images.
- Construct the spider kernel matrix A.
- Compute eigenvectors $A\psi_i = \lambda_i \psi_i$.
- Embed the data into the three linear eigenvectors (ψ_1, ψ_2, ψ_3)

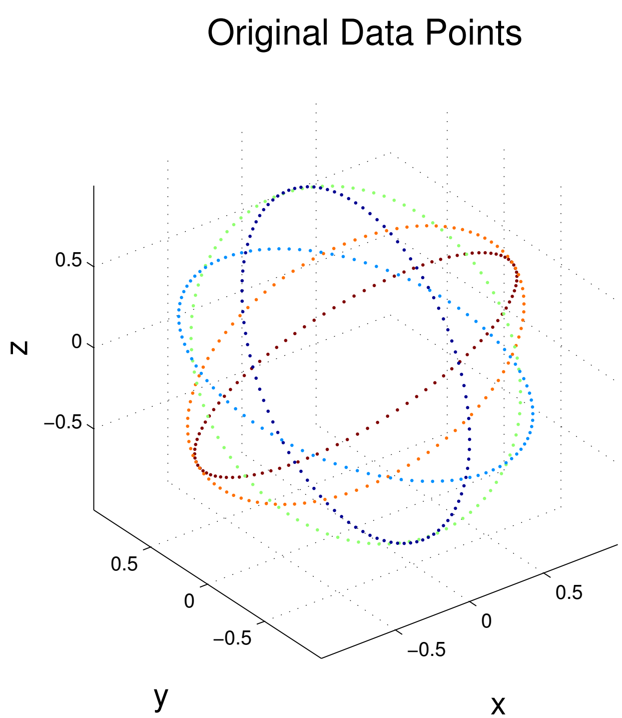
$$(k,l) \mapsto (\psi_1(k,l), \psi_2(k,l), \psi_3(k,l)).$$

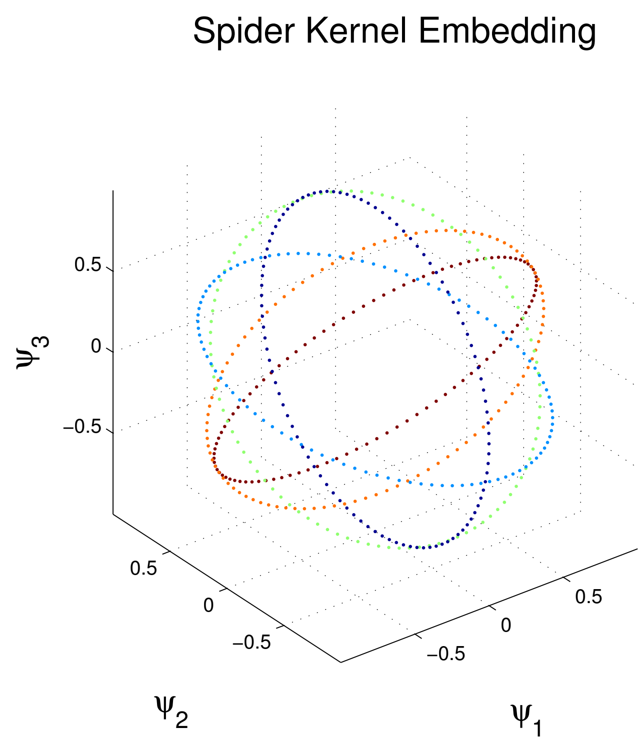
- Reveals molecule orientations up to rotation and reflection.
- Final cosmetics:
  PCA same image radial lines and equally space them.

Numerical Spectrum

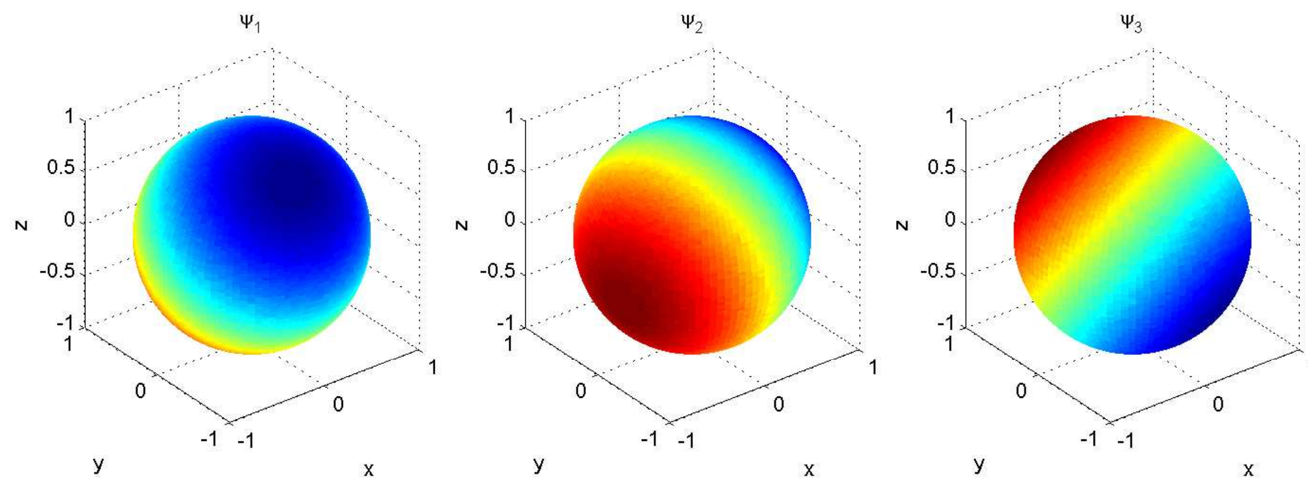


Data vs. Embedding (only 5 circles are shown)

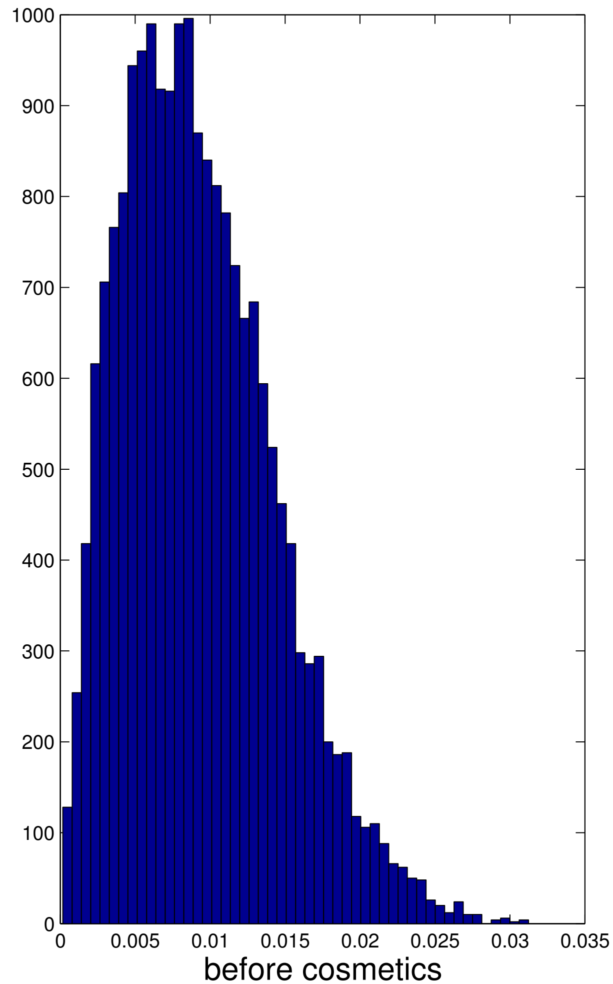


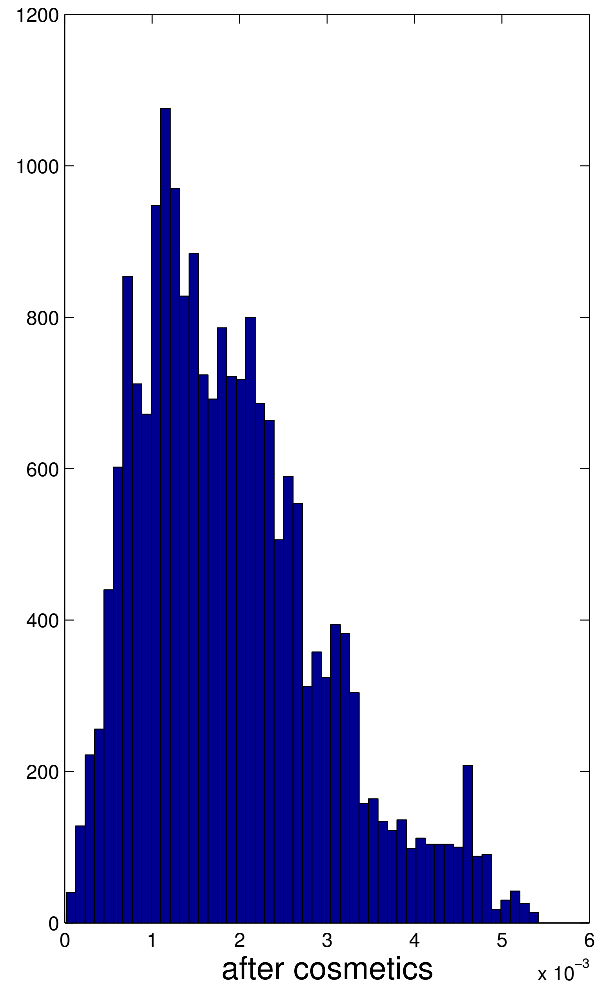


Linear Eigenfunctions



Angle difference histogram

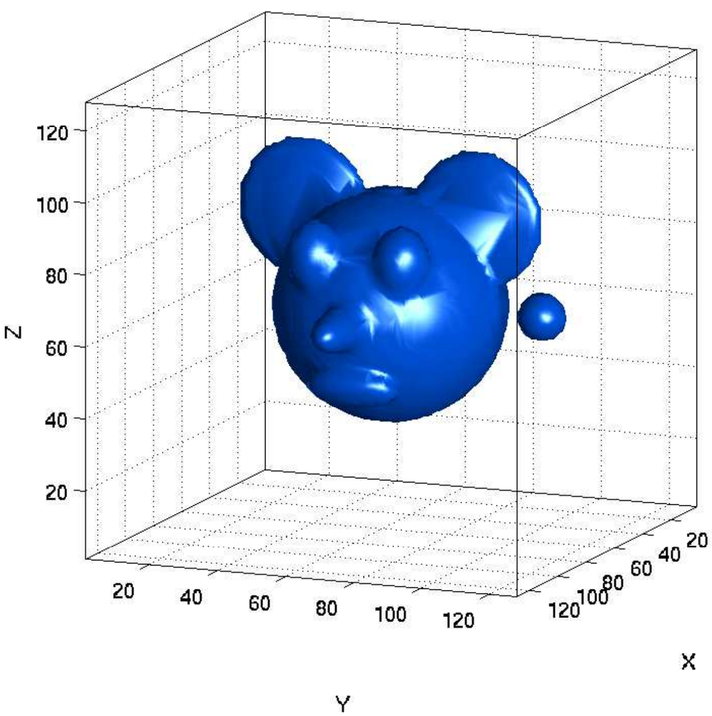




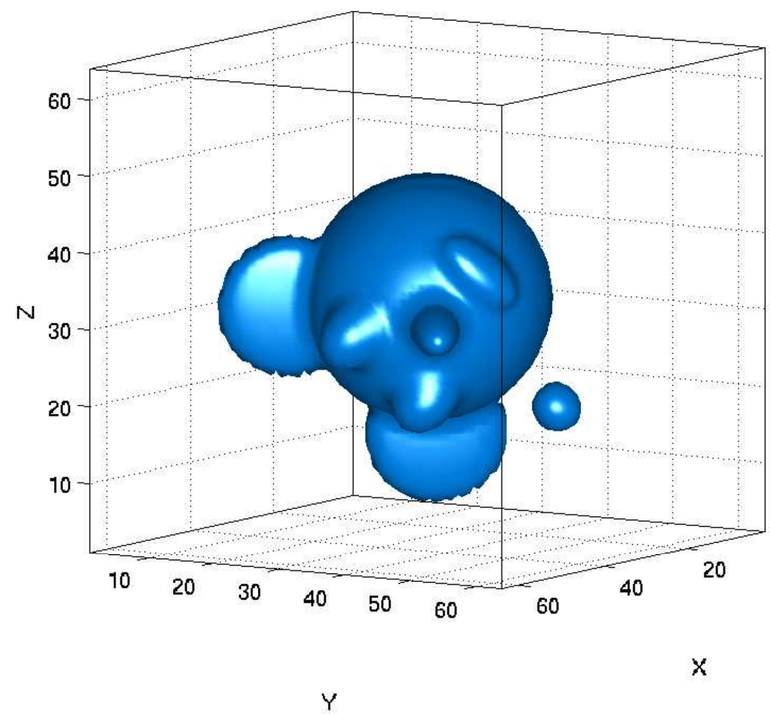
Spider kernel advantages

- Global: all radial lines are linked together.
- Fast: linear in data size KL and intersection points $\binom{K}{2}$.
- Averaging: all geometric information is averaged.
- Robust: errors due to false detections of common lines are smoothed out (can be viewed as matrix perturbation).
- Embedding error decreases like $1/\sqrt{K}$.
- Optional: omit uncertain common lines (fewer legs).

Toy Example



(a) original



(b) reconstructed