UNIVERSITY of WISCONSIN

"Crystallography without Crystals"

Determining the Structure of Individual Biological Molecules & Nanoparticles

Abbas Ourmazd
ourmazd@uwm.edu

Acknowledgments

Collaborators: Russell Fung

Dilano Saldin

Valentin Shneerson

Discussions: Len Feldman

Paul Fuoss

Eric Isaacs

Qun Shen

John Spence

Dmitri Starodub

Brian Stephenson

Why Single Molecules?

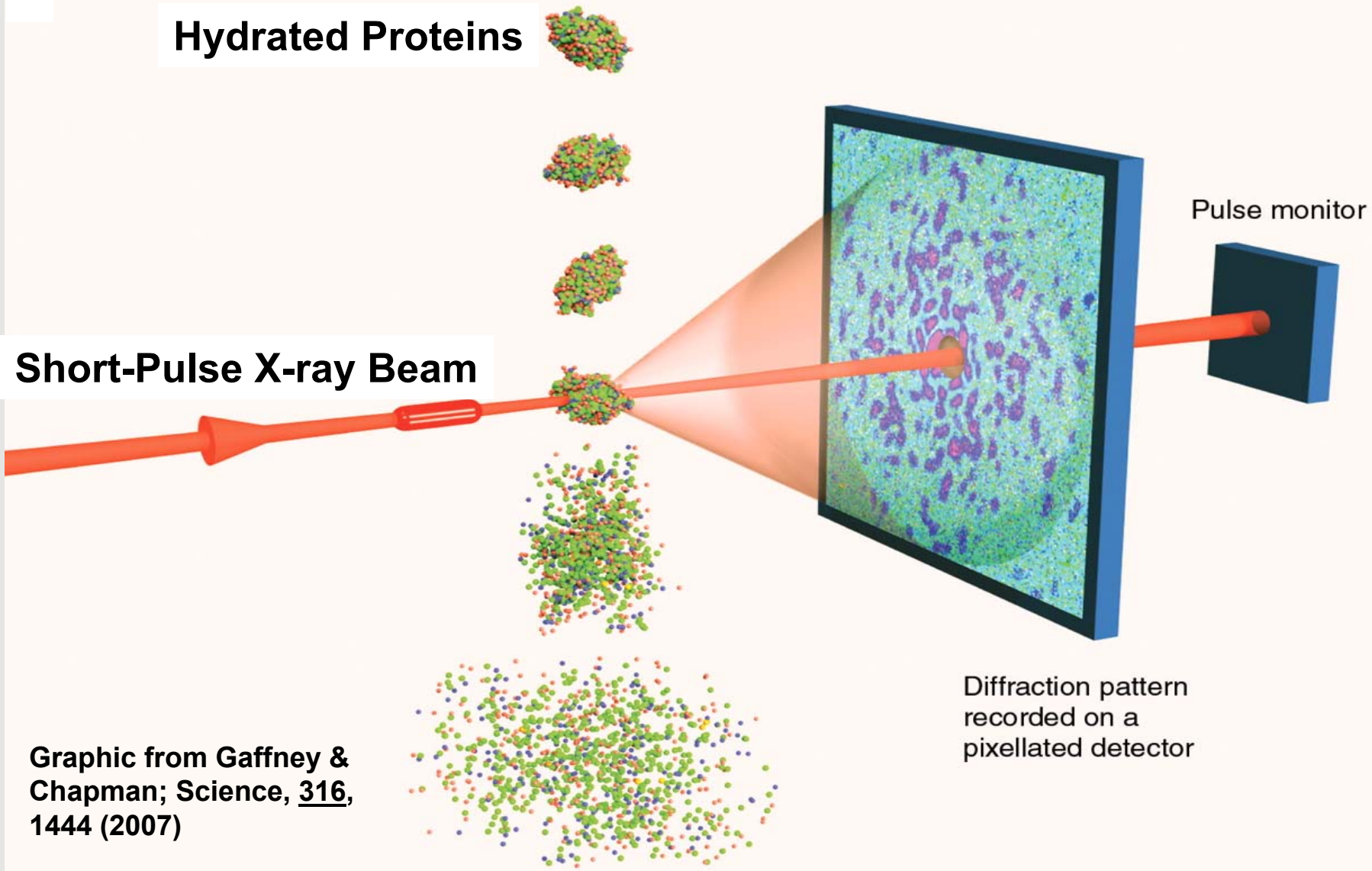
The Scorecard	Number	Percent
Proteins sequenced	>750,000	
Protein structures determined	44,700	<6%
Membrane protein structures	460	<0.1%

Source: Protein Data Bank, July '07

- 70% of today's drugs aimed at membrane proteins
  - Notoriously difficult to crystallize
- Purification and crystallization major bottlenecks
  - Crystals complicate "inversion problem"

Proposed Experiment

[E.g., Neutze et al, Nature 406, 752 (2000)]



Graphic from Gaffney & Chapman; Science, 316, 1444 (2007)

Key Challenges

- **Synchronized beam of hydrated proteins**
  - **In native state, not too much water**
- **Reconstitute 3-D intensity distribution**
  - **Each 2-D "snapshot" from unknown random orientation**
- **Very few photons scattered "per shot"**
  - **Next-generation synchrotrons (XFELs): ~ 10^3 photons/shot**
  - **Current-generation synchrotrons: ~ 10^{-2} photons/shot**
  - **XFEL shot blows molecule apart**
- **Collect data within 20fs after pulse arrival**
  - **"After the molecule is blown up, before it has flown apart"**

Executive Summary

- **Single-molecule scattering "Grand Challenge"**
  - **Opens research into all macromolecules & nanoparticles**
  - **Including non-crystallizing proteins and fuels**
- **Single 500 kDa protein molecule in XFEL scatters 10^7 photons/sec**
  - **More than enough photons to reconstruct structure**
- **But only 4.10^{-2} photons/pixel per shot**
- **Each diffraction pattern from unknown orientation**
  - **Snapshot of rotating molecule**
- **Dose to orient snapshot at least 100x more than XFEL can deliver**
  - **Using proposed orientation techniques**

Executive Summary: Results

- Succeeded in orienting dp's down to ~10^{-2} ph/pixel
  - First results; many improvements needed
  - Threshold for XFEL reached
- Using only ≤ 10^5 photons
  - XFEL delivers 10^9 photons in minutes
- Single-molecule crystallography now possible in principle
  - "Scatter & destroy" mode; each pulse blows up molecule
- Can per-shot dose be reduced significantly?
  - Would make XFEL experiments much easier
  - Single-molecule crystallography on 3rd Generation sources??

Single-Molecule X-ray Scattering: Orders of Magnitude

- **Assumptions:**
  a. **Macromolecule with N atoms scatters as N carbon atoms**
  b. **Pixel area: $(1/2L)^2$**
  c. **Need 10^3 scattered photons per pixel**
  d. **Scattered amplitude: low-angle ~ N^2; high-angle ~ N**
  e. **0.1nm radiation (12.4 keV)**
  f. **500 kDa (globular) molecule**
     - **Yeast proteins: ~ 50kDa**
     - **Largest known proteins (titins) ~ 3000 kDa**

- **Number of scattered photons/pulse/pixel:**

$$n \sim \Omega_{pixel} W \sigma_C N_{atoms} = \frac{\lambda}{4a^2} W \sigma_C N_{atoms}^{1/3}$$

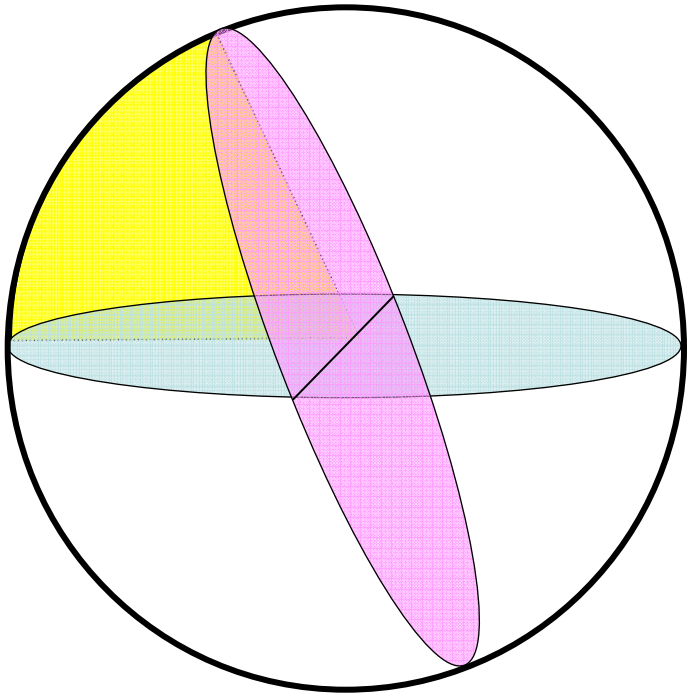
Single-Molecule X-ray Scattering: Orders of Magnitude

A. Ourmazd X-ray Beam		Flux per mm^2	Counts per pulse per pixel		No. of Pulses for 10^9 scattered photons		Time (sec) for 1E9 scattered photons	
Source	Ø (μm)	per pulse	Small Angle	Large Angle	Small Angle	Large Angle	Small Angle	Large Angle
XFEL	0.1	3.10^{20}	10^4	4.10^{-2}	0.1	2.10^4	10^{-3}	2.10^2
APS	0.01	10^{15}	4.10^{-2}	2.10^{-7}	3.10^4	6.10^9	3.10^2	6.10^7

1. XFEL scatters 10^9 photons from a 500 kDa protein in minutes
2. PLENTY of scattered photons; VERY FEW scattered per shot
3. Orienting Diffraction patterns is KEY

Aligning the 2-D Snapshots: Common-Line Approach

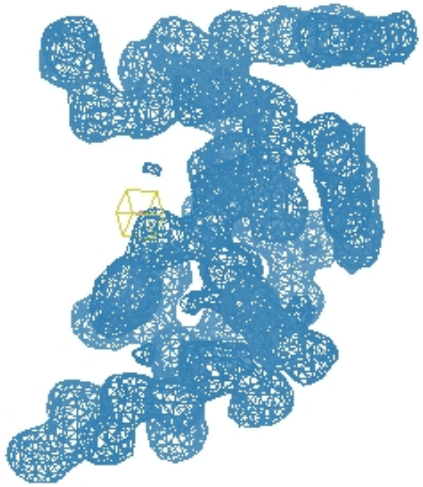
- Diffraction patterns of same object share “common line” of diffracted intensity
  - “Central Section Theorem”
- Three planes fix relative orientations
  - Two with Ewald-sphere curvature
- No phase information available
  - “Friedel ambiguity”
  - Key difference with cryo-EM
- Friedel ambiguity can be resolved
  - Using “consistency restriction”
  - “Handedness” ambiguity remains



Electron Density Recovery

Model of protein Chignolin
(From atom coordinates in PDB)

Recovered Solution
(From DPs of random orientations)



- **1Å photons; ~ 1 Å resolution (collect semi-∠ ~ 32º); Low-angle data excluded**
- **Correlation coefficient ~ 0.8**
- **Shneerson, Ourmazd & Saldin, Acta Cryst, A64, 303 (2008) (arXiv:0710.2561)**

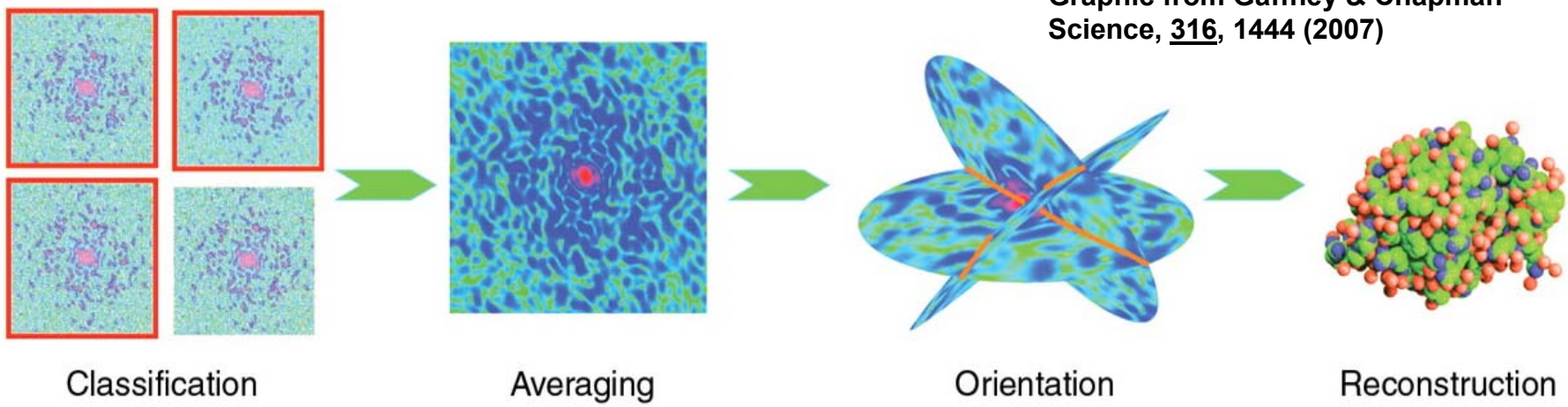
Common-Line Method

- Can align dp's and recover structure in absence of noise
  - RMS alignment accuracy < 0.5˚
- Works with ≥ 10 photons/pixel + shot noise
  - 3 orders of magnitude from expected signal levels
  - Significant performance degradation below 100 ph/pixel
- Cannot be fixed by orientational classification & averaging
  - Flux for reliable classification 100x higher than focused XFEL beam
  - [Bortel & Faigel, J. Structural Biology 158, 10 (2007)]
- Common-line makes poor use of available information
  - Uses correlations between lines of diffracted intensity
  - Highly susceptible to noise
- Must use correlations in entire diffracted photon ensemble
  - From diffraction pattern alignment to photon assignment

Proposed "Algorithm"

[E.g., Huldt et al, J. Structural Biology 144, 219 (2003)]

Graphic from Gaffney & Chapman
Science, 316, 1444 (2007)



- **Averaging over "similar patterns" needed to orient diffraction patterns**
  - **Requires classifying single-shot patterns containing few photons**
- **Needs single-shot fluence ≥10^{22} photons/mm^2**
  - **XFEL delivers ~10^{20} photons/mm^2 into 100nm Ø probe**
  - **[Bortel & Faigul, J. Structural Biology 158, 10 (2007)]**
- **Insufficient flux for orientational classification (& averaging)**

Common-Line Method

- Imagine classification could be done (somehow)
  - DP's could be averaged to enhance signal/noise
- Common-line needs 10 ph/pixel; 10^{-2} available in each dp
  - Must average 10^3 dp's $\Rightarrow$ need 10^3 dp's per orientation class
  - For 100Å particle, need 10^6 orientational classes [B&G]
  - Must collect 10^9 dp's
- One experiment would take > 4 months of beam time at LCLS
  - 100 patterns collected per second
- Going to larger molecules does not help
  - 300Å particle gives 3x more signal, needs 20x more classes
- Move from dp alignment to photon assignment
  - Use correlations in entire diffracted photon ensemble

Reconstructing the 3D Diff. Intensity: New Approach

- How do you put a broken glass back together?
  - Like a 3-D jigsaw puzzle
  - Based on correlations between the pieces
- Reconstructing unseen vase broken into 10^6 pieces
  - About the number of orientations of the molecule
  - I.e., the number of diffraction snapshots
- Can you put it back together?
  - I.e., reconstruct the 3-D diffracted intensity distribution
  - Like tomography with no orientational information
- Under a light delivering 10^{-2} photons per detector pixel

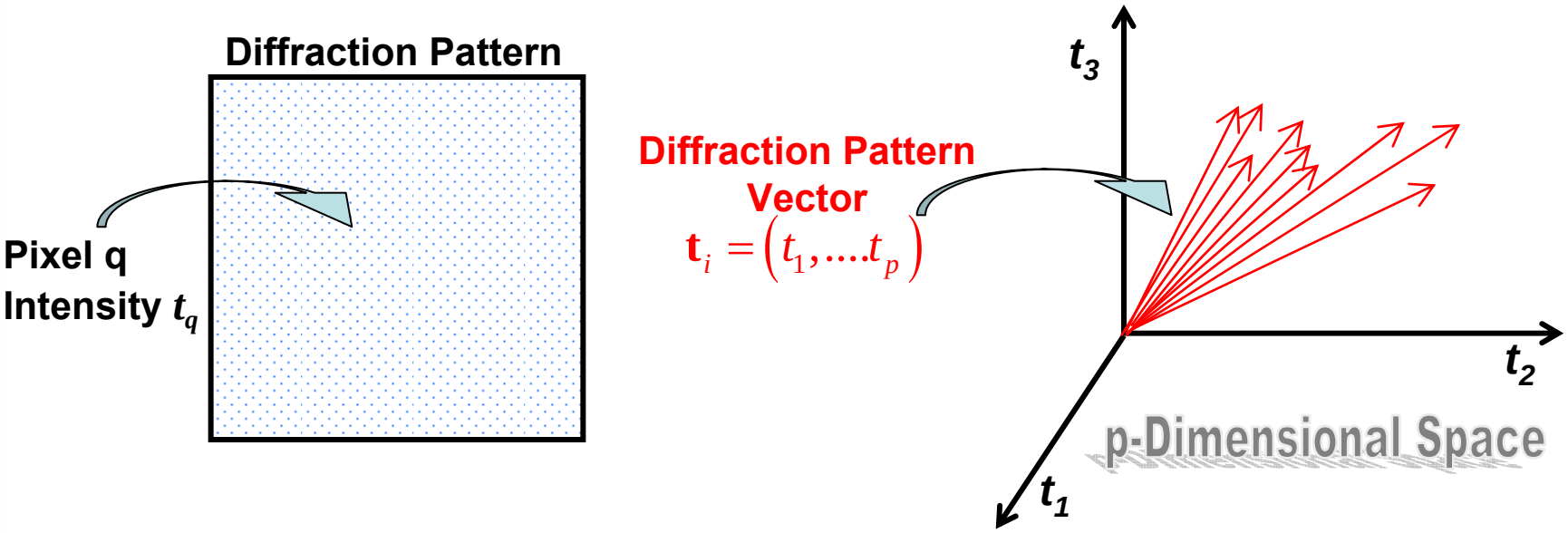
That's what we are trying to do!

New Approach: Summary

- Uses ensemble of scattered photons
  - To first order, does not rely on photons scattered per shot
- Reconstructs diff. intensity distribution from correlations
  - Within scattered photon ensemble
- Based on generative Bayesian mixture modeling
  - Developed originally for data visualization & neural networks
- Can align diffraction patterns down to MPC ~ 0.01 ph/pixel
  - Anticipated MPC for 500kDa protein with LCLS
  - 1000x improvement over previous techniques
  - Uses 10^5 scattered photons only (compared with 10^9 from LCLS)
  - Anticipate significant room for improvement

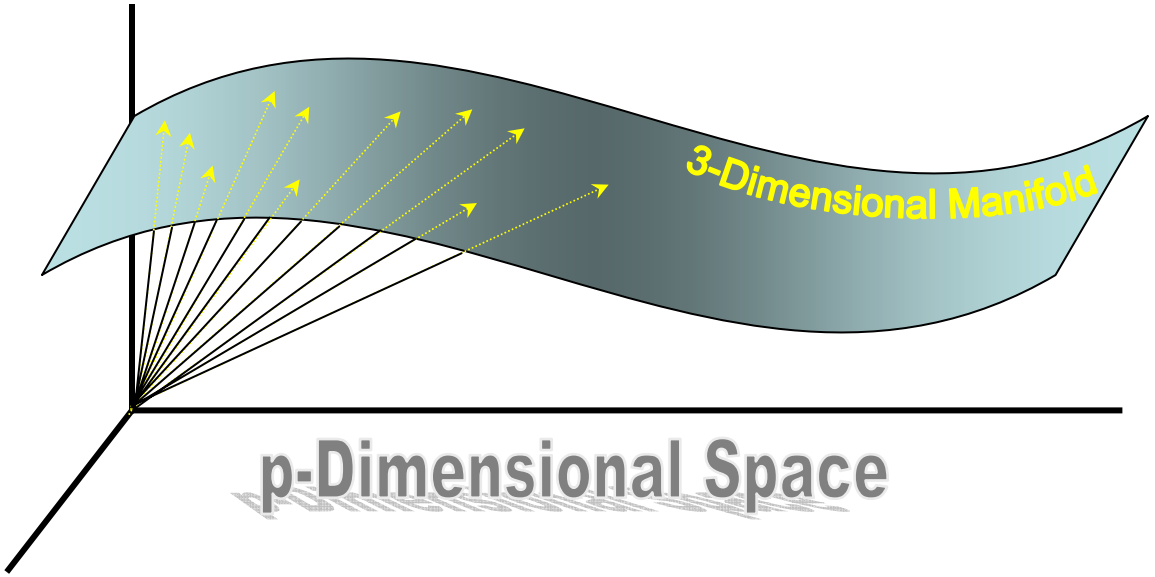
New Approach: Data Representation

- All we have is ensemble of diffracted intensities
  - A diffraction pattern is $\mathbf{t}_i = (t_1, \ldots . t_p)$
  - A vector in p-dimensional "intensity space"
  - Total dataset is collection of vectors $\mathbf{T} = (\mathbf{t}_1, \ldots . \mathbf{t}_d)$

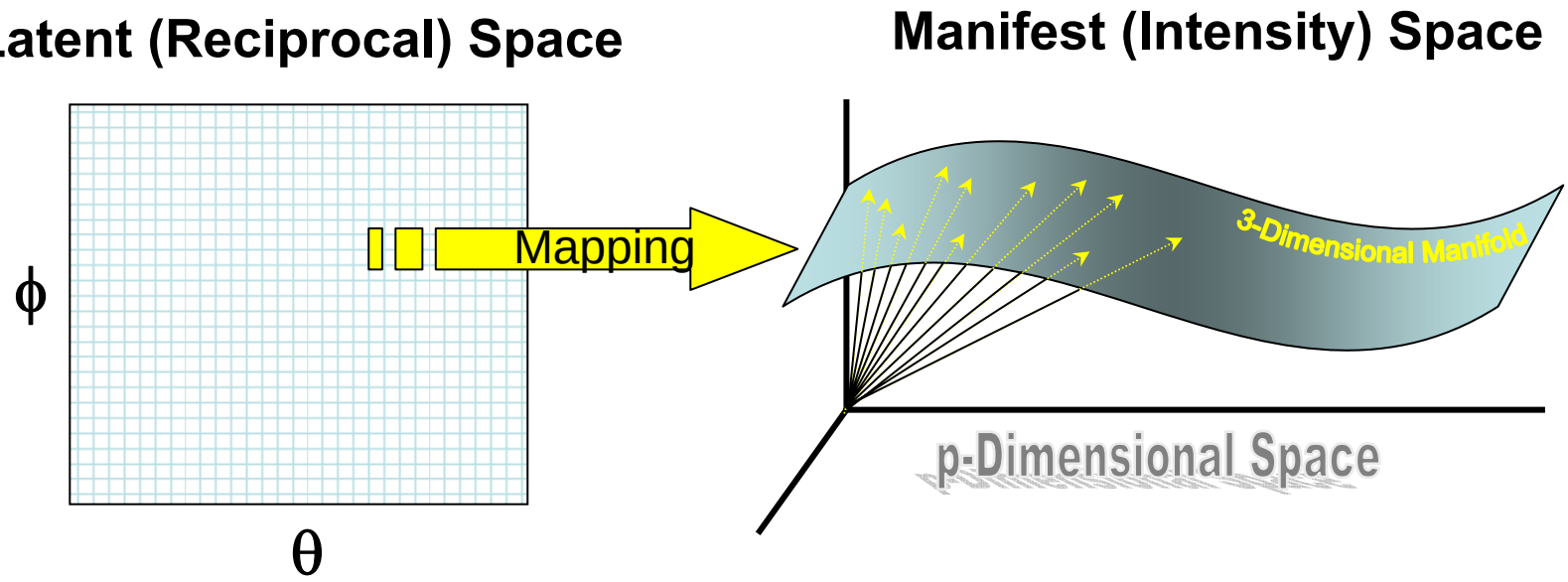


Reconstituting the 3-D Diffracted Intensity Distribution

- Diffracted intensity vectors live in p-dimensional space
- But intensities (& vector) function of only three variables
  - Angles (θ, ϕ, ψ) defining molecular orientation
- Vectors define a 3-D manifold in p-dimensional space



Manifest & Latent Spaces



- **Diffraction pattern vectors function of three latent (hidden) variables**
  - **Confines vectors to 3-D manifold in p-dimensional space**
- **Mapping between two spaces nonlinear**
  - **Maps 3-D reciprocal space to 3-D manifold in intensity space**
- **Maps 3-D intensity distribution to p-D vector distribution**
  - **Links distributions in "latent" reciprocal and "manifest" intensity spaces**

Generative Topographic Mapping

[C.M. Bishop, *Neural Networks for Computation*, OUP (1995)]

- Type of (nonlinear) factor analysis
  - Developed for data visualization, neural network applications
  - Linear factor analysis used in bio- & psychometrics
- Fits low-D manifold to data to determine mapping function
  - "Principled" probabilistic approach (Bayesian statistics)
- Allows reconstruction of 3-D intensity distribution
  - Links 3-D reciprocal space to p-D intensity space
  - Based on maximum likelihood, Bayesian statistics
  - Uses correlations in entire diffracted photon ensemble
- Might allow direct connection to electron density

Generative Topographic Mapping (GTM)

- **Mapping between (3-D) latent and (p-D) manifest spaces nonlinear**

$$\mathbf{y} = \mathbf{W}\phi(\mathbf{x})$$

$$\phi(\mathbf{x}) = \{\phi_j(\mathbf{x})\}, \qquad 0 \le j \le M$$

$\mathbf{y}$: Mapping function; $\mathbf{x}$: Latent space coordinate

$\phi(\mathbf{x})$: Basis set; $\mathbf{W}$: Free parameters

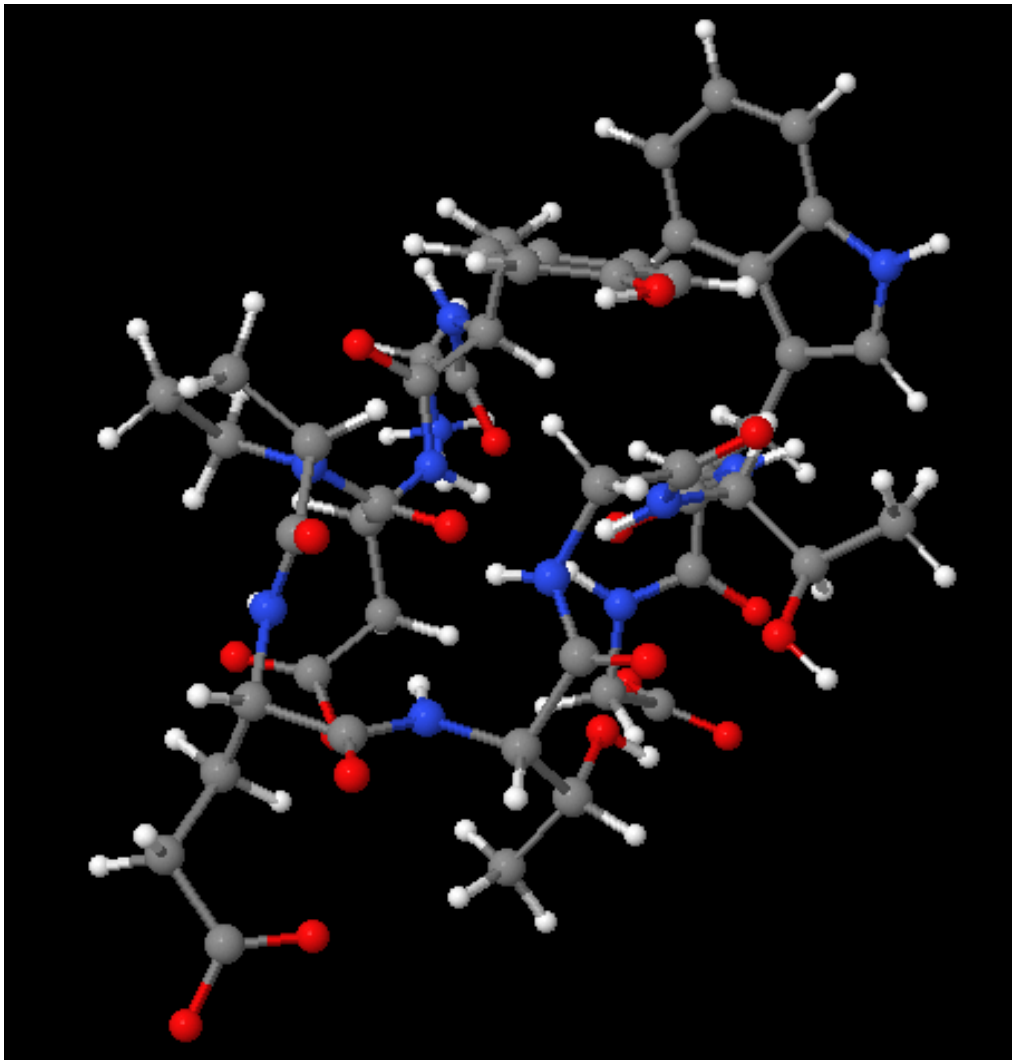
- **Determine nonlinear function by fitting 3-D manifold to data**
  - **In data space, by adjusting weights W**
  - **Use maximum likelihood (EM) algorithm**
  - **[C.M. Bishop, *Neural Networks for Computation*, OUP (1995)]**

- **Map vector distribution to diffracted intensity distribution**
  - **From "manifest" intensity space to "latent" reciprocal space**
  - **Through nonlinear function $\mathbf{y}$, Bayesian statistics**

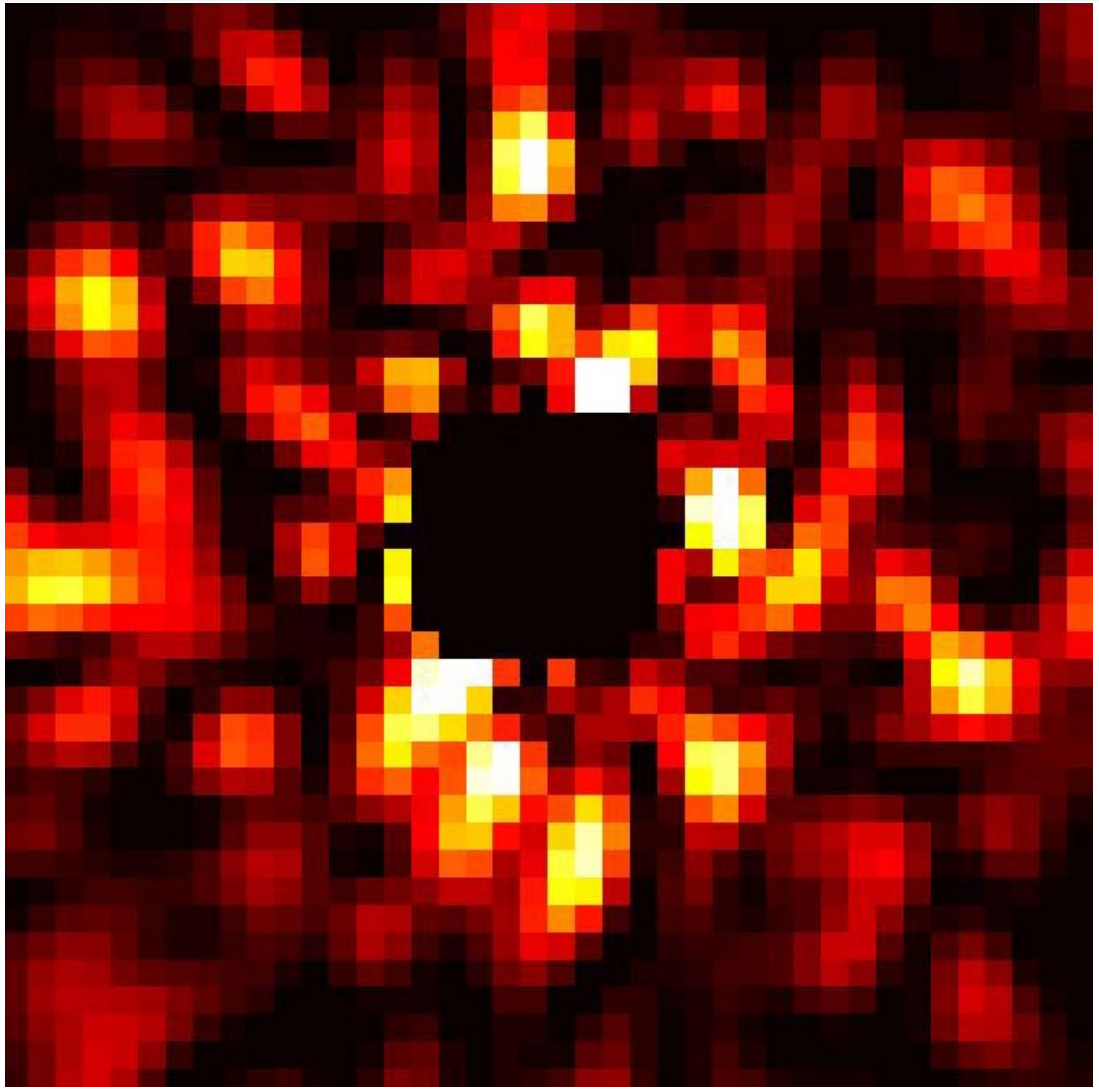
Reconstructing a Protein

- Take small protein
  - Chignolin, 10 residues, ~ 100 atoms
- Simulate diff. patterns at random molecular orientations
  - Each one corresponding to a diffraction snapshot
- Signal ~ 10^{-2} photons per pixel + shot noise
  - Signal/noise expected for 500kDa molecule
- Determine orientations with no prior information
  - Other than dimensionality of rotation space (1-D or 3-D)
- Compare with correct orientations

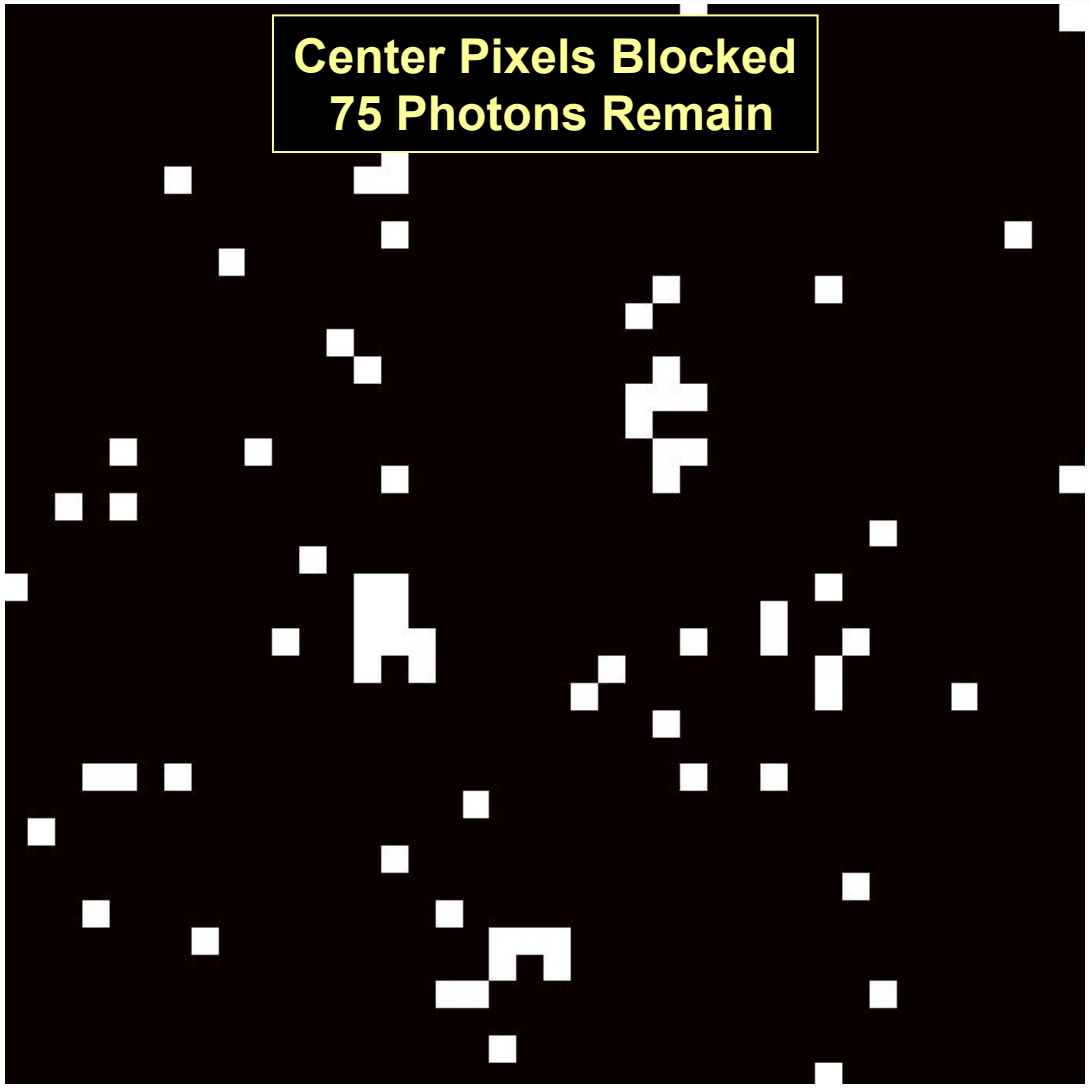
Model Protein: Chignolin



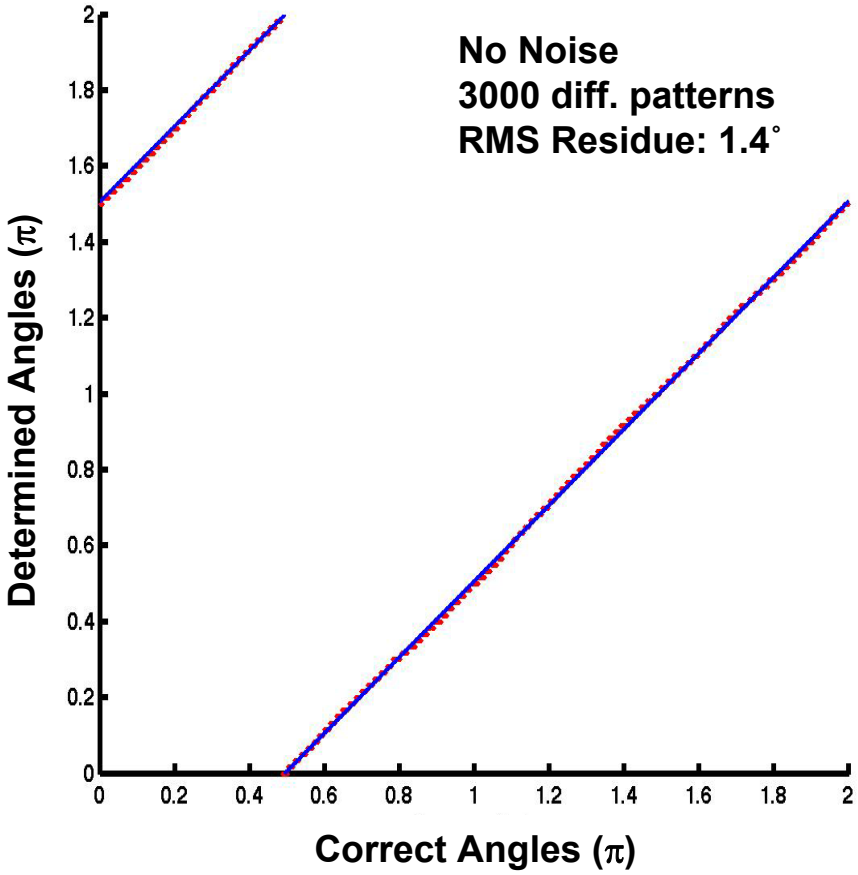
Diffraction Snapshot
No Noise

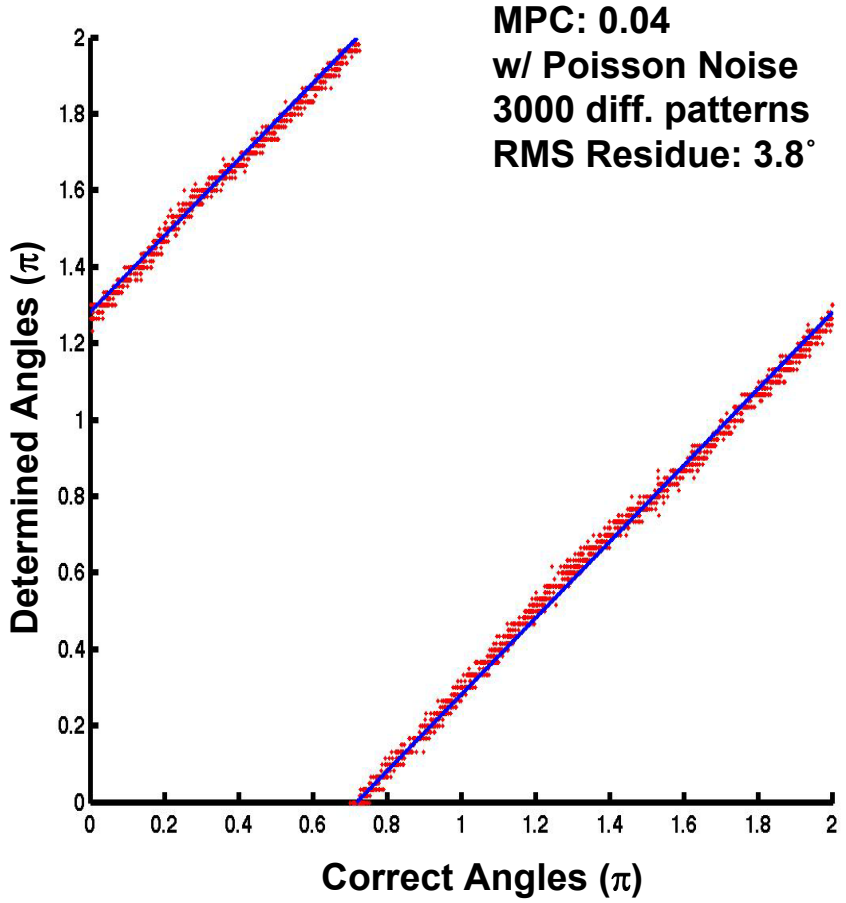


Diffraction Snapshot
$4x10^{-2}$ Photon/Pixel + Shot Noise

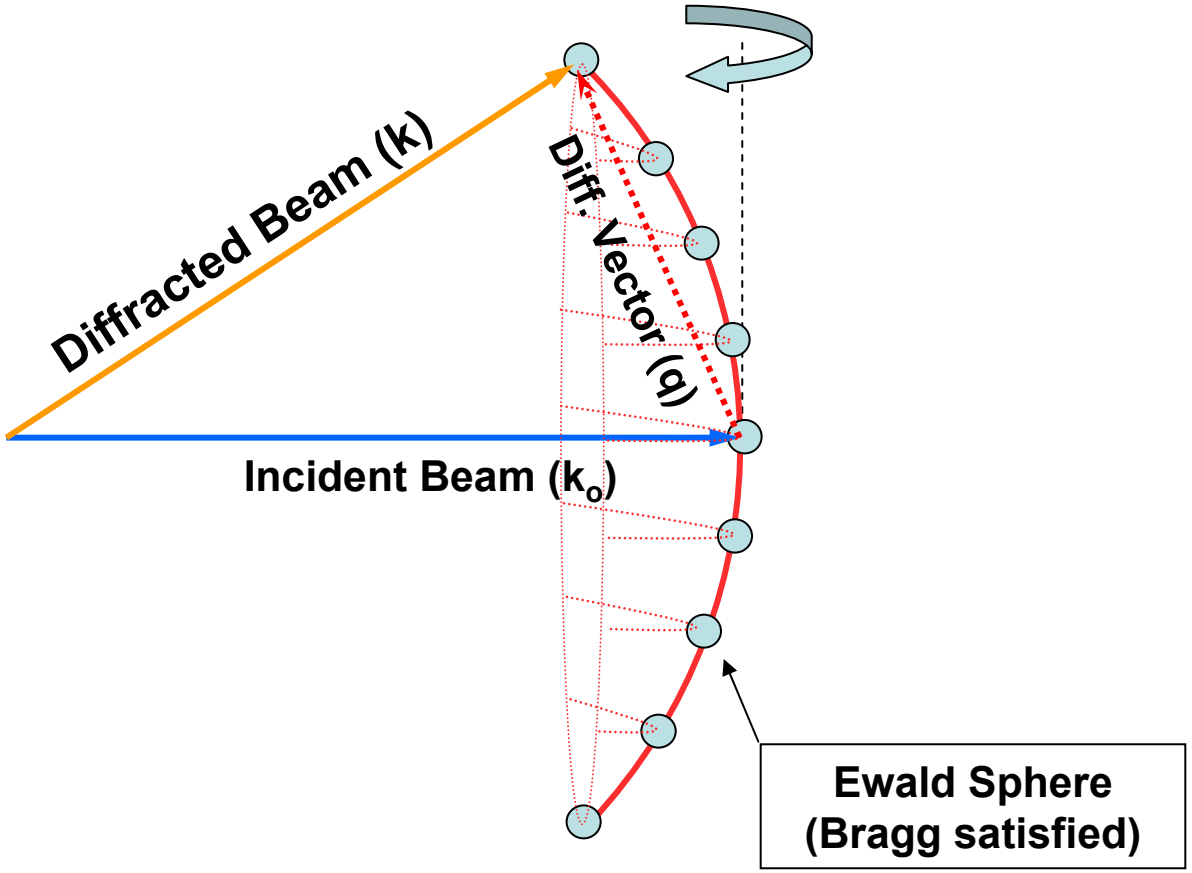


Angles Determined by GTM
Molecule Rotating About One Axis

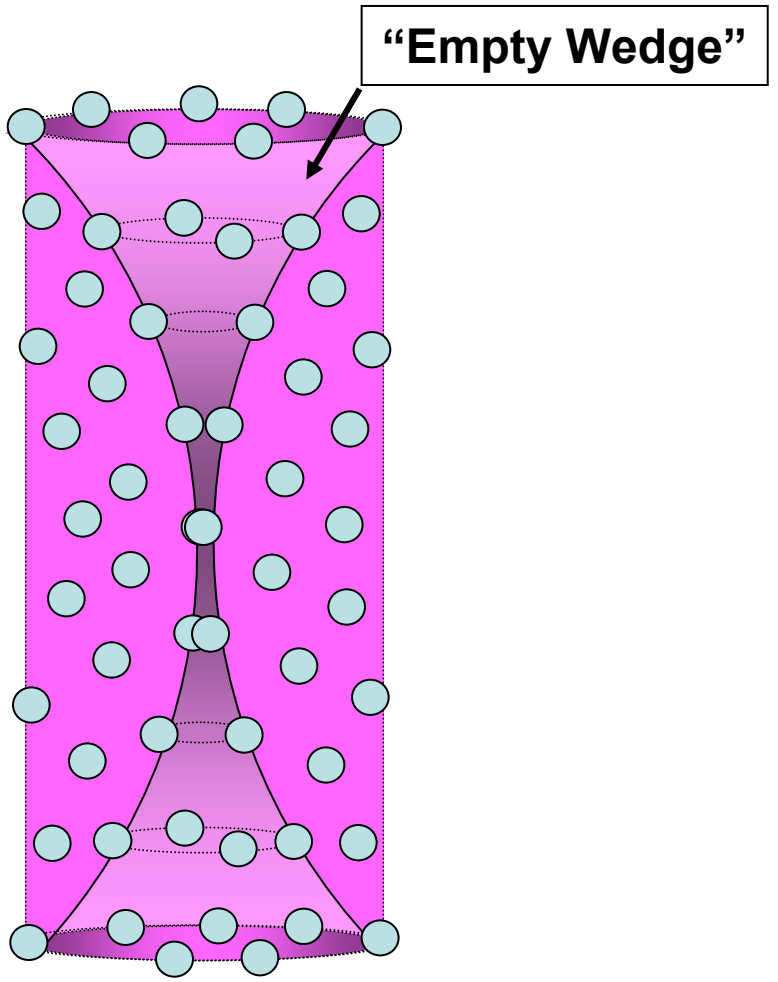




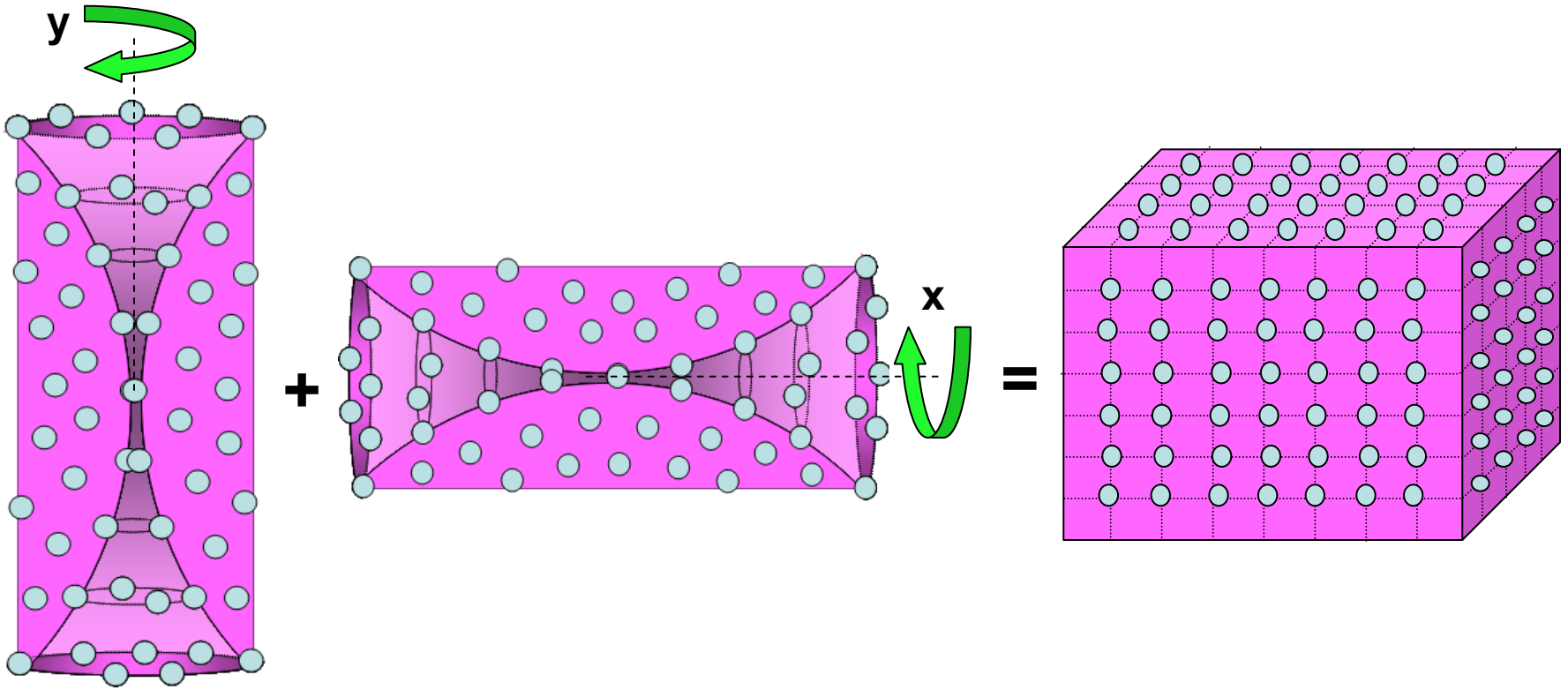
Diffraction Geometry



Reciprocal Lattice Filling
Rotation About One Axis



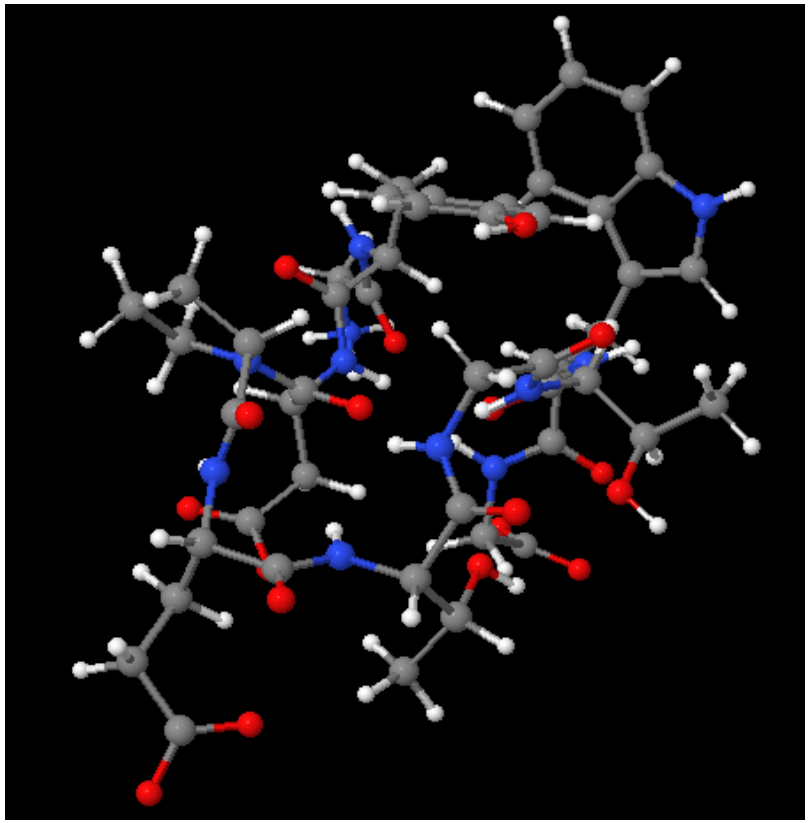
Reciprocal Lattice Filling
Rotation About Two Axes



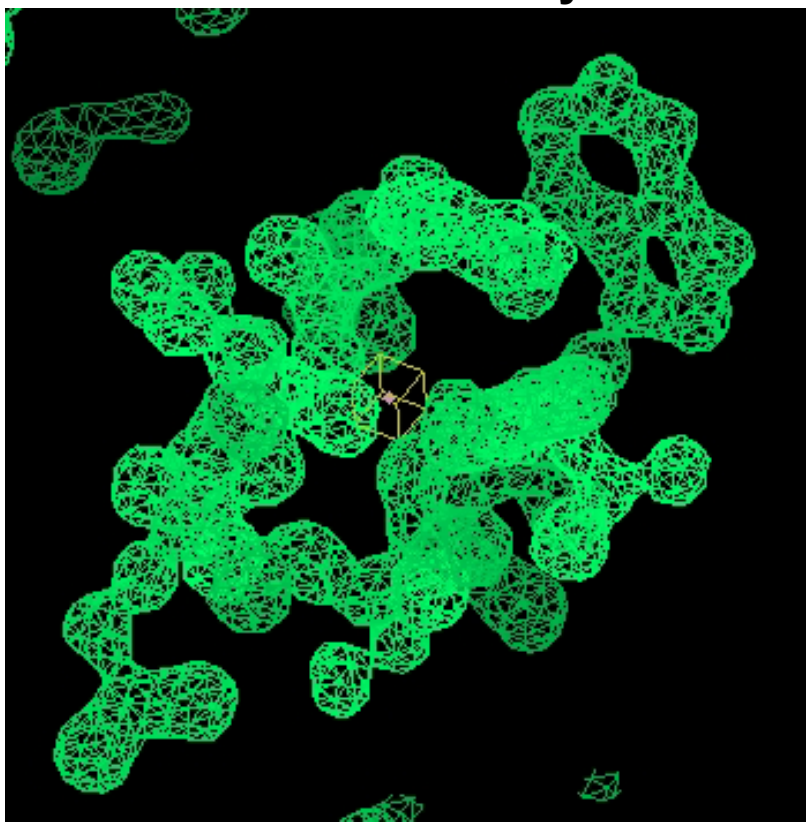
Produce uniform gird of points in reciprocal space for “Phasing”

Model Protein: Chignolin

Ball-and-Stick Model

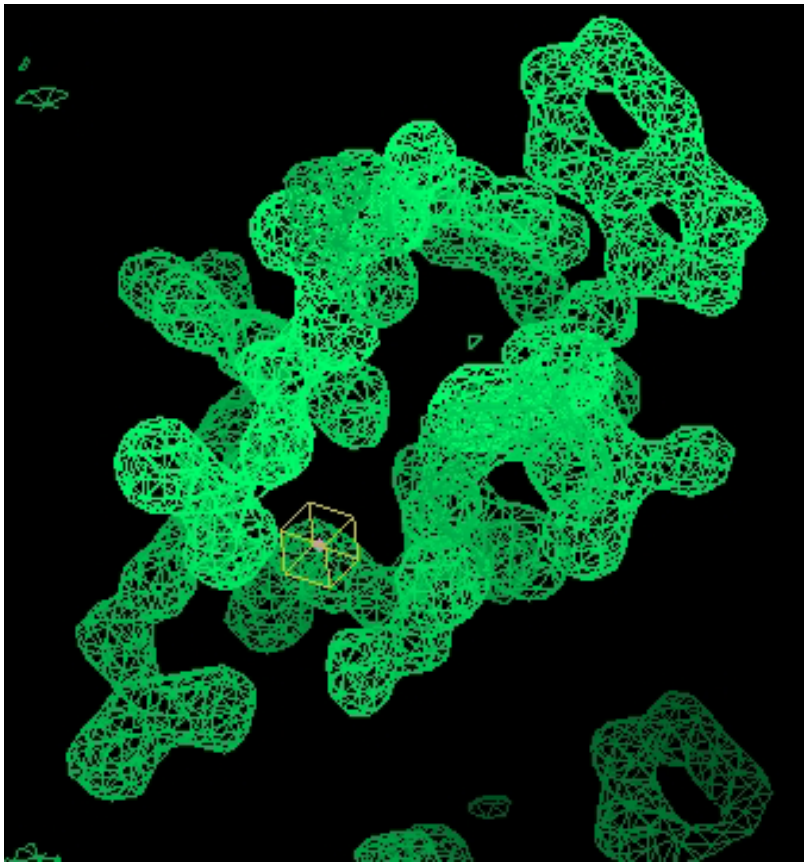


Electron Density

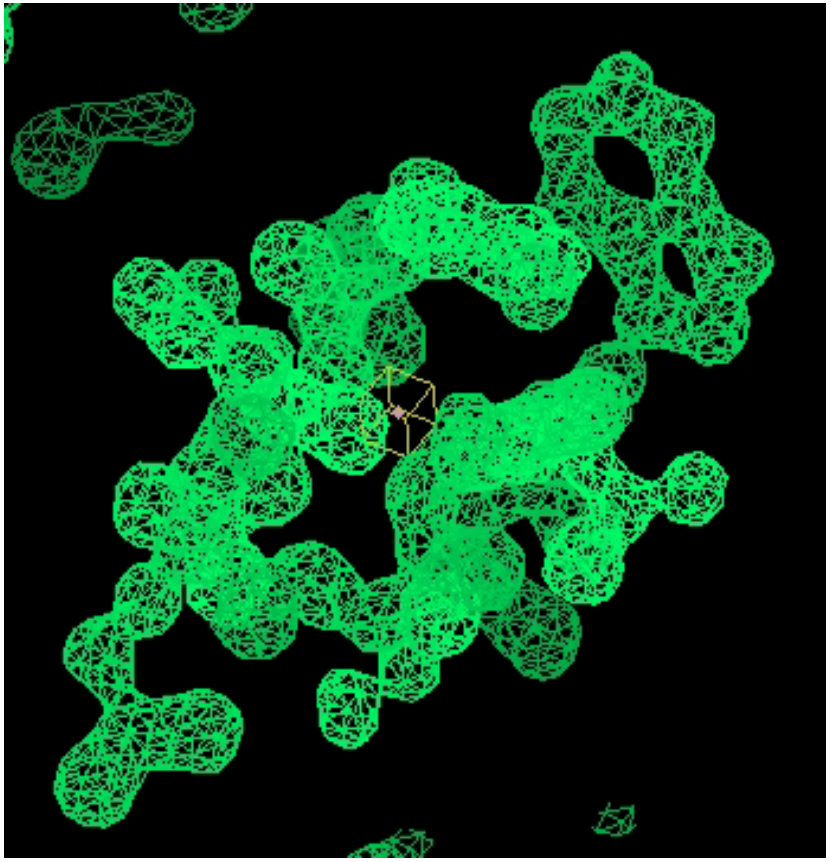


Reconstructed Electron Density
Noise-Free

Reconstructed with GTM Angles



Actual Electron Density

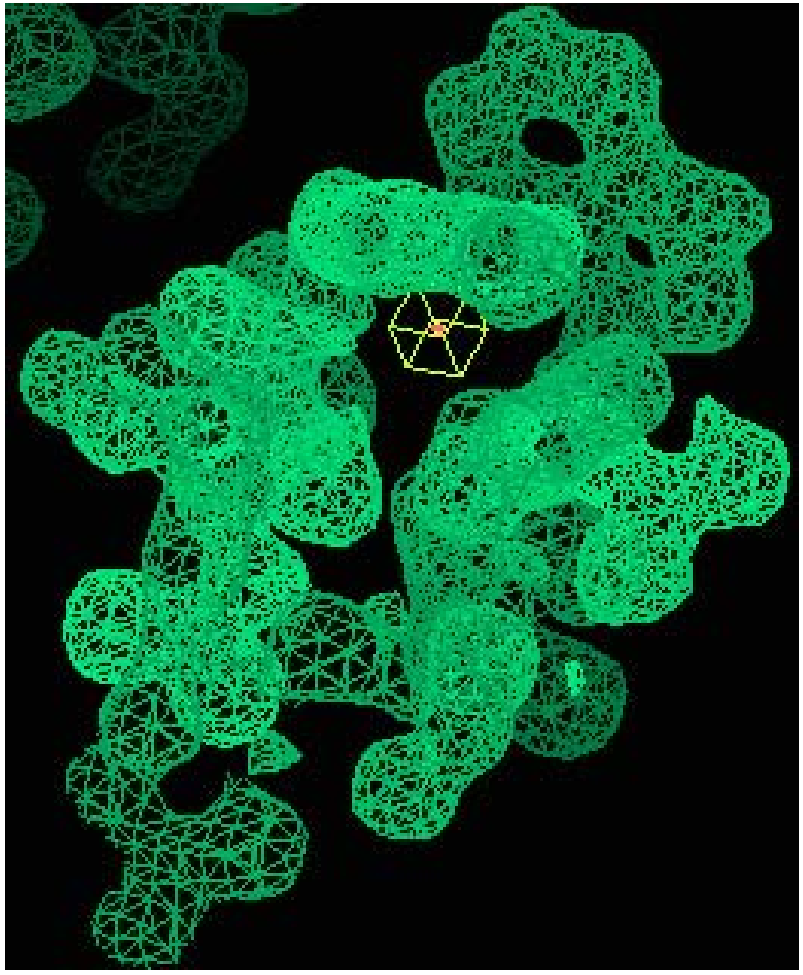


Reducing Mean Photon Count

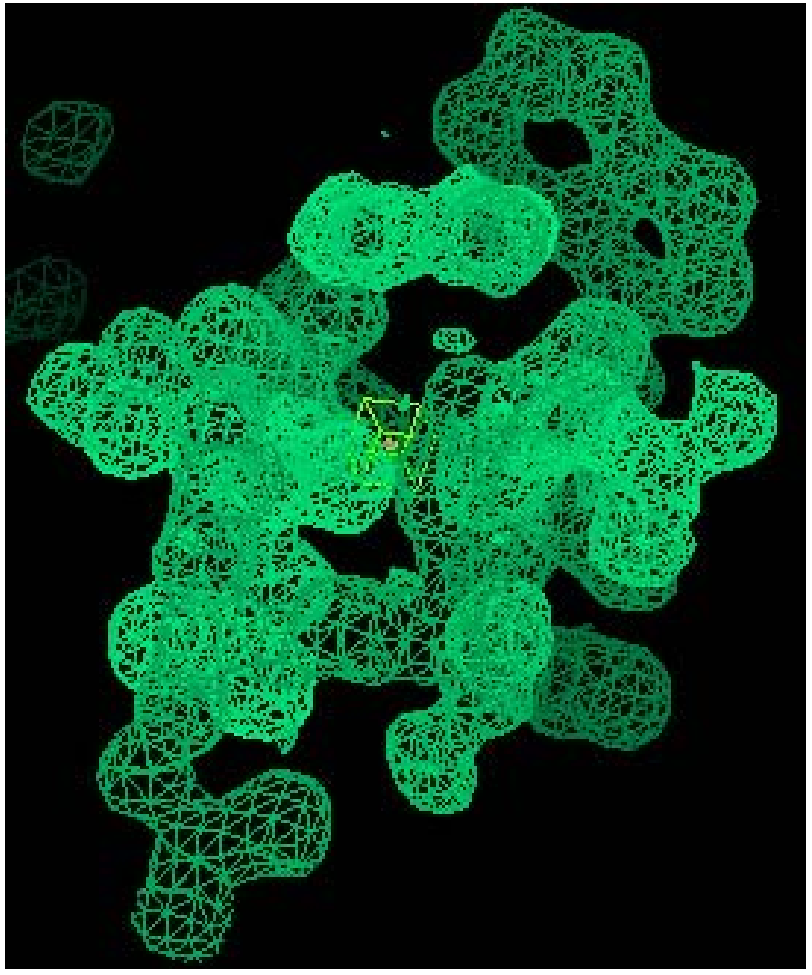
- Shot noise increases
  - Modeled as Poisson statistics
- Need ~ 5 photons/pixel for “phasing”
  - Iterative recovery of electron density from intensities
- Need ~ 100 ph/pixel for gridding
  - Due to inadequacies of gridding algorithm?
- Reconstruction at 0.04 MPC needs ~30 million dp’s
  - Average patterns to reach 100 ph/pixel (1-D rotation axis)
  - GTM of this magnitude beyond our desktop CPU/memory capacity
- Distribute dp’s according to GTM accuracy @ 0.4MPC
  - Simulated 300,000 dp’s, distributed to mimic GTM error
  - Gridding and phasing

Reconstructed Electron Density
Mean Photon Count: 0.4 per Pixel

Reconstructed with GTM Angles

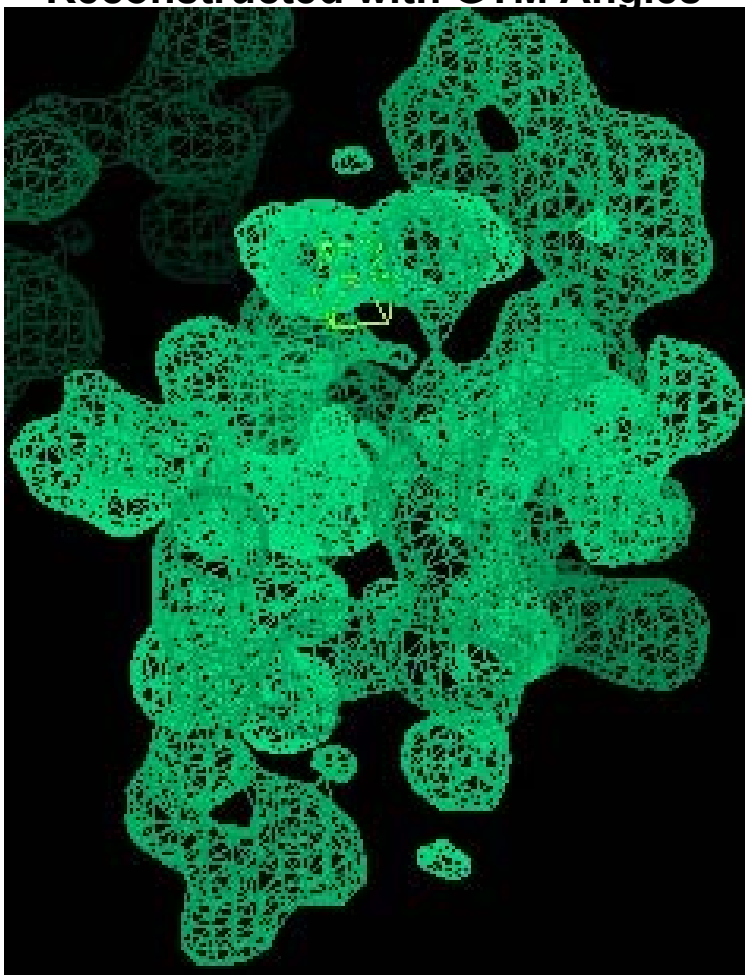


Actual Electron Density

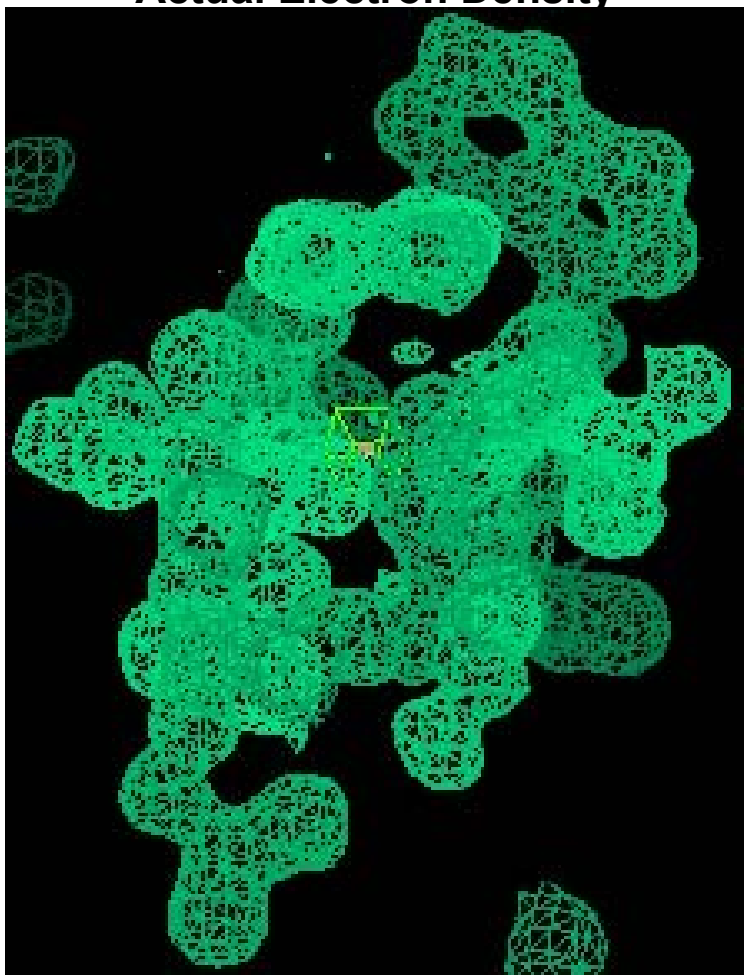


Reconstructed Electron Density
Mean Photon Count: 0.04 per Pixel

Reconstructed with GTM Angles



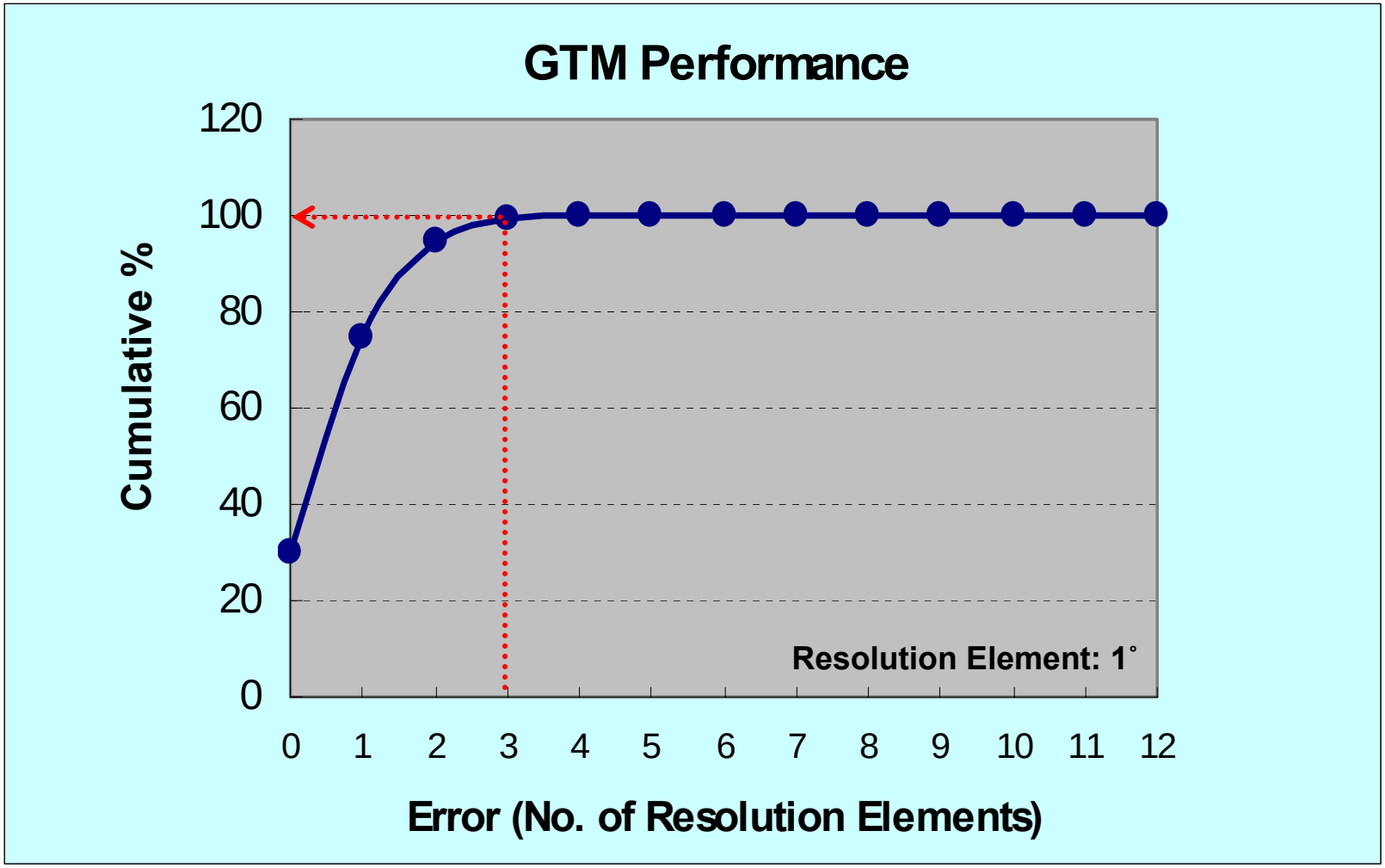
Actual Electron Density



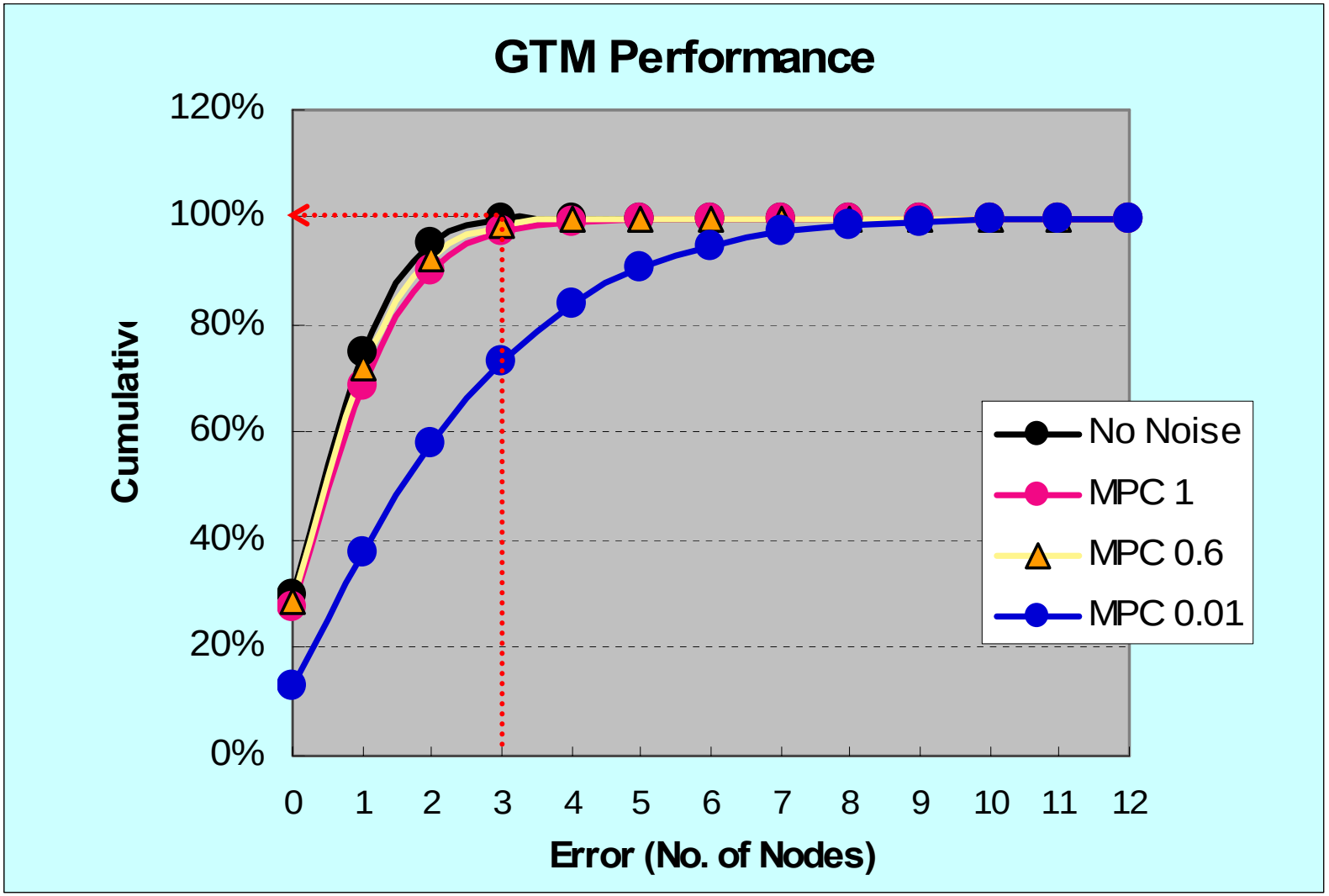
Alignment
3-D Rotational Freedom

- **Orientational distance metric**
  - **How do you define orientational "proximity" in SO3?**
  - **Quaternions**
- **Figure of Merit**
  - **How well has the orientation been determined?**
  - **To within two or three latent space nodes**
- **Effect of noise**
  - **How low can we go in mean photon count per pixel?**
  - **Demonstrated performance down to 0.04 ph/pixel with Poisson noise**
- **Computational load**
  - **Memory is primary limitation**
  - **Present limit: 10^4 data vectors, each a 4x40 pixel diffraction pattern**
  - **~30˚x 30˚x30˚ patches of orientational angles**

Aligning in 3D: Interim Results
No Noise



Aligning in 3D with Poisson Noise



Aligning in 3D: Summary

- Alignment possible to within 2-3 resolution elements
  - Each element corresponds to ~ 1˚- 3˚
- Alignment possible down to 0.01 photons/pixel
  - Using ensemble of only ~ 10^5 scattered photons
- Anticipate significant room for improvement
  - Replace Gaussian noise model in GTM with Poisson
  - Provide more photons
  - Can collect 10^9 scattered photons in an hour with LCLS
- Encouraging preliminary results

What Does It All Mean?

- Can reconstruct diffracted intensity distribution down to MPC 0.04
  - From correlations within diff. photon ensemble from small protein
  - Mean photon count (MPC) 0.04 / pixel expected from 500 kDa protein
- Can trade single-shot flux for total number of shots?
  - Such that enough photons are scattered in experiment
- Reduce single-shot flux below damage threshold?
  - Provided experimental times remain reasonable
- What is the damage threshold for single molecule?
  - Indications it might be 100x higher than Henderson limit
  - If so, "sweet spot" is 10^{18} photons/mm^2/shot
  - Molecule not destroyed by shot
  - Data collection window extended to ps-ns regime

Conclusions

- **Can reconstruct 3-D intensity distribution down to ~10^{-2} ph/pixel**
  - **Applicable to single molecules, single particles, colloids, etc.**
  - **Removed the tyranny of single-shot dose requirement**
  - **Using correlations within entire scattered photon ensemble**
- **Could be used for range of other important problems**
  - **Should allow direct access to electron density**
  - **Adaptive digital energy filter**
- **Critical issues remain**
  - **Minimum photon count needed for structure recovery?**
  - **Radiation damage threshold; suitable operating regime, etc.**
- **Success would have significant & broad impact**
  - **Access to all macromolecules, possibly different conformations**
  - **Implications for physics, materials, biochemistry, drug design**