NEW BRAINS FOR THE 'SCOPES

Next-generation intelligent methods for biological image processing

Christoph Best
Max Planck Institut für Biochemie
Martinsried, Germany

Schloss Ringberg, March 5, 2007

Informatics for Images

Classical image processing:
Fourier filtering methods
Correlation
Deterministic algorithms

Image file formats

Next-generation bioimage informatics:
Non-linear filters
Pattern matching
Probabilistic algorithms
Machine-learning methods
Monte Carlo randomized algorithms
Information integration
Image databases
Data mining

Talk outline

Three problems in single-particle analysis

- ▷ Clustering and classifying difficult data
- ▷ Automated picking/boxing of particles
- ▷ Model-free determination of view directions

Graphical interlude:

- ▷ Visualizing tomograms using UCSF Chimera

Two exercises in tomography:

- ▷ 3D reconstruction using level sets
- ▷ Automatic triangulation of marker positions

Single-particle analysis

Three problems in single-particle analysis

- Clustering and classifying difficult data
- Automated picking/boxing of particles
- Model-free determination of view directions

Lessons learned

Alignment is dangerous

- Can generate class averages that are not supported by the class members individually
- Requires careful analysis of the contents and variation of each class
- Preferable: automatic quality assurance to quantify how probable a model is given the observations
- and how much more probable it is than other models (null hypothesis)

Future:

- Bootstrap, Monte Carlo sampling of model space...

Particle picking

Problem:

- ▹ Views are randomly distributed on images
- ▹ Must pick regions with particles from image

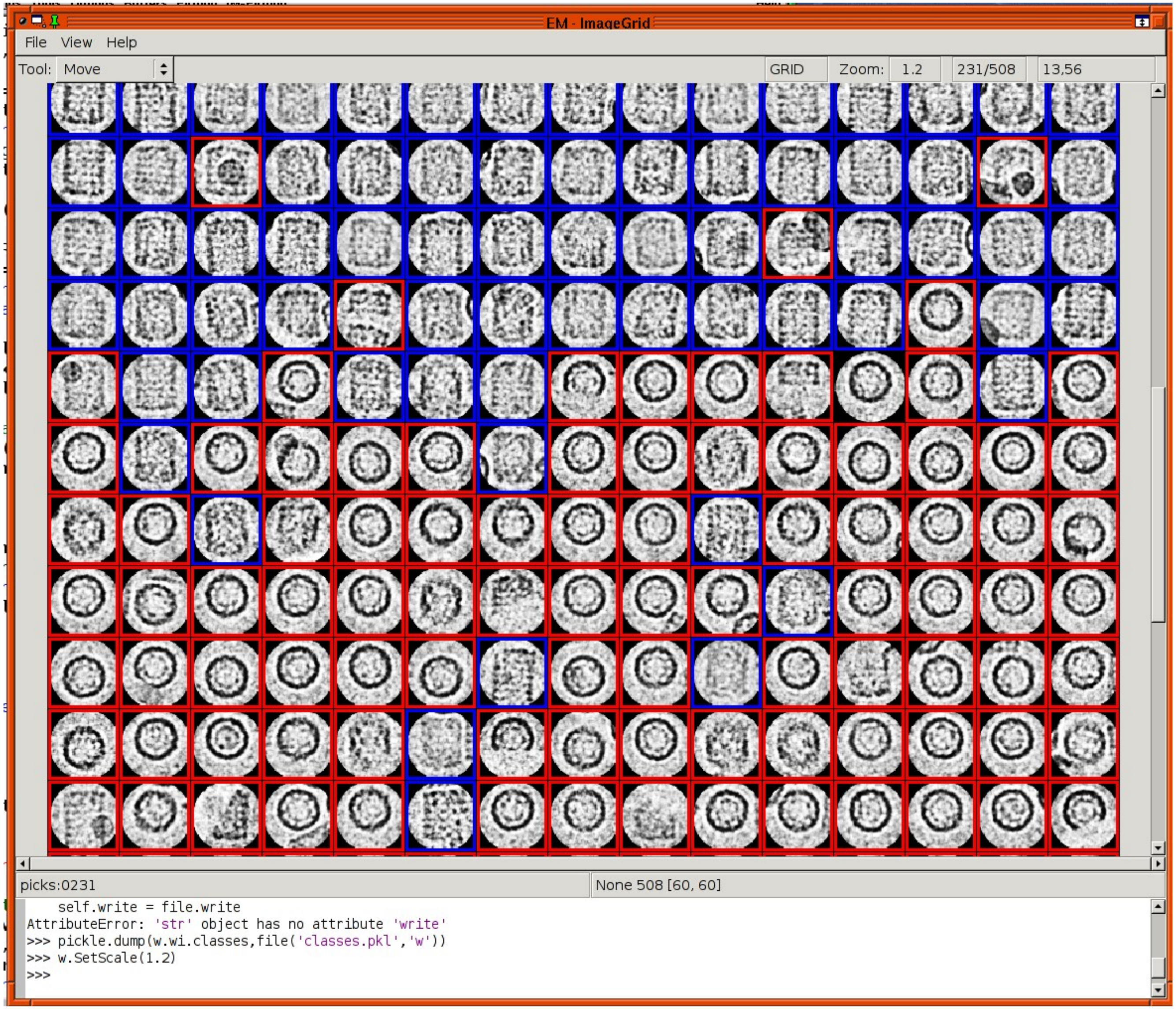
Difficulty: high noise ➜ simple template matching does not work

Approach:

Initial picks by linear correlation

Use a Support Vector Machine (SVM) to select for correct particles according to a manually chosen data set

Picking by template matching



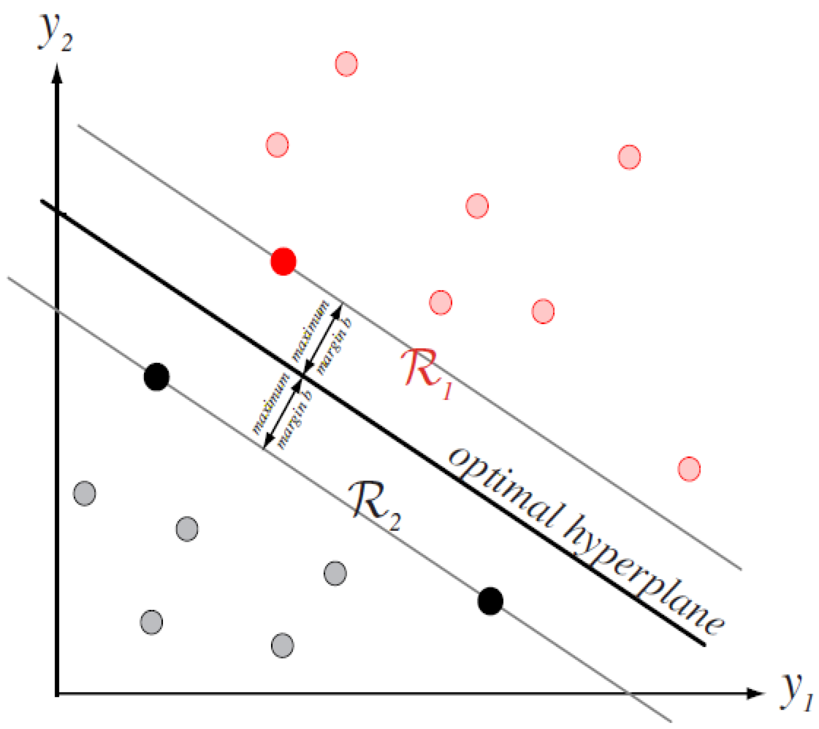
Picking by
linear correlation

many
mis-picks

Apply SVM
to pixel vector
(reduced) of
the images

Coloring: training data set

Support Vector Machines



From Duda *et al.*, *Pattern Classification*

Machine Learning:

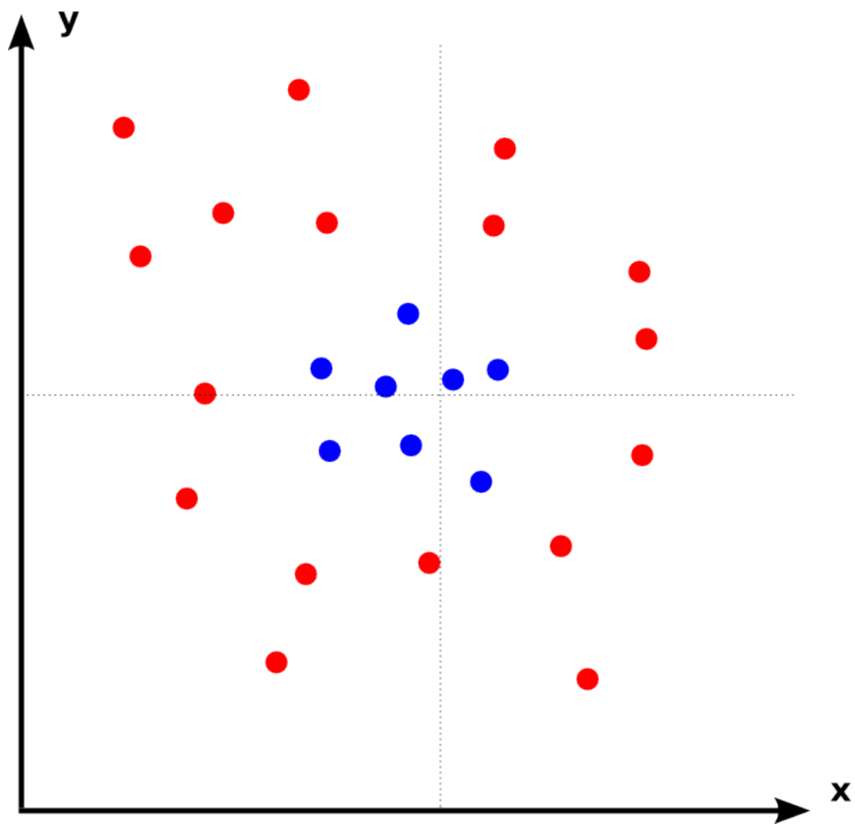
- Training (vs rules)

Support Vector Machine:

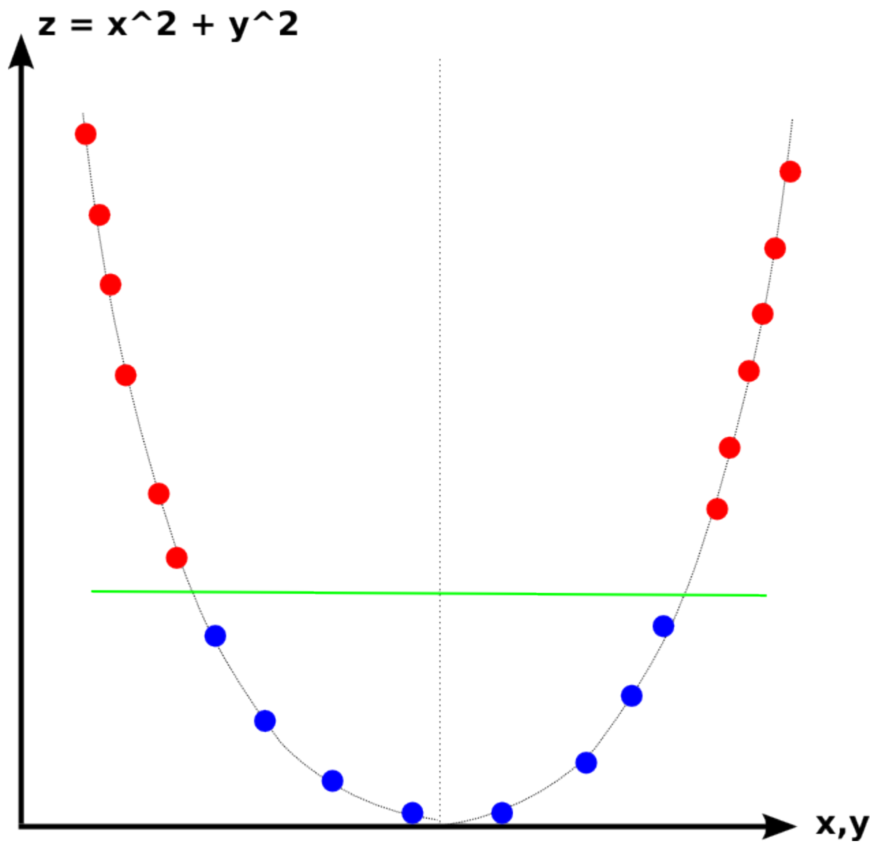
- Linear classifier
- Extended to higher polynomials
- Efficient calculation of the separating hyperplane by duality transform

Non-linearity

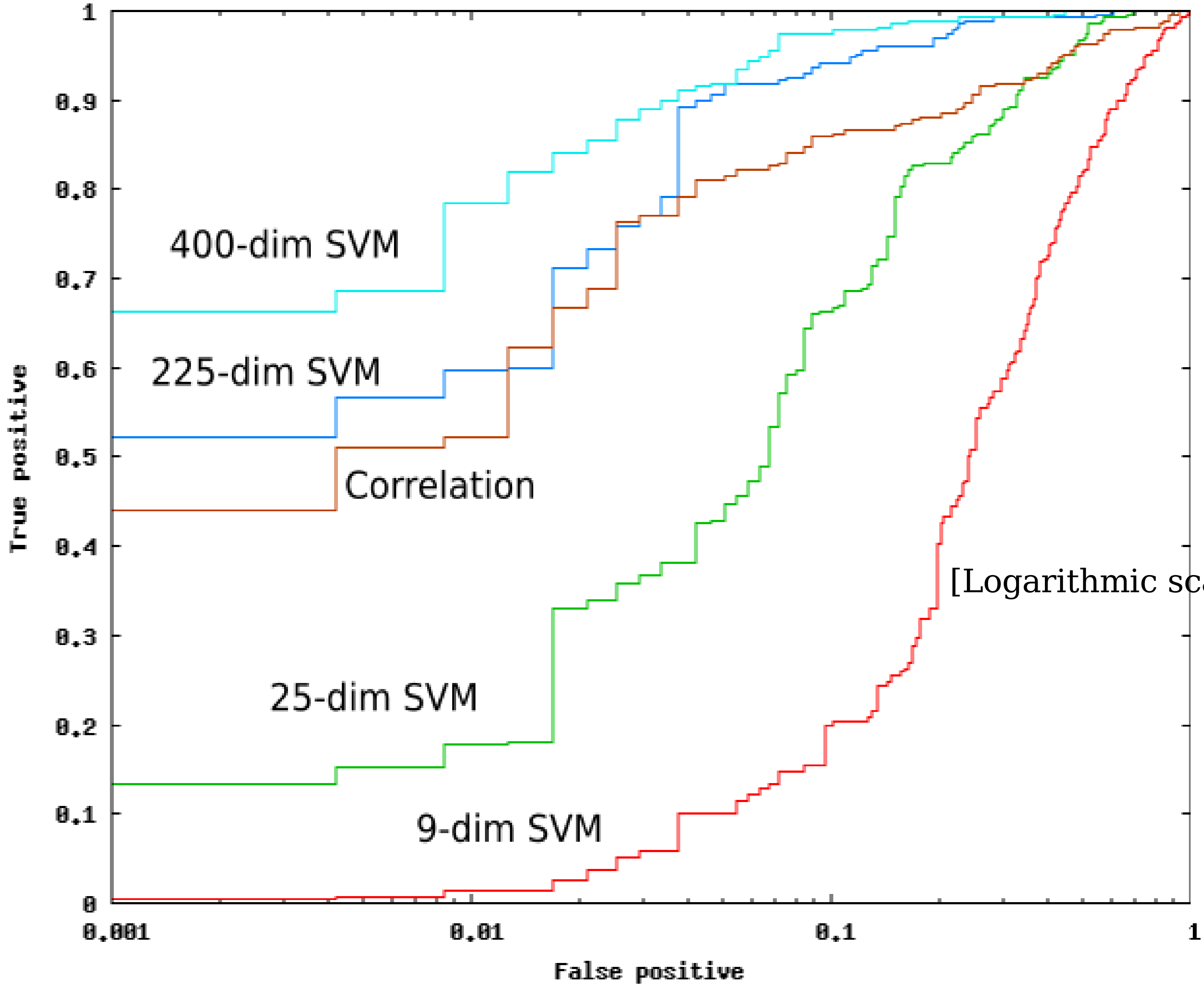
Linearly inseperable



Linearly separable
after introduction of pseudo-variable



Improving picking using SVMs

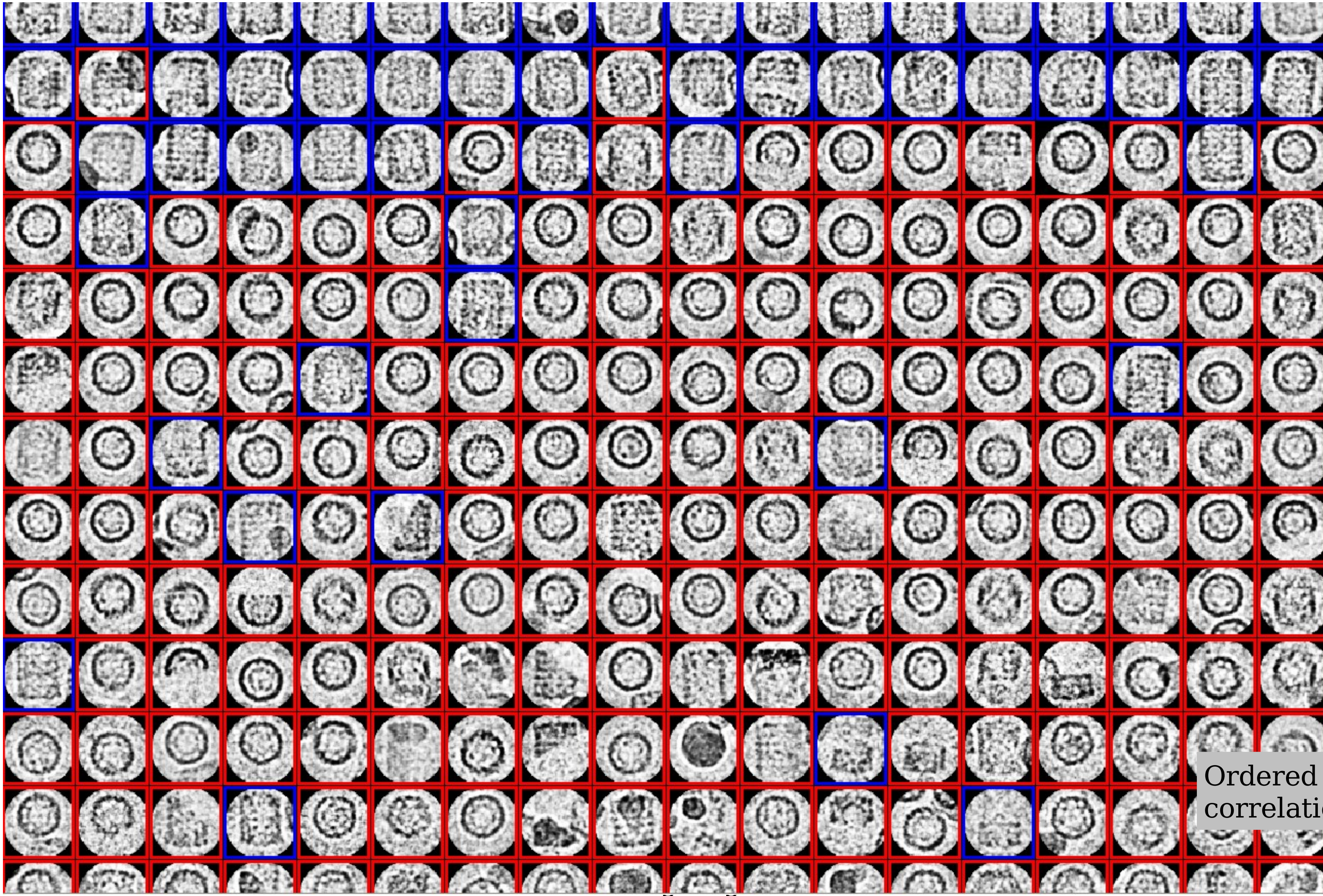


[Logarithmic scale]

tics
t

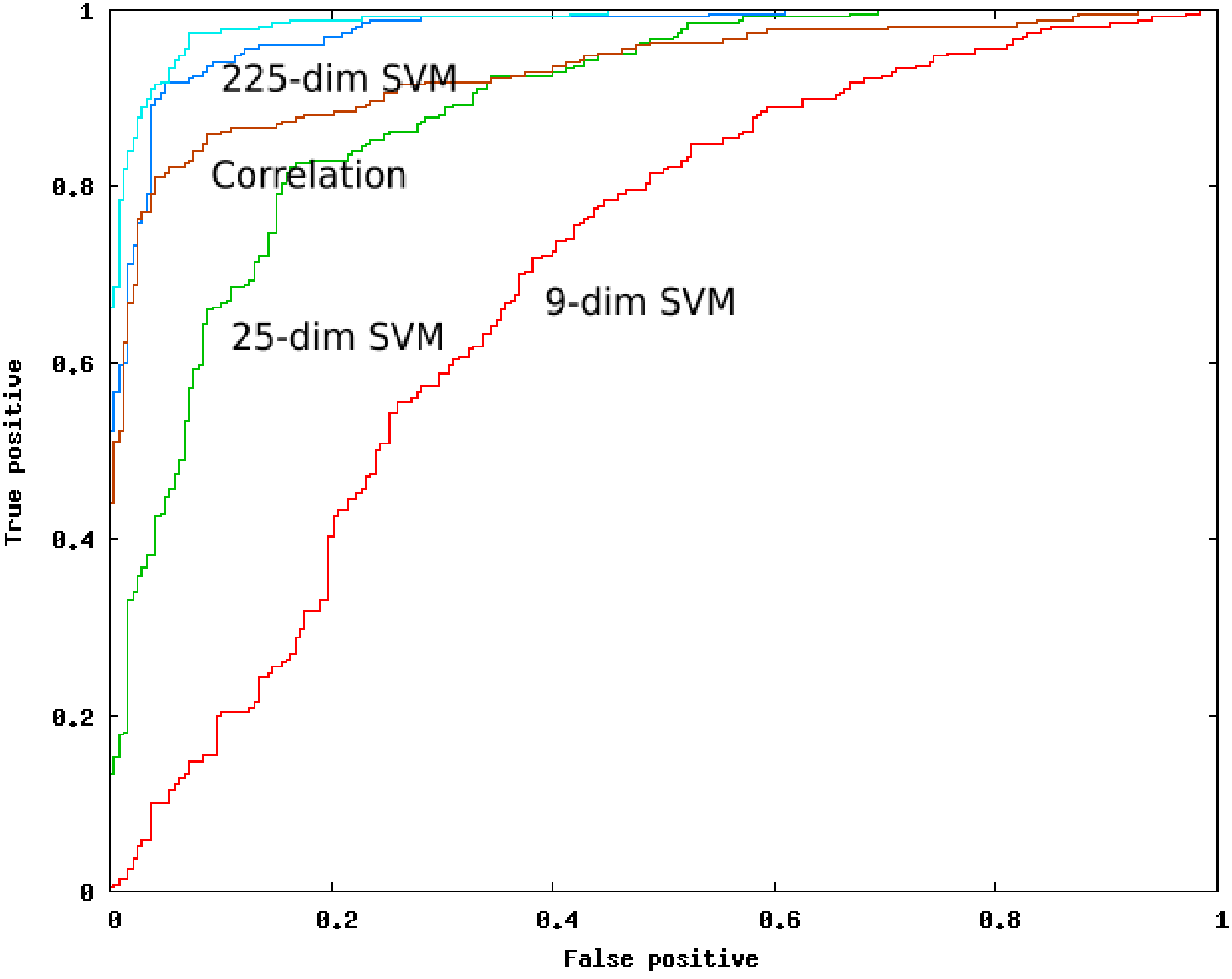
Best

Picking result



Ordered by
correlation value

Improving picking using SVMs



Receiver operating characteristics for different feature set sizes

M. Tacke, C. Best 2006

Summary: Picking

- Beyond template matching
  - SVM recognizes feature beyond simple comparison

- Future:
  - Rotational invariance? -> need no alignment
  - Real data: TPP2 on film

Sorting images into views

Basic problem in single-particle analysis:

- We do not know the projections angles
- Nor do we know the 3D structure of the object

If we knew one, it would be simple...

Approach:

- Harvest the only information we have:
  - The similarity between images
  - The knowledge that somehow these images can be arranged on the observation sphere

Sorting views into angles

- Problem:

  How can we sort the views of a particle according to the viewing angle (elevation, azimuth) ?

- Answer:

  Similar angles ➜ similar images

- Does not require any knowledge about the actual 3D model!
- HOW?

  Parameter estimation in a probabilistic model

Parameter estimation

Problem:

We **do** know the images – why would we care about their probability distribution?

Bayesian parameter estimation:

$$\mathrm{P}(M|\phi) \Leftrightarrow \mathrm{P}(\phi|M)$$

Images

Viewing angles

This is done using **Bayes' formula**

Simplified version: **Maximum-likelihood estimation**

$$\phi = \max \mathrm{P}(M|\phi)$$

The best angular assignments are those which make the images most probable

Self-organizing point map

Joint probability distribution:

$$P(\{M^{(n)}\}|\{\phi^{(n)}\}) = \prod_{i=1}^{N} P(M^{(i)}|\{M^{(i')}, \phi^{(i')}\})$$

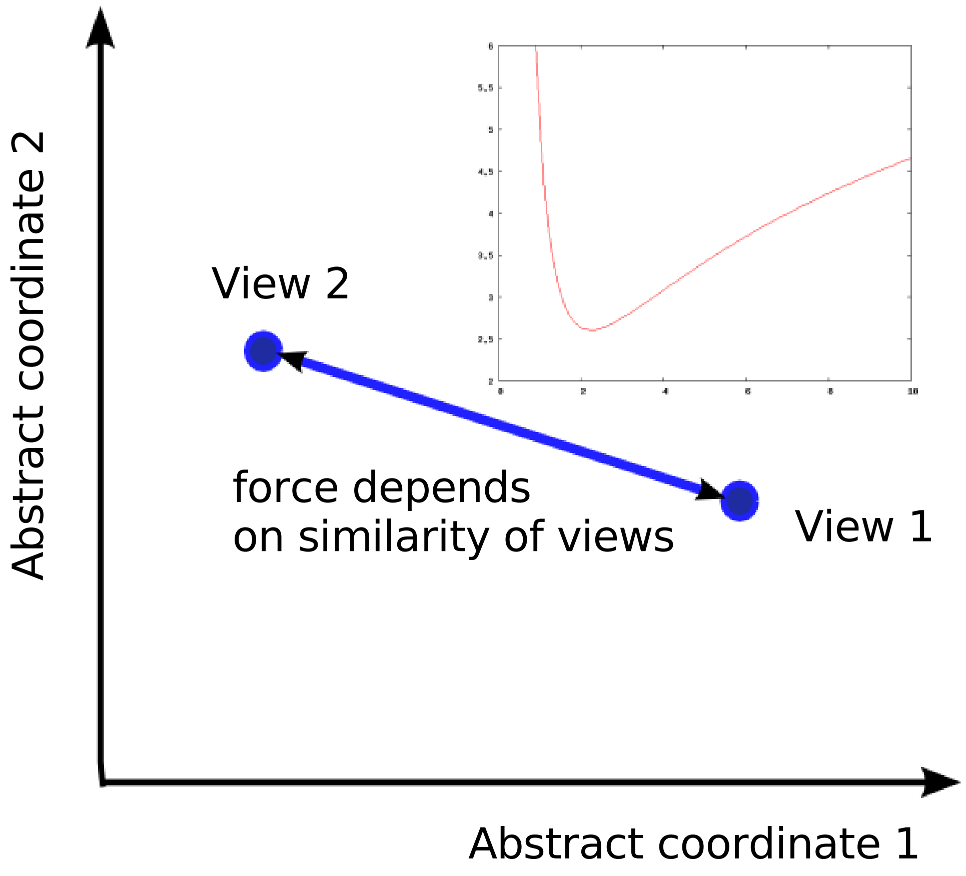
Maximum-likelihood principle ➜ Hamiltonian:

$$\boxed{\begin{aligned} &- \ln L(\phi) = \\ &\quad \sum_{n,m} \left(\underbrace{\frac{D}{2} \ln 2\pi\kappa(|\phi^{(n)} - \phi^{(m)}|)}_{\text{Attractive force}} + \underbrace{\frac{|M^{(n)} - M^{(m)}|^2}{2\kappa(|\phi^{(n)} - \phi^{(m)}|)^2}}_{\text{Repulsive force}} \right) \end{aligned}}$$

Point-to-point potential ➜ multidimensional scaling

Gradient descent solution

Optimization process

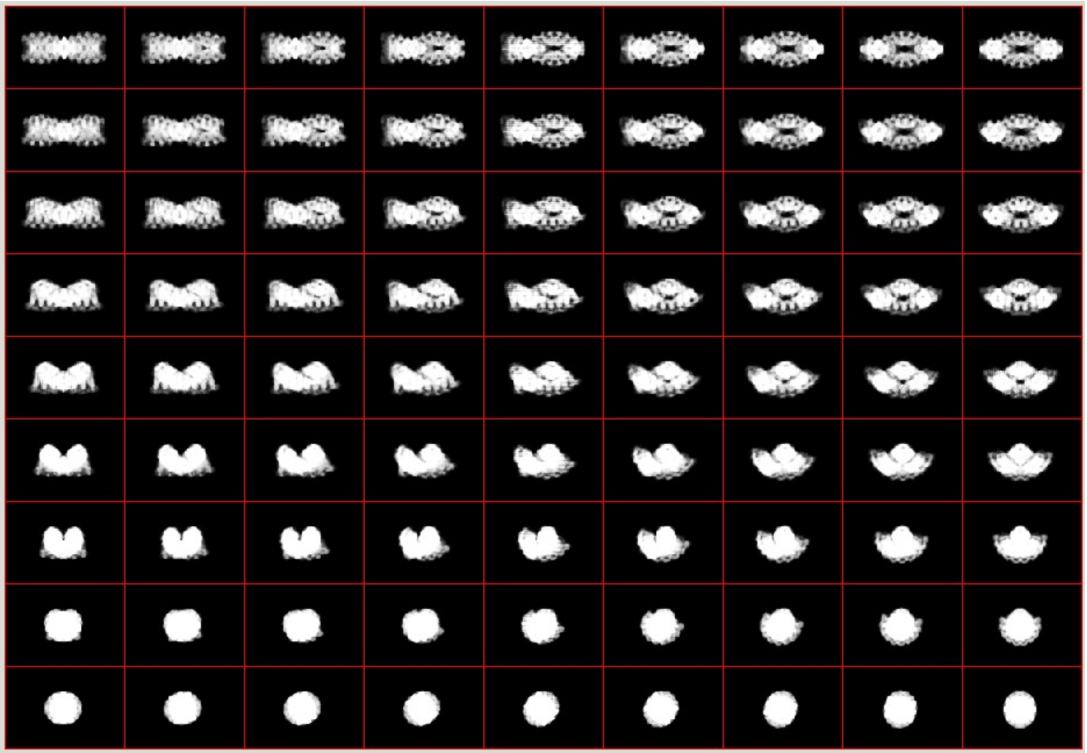


"Spring embedding"

Attractive and repulsive "force" between points (=images)

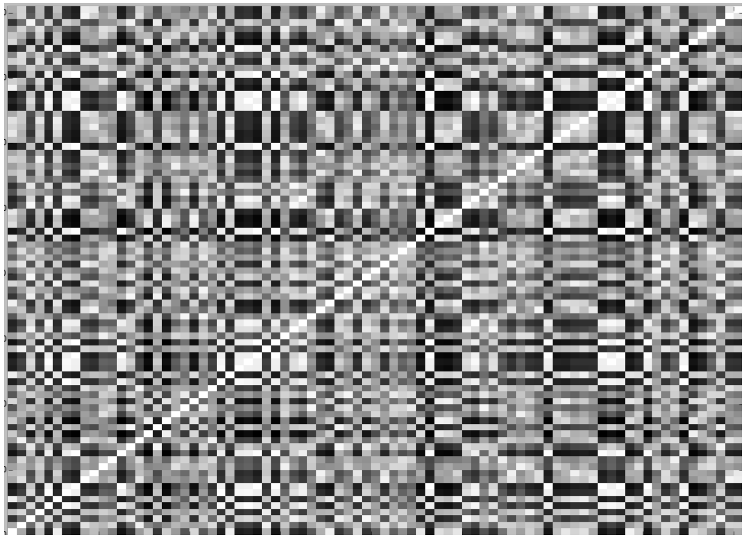
Minimum (=optimum) is determined by how similar the images appear

Similarity matrix



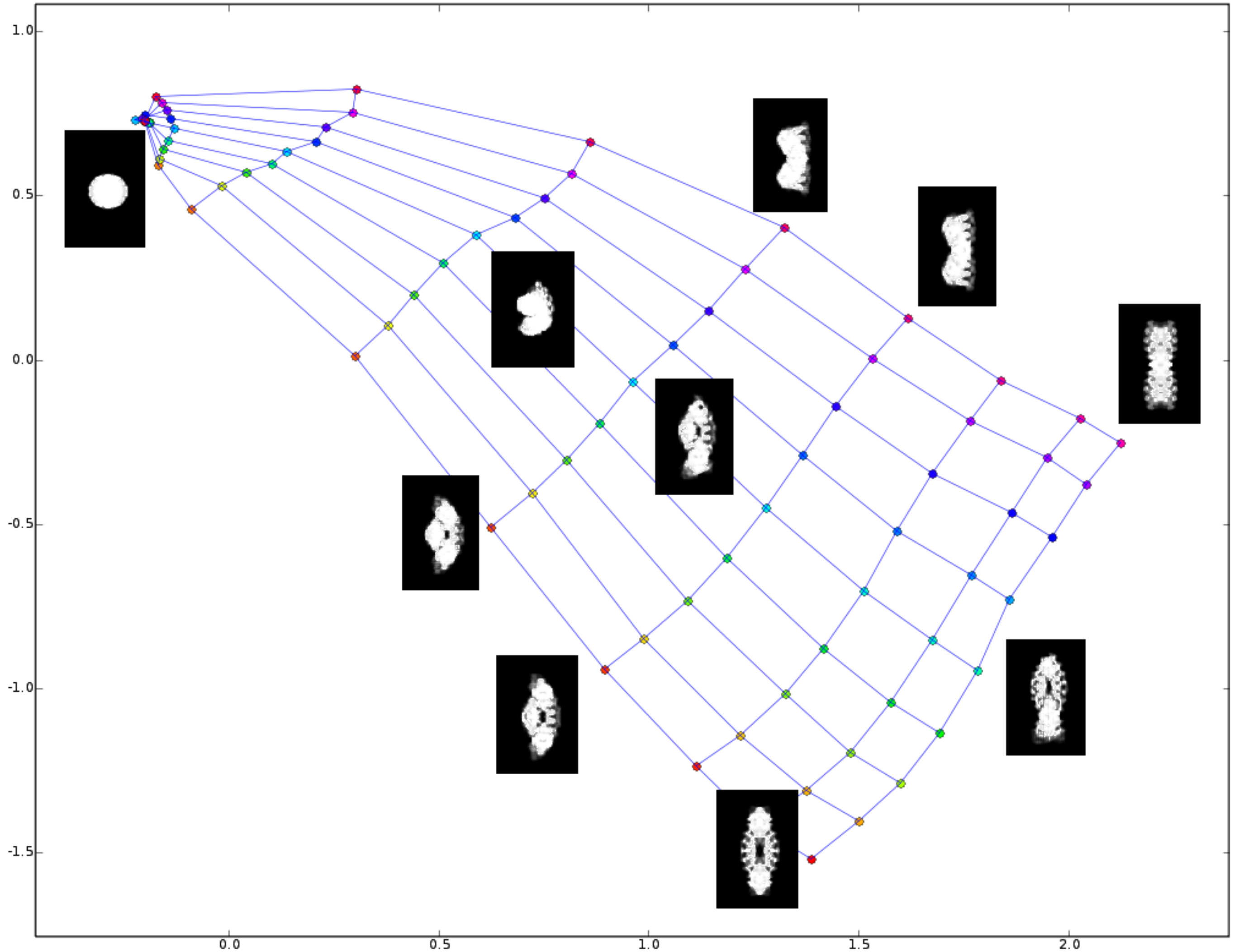
9x9 projections of TPP2

Correlation matrix:



Pairwise correlation max. over translations and rotations

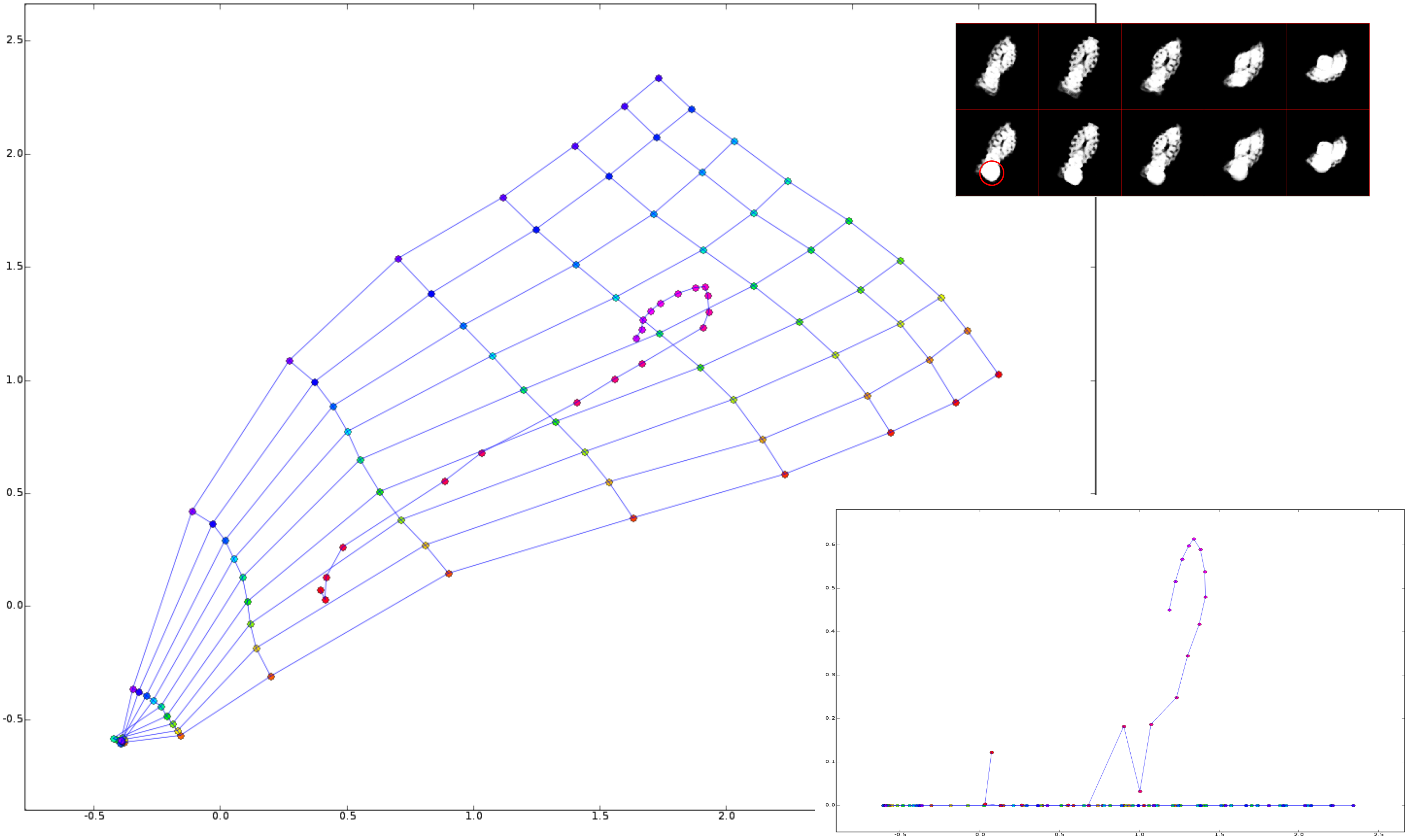
Result



Good representation of original distribution of viewing angles

Good as an initial model for iterative refinement

Tomographic classification



Outlook

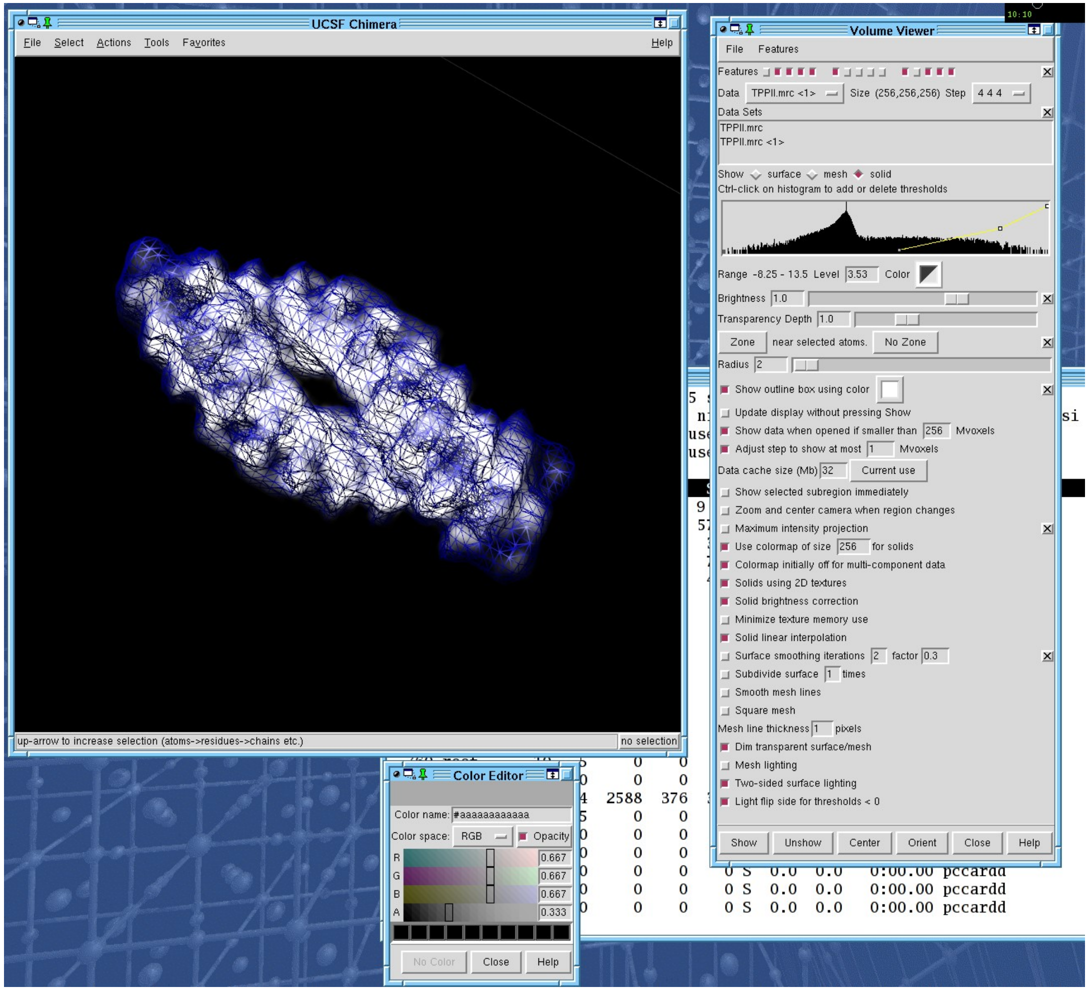
Unified view of single-particle analysis:

- ▷ Simultaneous estimation of two unknowns:
  - ▹ View angles
  - ▹ 3D density distribution / shape
- ▷ Include additional information:
  - ▹ Random conical tilt
  - ▹ Tomography
  - ▹ Anything in between
- ▷ Combine with advanced shape reconstruction (2^{nd} part of talk)

Graphical interlude

Visualizing tomograms using UCSF Chimera

- UCSF Chimera
  - Visualization platform for structural biology
  - NIH Resource for Biocomputing, Visualization, and Informatics
  - Mature platform, >10 yrs
  - Python/C++ ➔ easily extensible



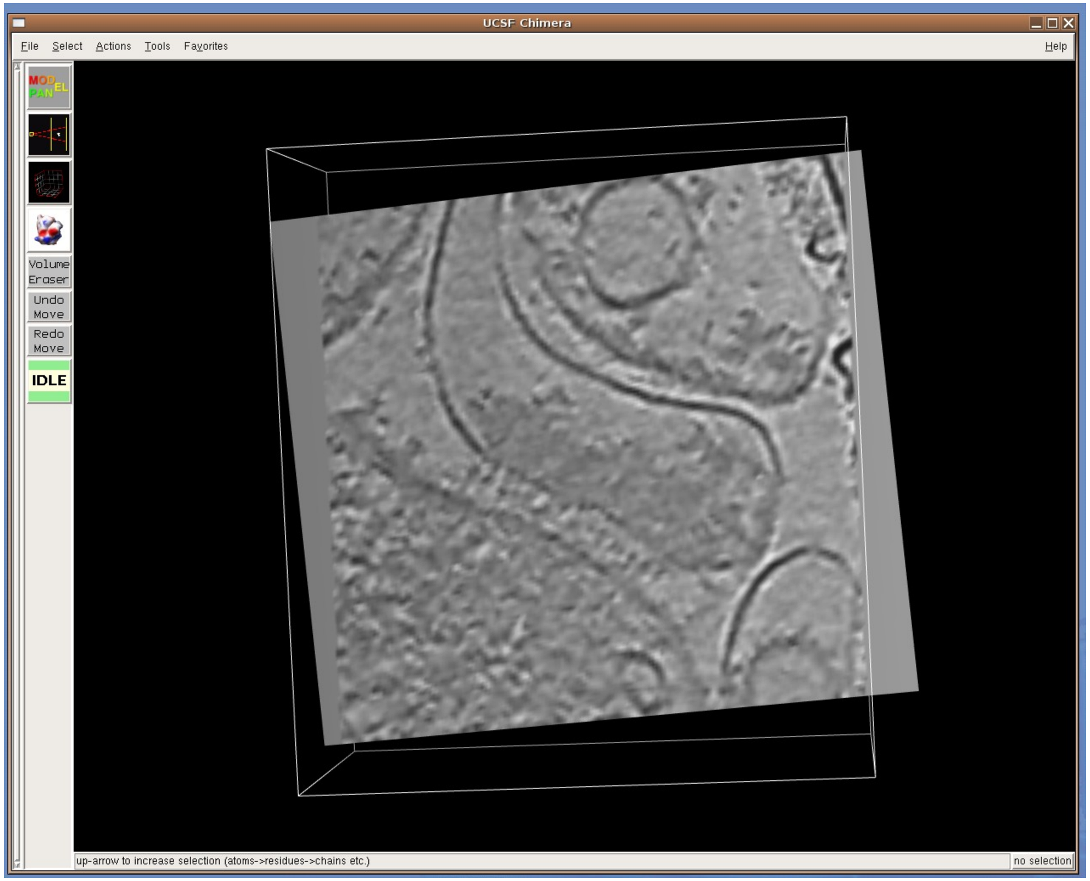
Tomography & Chimera

Tomography tools for Chimera

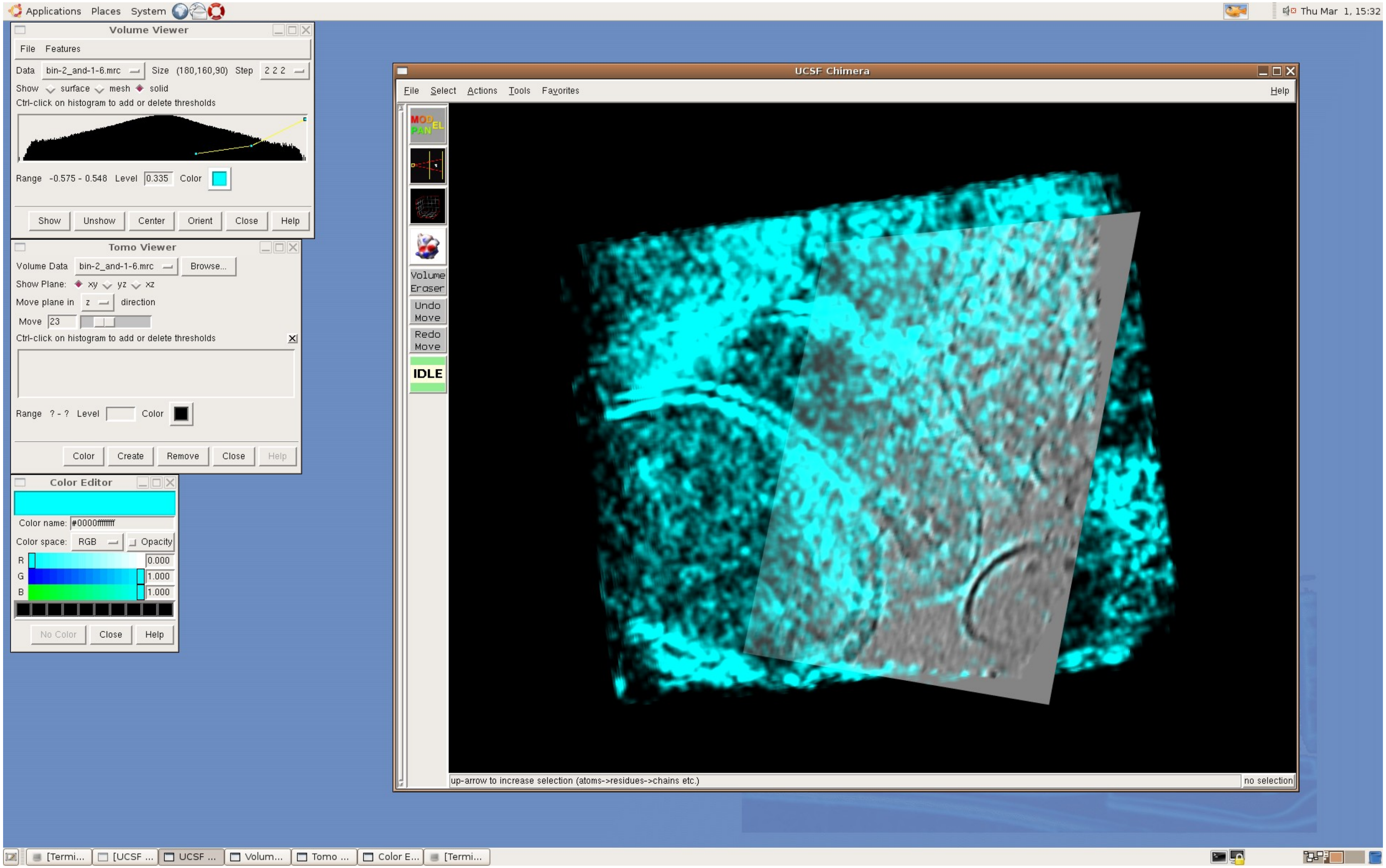
Master's thesis Karin Gross
(FH Weihenstephan/Bioinformatik)

- ▷ Add slicing capabilities to Chimera
- ▷ User interface for adjusting viewing parameters
- ▷ Oblique sections

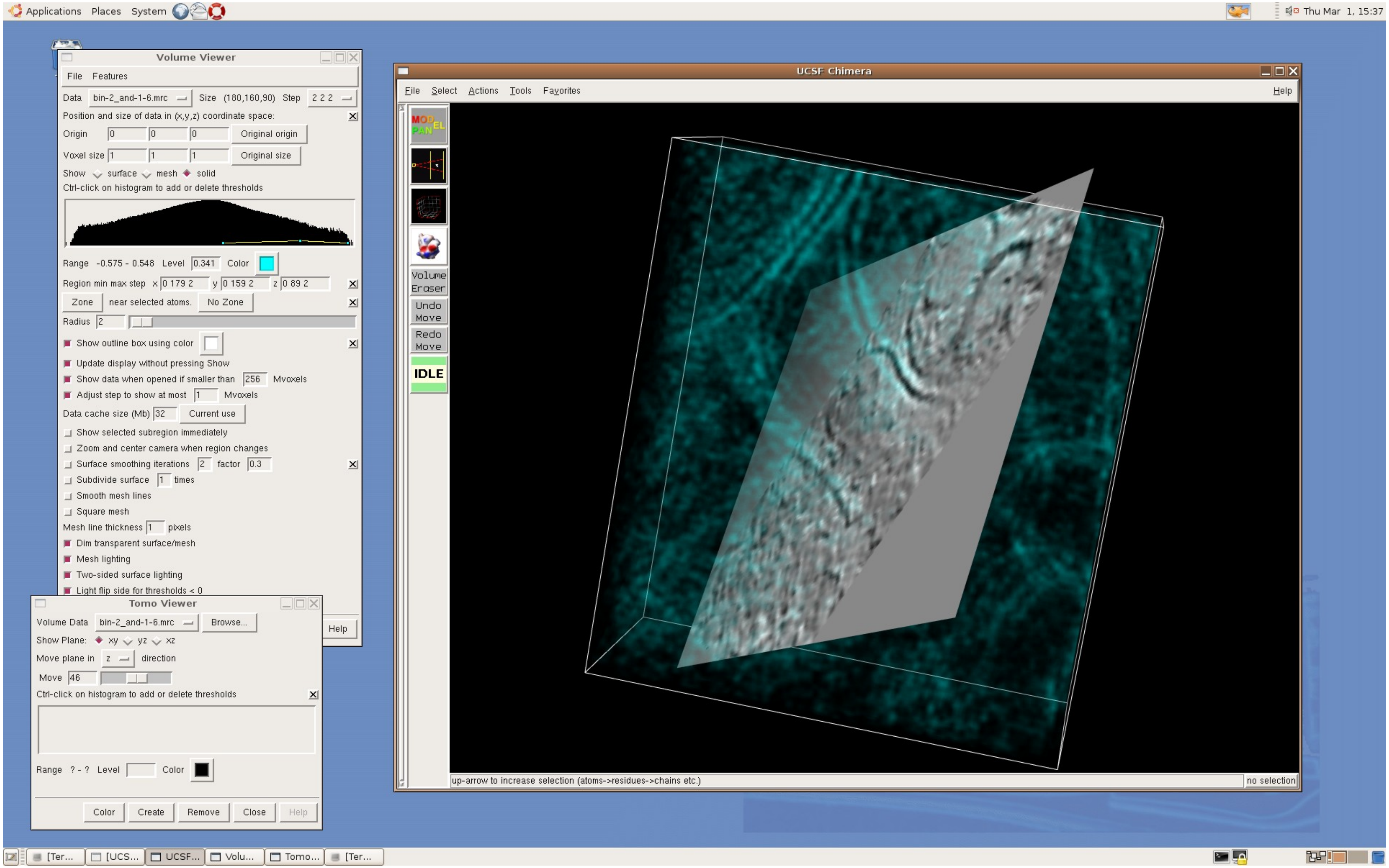
- ▷ Markers and segmentation on slices
- ▷ Integrate semi automatic segmentation (level set)
- ▷ Tools for 2D/3D rendering of label sets



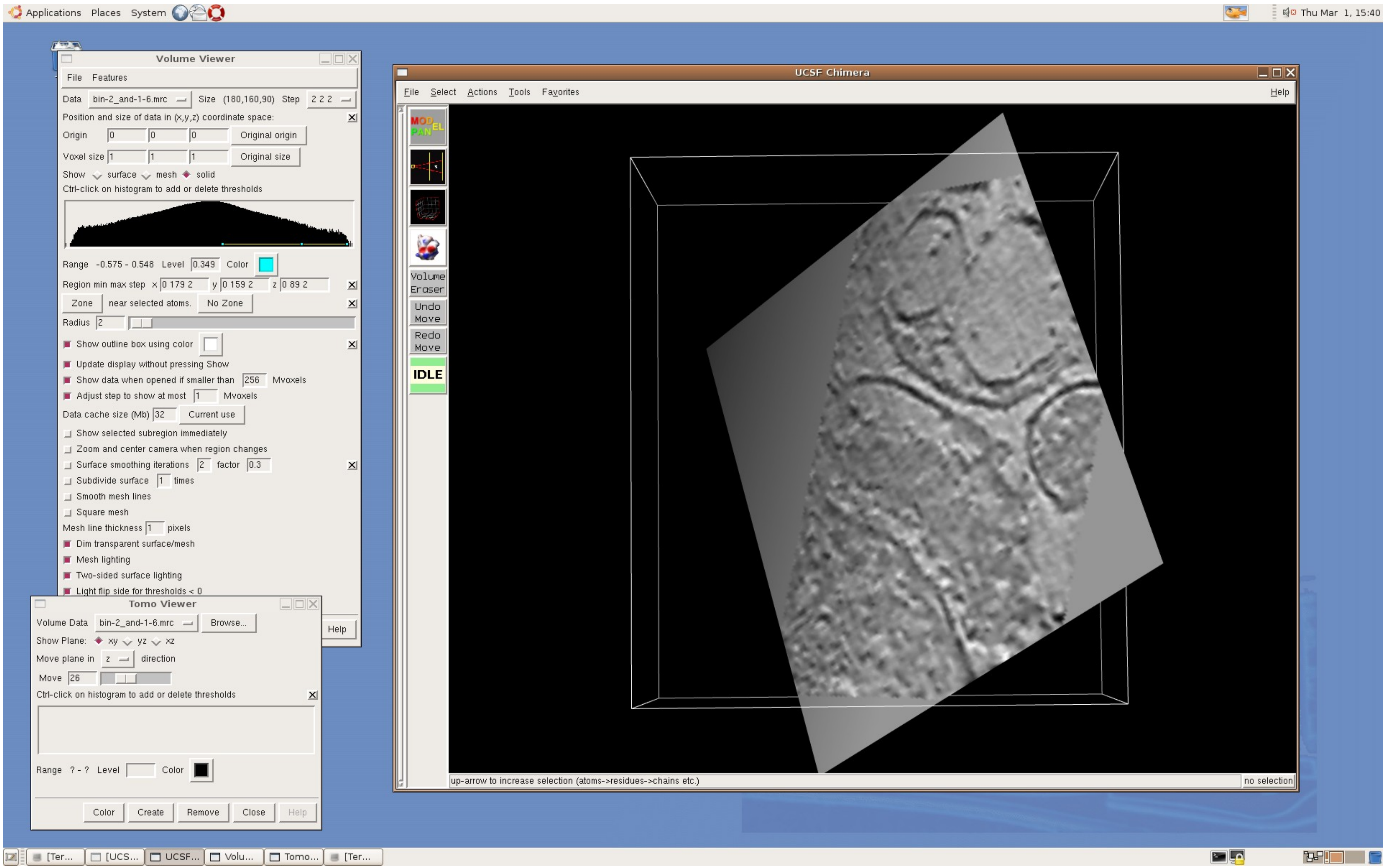
Tomography & Chimera II



Tomography & Chimera III



Tomography&Chimera IV



Two Exercises in Tomography

Two exercises in tomography:

- ▷ 3D reconstruction using level sets
- ▷ Automatic triangulation of marker positions

General goals:

- ▷ Less projections
- ▷ More resolution

3D reconstruction using level sets

"And now for something completely different..."

<u>Tomographic reconstruction:</u>

- Reconstruct densities
- No further information about which densities and structures are to be expected

<u>Level-set tomographic reconstruction:</u>

- Reconstruct shapes
- Makes use of additional information:
    - Object has a surface (with certain properties e.g. curvature)
    - Object has constant density
    - Object is connected

3D reconstruction using level sets II

Expected advantages:

- More robust against missing wedge
- Works well when only a few projections are given
    - e.g. Initial guesses for single particle reconstruction
- Combines segmentation and 3D reconstruction in one process

Level set methods

IDEA: Form a shape until its set of projections fits the observed projection

PROBLEM:
How do you manipulate a geometric shape in a computer?

Classical solution: triangulation of surfaces

- Good for visualization, but difficult to manipulate

Level set methods:

- Describe surfaces as the zero levels of continuous scalar functions
- Control surface shape by differential equation governing the scalar field
- Computationally expensive, but simpler and more

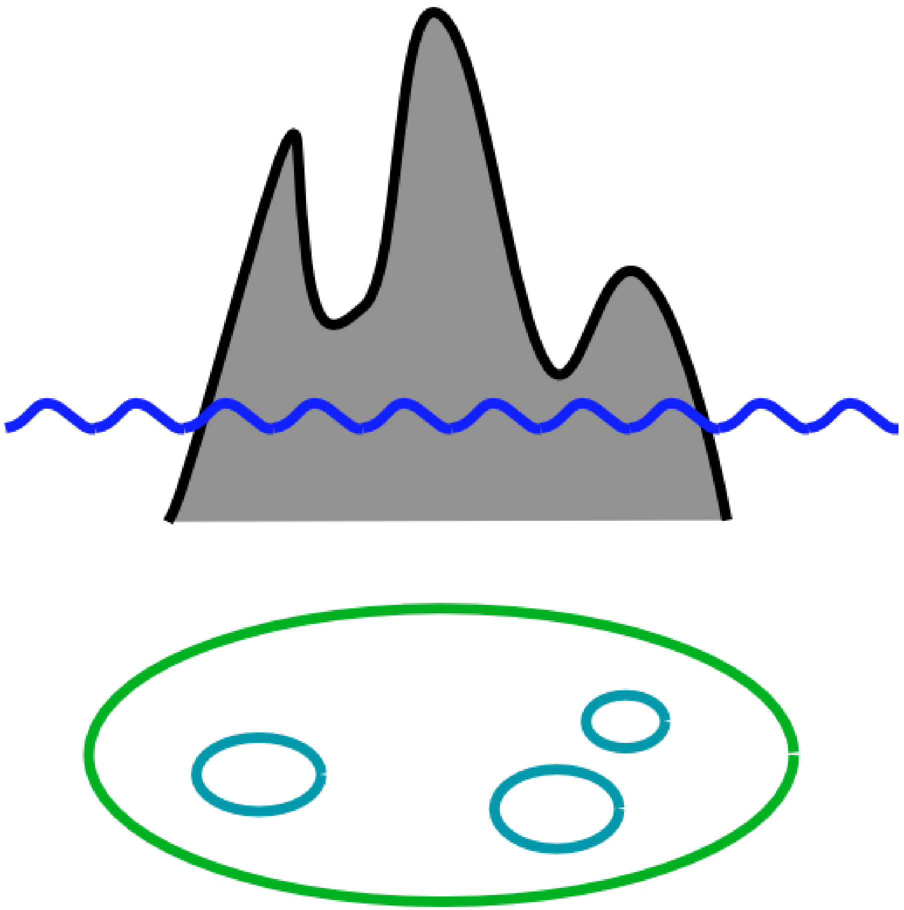
Level set reconstruction II

So what does this mean?

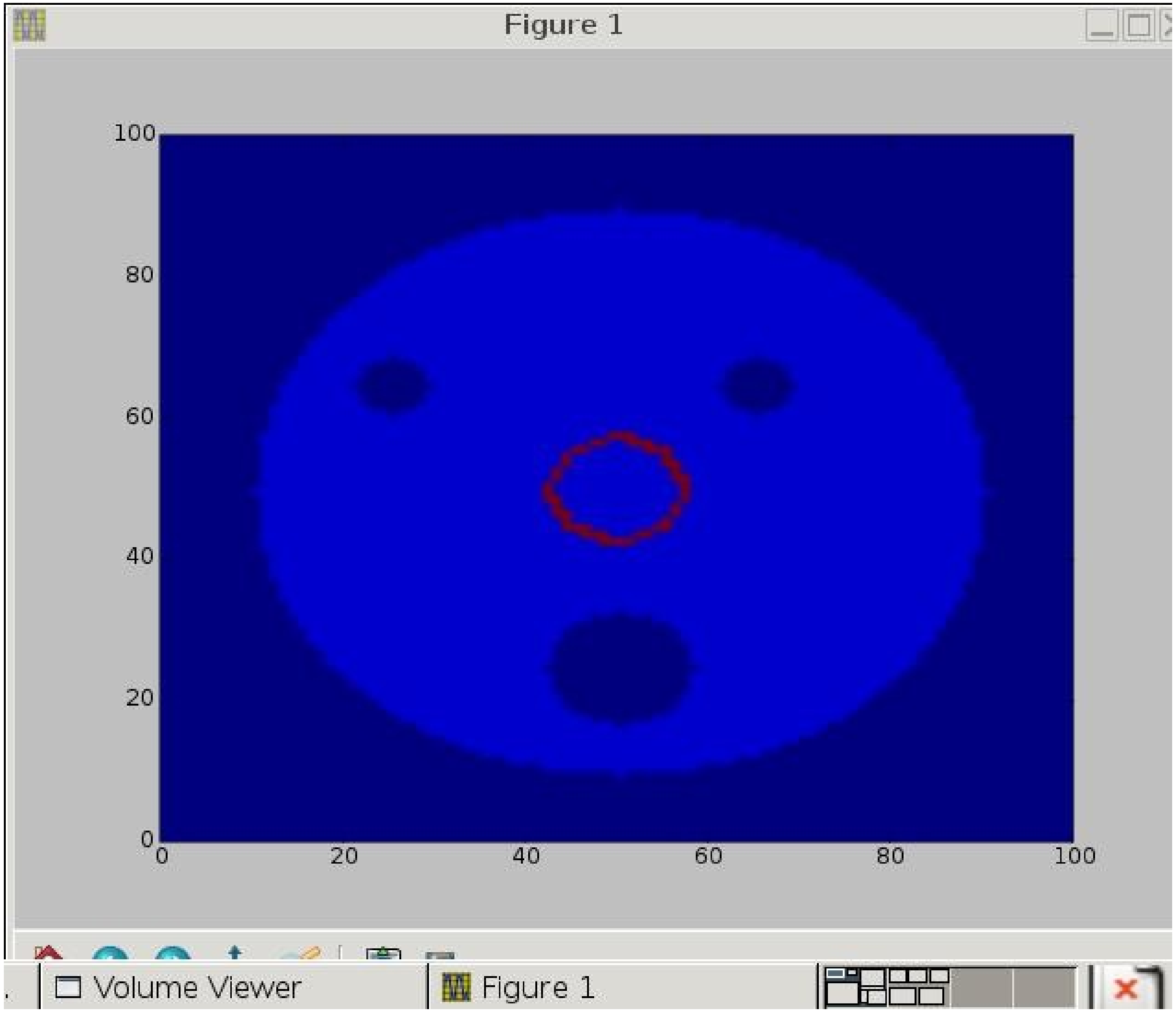
Imagine a hilly island in the sea

Different water levels generate different coastlines

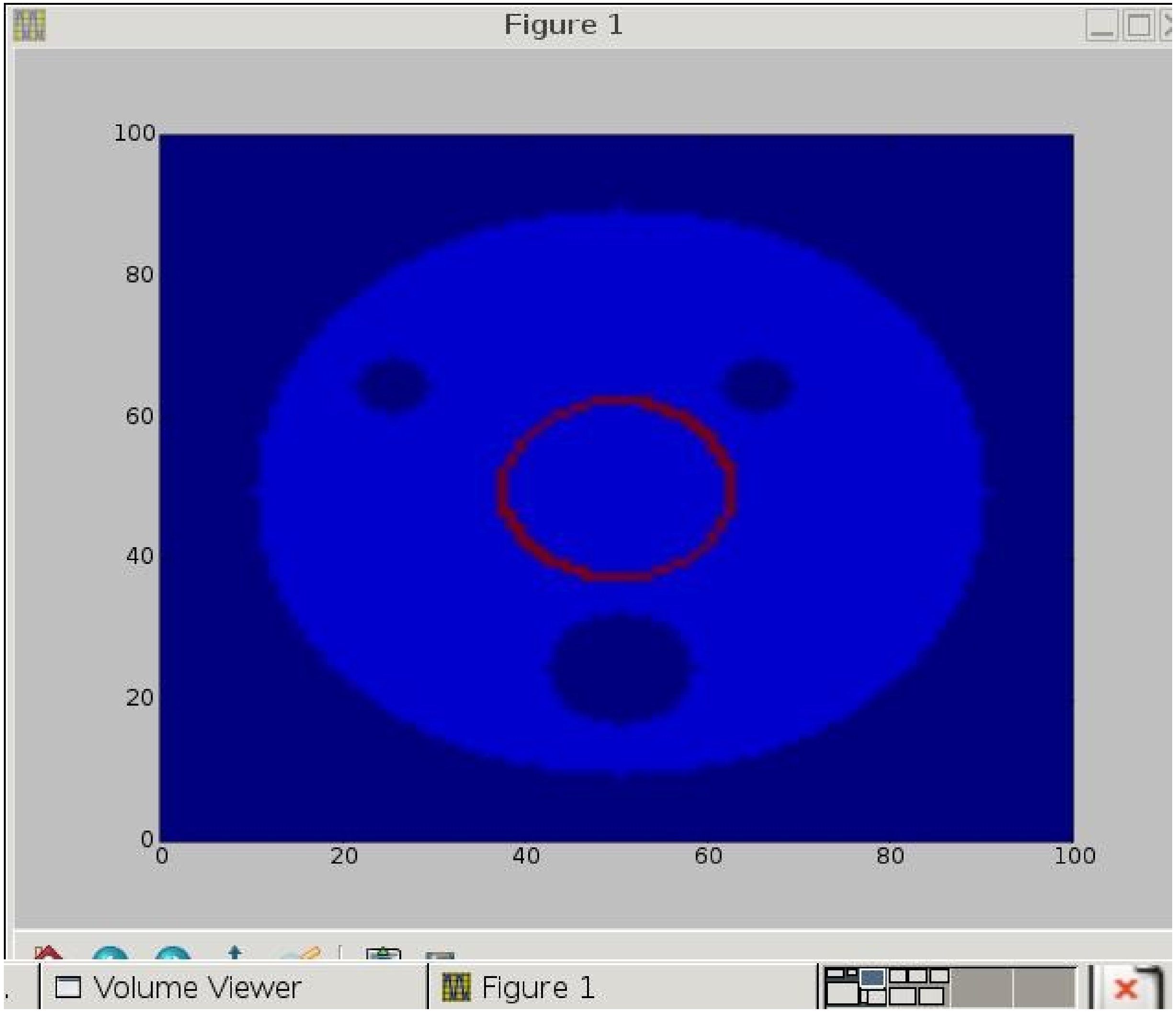
Continuous deformations of the island can generate very complex geometric changes, e.g. dissociation



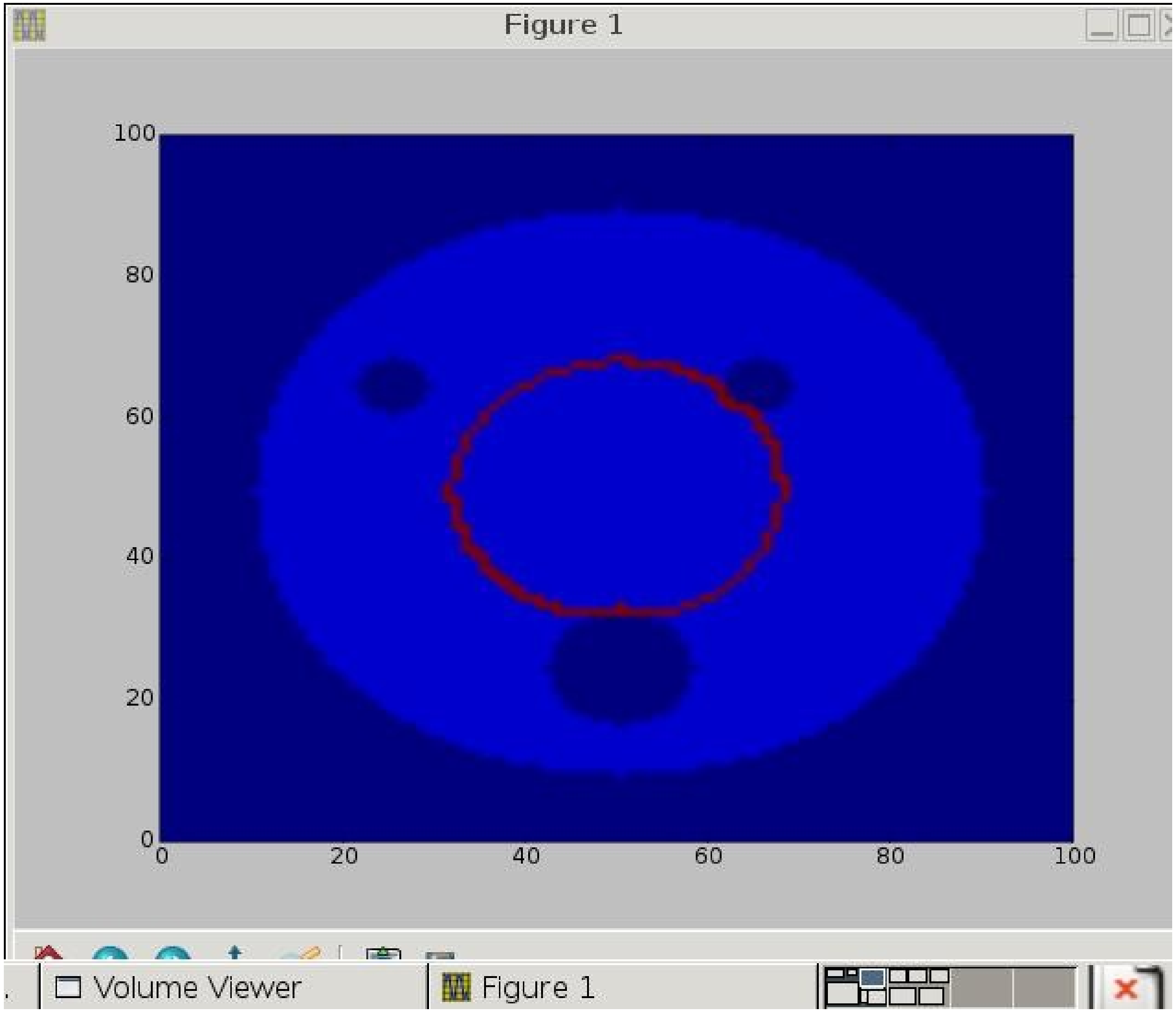
Level set reconstruction



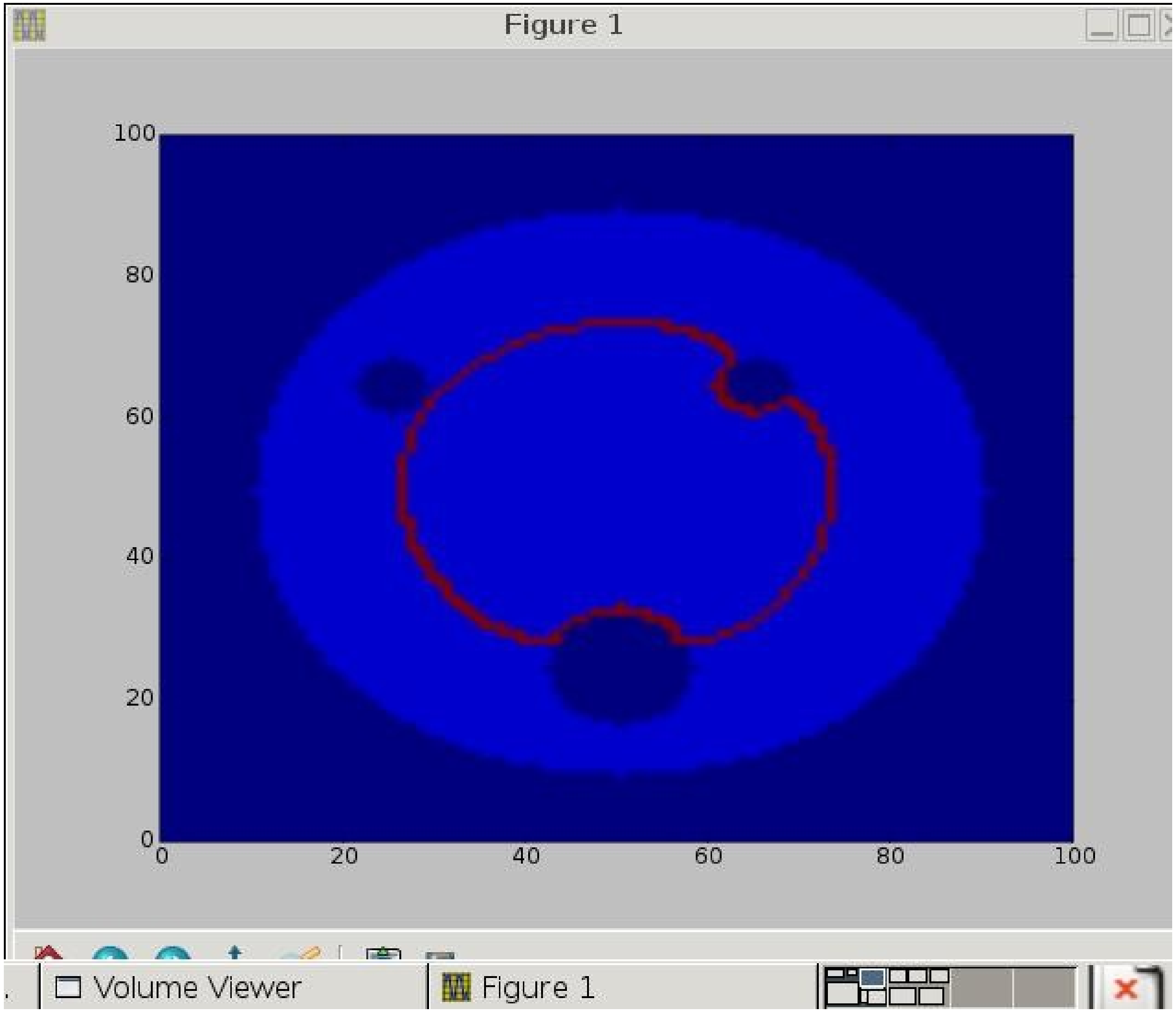
Level set reconstruction



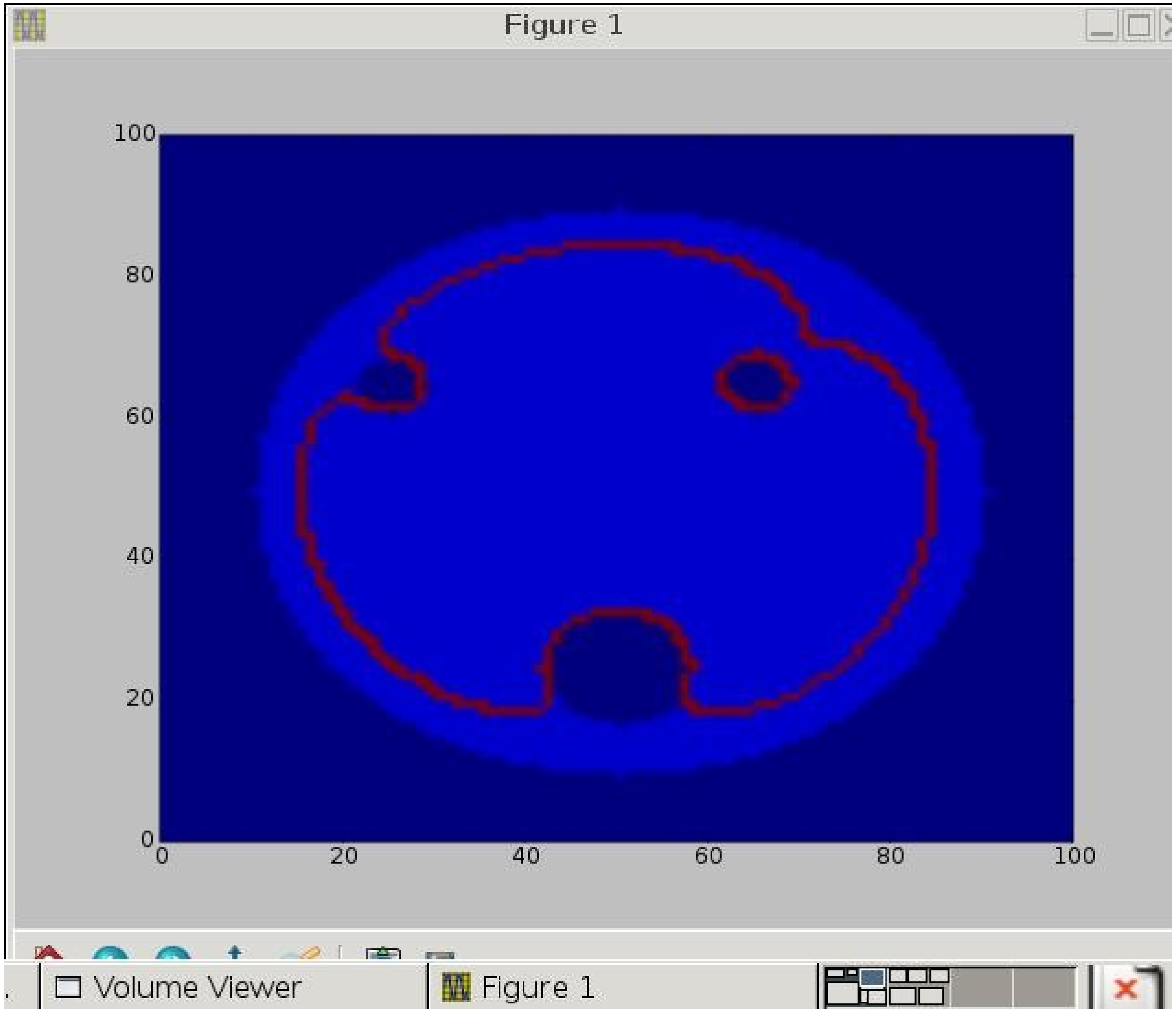
Level set reconstruction



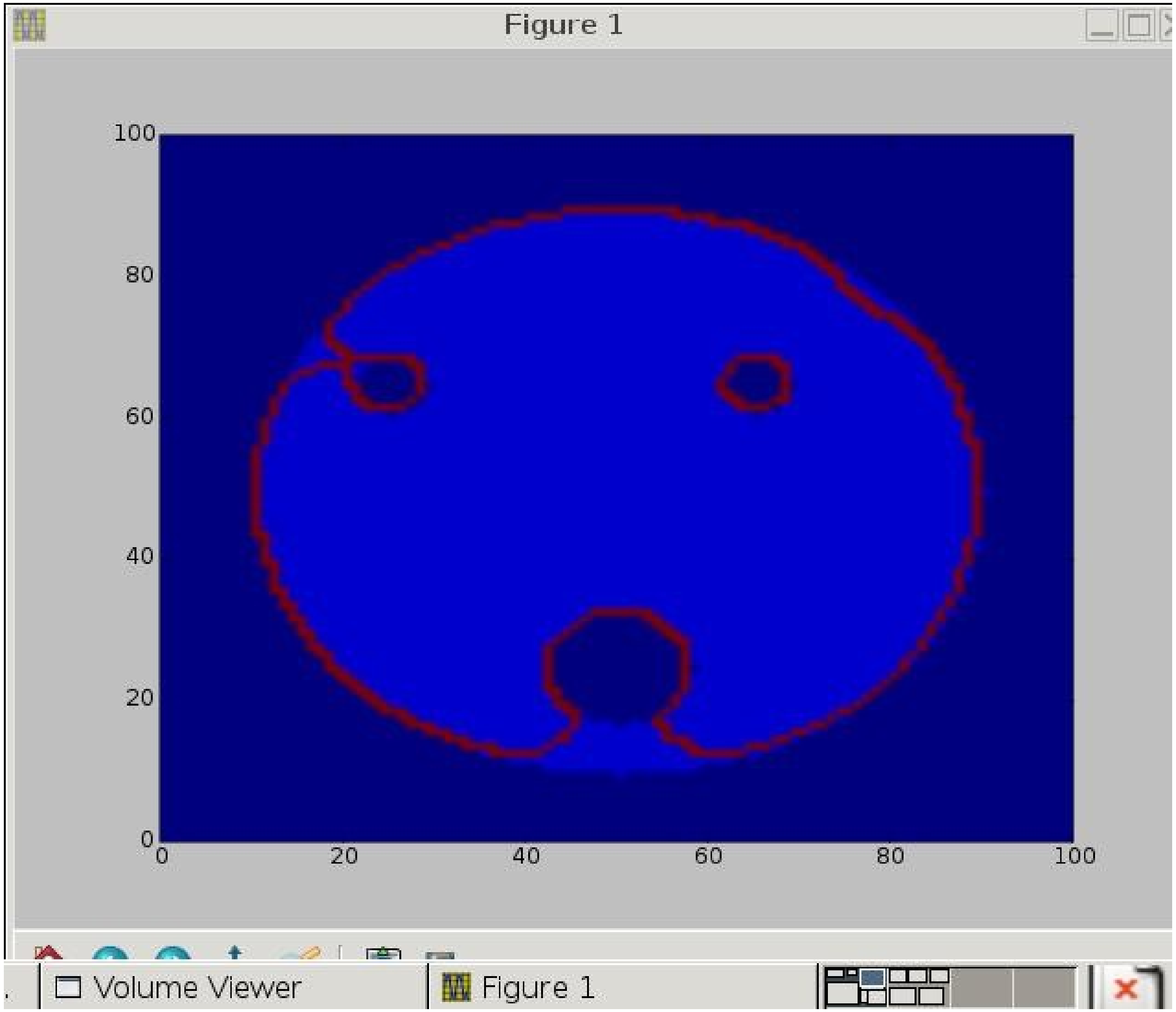
Level set reconstruction



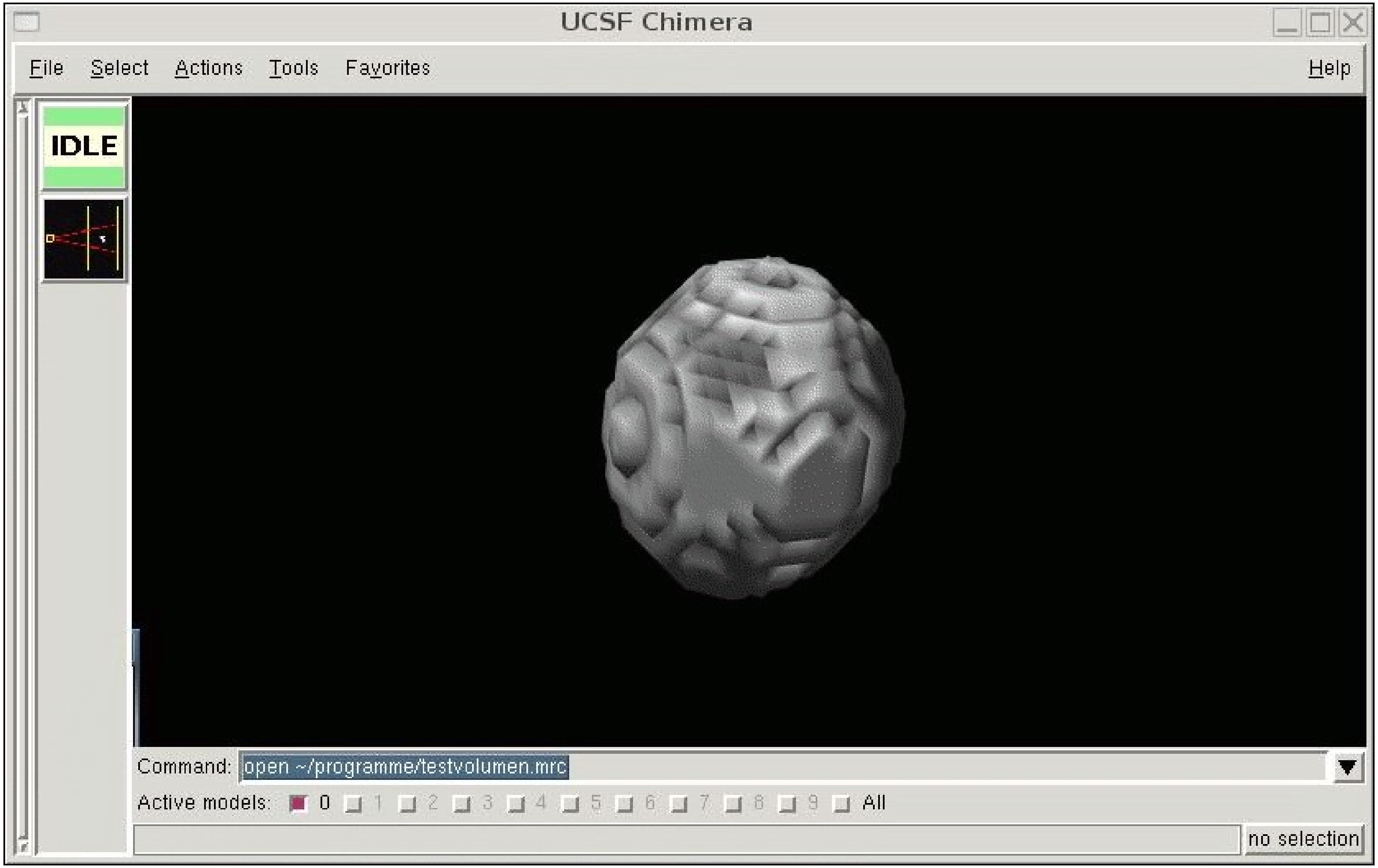
Level set reconstruction



Level set reconstruction



Manipulating 3D shapes



Reconstruction Algorithm

Algorithm:

Repeat:
- Calculate model projection
- Compare to observation
- Expand/contract model accordingly

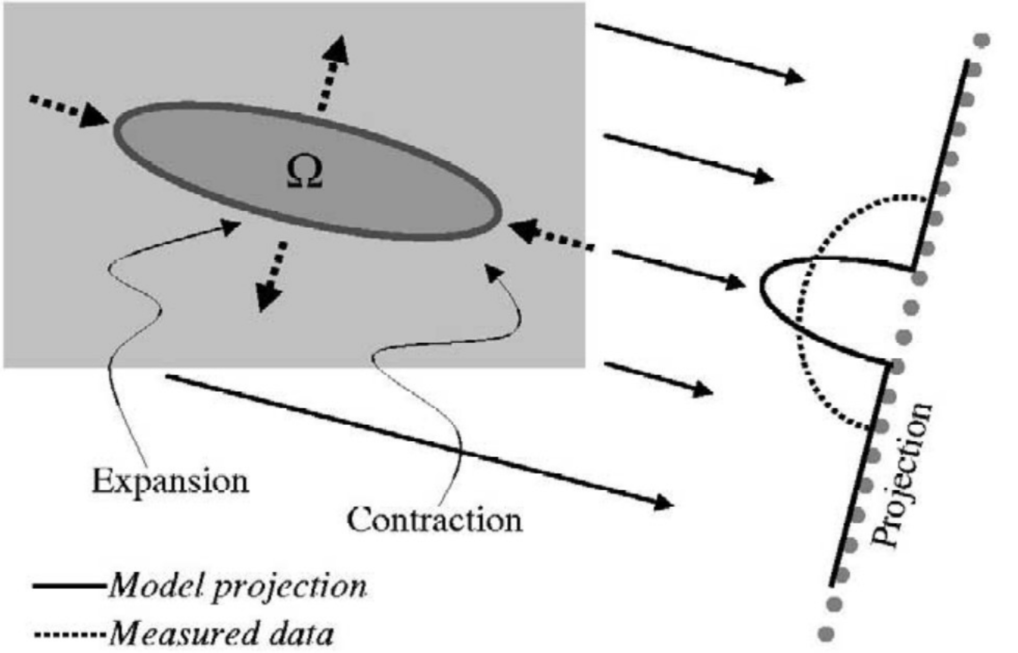


Fig. 3. The model expands or contracts based on the difference in the sinograms between the projected model and the measured data.

Ross T. Whitaker, Vidya Elangovan*, *A direct approach to estimating surfaces in tomographic data*, Medical Image Analysis 6 (2002) 235–249

Reconstruction algorithm

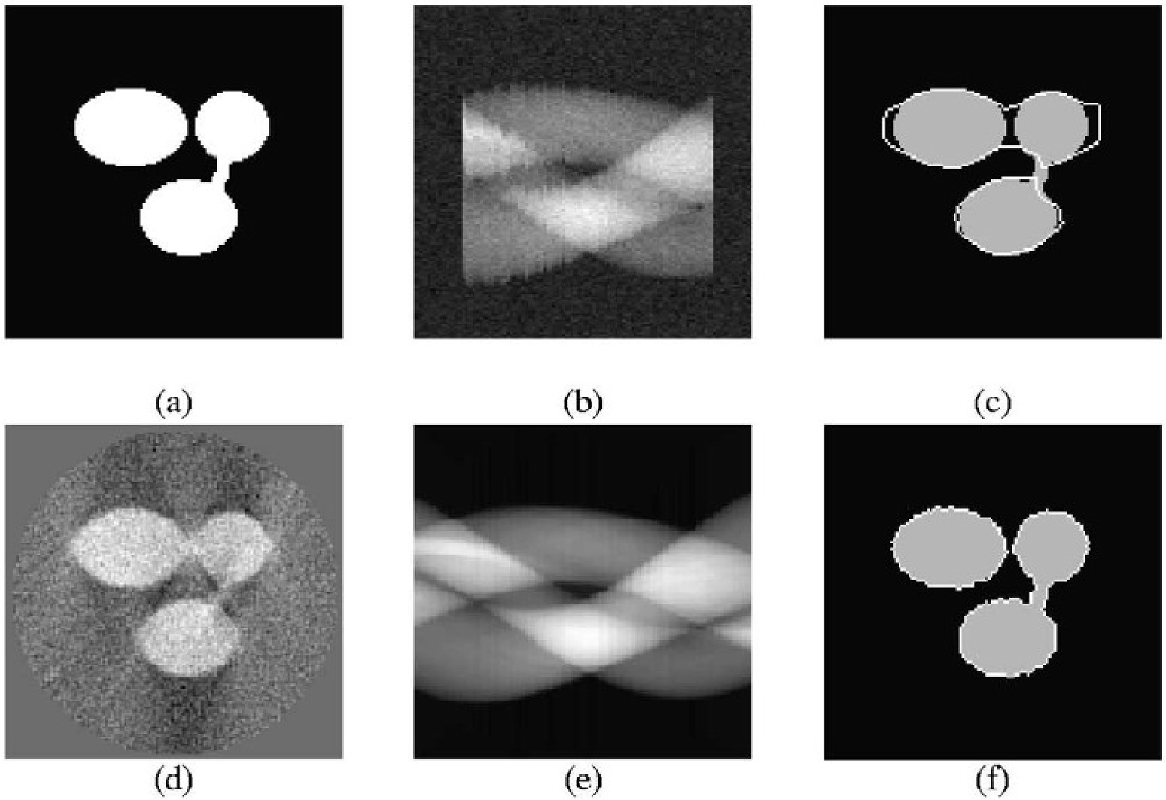
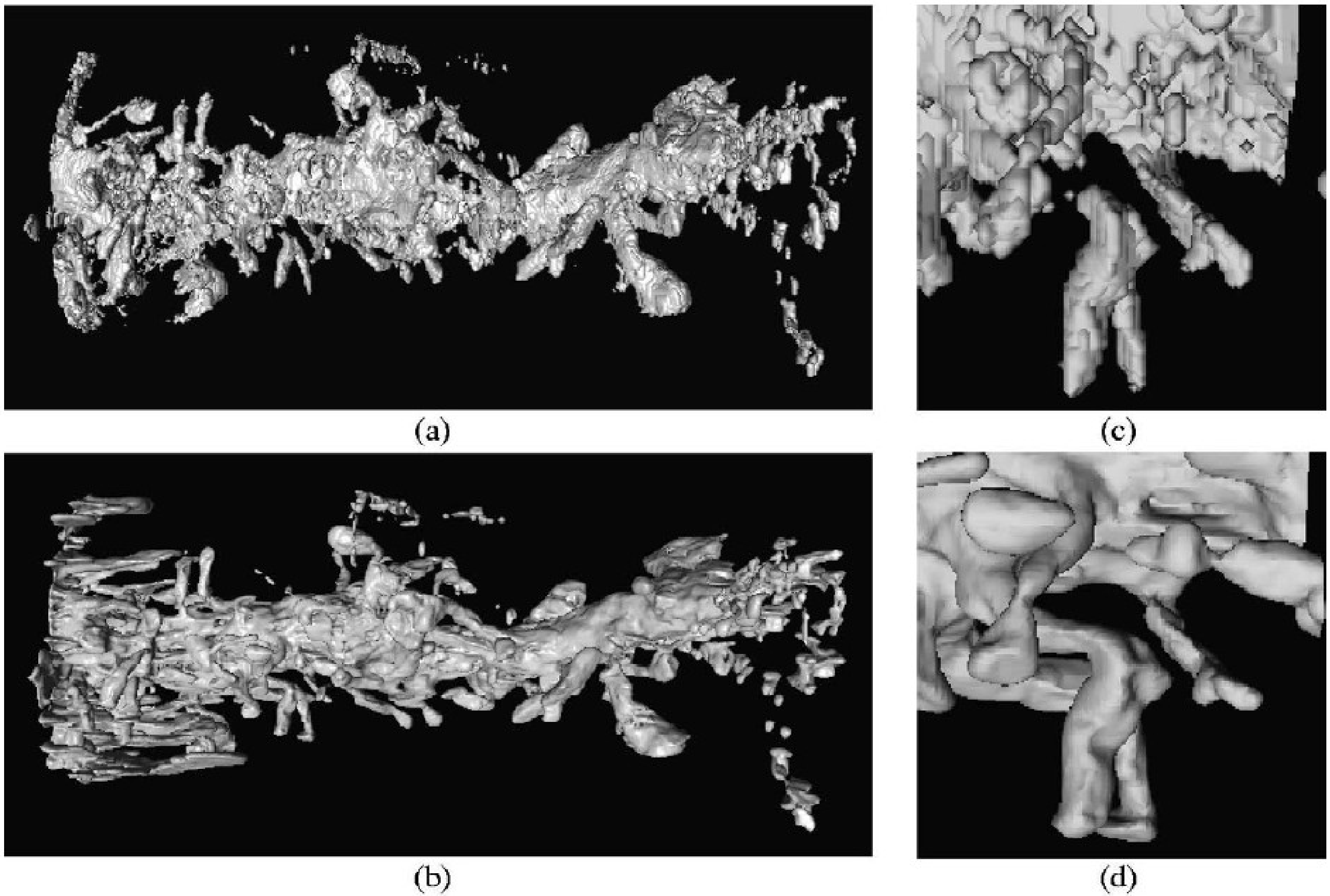


Fig. 8. Results of a 2D simulation. (a) Digitally simulated input image. (b) Limited-angle, noisy, misaligned sinogram created by projecting the input image. (c) Initial model obtained by thresholding the back projection. (d) Back projection showing artifacts. (e) Sinogram estimated by the proposed method. (f) Final model showing the correct segmentation of the input (note: the initial and final models are white contours overlaid on the input data).

Ross T. Whitaker, Vidya Elangovan*, *A direct approach to estimating surfaces in tomographic data*, Medical Image Analysis 6 (2002) 235–249

Expected results



Ross T. Whitaker, Vidya Elangovan*, *A direct approach to estimating surfaces in tomographic data*, Medical Image Analysis 6 (2002) 235–249

Current status

Diploma thesis in bioinformatics of Andreas Grimm

Cooperation with Ralf Zimmer (LMU)

- Implemented 2D and 3D algorithm
- Simple structures can be reconstructured using <10 projections
- Little distortion from missing wedge

Outlook:

- Combine with single-particle algorithm
- Unify tomography and single-particle shape reconstruction

3D reconstruction by triangulation

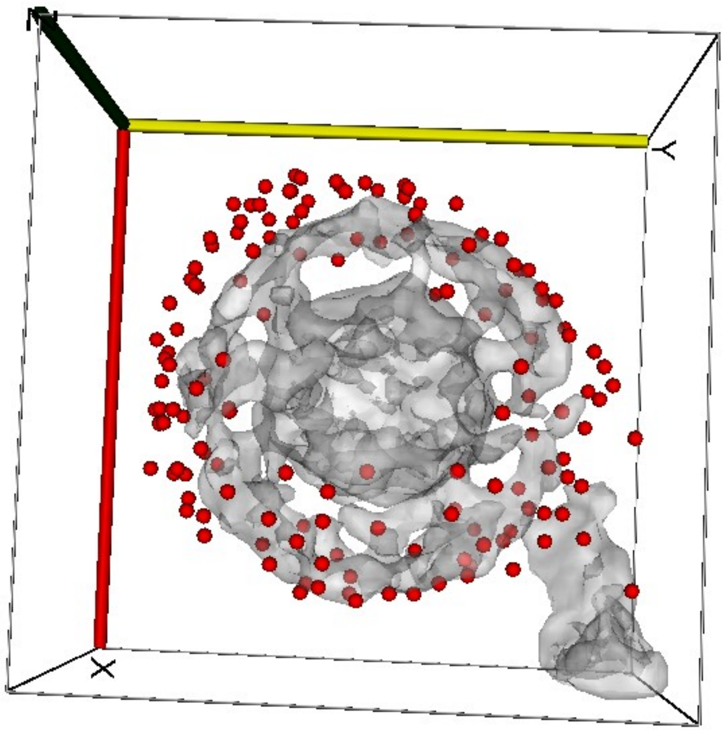
"And now for something completely different"

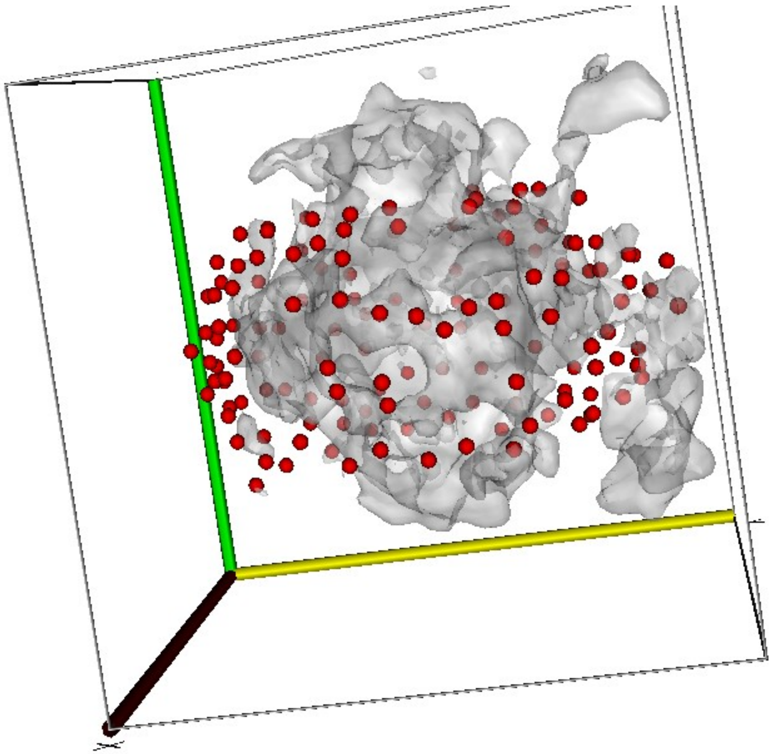
3D reconstruction of marker positions
- ▹ Locate markers automatically in individual projections
- ▹ Calculate 3D positions by triangulation

Advantages:
- ▹ Need only a few projections
- ▹ Less chance for obscurations

Applications:
- ▹ K. Grünewald: Locating immunogold markers for glycoprotein spikes
- ▹ V. Lucic: Estimating alignment errors in tomograms

Automatic localization of markers





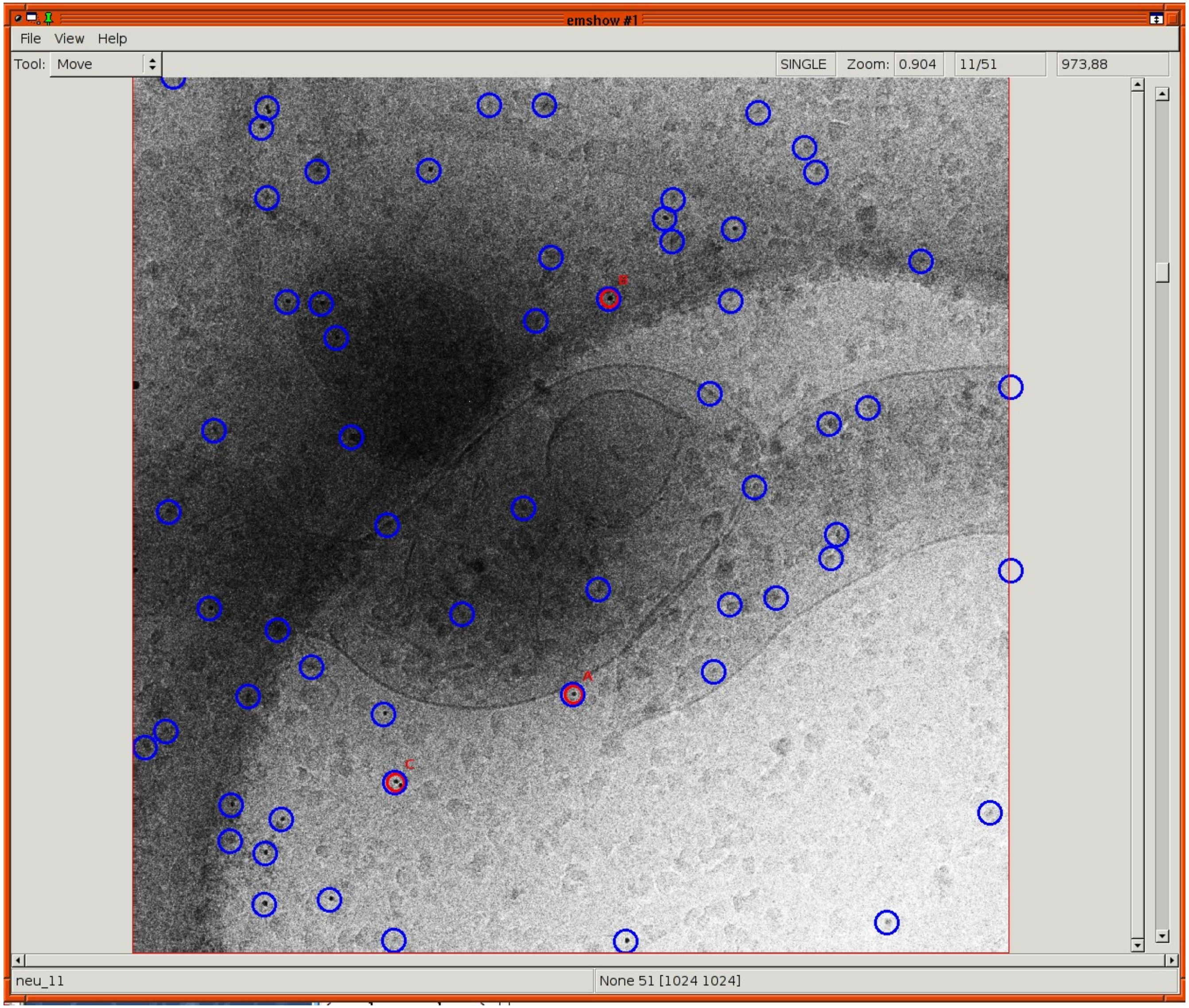
Automatic localization of glycoprotein spikes by immuno-labeled gold
Problem: Markers at the top and bottom are obscured by the capsid

Automatic localization of markers

Example tomogram (V. Lucic)

Markers located automatically and aligned using three manually picked markers

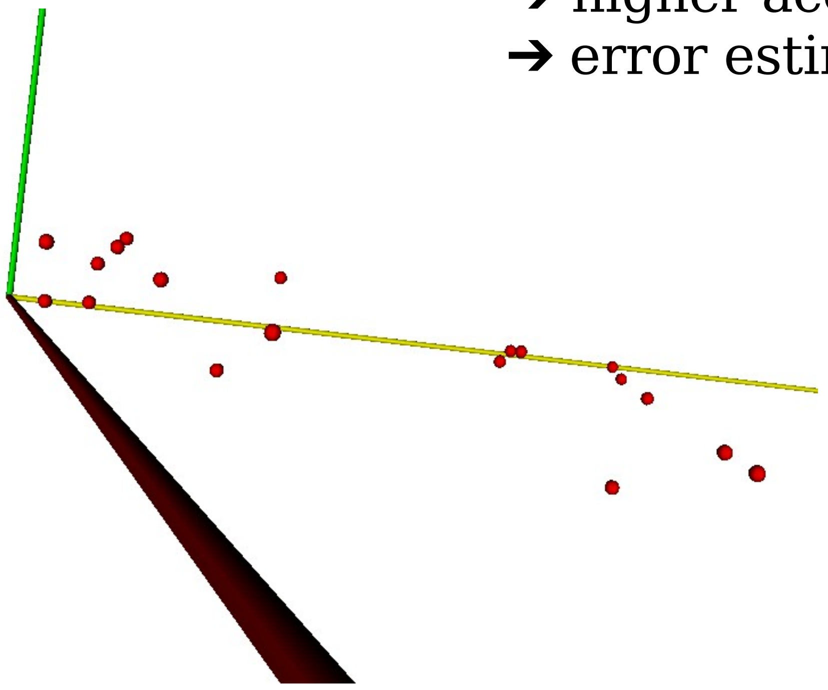
Re-alignment using 18 markers that are visible in most projections



3D localization

Triangulation: Z-position of markers can be derived by combining at least two views

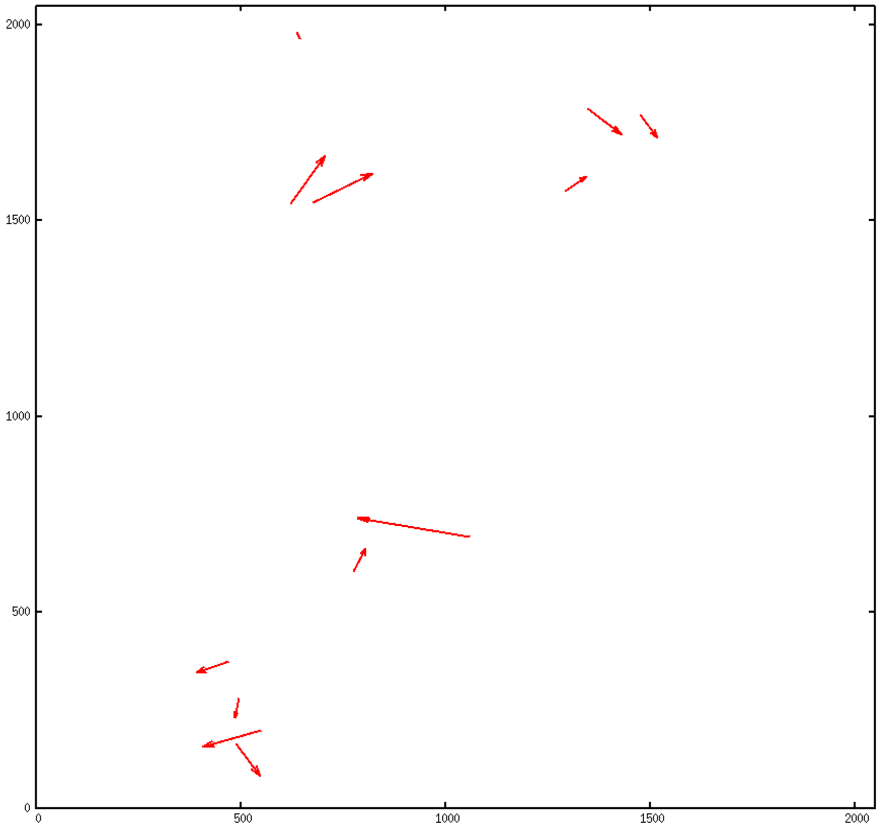
Having more views:
➔ higher accuracy
➔ error estimates



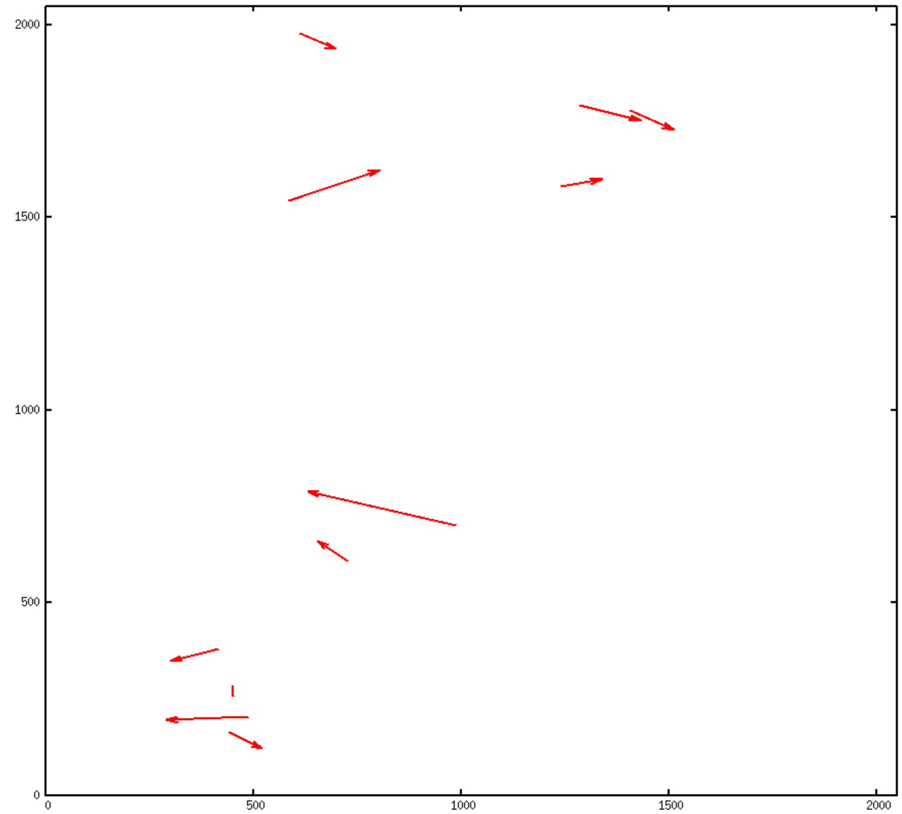
Patterns of marker motion

Displacements of automarkers against expected position

In projection 3

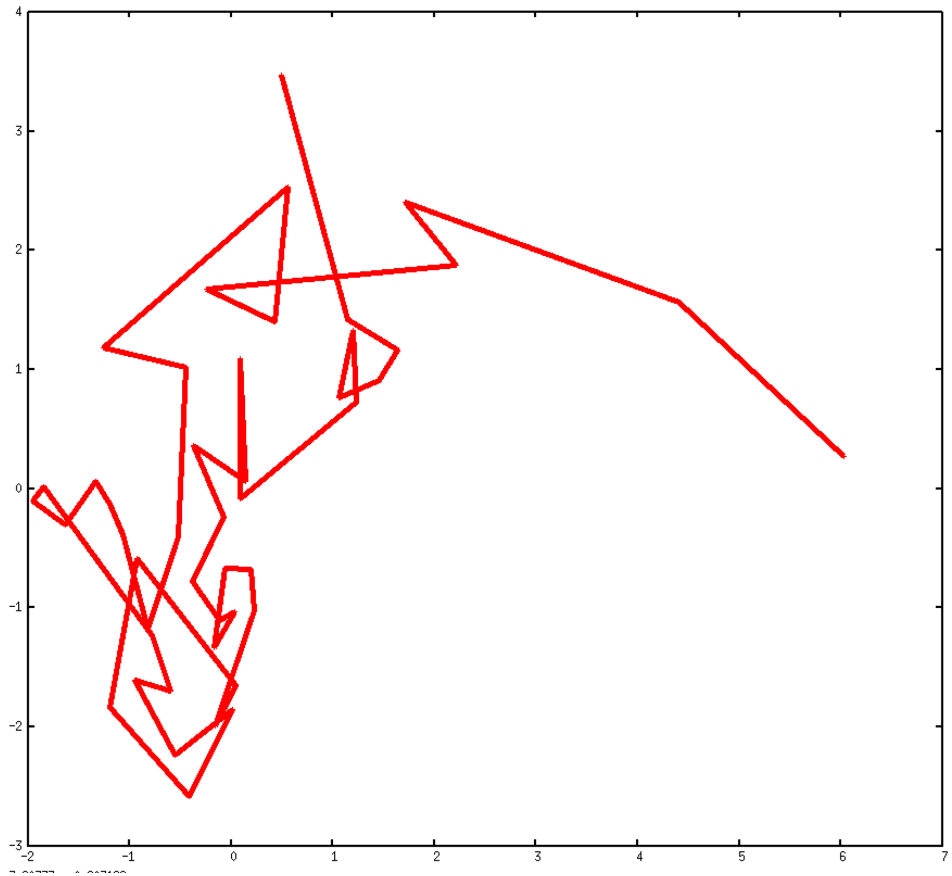


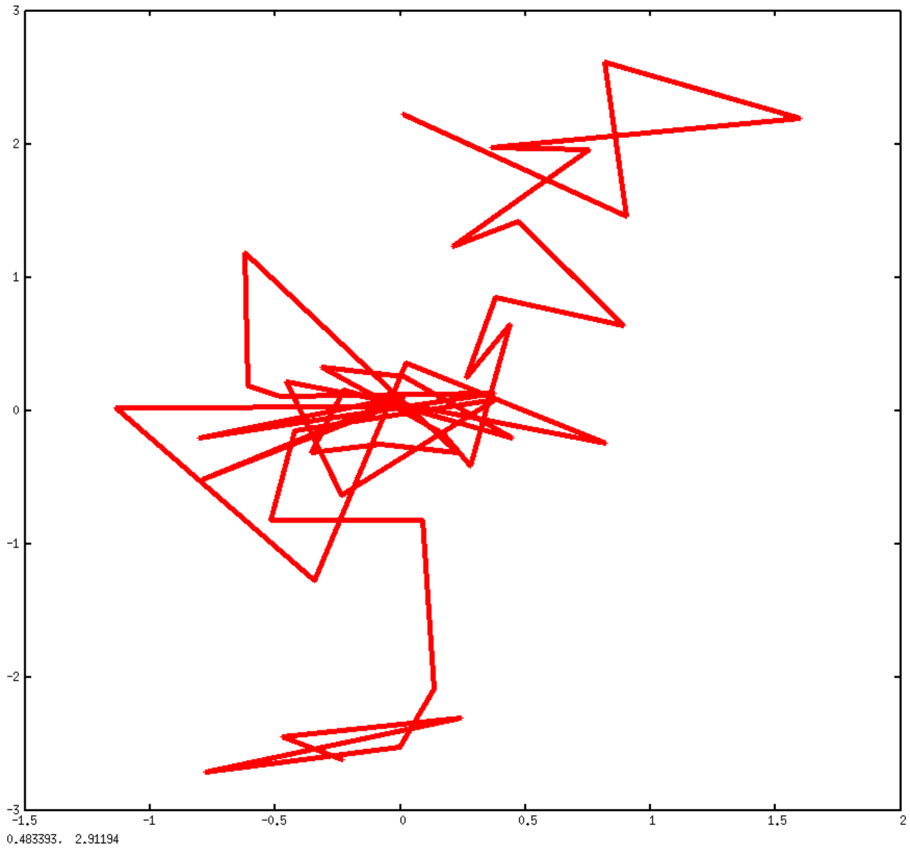
In projection 4



Apparent motion of markers

Motion of marker 17 and 11 over all projections
Evidence for systematic deviations





Future steps

- Fully automatic acquisition and alignment of marker points
- Acquisition of other prominent features
- Correction of distortions

- Analysis of virus glycoprotein spike distributions
- With small number of projection views
- Many images automatically analyzed -> statistically significant statements about marker distributions

Outlook

Central topic: Integrate disparate approaches

Why? To stay as close as possible to the original data.

- Picking / selecting
- Clusterting / classification
- Angular assignment
- Shape / density reconstruction
- Denoising / segmentation

Future: Probabilistic denoising

IDEA: Nonlinear/anisotropic diffusion and bilateral filtering can be seen as specializations of a probabilistic model

Image and Vision Computing 22 (2004) 73–81

A common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean shift

Danny Barash[a,*], Dorin Comaniciu[b]

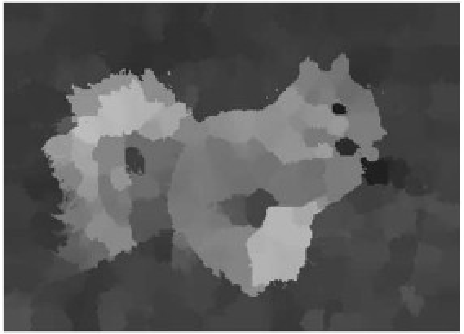
[a] *Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and Howard Hughes Medical Institute, 31 Washington Place, Main 1021, New York, NY 10003, USA*

[b] *Real-Time Vision and Modeling Department, Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, USA*

Similar to clustering pixels

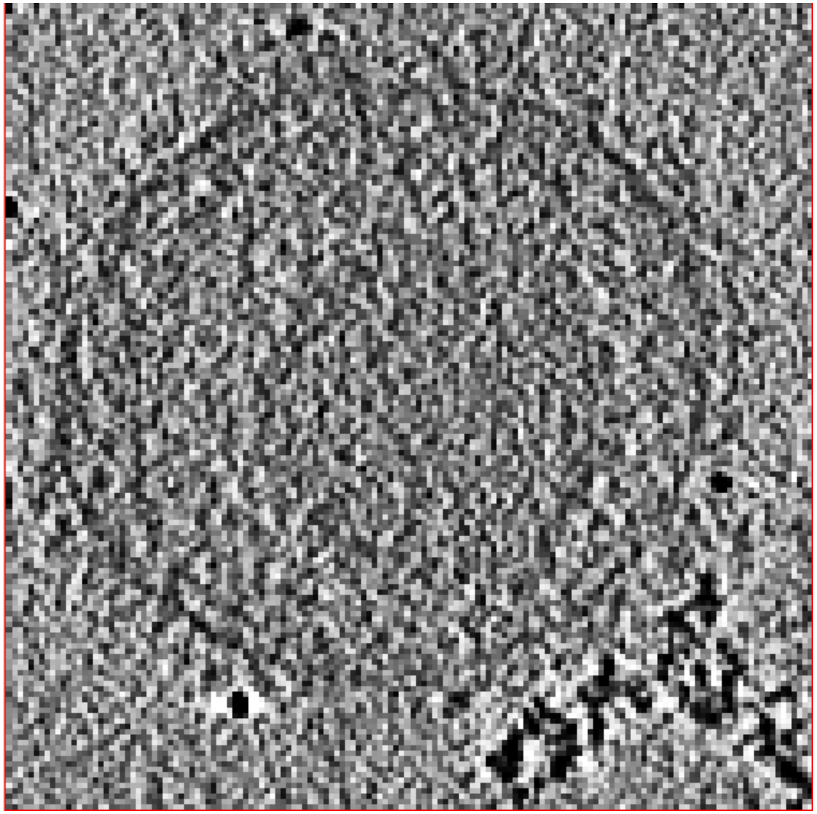
Segments/denoised regions are found by a clustering algorithm similar to *k*-means

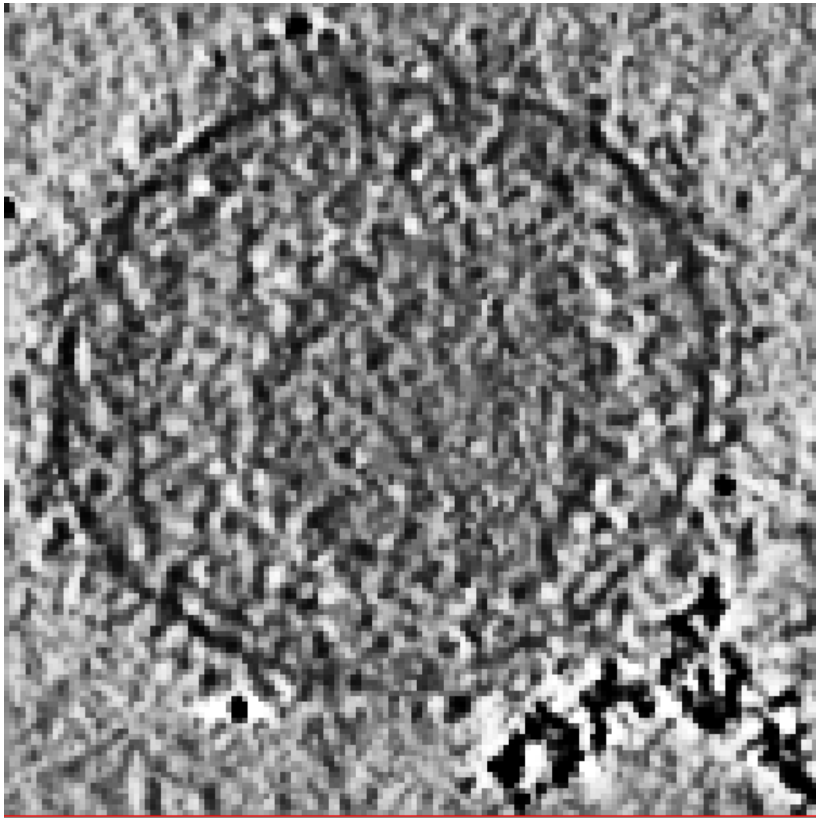
Fig. 1. Original squirrel image.



Bilateral filter

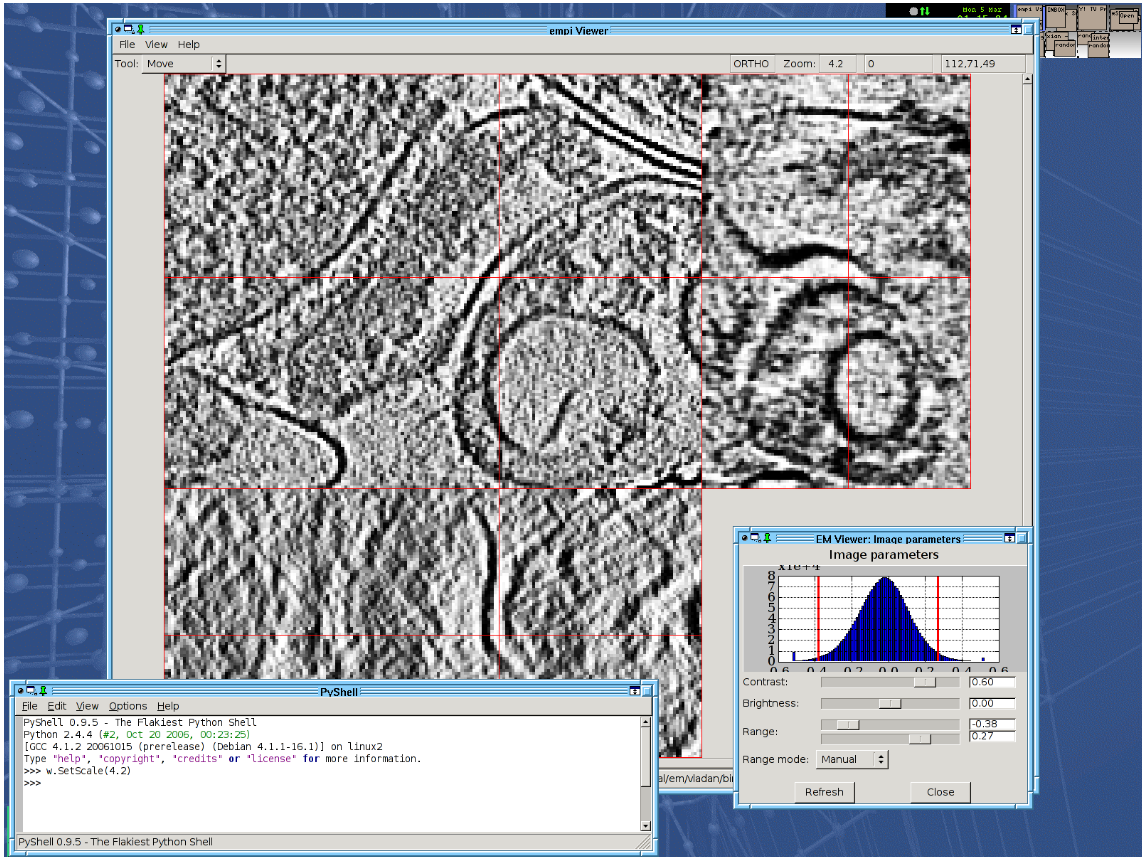
Very simple algorithm: Smoothing in combined position-value space
Only pixels that are close in position and value averaged





Software infrastructure: empi

- Python module **empi** for EM image processing
  - Basis for new algorithm development
  - Rapid prototyping
  - Easy integration of C++ modules
  - Supports Itanium supercomputing platform
  - Integrated programmable 2D/3D viewer



- Future:
  - Fully parallelized alignment and projection routines
  - Integration with SPARX environment

Conclusions & Outlook

"Intelligent Systems for Molecular Biology"

Advanced informatics:

- Learning systems
- Non-linear methods
- Probabilistic models

Benefits:

- More reliable reconstructions
- Increased resolution
- Less dependence on human interaction -> less bias
- Higher throughput