

Image Informatics Tools in Support of Systems Biology

Robert F. Murphy

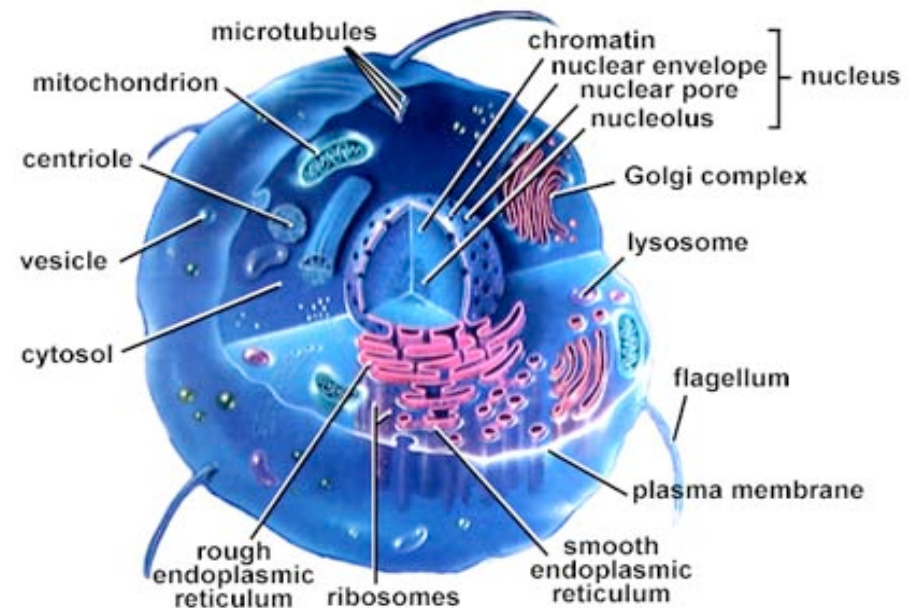
Departments of Biological Sciences, Biomedical
Engineering and Machine Learning and



Carnegie Mellon

Systems Biology and Location Proteomics

- All systems biology must be data driven
- Key to progress
 - identification of aspect that needs to be analyzed “ome-wide”
 - development of assays and automated analysis approaches
- Systems biology needs systematic information on high-resolution subcellular location
 - Eventually, for every expressed protein for all cell types under all conditions
- Providing this information is the goal of Location Proteomics





Automated Interpretation

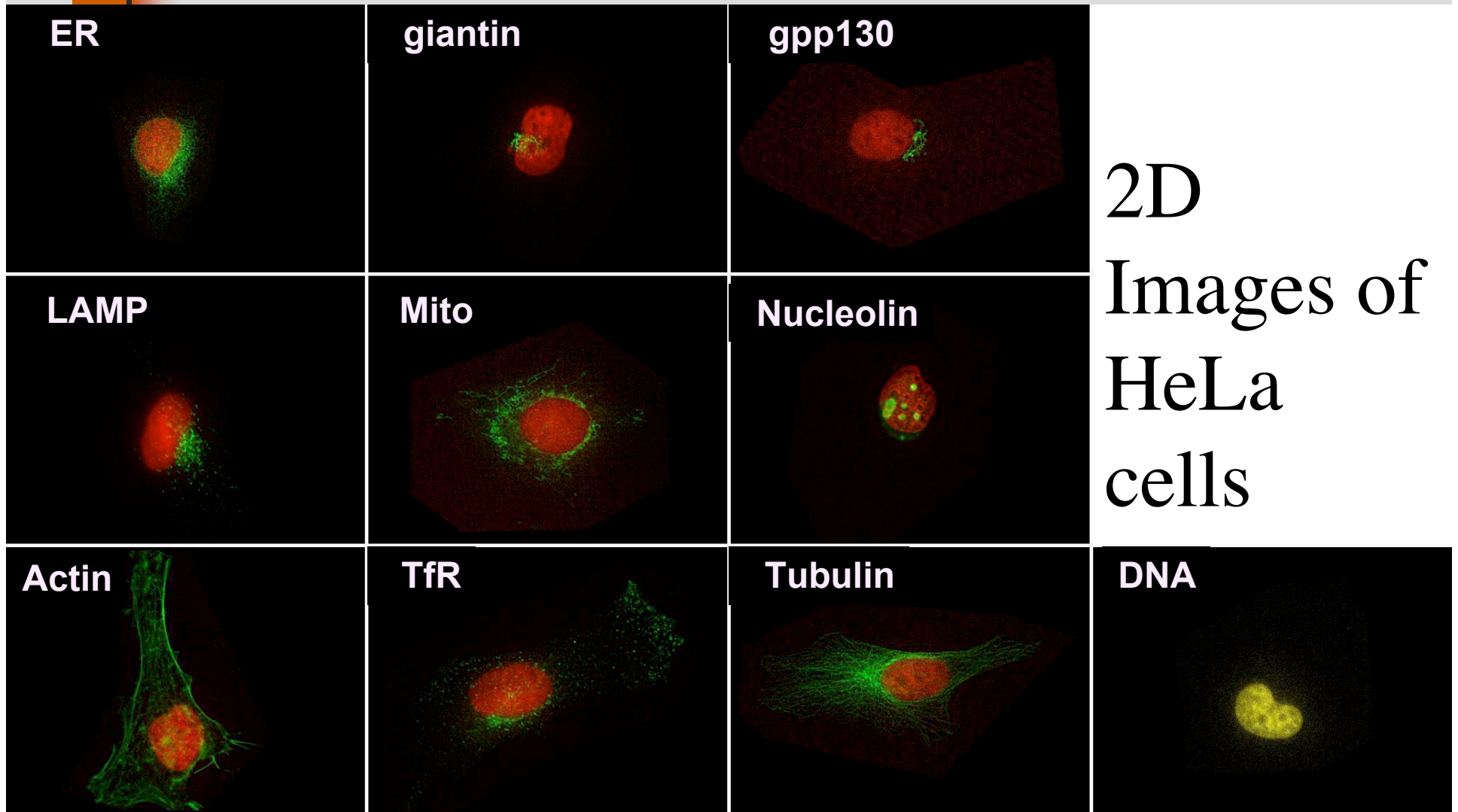
- Traditional analysis of fluorescence microscope images has occurred by visual inspection
- Our goal over the past ten years has to been automate the interpretation, to yield better
 - Objectivity
 - Sensitivity
 - Reproducibility

Supervised Learning of High-Resolution Subcellular Location Patterns



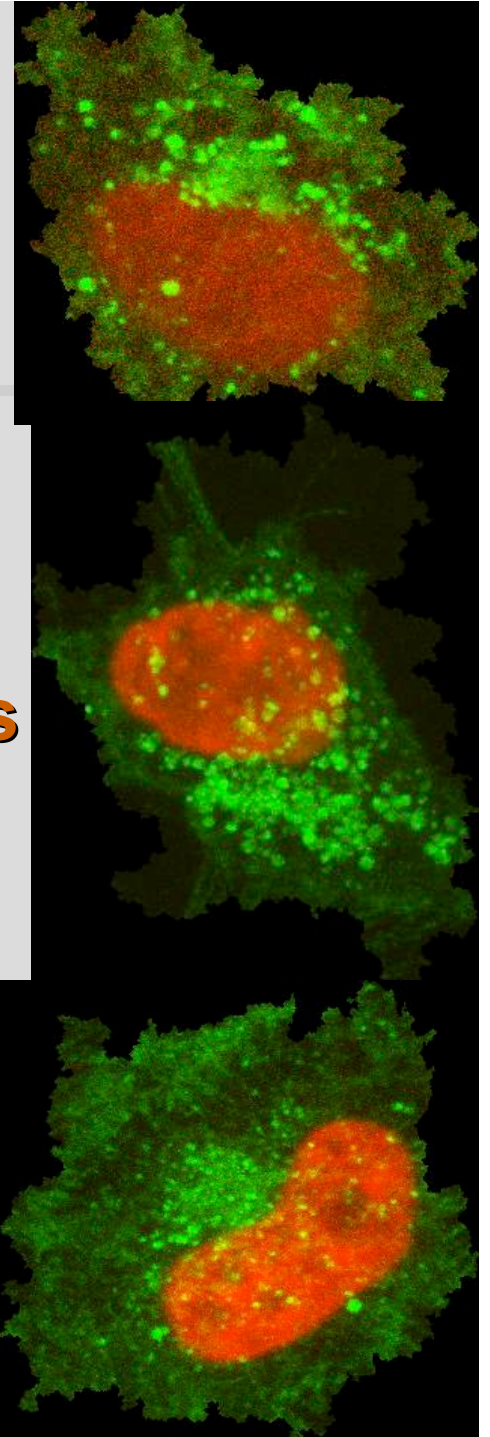
Carnegie Mellon

The goal: Learn to recognize all major subcellular patterns



The Challenge

- Pixel-by-pixel or region-by-region matching will not work for cell patterns because different cells have different **shapes, sizes, orientations**
- Organelles/structures within cells are **not found in fixed locations**
- ***Instead, describe each image numerically and compare the descriptors***



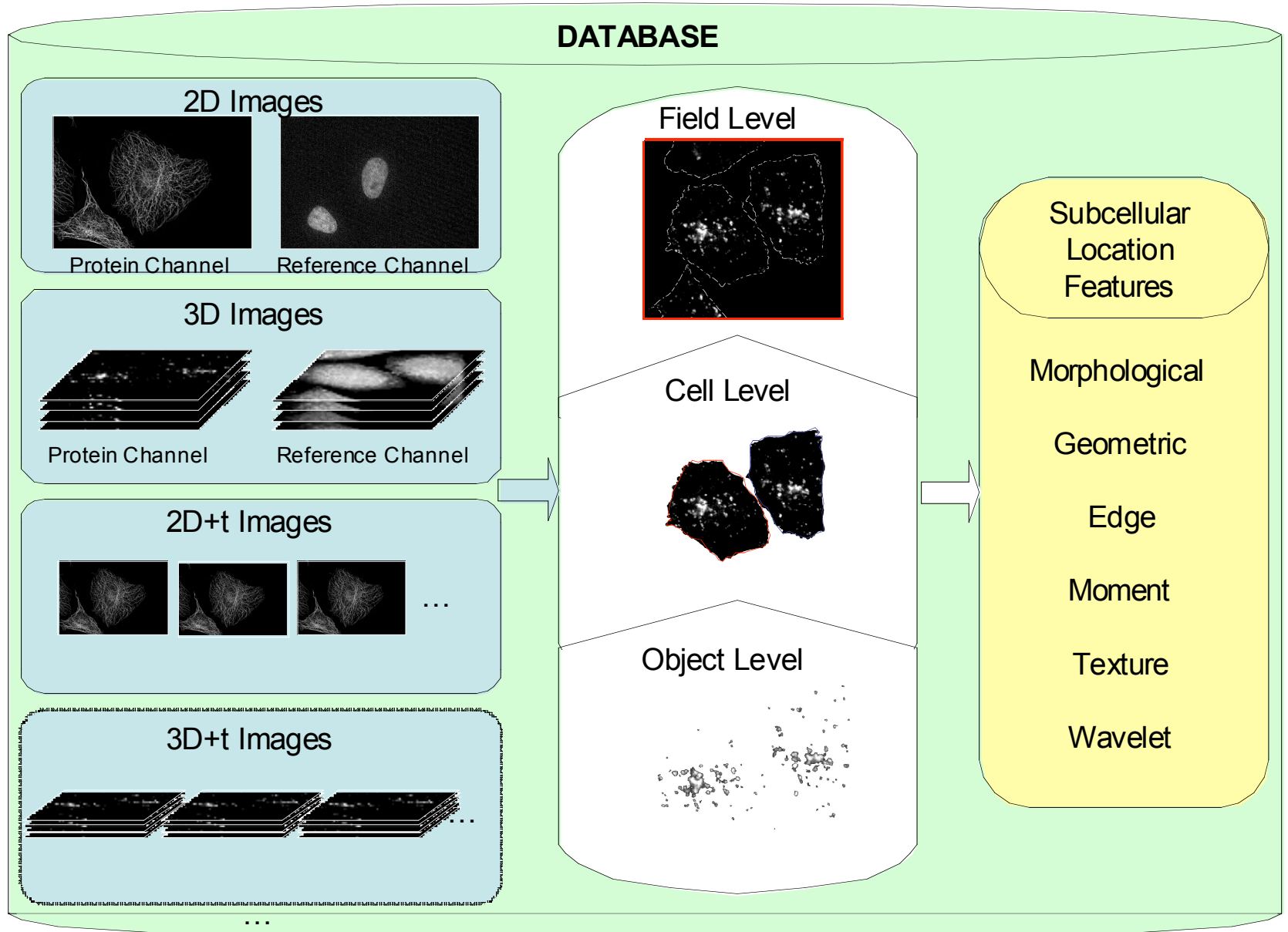
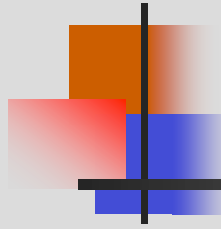


Feature-based, Supervised learning approach

1. Create sets of images showing the location of many different proteins (each set defines one **class** of pattern)
2. Reduce each image to a set of numerical values (“**features**”) that are insensitive to position and rotation of the cell
3. Use **machine learning methods** to “learn” how to distinguish each class using the features

Boland et al 1996; 1997; 1998;
Boland & Murphy 2001; Huang &
Murphy 2004

**Boland et al 1997; 1998;
Boland & Murphy 2001;
Huang & Murphy 2004**



Murphy et al 2000;
 Boland & Murphy 2001;
 Huang & Murphy 2004

2D Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	99	1	0	0	0	0	0	0	0	0
ER	0	97	0	0	0	2	0	0	0	1
Gia	0	0	91	7	0	0	0	0	2	0
Gpp	0	0	14	82	0	0	2	0	1	0
Lam	0	0	1	0	88	1	0	0	10	0
Mit	0	3	0	0	0	92	0	0	3	3
Nuc	0	0	0	0	0	0	99	0	1	0
Act	0	0	0	0	0	0	0	100	0	0
TfR	0	1	0	0	12	2	0	1	81	2
Tub	1	2	0	0	0	1	0	0	1	95

Overall accuracy = 92%

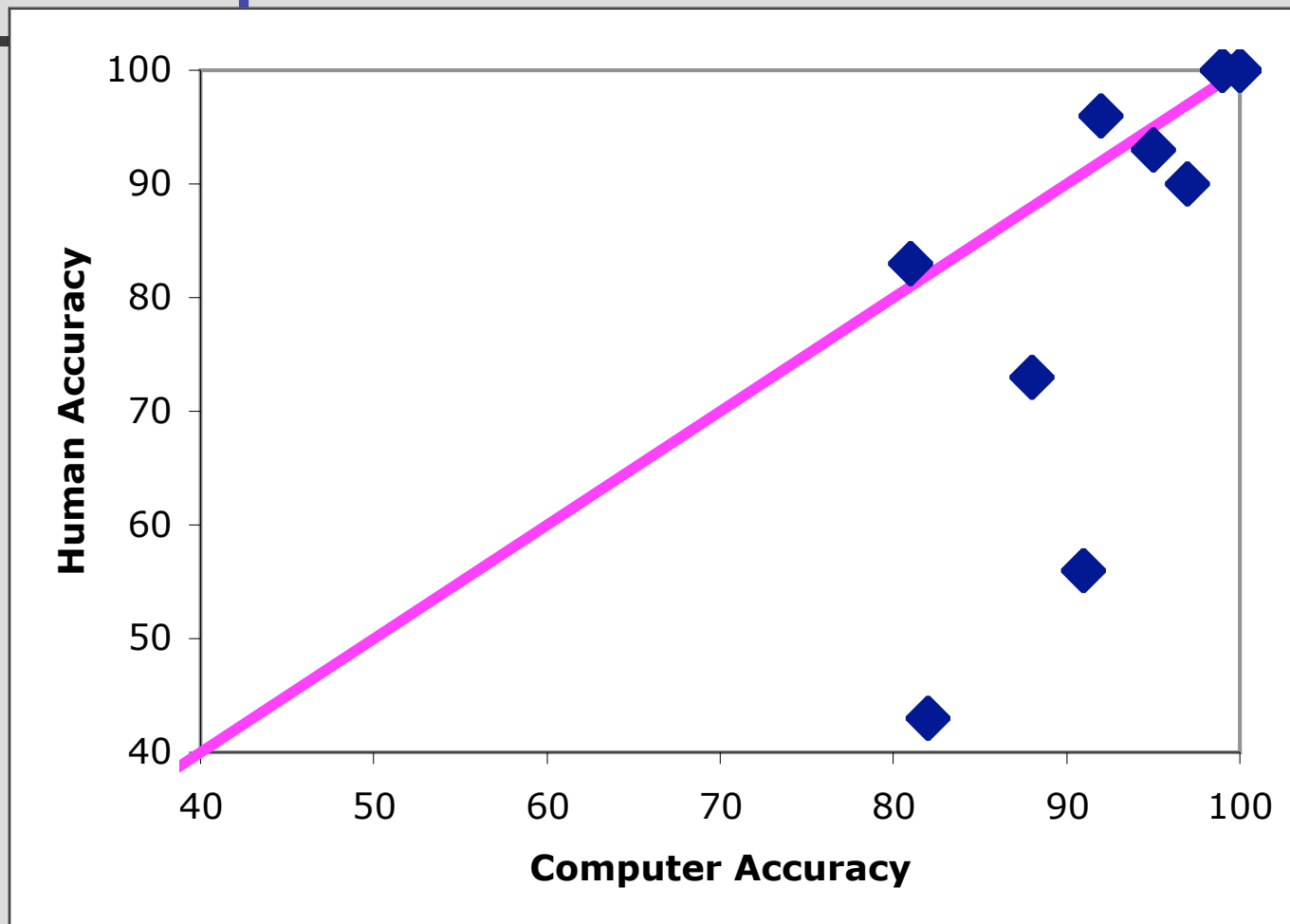


Human Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	100	0	0	0	0	0	0	0	0	0
ER	0	90	0	0	3	6	0	0	0	0
Gia	0	0	56	36	3	3	0	0	0	0
Gpp	0	0	54	33	0	0	0	0	3	0
Lam	0	0	6	0	73	0	0	0	20	0
Mit	0	3	0	0	0	96	0	0	0	3
Nuc	0	0	0	0	0	0	100	0	0	0
Act	0	0	0	0	0	0	0	100	0	0
TfR	0	13	0	0	3	0	0	0	83	0
Tub	0	3	0	0	0	0	0	3	0	93

Overall accuracy = 83%

Computer vs. Human





Conclusions (1996-2004)

- Automated classification of subcellular patterns possible without colocalization
- Accuracy better than visual examination
 - Similar for basic patterns
 - Better for similar patterns
- 3D images give better accuracy than 2D
- **>> SLFs capture essence of patterns**

Unsupervised Learning to Identify High-Resolution Protein Patterns



Carnegie Mellon

Location Proteomics

- **Tag** many proteins
 - We have used **CD-tagging** (developed by **Jonathan Jarvik** and **Peter Berget**): Infect population of cells with a retrovirus carrying DNA sequence that will “tag” in a random gene



Jarvik
et al
2002

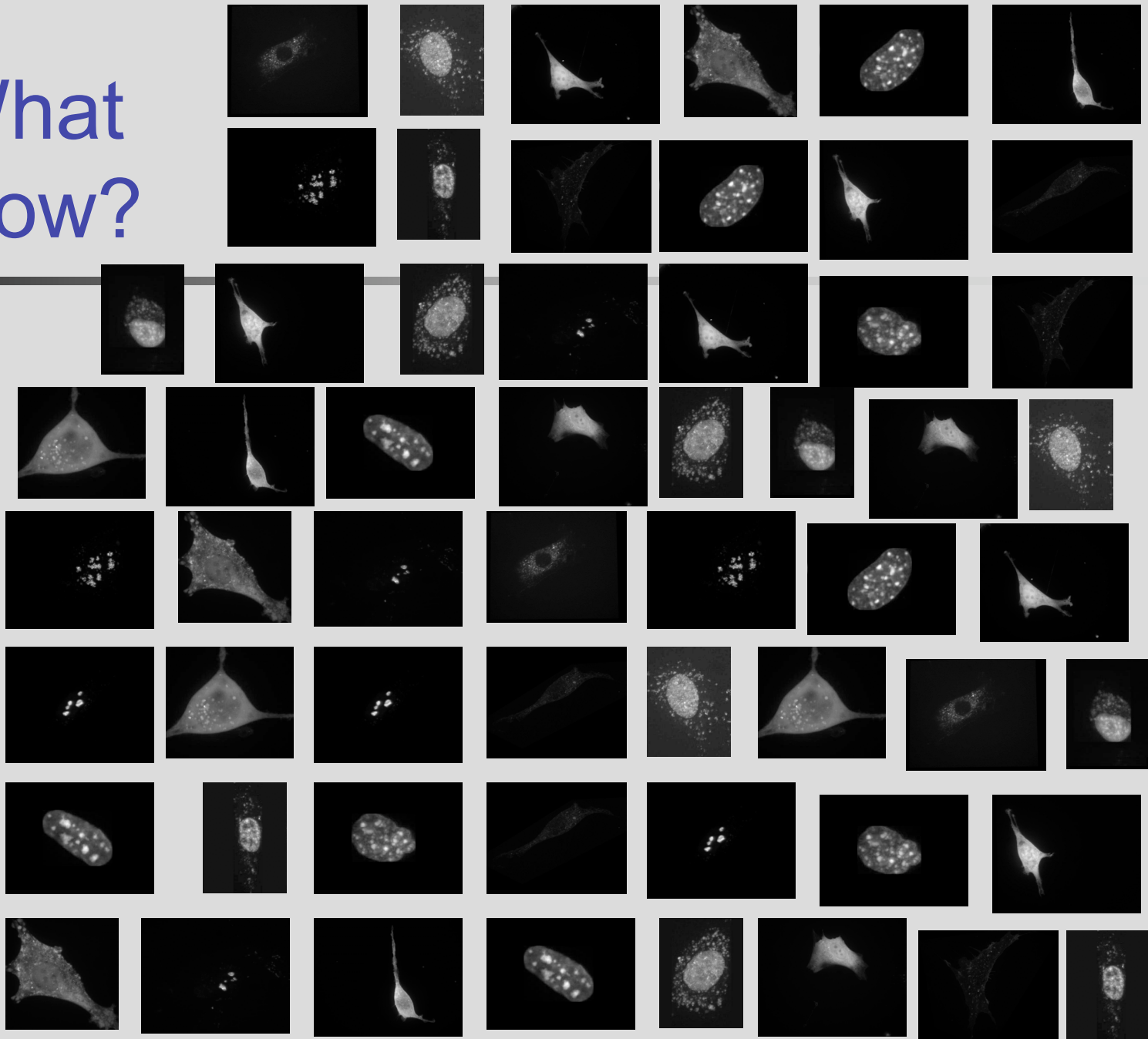
Isolate separate **clones**, each of which produces express one tagged protein

Use RT-PCR to **identify tagged gene** in each clone

- Collect **many live cell images** for each clone using spinning disk confocal fluorescence microscopy

What Now?

Group
~90
tagged
clones
by
pattern



Generative Models for Subcellular Location Patterns



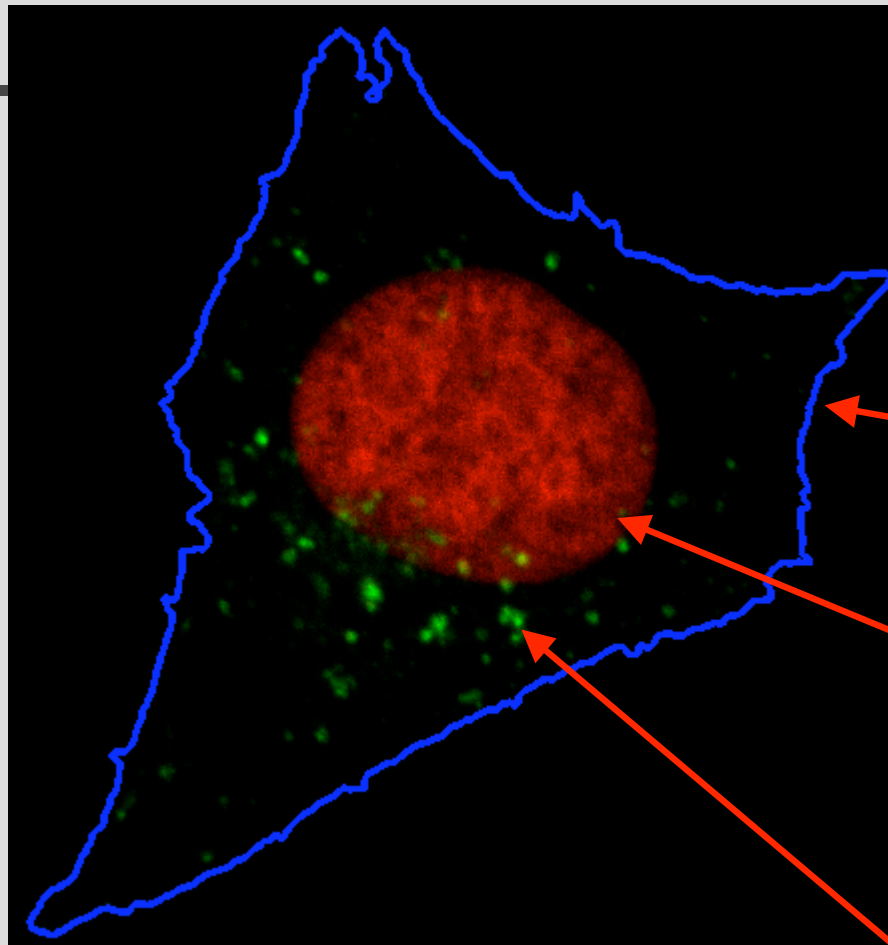
Carnegie Mellon



Need

- How do we communicate results of clustering patterns?
- Show all images from a given cluster?
 - Long download
 - No ability to generalize
- Proposal: Use generative models

LAMP2 pattern



Cell membrane

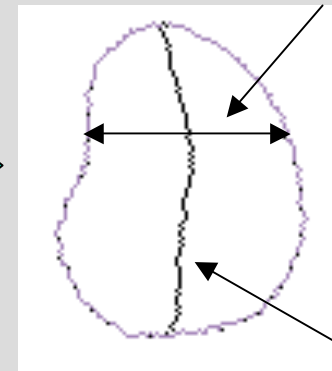
Nucleus

Protein

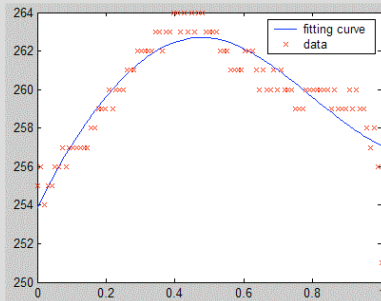
Nuclear Shape - Medial Axis Model



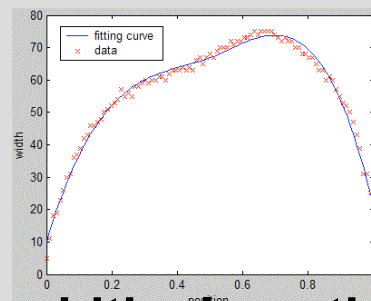
Rotate



Represented by two curves



the medial axis



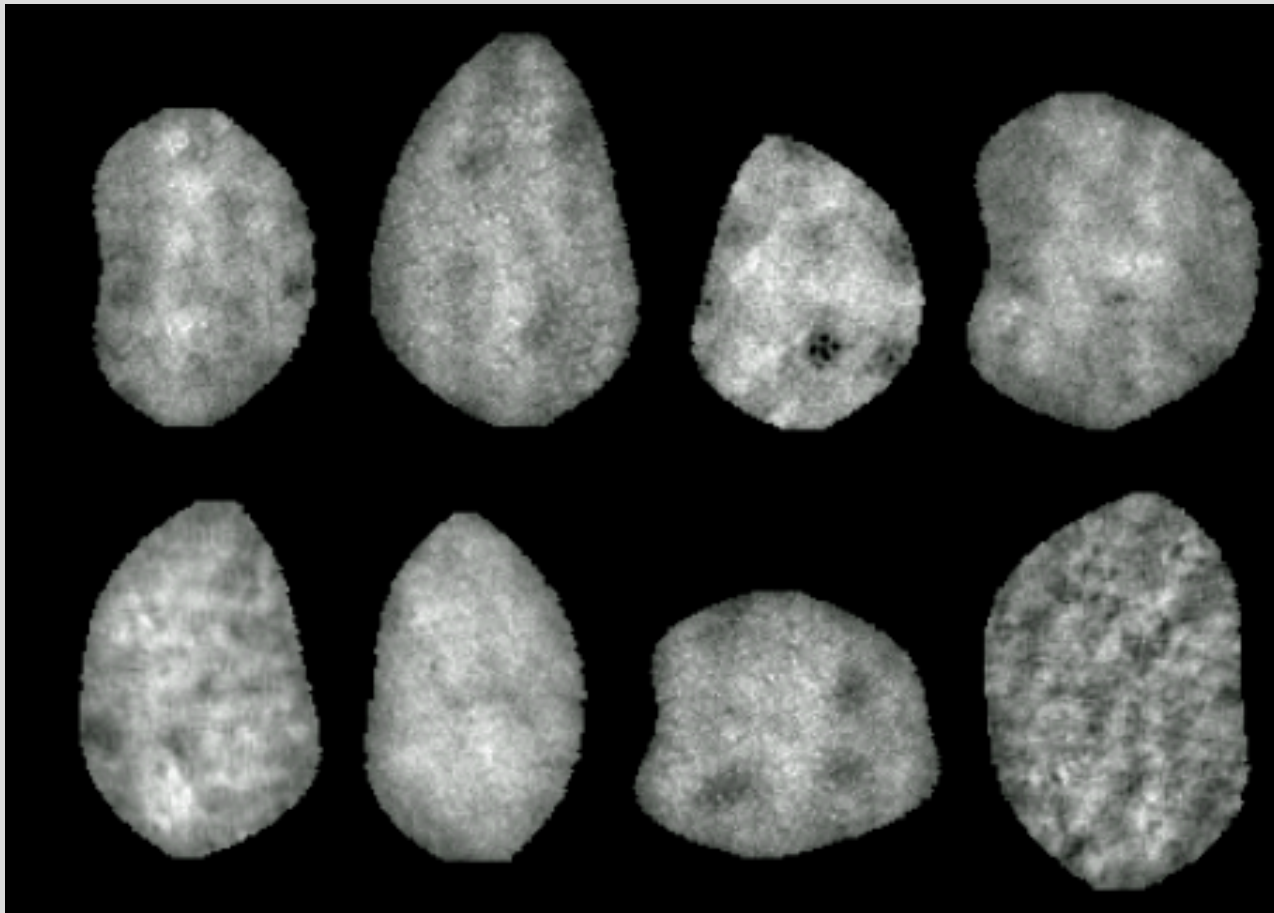
width along the medial axis



Synthetic Nuclear Shapes

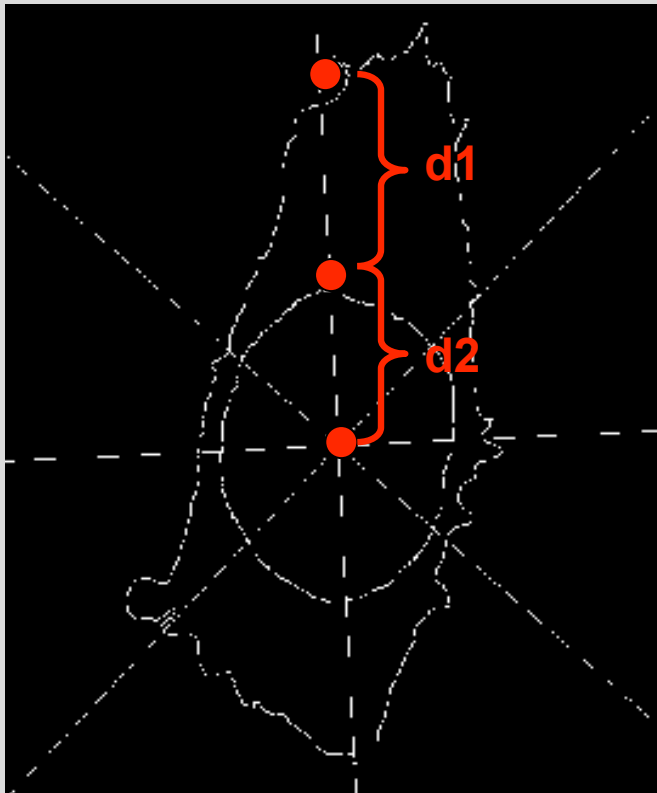


Synthetic nuclei generated by learned model



Cell Shape

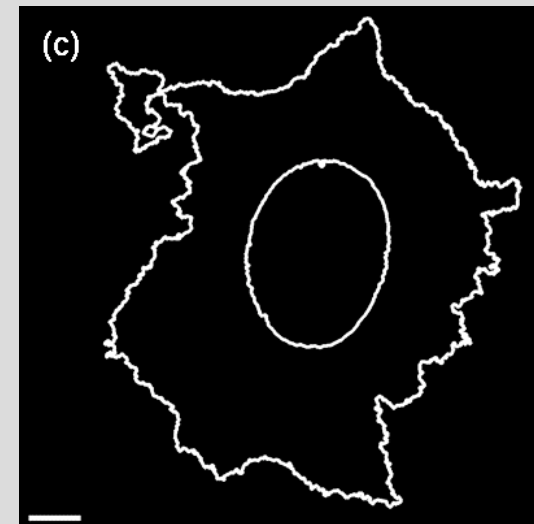
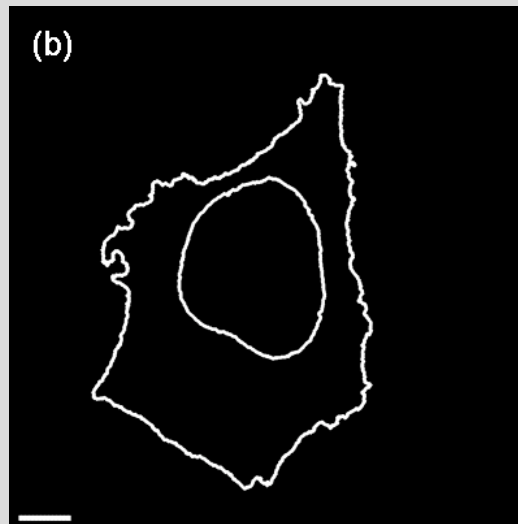
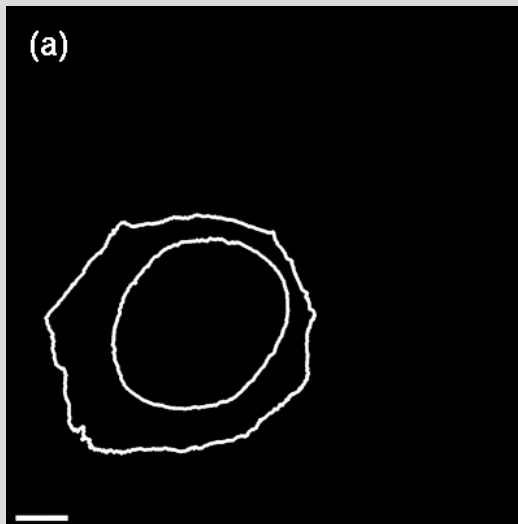
Description: Distance Ratio



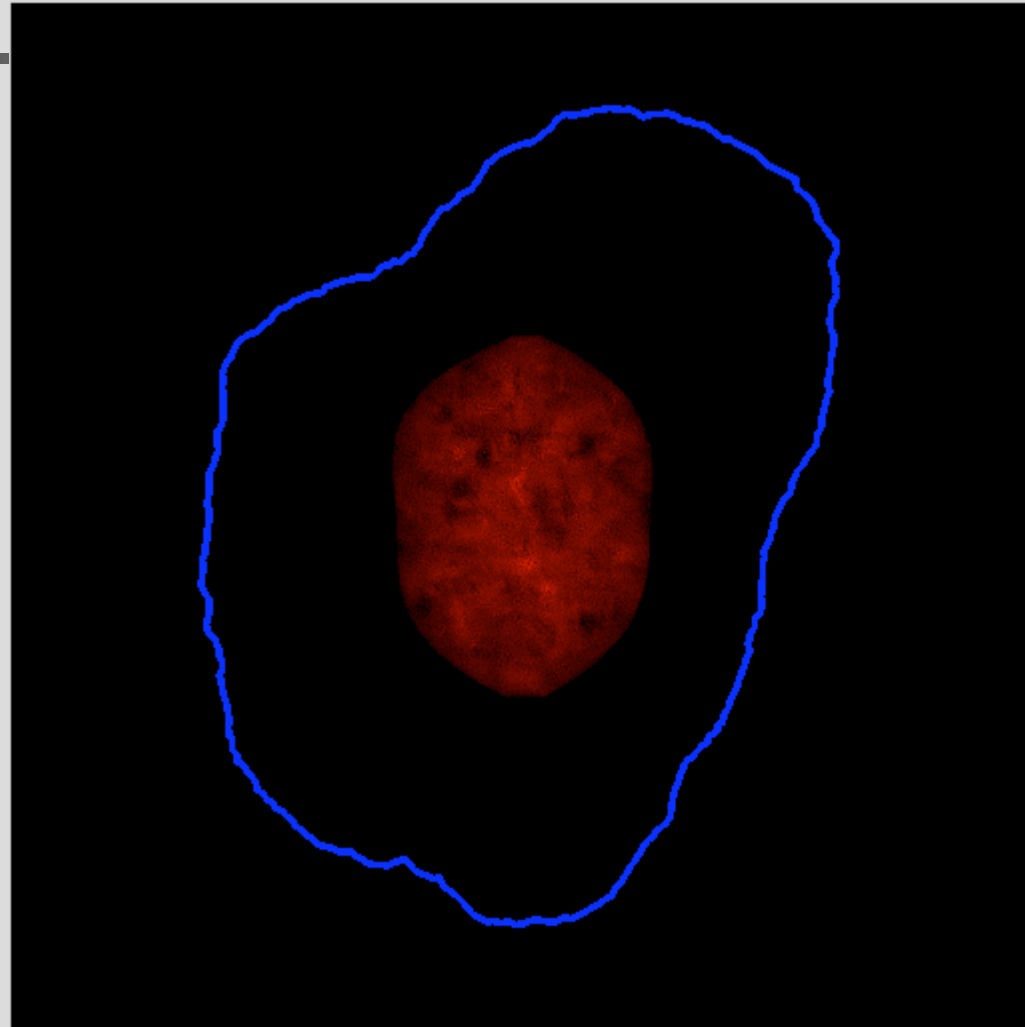
$$r = \frac{d_1 + d_2}{d_2}$$

Capture variation as a principal components model

Examples of natural variation in cell shape

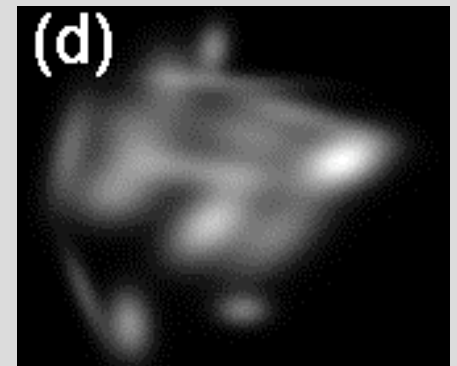
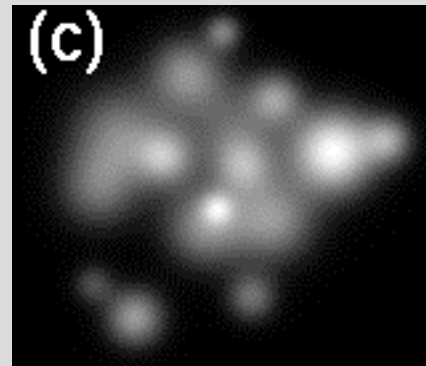
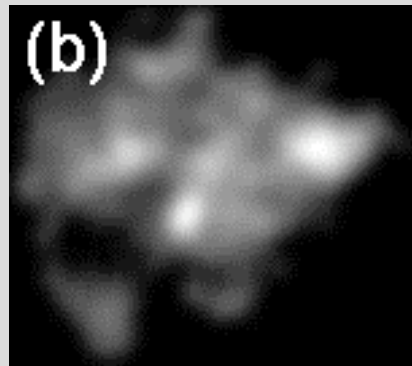
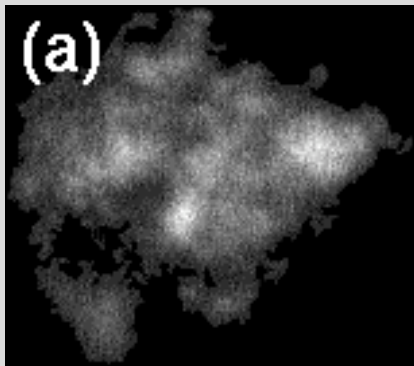


Example cell boundary generated from learned model

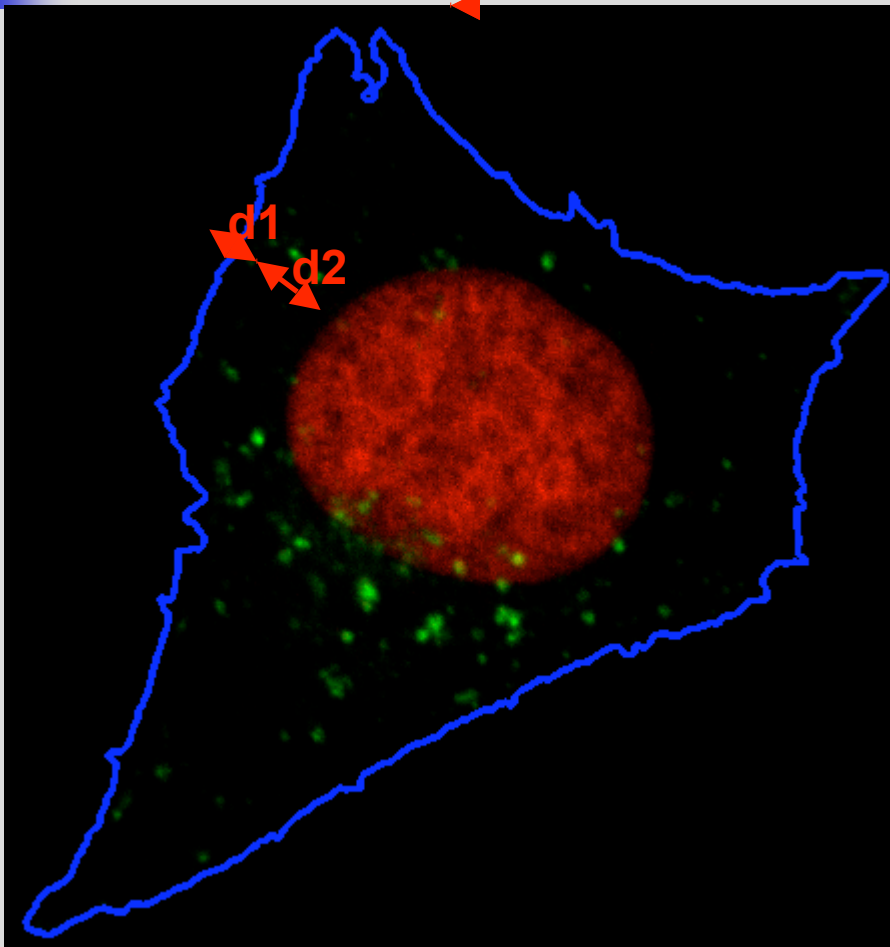


Modeling Vesicular Organelles

- Original image and fitted Gaussians of increasing complexity

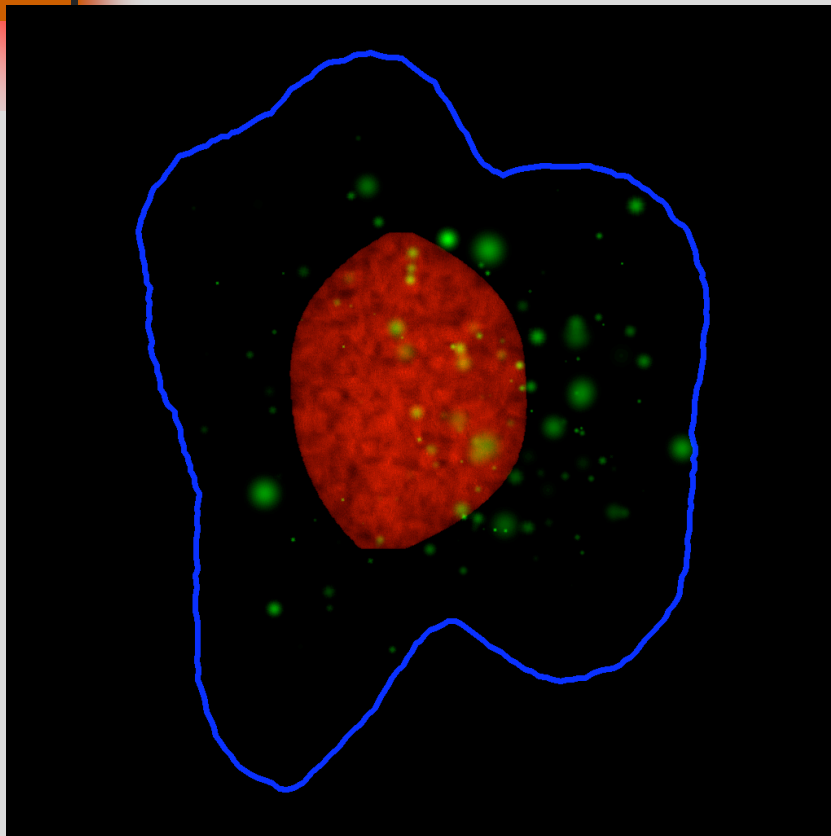


Object Positions

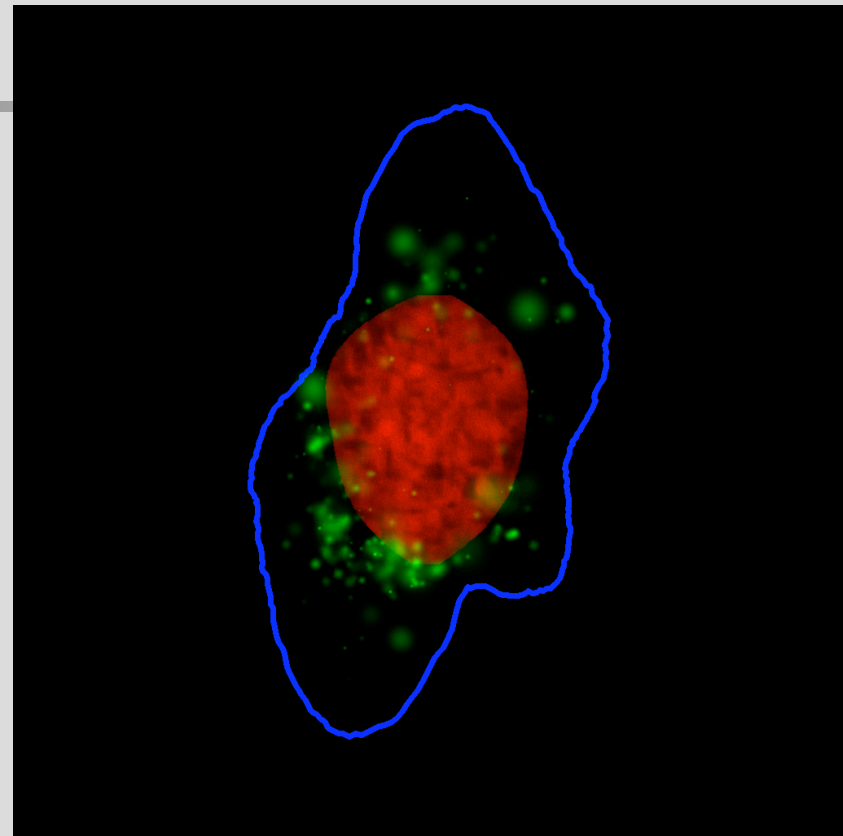


$$r = \frac{d_2}{d_1 + d_2}$$

Synthesized Images

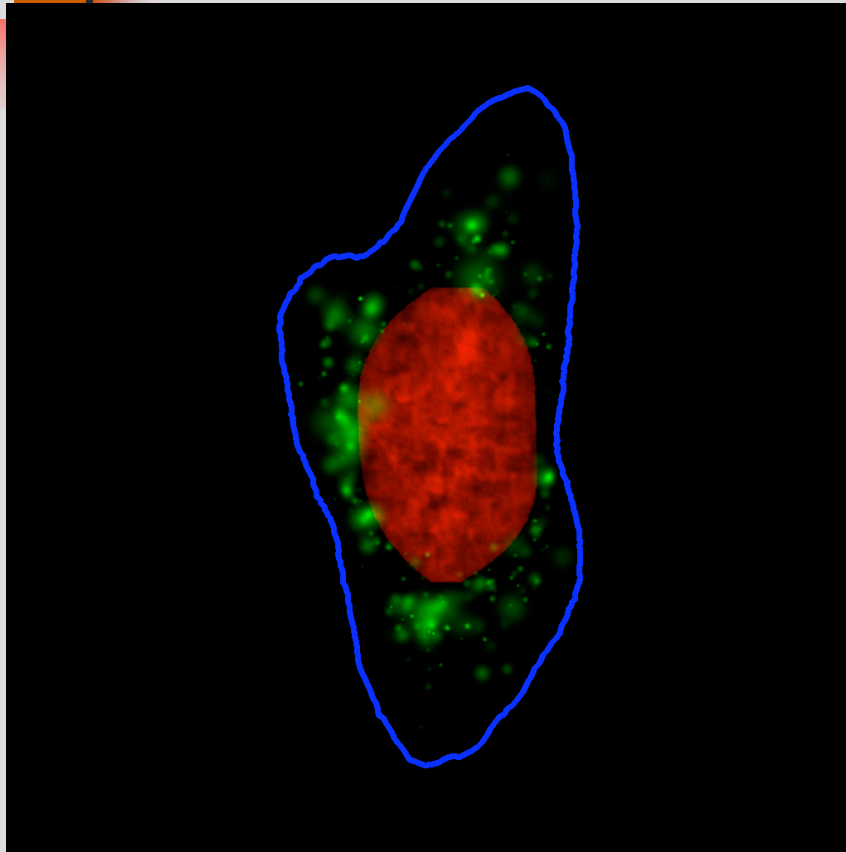


Lysosomes

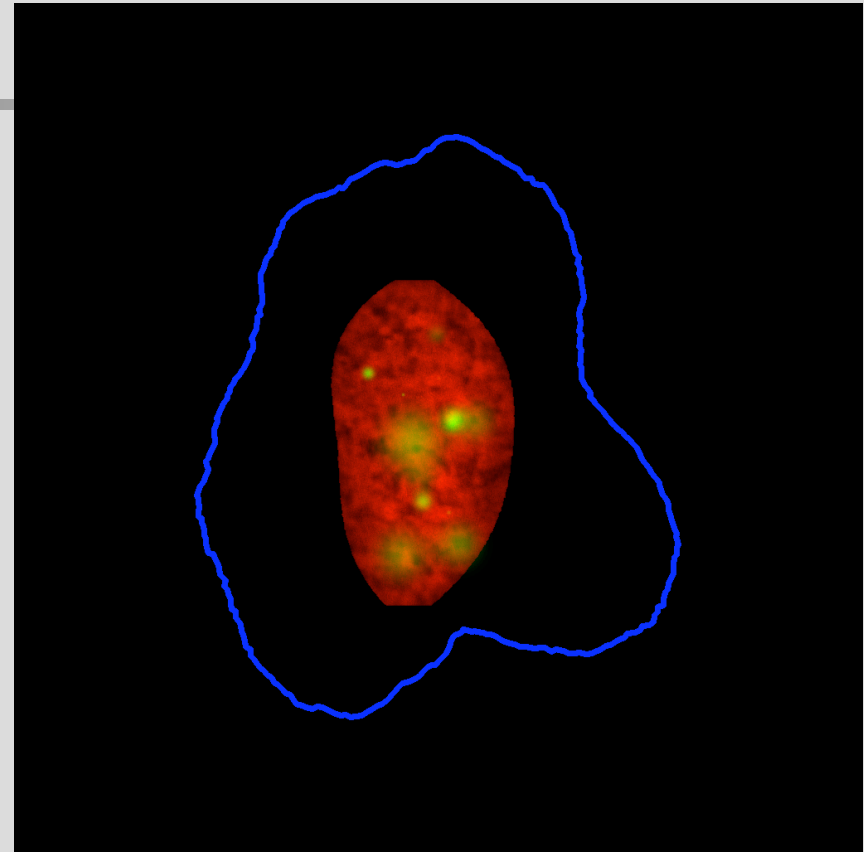


Endosomes

Synthesized Images



Mitochondria



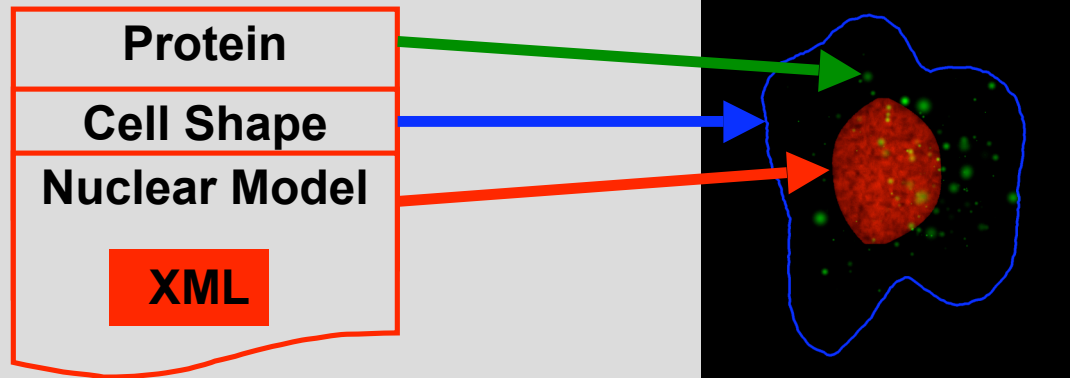
Nucleoli



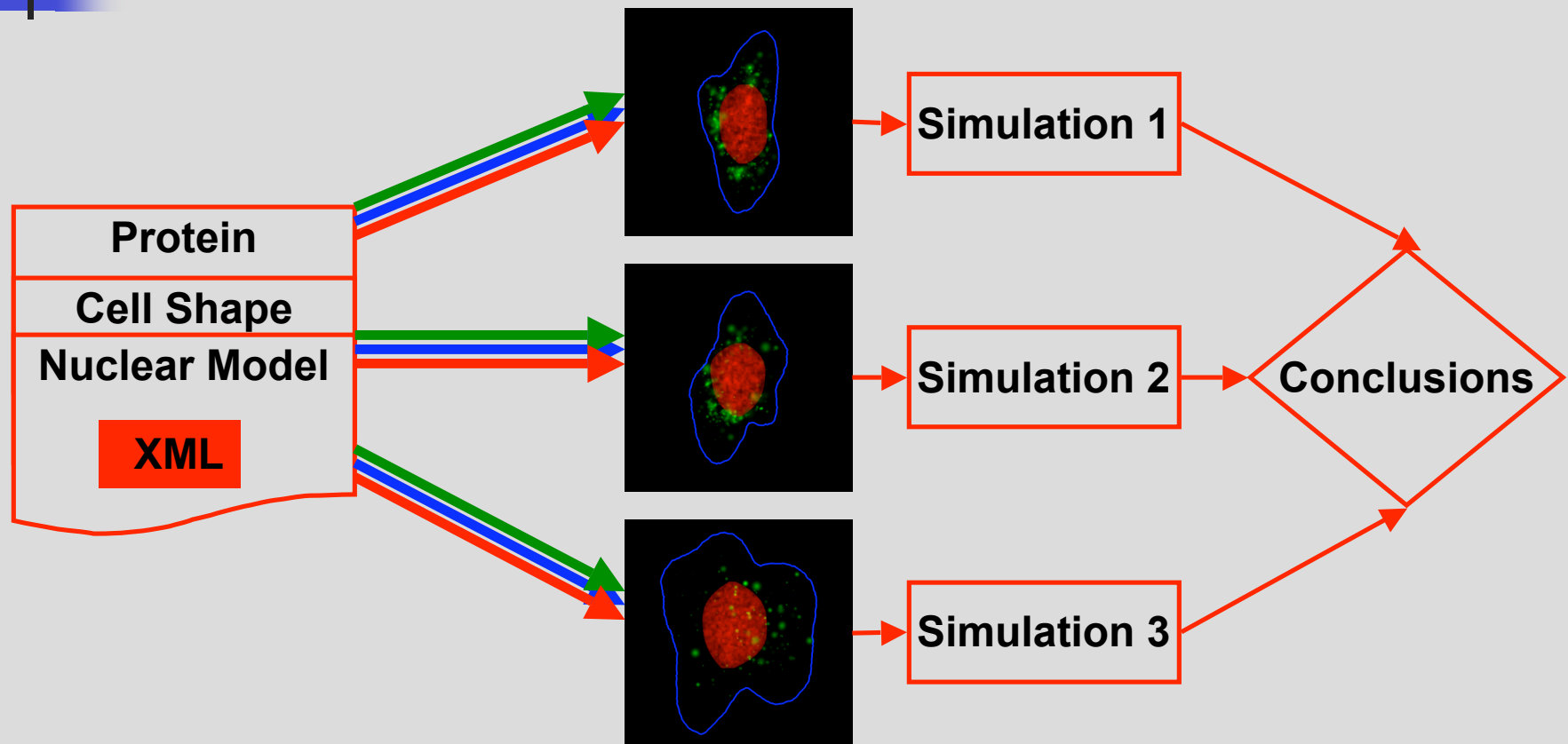
Model Distribution

- Generative models provide better way of distributing what is known about “subcellular location families” (or other imaging results, such as illustrating change due to drug addition)
- Have initial XML design for capturing the models for distribution
- Have portable tool for generating images from the model

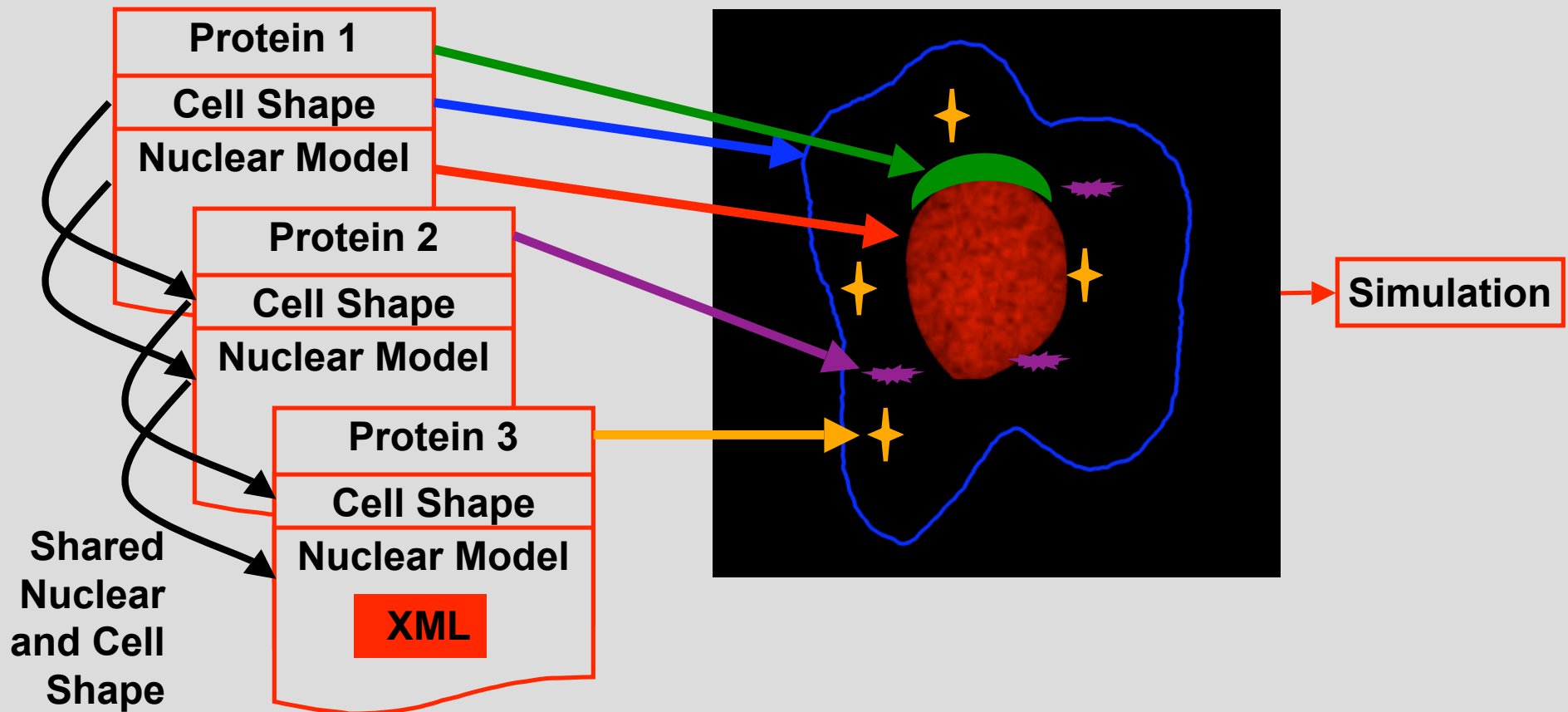
Generation Process



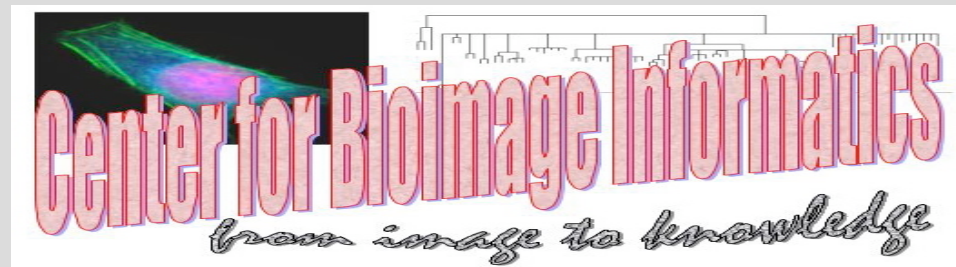
Generating Multiple Distributions for Simulations



Combining Models for Cell Simulations



The Protein Subcellular Location Image Database (PSLID)



Carnegie Mellon

PSLID: Protein Subcellular Location Image Database

- A publicly accessible image database at <http://pslid.cbi.cmu.edu>
 - Version 3 released February 2, 2007
 - 2D and 3D images (single cell regions defined)
 - Two cell types, HeLa and 3T3
 - Over 120,000 images/3000 unique fields/14,000 cells
 - 111 classes; 55 known proteins; 11 targeting mutants of a single protein
 - Programmatic search via URL (SOAP in the works)

PSLID: Protein Subcellular Location Image Database

- A downloadable open source system for creating local databases
 - Version 3 of software released February 13, 2007
 - Focused on subcellular pattern analysis
 - SLF features integrated into database
 - Integrated comparison, classification, clustering tools
 - Designed for high-throughput microscopy
 - Interface to OME in the works
 - Large ITR project with UCSB for distributed system

See poster by Estelle Glory



External search

- The programmable search has the following format:
`http://pslid.cbi.cmu.edu/public3/search.jsp?arguments`
- The following search arguments are supported:
 - `protein=protein name`
 - `cell_type=cell type`



External search

- <http://pslid.cbi.cmu.edu/public3/search.jsp?protein=calponin-2>

<http://pslid.cbi.cmu.edu/public3/search.jsp?protein=calponin-2>

<http://pslid.cbi...otein=calponin-2>

Search results for Image Type: 2D Static, Target: calponin-2

10 regions returned (30 regions shown) from the query.

View the [summary](#) of set **temp8_710B35DB64C10A8CF219992B3A193B57**.


Click  besides a given image to retrieve similar images in the database.


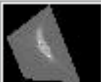







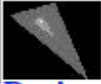

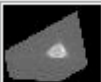




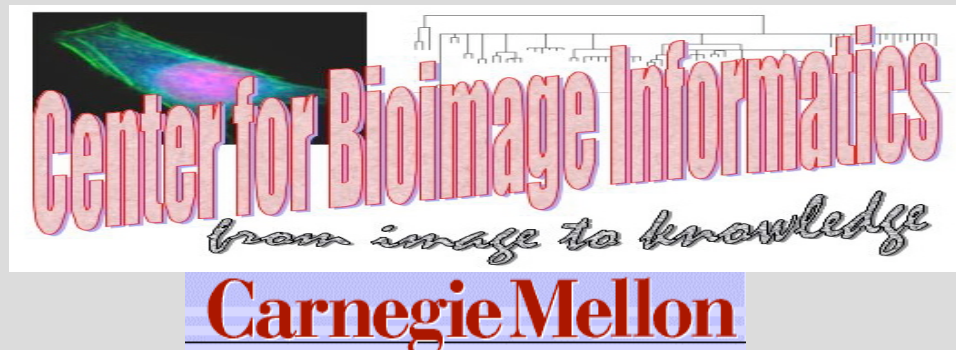
	Image	Cell Name	Organism	Segmenter	Experiment	Protocol	Target	Microscopy & Filter
	 Region 68249	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
	 Region 68280	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
	 Region 68311	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
	 Region 68342	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
	 Region 68373	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
	 Region 68404	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
	 Region 68435	3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500
		3T3	Mus musculus	External	Cyto039	GFP Live	Calponin-2	Olympus IX500

Image Content-based Retrieval and Interpretation of Micrographs from On-line Journal Articles

The Subcellular Location Image Finder (SLIF)





Ultimate Goal of the field

- Machine understanding of biological journal articles (text and image)
- Criteria for success: Turing test - have machine be able to answer questions about an article as well as a human scientist



Intermediate Goal

- Extract information from combination of text and any kind of image in biological journal article
- Criteria for success: Achieve high precision and recall for extracted assertions (compared to expert scientist)



Immediate Goal (SLIF)

- Extract information about subcellular location from captions and figures containing fluorescence microscope images in biological journal articles
- Criteria for success: Achieve high precision and recall for extracted assertions (compared to expert scientist)



State of art: Bio Journal Information Extraction

- A number of systems to index literature via extracted terms
- A few systems to index image content in literature
- A few systems for document classification

Overview: Image processing tasks

Segment
into
“panels”

Detect & remove
annotations

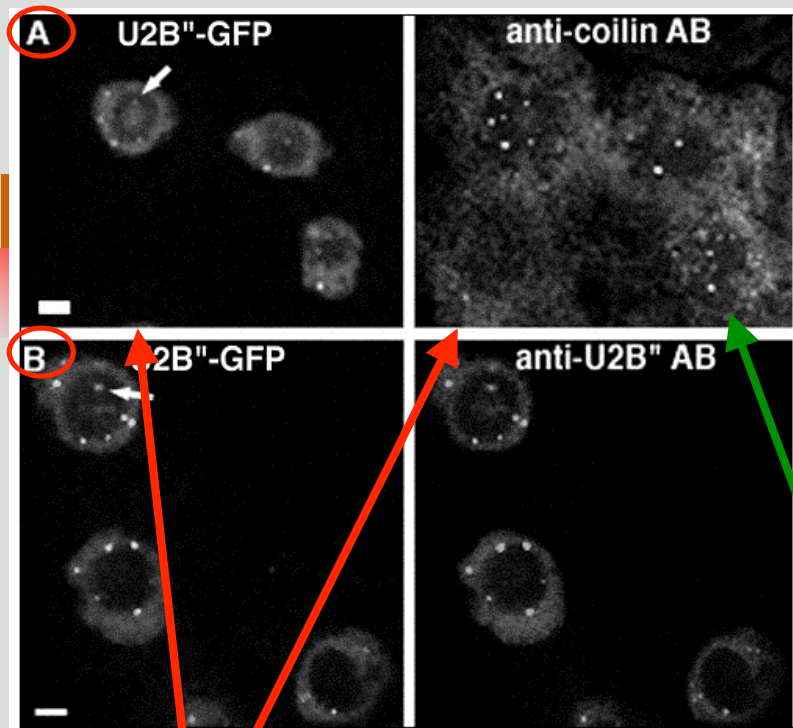
Classify
panels

Find scale bars

A U2B''-GFP anti-coilin AB

B U2B''-GFP anti-U2B'' AB

FMI+ FMI+ FMI+ FMI+



Overview: Text processing tasks

- Find *entity names* in text, and *panel labels* in text and the image.
- *Match* panels labels in text to panel labels on the image.
- *Associate* entity names to textual panel labels using *scoping* rules.

Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B0-GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies are in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B0 also express the U2B0-GFP fusion. Bars, 5 μ m. Movement of Coiled Bodies Vol. 10, July 1999 2299

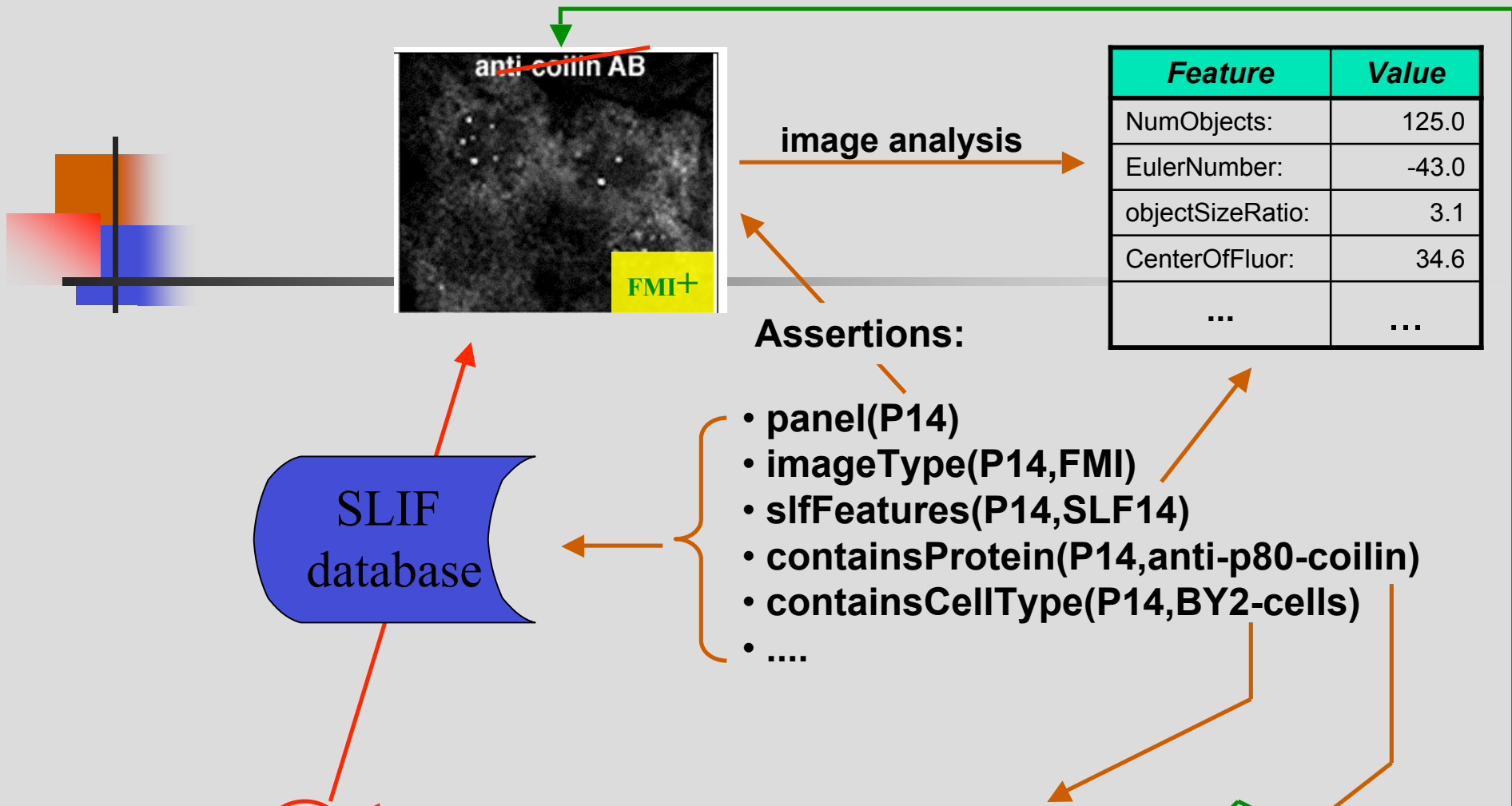


Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B₈-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown....



Linking to SLIF from another website

- Queries against the database can be made from outside the website using <http://slif.cbi.cmu.edu/SLIF/search.jsp?arguments>
- The arguments are:
 - protein=<protein name>
 - level=figure OR level=panel
 - type=FMI (NOTE that BOTH level and type must be present if either is present)
 - pixel_size_lo=<lower bound>
 - pixel_size_hi=<upper bound> (NOTE that both upper and lower bounds must be specified)
 - location=<subcellular location>



SOAP interface - query DTD

```
<!ELEMENT slif_query  
  (protein_name,  
   fmi_level,  
   pixel_res_lower_bound,  
   pixel_res_upper_bound,  
   subcellular_location)>  
<!ELEMENT protein_name (#PCDATA)>  
<!ELEMENT fmi_level EMPTY>  
<!ATTLIST fmi_level figure_or_panel (figure|panel)  
  #REQUIRED>  
<!ELEMENT pixel_res_lower_bound (#PCDATA)>  
<!ELEMENT pixel_res_upper_bound (#PCDATA)>  
<!ELEMENT subcellular_location (#PCDATA)>
```



SOAP interface - results DTD

```
<!ELEMENT slif_result
(slif_row*)>
<!ELEMENT slif_row
(caption,
figure_url,
panel_url,
protein_name,
cell_name,
subcellular_location,
pixel_resolution)>
<!ELEMENT caption (#PCDATA)>
<!ELEMENT figure_url (#PCDATA)>
<!ELEMENT panel_url (#PCDATA)>
<!ELEMENT protein_name (#PCDATA)>
<!ELEMENT cell_name (#PCDATA)>
<!ELEMENT subcellular_location (#PCDATA)>
<!ELEMENT pixel_resolution (#PCDATA)>
```

WELCOME TO



Subcellular Location Image Finder

SLIF (Subcellular Location Image Finder) automatically extracts information about protein subcellular locations from figure-caption pairs in biological literature. SLIF separates figures into panels and decides which panels contain fluorescence microscope images (FMI). It applies image processing methods to analyze the FMI and extract a quantitative description of the localization patterns they display. The associated captions are also processed to identify which portions of the caption refer to which panels and to identify the names of proteins contained in the captions. The results of this analysis are stored in the SLIF database.

Our long-term goal is to develop a large library of annotated and analyzed fluorescence microscope images, in order to support data-mining.

[PNAS, version 3.0](#)

The current version of the database contains records for 15180 papers from volumes 94-99 of the Proceedings of the National Academy of Sciences (USA), generously made available by the Academy for demonstration purposes.

BioMed Central, version 1.0

Due for release March 5, 2007

Pubmed Central, version 1.0

The database will be expanded shortly to include all open access articles in Pubmed Central, including BMC papers but not PNAS papers (approximately 45,000 as of 31 December 2007).

A service of the Robert F. Murphy laboratory
Departments of [Biological Sciences](#), [Biomedical Engineering](#), and [Machine Learning](#)
and [Center for Bioimage Informatics](#)
[Carnegie Mellon University](#), Pittsburgh, Pennsylvania, U.S.A.



Conclusions

- Methods well worked out for classifying and learning protein patterns - better than visual examination
- Temporal information improves discrimination
- Progress on decomposing complex patterns and generative models
 - High-resolution, reliable data for bottom-up systems modeling
- Graphical models provide improved classification of single cells in fields (and potentially tissues)
- Image database integrated with interpretation tools (PSLID)
- Information extractor for online text and images (SLIF)

Acknowledgments



Michael



Mia



Greg



Meel

■ Students

- Dr. Michael Boland
- Dr. Mia Markey (ugrad)
- Gregory Porreca (ugrad)
- Dr. Meel Velliste
- Dr. Kai Huang
- **Dr. Xiang Chen**
- **Ting Zhao**
- **Shann-Ching Chen**
- **Juchang Hua**



Kai



Xiang

■ Funding

- NSF, NIH, Commonwealth of Pennsylvania

■ Collaborators/Consultants

- David Casasent, Simon Watkins, **Jon Jarvik, Peter Berget, Jack Rohrer**, Tom Mitchell, Christos Faloutsos, Jelena Kovacevic, **William Cohen, Geoff Gordon**, **Simon Watkins, Alan Waggoner**



Juchang

Sam Ting

The Future of Subcellular Pattern Analysis



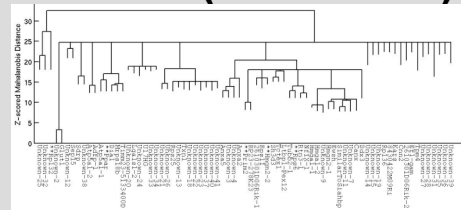
Carnegie Mellon

The problem

Cell Type
(Order 10^2)

Condition
(Order 10^2)

Protein (Order 10^4)



Plus: Time scale from subsecond
to years⁵⁸

Other subcellular location projects

- O'Shea group - Yeast
 - GFP-tagged cDNAs
 - GFP and DNA images with some additional markers
- Pepperkok group - human (MCF7 cells)
 - GFP-tagged cDNAs
 - GFP and DNA images
- Uhlen group (Protein Atlas) - human
 - Immunohistochemistry with monospecific antibodies
 - DAB and hematoxylin images
 - Fixed tissues
- Schubert group (MELK technology)
 - Cycles of immunofluorescence, imaging and bleaching
 - Fixed tissues

Automated Analysis - see poster by Shann-Ching Chen



Orthogonal data sources

- Cytochemical images like Protein Atlas (fixed cells, one color)
- Sequential multicolor immunofluorescence like MELK (fixed cells, many colors)
- GFP-tagged proteins (live cells, one to few colors)



How do we really analyze subcellular location?

- Classification and comparison good for focused questions but there are too many questions to ask
- Need intelligent (optimized) data collection: probabilistic methods to integrate available data, make predictions and suggest experiments



Human Cytome Project?

- Scope of problem argues for cooperation on grand scale
- New inference and synthesis methods

home

you are here: home

navigation

- Home
- About NCIBI
- Computational Technology
- Driving Biological Problems
- Resources and Software
- Education and Training

National Center for Integrative Biomedical Informatics

by plone — last modified 2005-09-29 09:29 AM

Mission

The mission of the NCIBI is to facilitate scientific exploration of complex diseases that are currently infeasible.

The Center develops and interactively integrates analytical and modeling technologies to provide contextually appropriate molecular biology information from emerging experimental

information access and data analysis workflow knowledge models of biological systems. Current problems are prostate cancer progression, prevalence of type 2 diabetes, and genetic susceptibility.

research, training, and education programs.

Collaborators



Collaborations with Bill Mohler, Ian Moraru, Les Loew, Paul Campagnola (U Conn)

Collaboration with Badri Roysam (RPI) and Sally Temple (Albany Med Coll), Stem Cell Patterning and FARSIGHT system

Collaboration with Dan Rines and Sumit Chanda (GNF San Diego) on high throughput location proteomics