March 9, 2018

Integrating Information from Diverse Microscope Images: Learning and Using Generative Models of Cell Organization

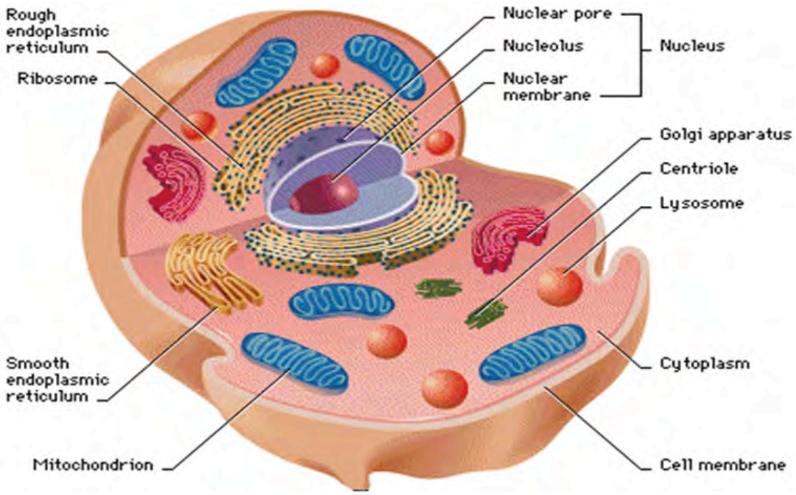
Robert F. Murphy
Ray & Stephanie Lane Professor of Computational Biology and
Professor of Biological Sciences, Biomedical Engineering and Machine Learning
External Senior Fellow, Freiburg Institute for Advanced Studies
Honorary Professor, Faculty of Biology, University of Freiburg, Germany

Carnegie Mellon University
Computational Biology Department

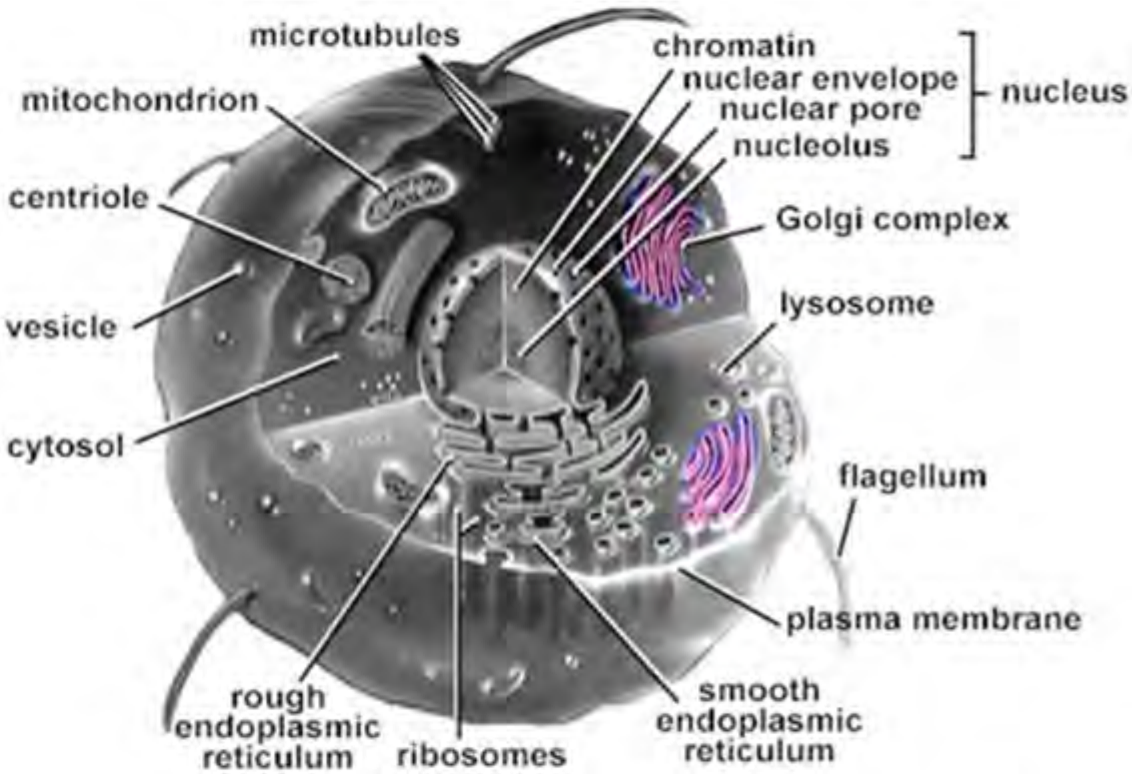
MMBioS
An NIH Biomedical Technology Research Center

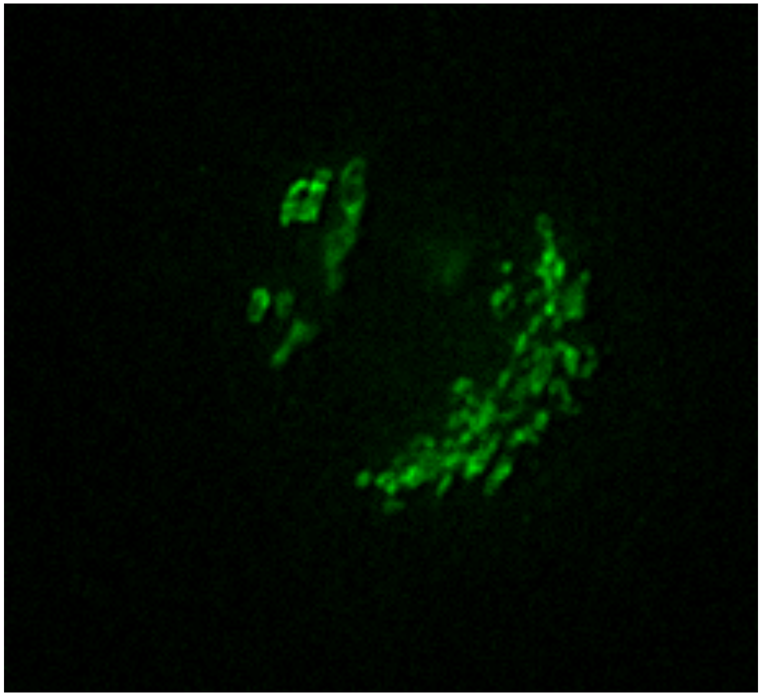
Classic problem in cell and developmental systems biology

- How do we learn and represent
  - sizes and shapes of different cell types
  - number, sizes, shapes, positions of organelles
  - the distribution of proteins across organelles
  - how organelles depend upon each other
  - how any of these vary
    - from cell to cell
    - from cell type to cell type
    - during development
    - in presence of perturbagens

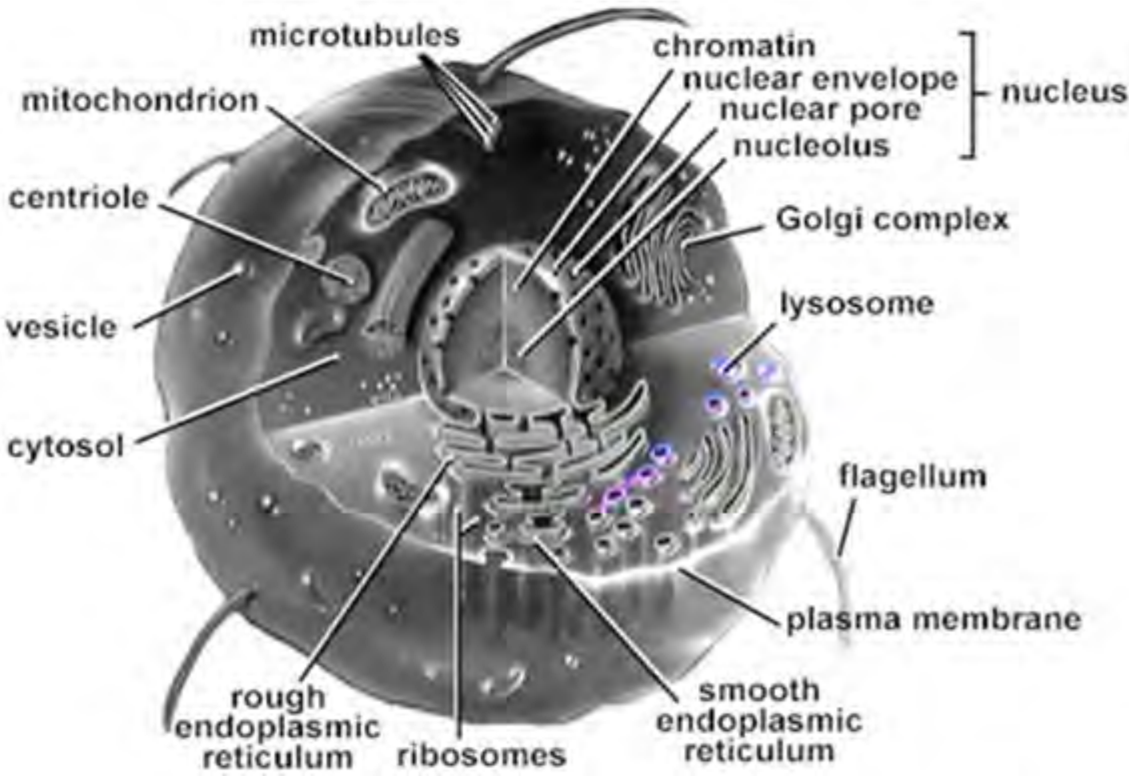


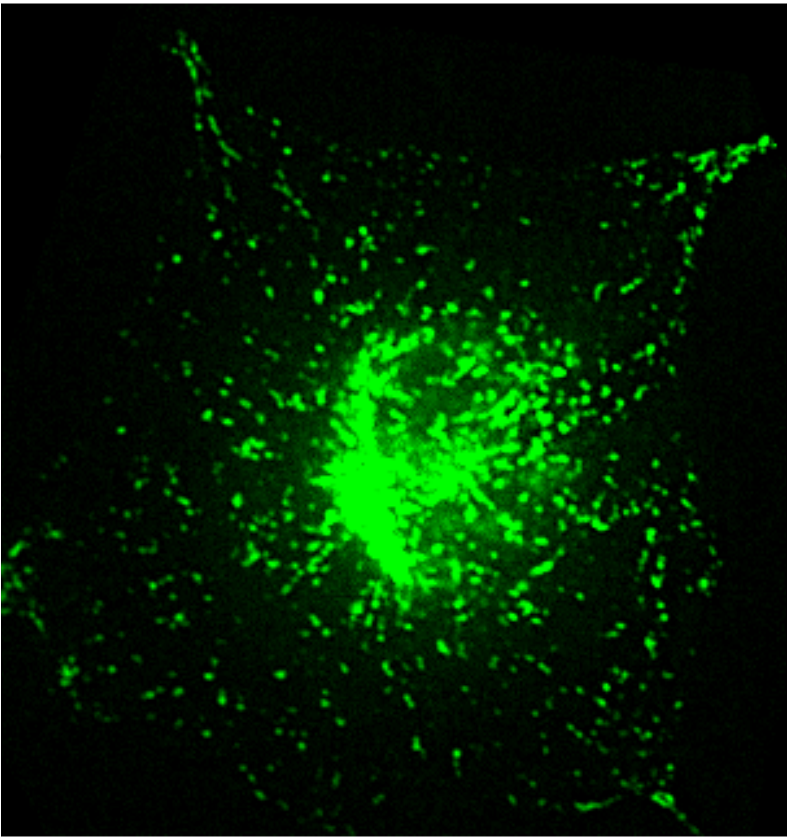
Subcellular Location



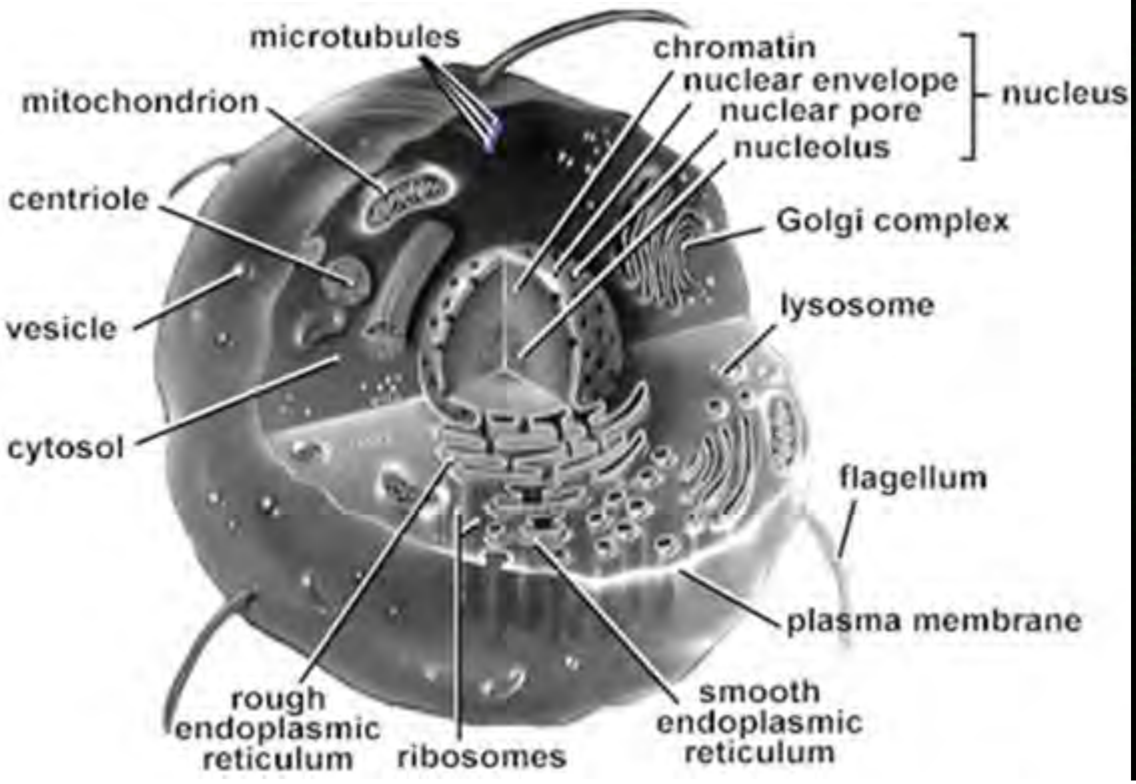


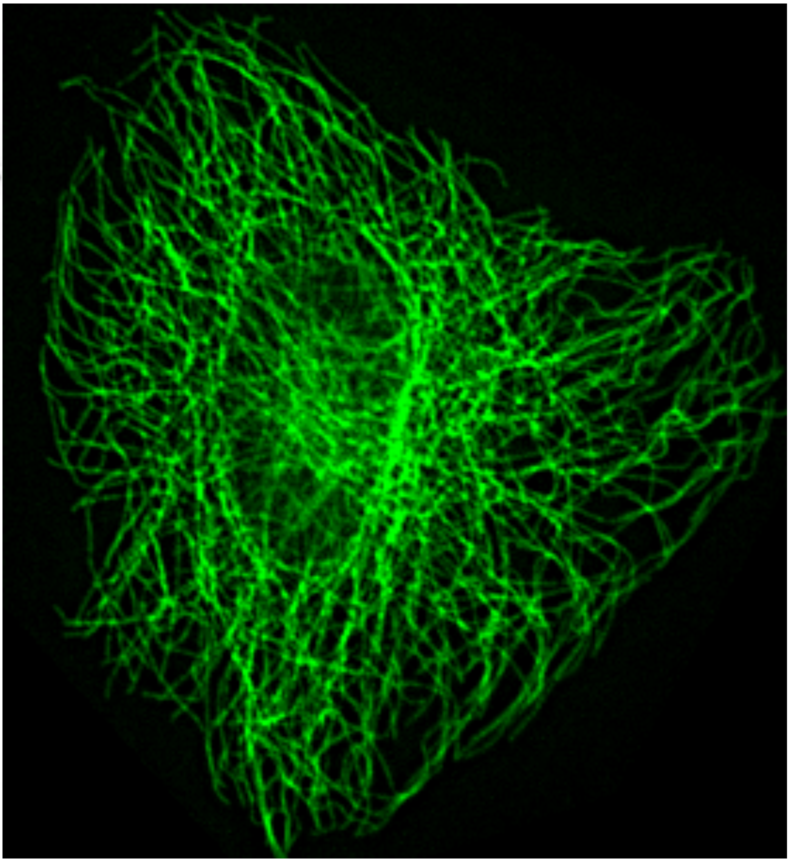
Subcellular Location



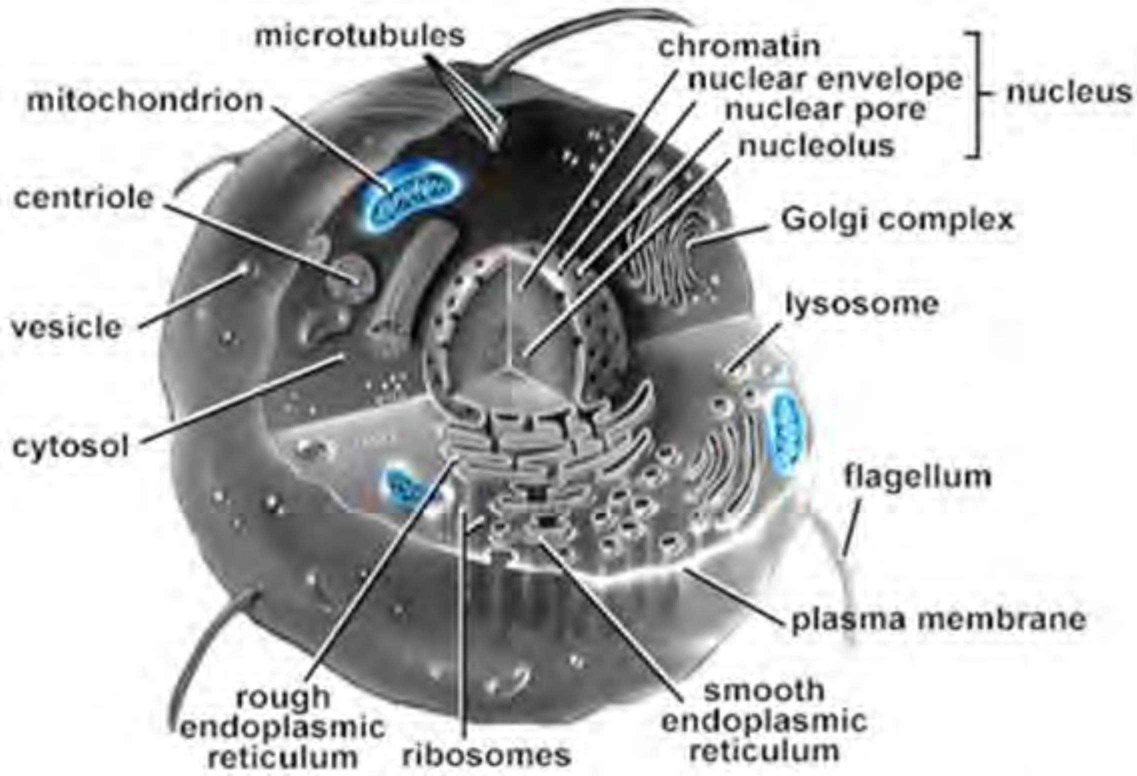


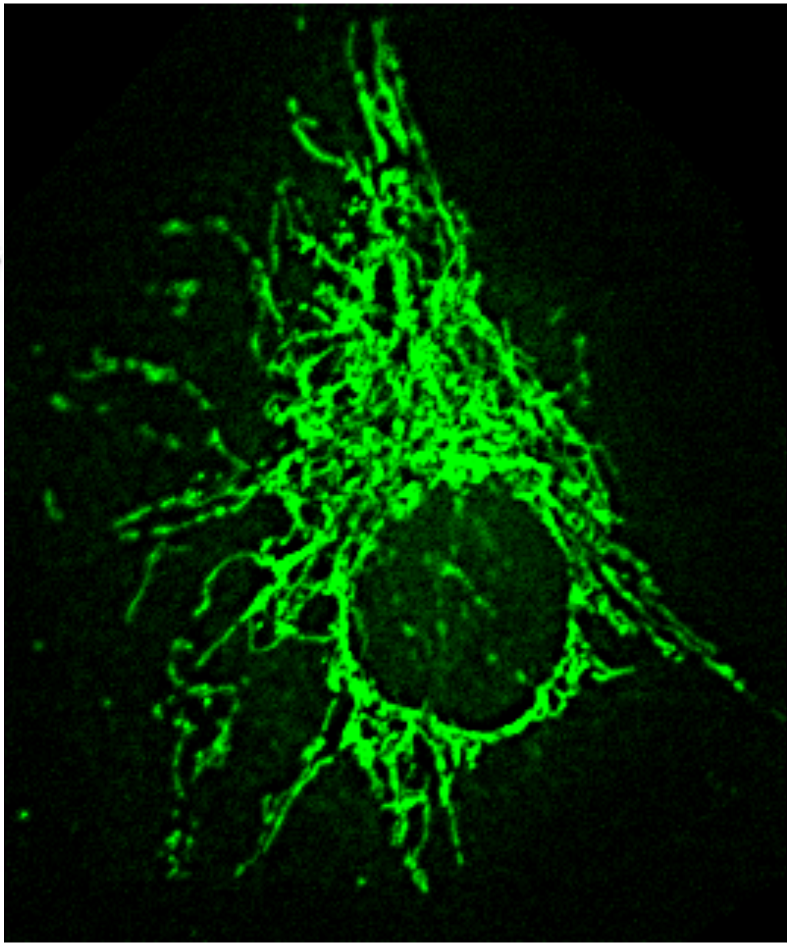
Subcellular Location





Subcellular Location

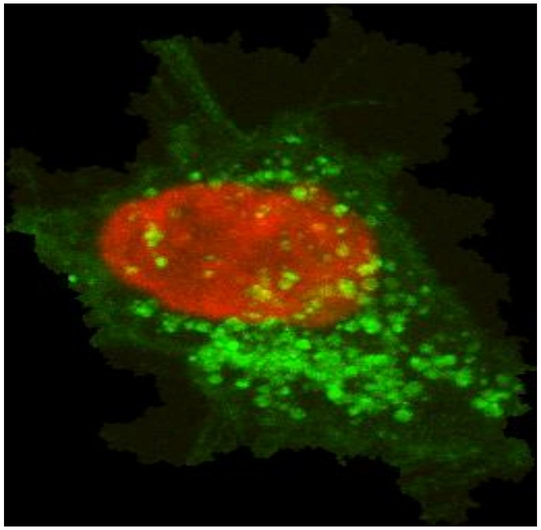




Classic approach

- Do biochemical or imaging experiments, capture relationships in words
    - “secretory vesicles bind to microtubules”
- Two problems
    - Difficult to establish these relationships from images
    - Does not adequately describe them
- Can we do better via machine learning?

Cellular Pattern Recognition

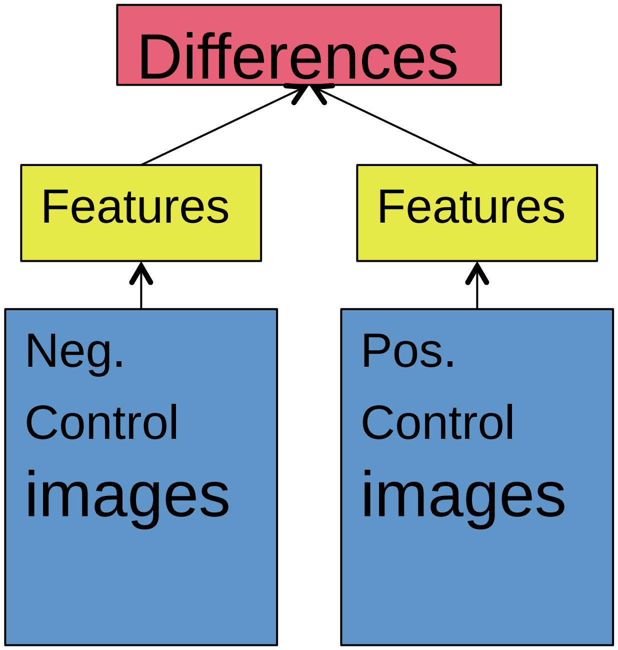


- Describe cell patterns using numerical features
- Do classification, etc. to assign terms
- First described in Boland, Markey & Murphy (1998) and Boland & Murphy (200
- Later popularized in packages such as CellProfiler, WND-CHARM, Ilastik, CellCognition, etc.

Drawback

- Image features are typically not transferable across images from different sources (widefield vs. confocal vs. superresolution, differences in magnification or camera pixel size, pixel bit depth, etc.)

Traditional High Content Screening/ Analysis



Differences

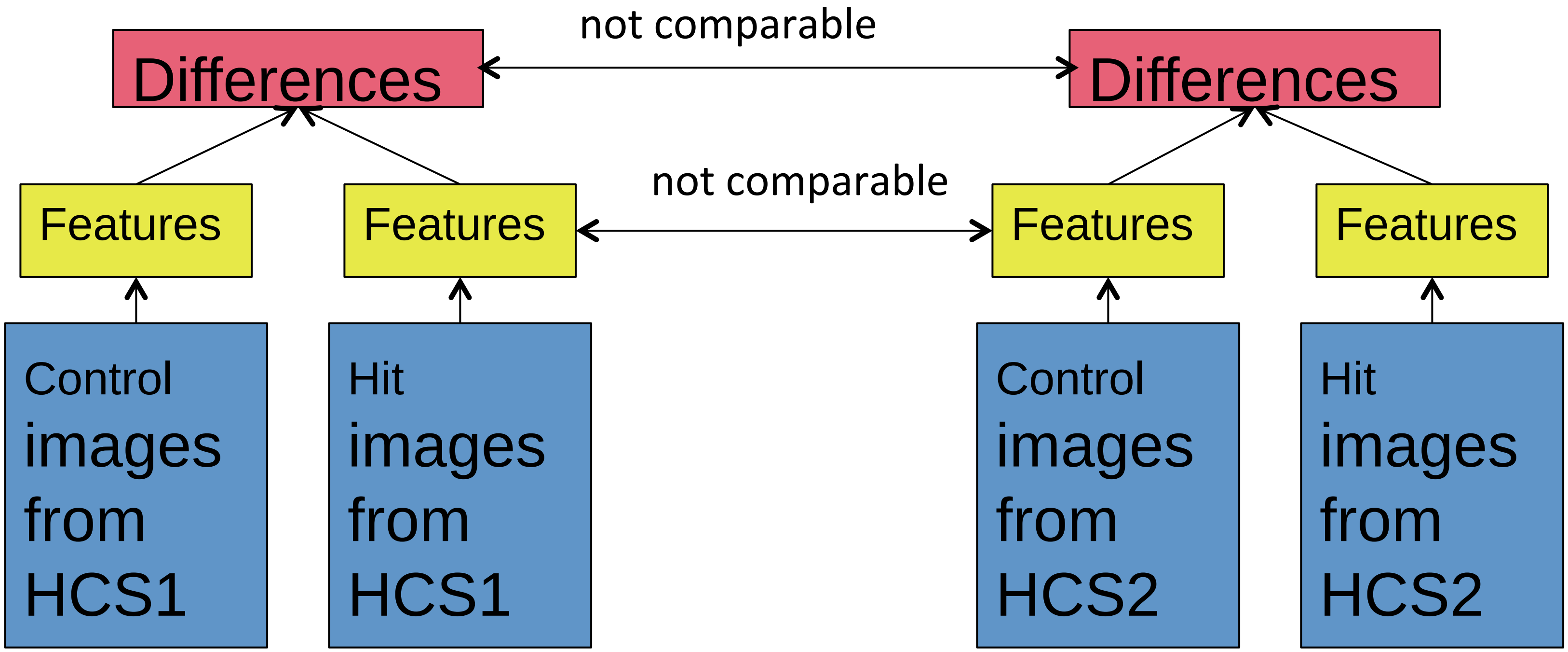
Features

Features

Neg. Control images

Pos. Control images

Different HCS systems or sites



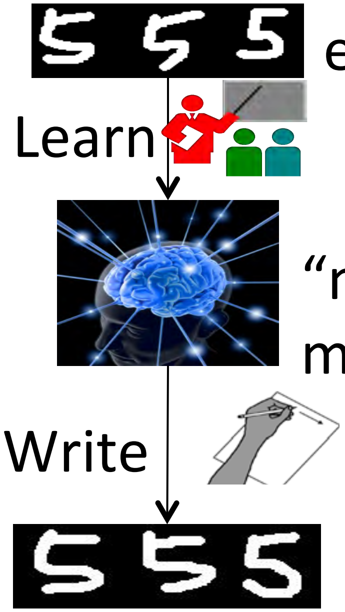
Traditional HCS design focused on finding hits *within* a given screen, not on comparing results *between* screens or learning *generalizable* effects

Another drawback

- Term assignment/classification approaches are incomplete and do not make full use of information in images
- “Is this an apple or an orange?” is a *discriminative* question; can be answered with 1 or 2 features
- “What does an apple look like?” requires a *generative model*

Generative models?

- Human cognition



examples

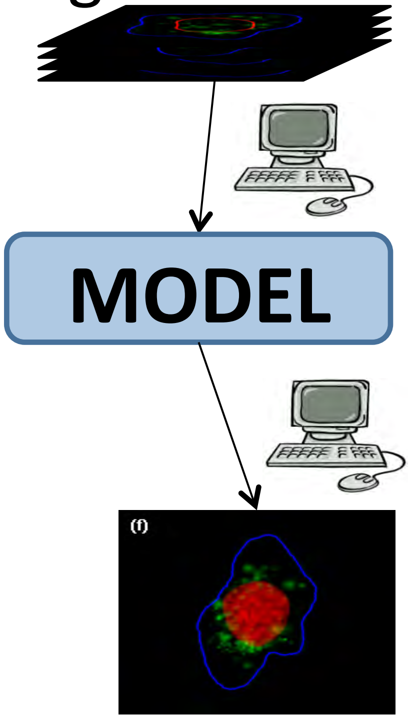
Learn

“mental model”

Write

Generated examples

- Image-based models



Training images

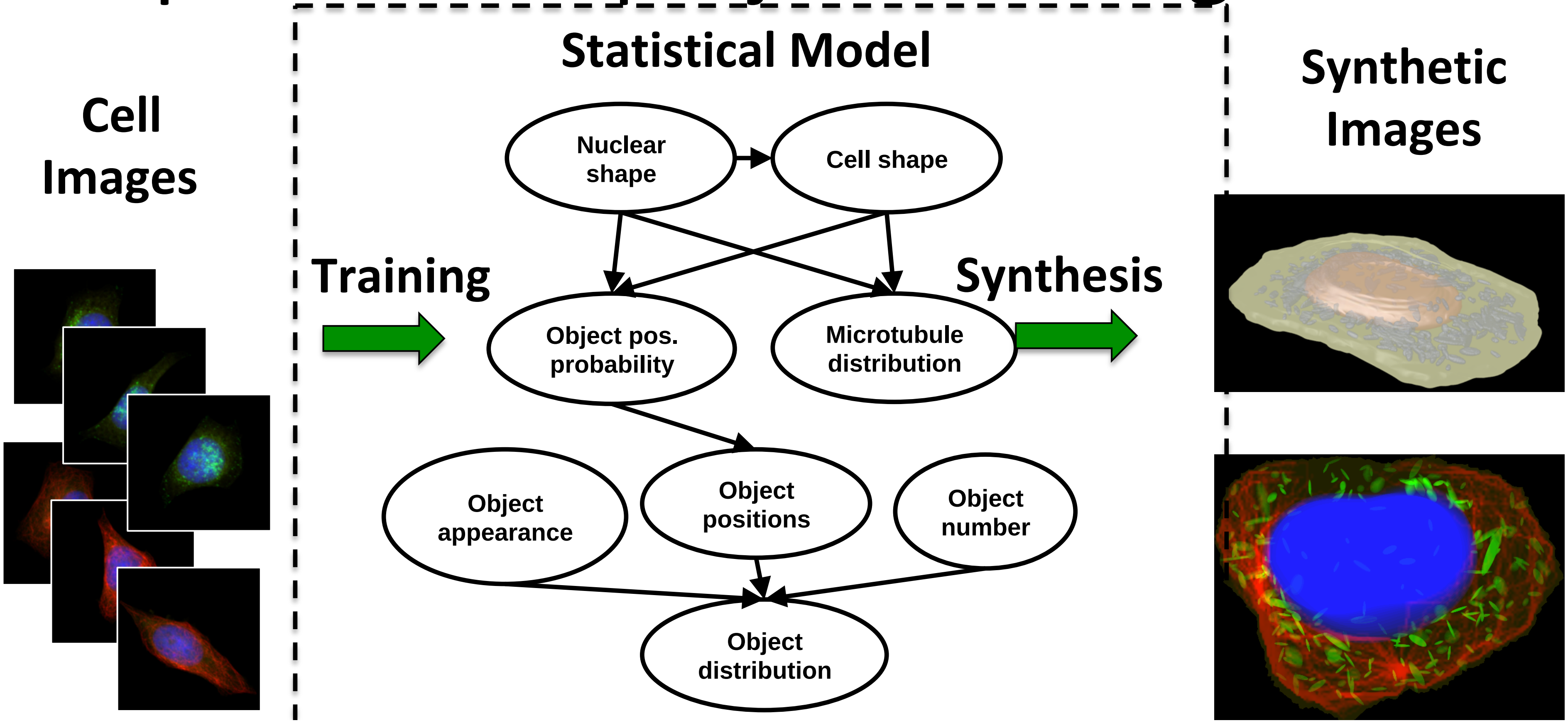
MODEL

Statistical generative model

Generated image

Zhao & Murphy, Cytometry 2007

Open source project: CellOrganizer



Cell Images

Training

Statistical Model

Nuclear shape

Cell shape

Object pos. probability

Microtubule distribution

Object appearance

Object positions

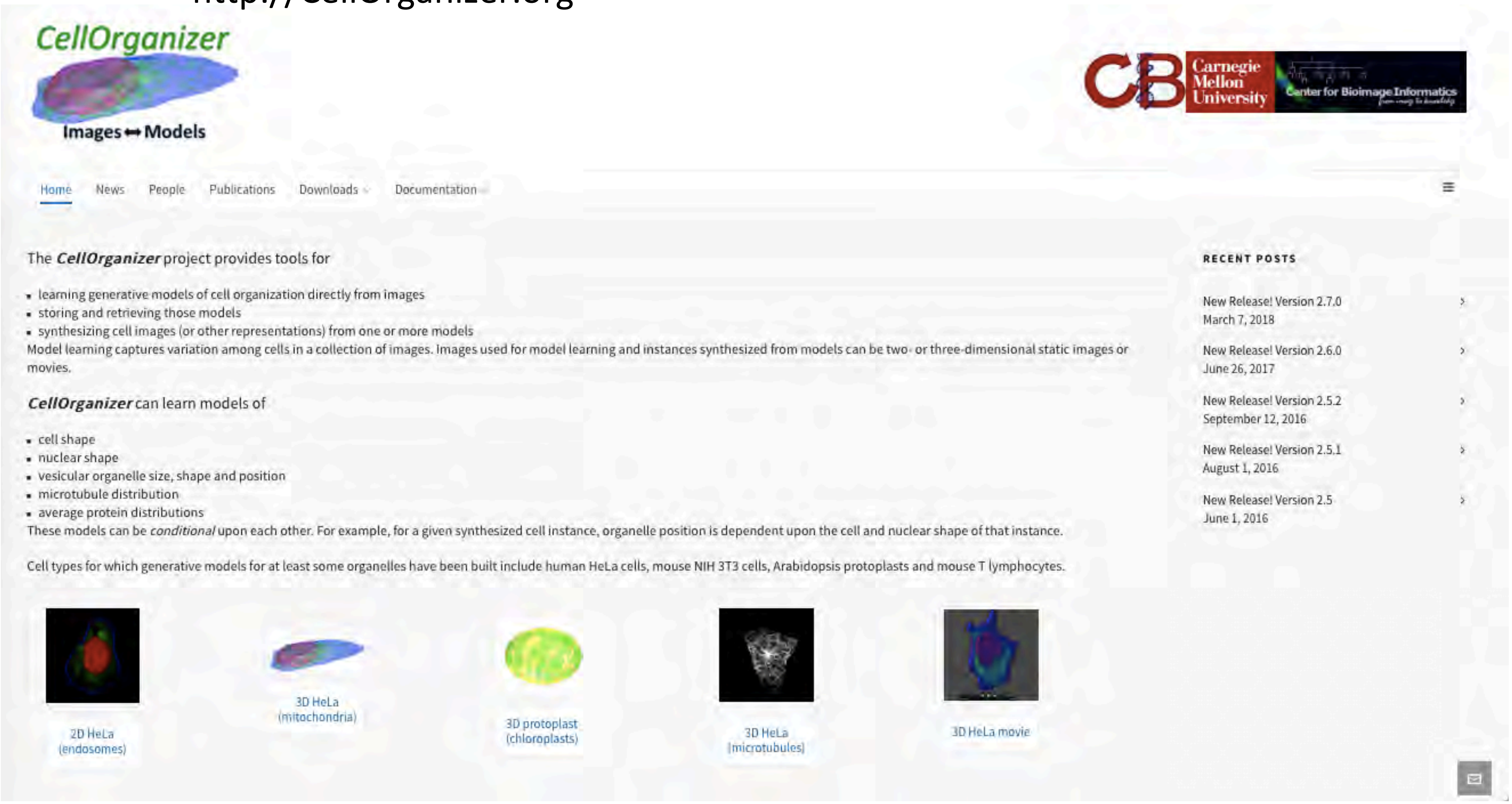
Object number

Object distribution

Synthesis

Synthetic Images

http://CellOrganizer.org



The *CellOrganizer* project provides tools for

- learning generative models of cell organization directly from images
- storing and retrieving those models
- synthesizing cell images (or other representations) from one or more models

Model learning captures variation among cells in a collection of images. Images used for model learning and instances synthesized from models can be two- or three-dimensional static images or movies.

CellOrganizer can learn models of

- cell shape
- nuclear shape
- vesicular organelle size, shape and position
- microtubule distribution
- average protein distributions

These models can be *conditional* upon each other. For example, for a given synthesized cell instance, organelle position is dependent upon the cell and nuclear shape of that instance.

Cell types for which generative models for at least some organelles have been built include human HeLa cells, mouse NIH 3T3 cells, Arabidopsis protoplasts and mouse T lymphocytes.

RECENT POSTS

New Release! Version 2.7.0
March 7, 2018

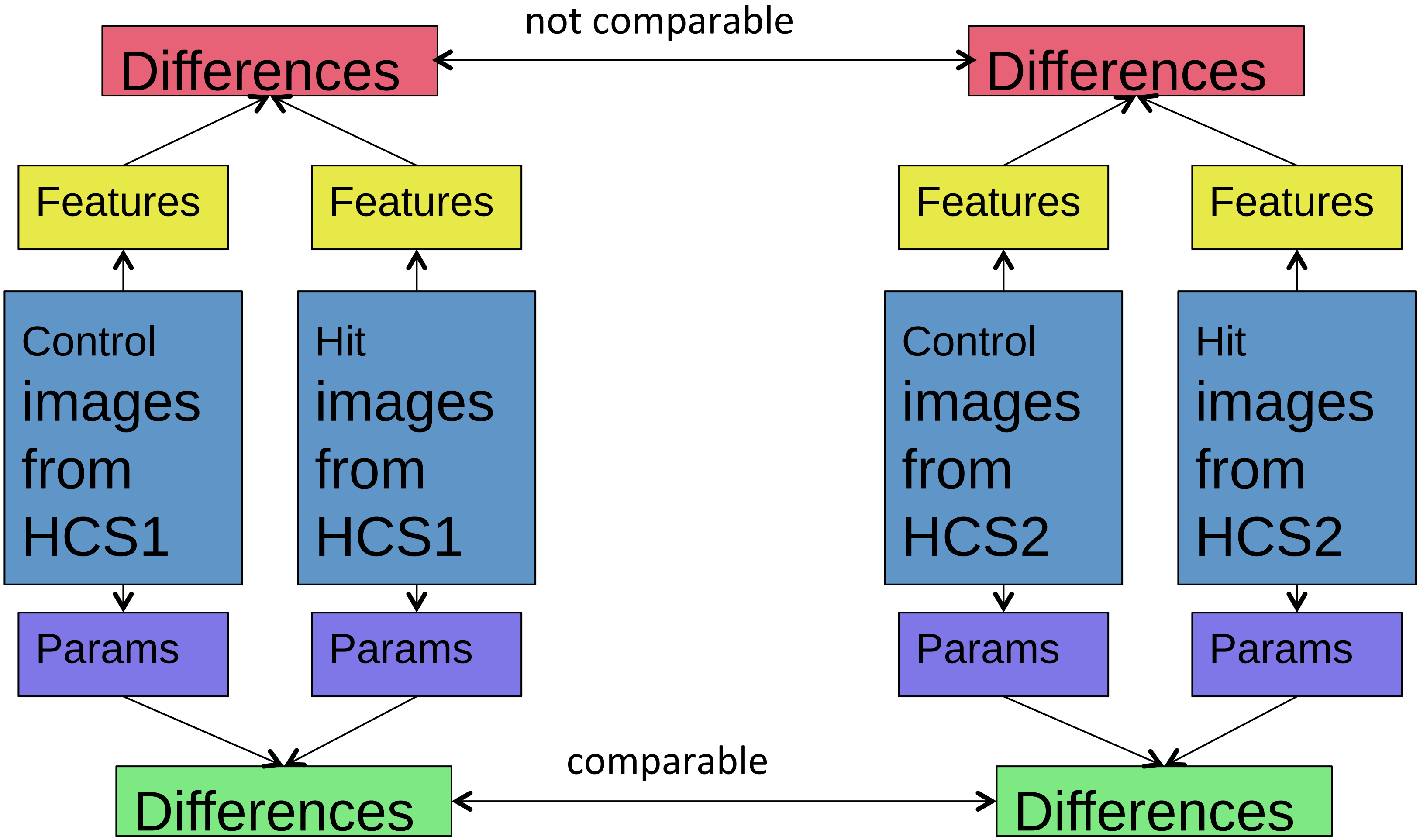
New Release! Version 2.6.0
June 26, 2017

New Release! Version 2.5.2
September 12, 2016

New Release! Version 2.5.1
August 1, 2016

New Release! Version 2.5
June 1, 2016

Generative vs. discriminative HCS



not comparable

Differences

Differences

Features

Features

Features

Features

Control images from HCS1

Hit images from HCS1

Control images from HCS2

Hit images from HCS2

Params

Params

Params

Params

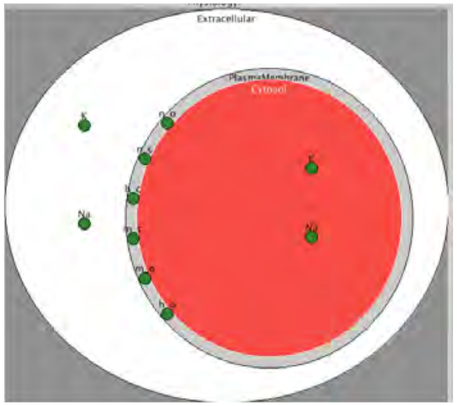
comparable

Differences

Differences

Compartmental models for cell simulations

- Use the assignments to put each protein "in" its compartment and
  - use a cartoon compartmental model
  - use real image to determine compartment volume/surface area
  - use PDEs for each pixel of a real image
- These geometries are not very realistic



Zhao & Murphy, Cytometry 2007

CellOrganizer modeling goals

- Cell models should be
  - **Automated**: learned directly from images,
  - **Generative**: able to synthesize new examples,
  - **Statistically accurate**: reflect variation from cell to cell,
  - **Compact**: can be communicated with significantly fewer bits than the training data.

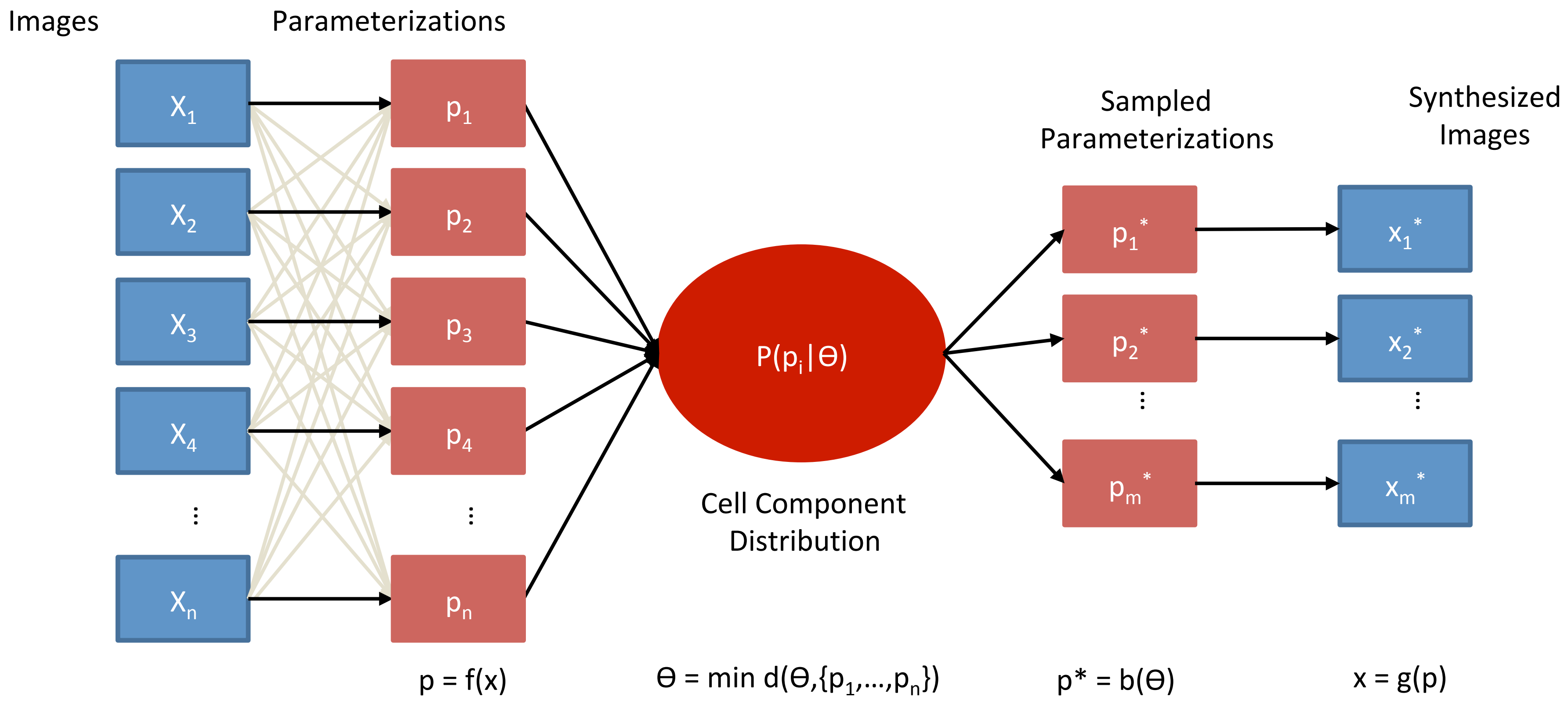
Classical inverse problem

- Learn underlying reality observed via imaging
- Extensive work on image **reconstruction** to create a (higher resolution?) model of a conserved structure (e.g., nuclear pore, ribosome) by *removing noise and variation*
- Our goal is learning *statistical*, ***generative*** model of reality sampled via imaging by removing noise but *keeping* variation

Parametric models

- Computer vision problems such as this have traditionally been tackled by hand-constructing models and learning their parameters from images

Parametric modeling (e.g., CellOrganizer)



Images

Parameterizations

Sampled Parameterizations

Synthesized Images

Cell Component Distribution

$p = f(x)$

$\Theta = \min d(\Theta,\{p_1,\ldots,p_n\})$

$p^* = b(\Theta)$

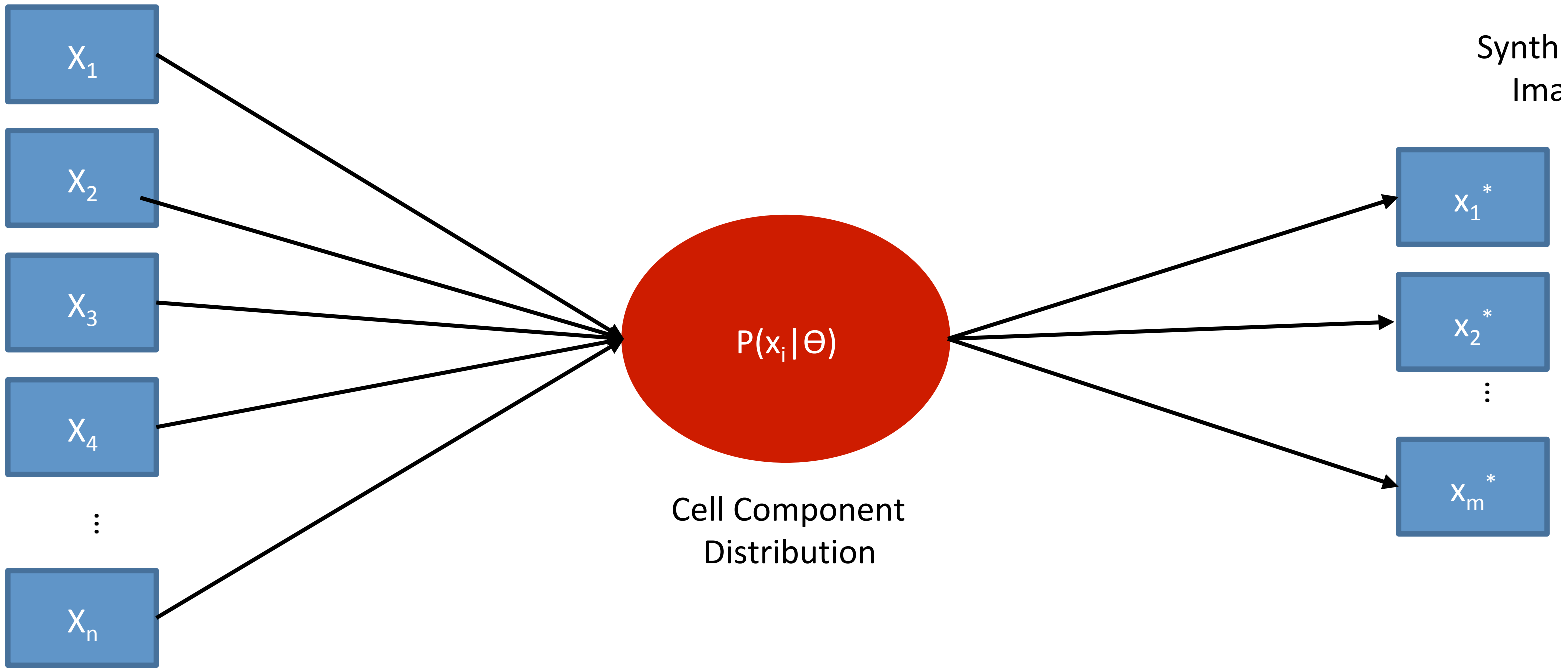
$x = g(p)$

“Deep” learning

- If large numbers of training examples are available, “deep learning” methods can learn directly from images without need for custom design

Deep learning models (e.g., autoencoders)

Images



X_1

X_2

X_3

X_4

⋮

X_n

$P(x_i|\Theta)$

Cell Component Distribution

Synthesized Images

x_1^*

x_2^*

⋮

x_m^*

Θ = min d(Θ,(x-b(Θ))

x = b(Θ)

But…

- Large numbers of “labeled” training images are often not available
- Deep learning models only understand pixels, not structures/objects
  - Not easily compared/combined across diverse images
  - Cells are not made of probabilistic “blobs” of macromolecules
  - Many organelles have discrete boundaries/structures

Challenge

- Fluorescent microscopy provides very useful information about cell organization and processes
- But the number of molecules that can be imaged at the same time in live cells is smaller than the number involved in many processes
- How do we combine information from different images to provide coherent picture?

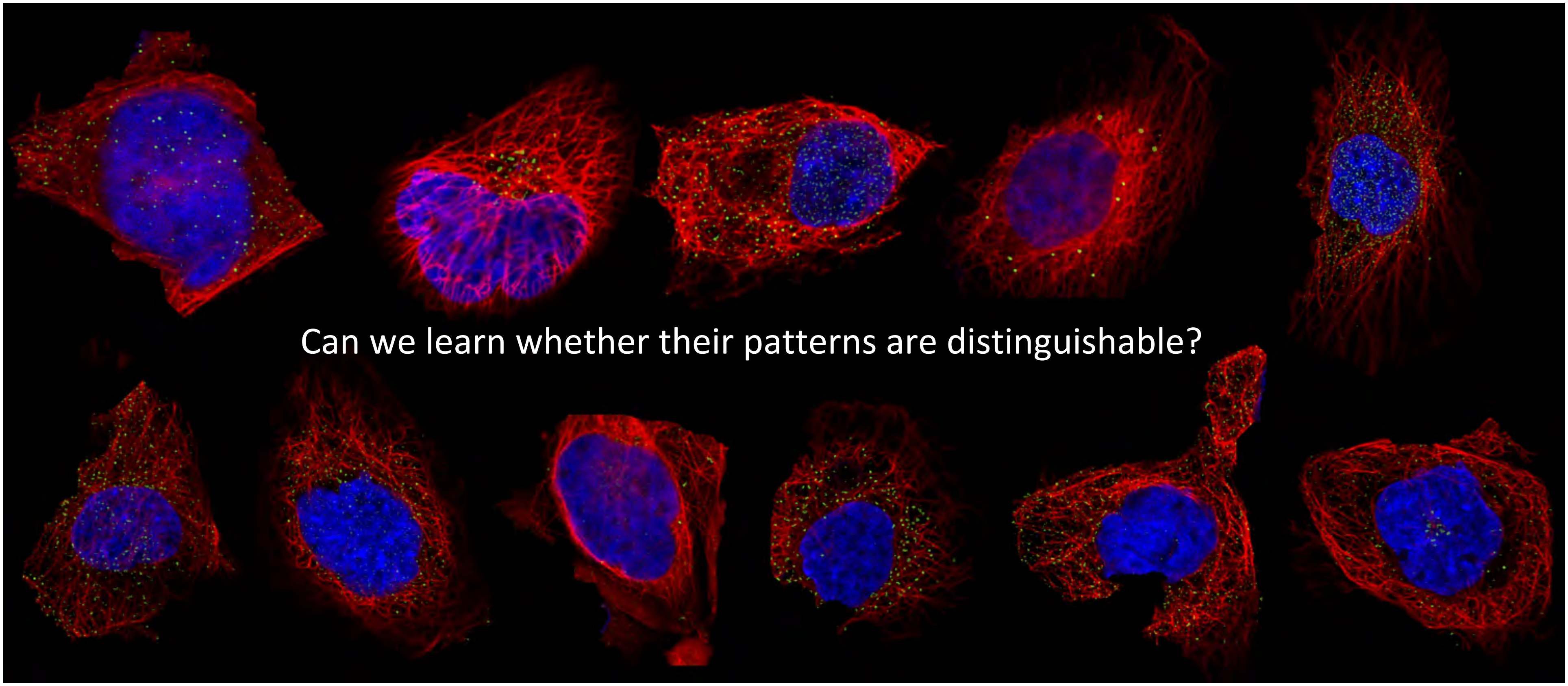
Solution?

- Merging information through generative models built upon a common reference
- Two examples:
  - Distinguishing different punctate structures from separate images
  - Learning potential spatial causal relationships involving in cell signaling

MODELING PUNCTATE ORGANELLE DISTRIBUTIONS

Johnson et al (2015) *PLoS Comp. Biol.*

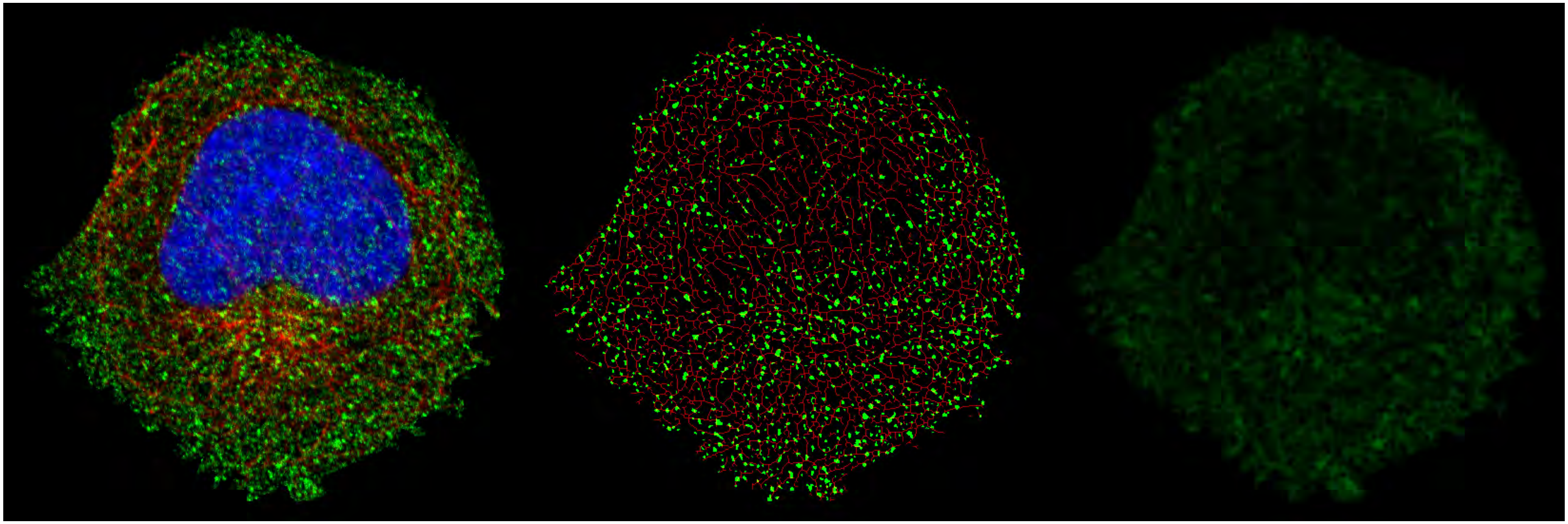
Images of 11 different “vesicle” proteins from Human Protein Atlas



Can we learn whether their patterns are distinguishable?

Segmentation of punctate organelles

- Use high pass filter



Original image

segmented puncta
and microtubules

remaining
fluorescence

Point process models

- Capture relationship between position of an organelle and positions of organelles of different types (**“inhomogeneous Poisson process”**)

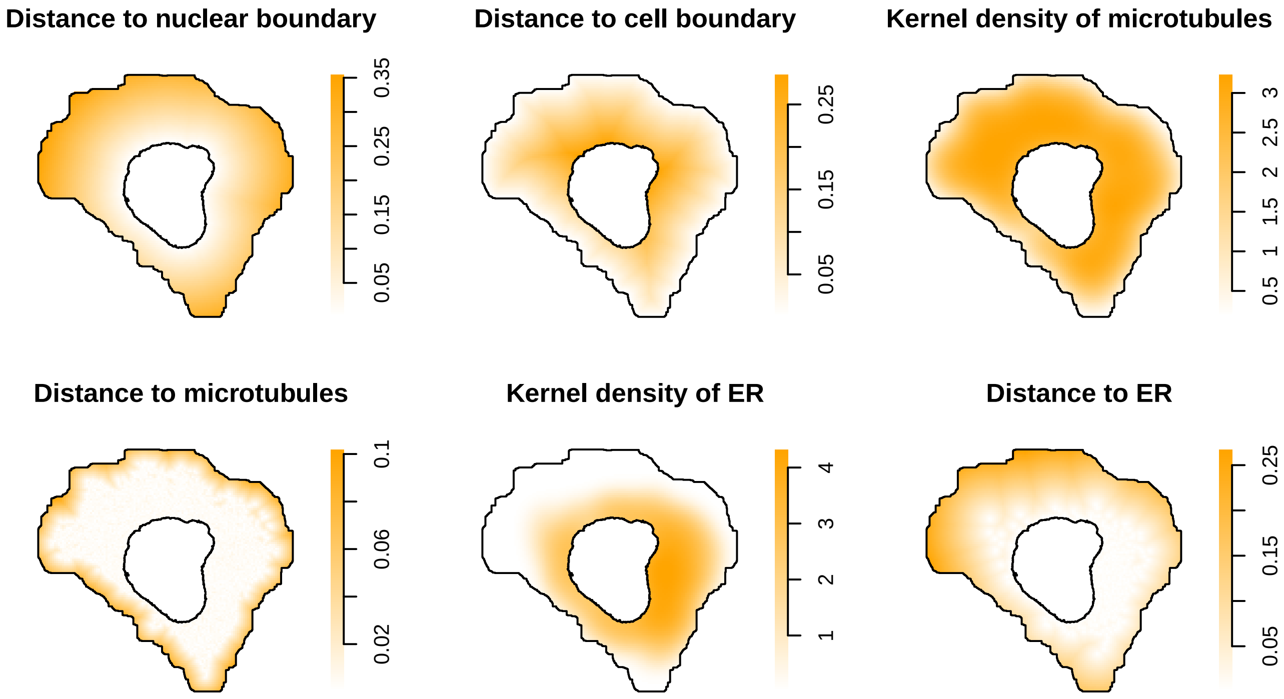
$$f\left(X^{(n)}|n\right) = \frac{1}{Z_\theta}\prod_{i=1}^{n} b_\theta(X_i)$$

- Positions of *n* organelles depend upon b_ϑ functions

Li et al (2016) *Cytometry*

Factors for point process models

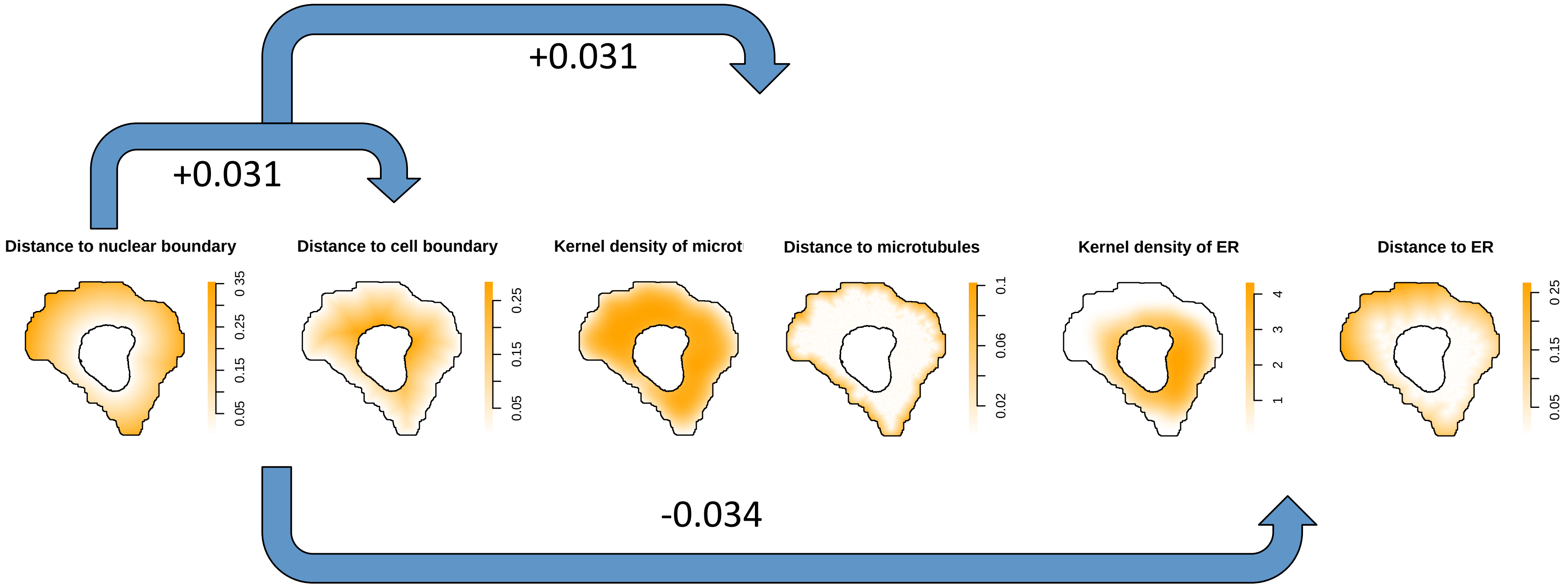
- The functions depend upon specified factors, variables for which values are known at all positions in the cell



Learning dependencies on factors

- An important question is to learn on *which* factors a particular pattern depends
- Can do this by cross-validation: for each combination of factors
  - Estimate parameters from training data
  - Estimate likelihood of test data being generated by that model
  - Average likelihoods

Contributions of different factors



Distance to nuclear boundary

Distance to cell boundary

Kernel density of microt

Distance to microtubules

Kernel density of ER

Distance to ER

How different are the 11 punctate patterns?

- Can also assess by cross-validation (only 2 images available in HPA!)
- Train 11 models using 1 image of each protein
- Assign remaining test image of each protein to the model that it has the highest likelihood of it having been produced by

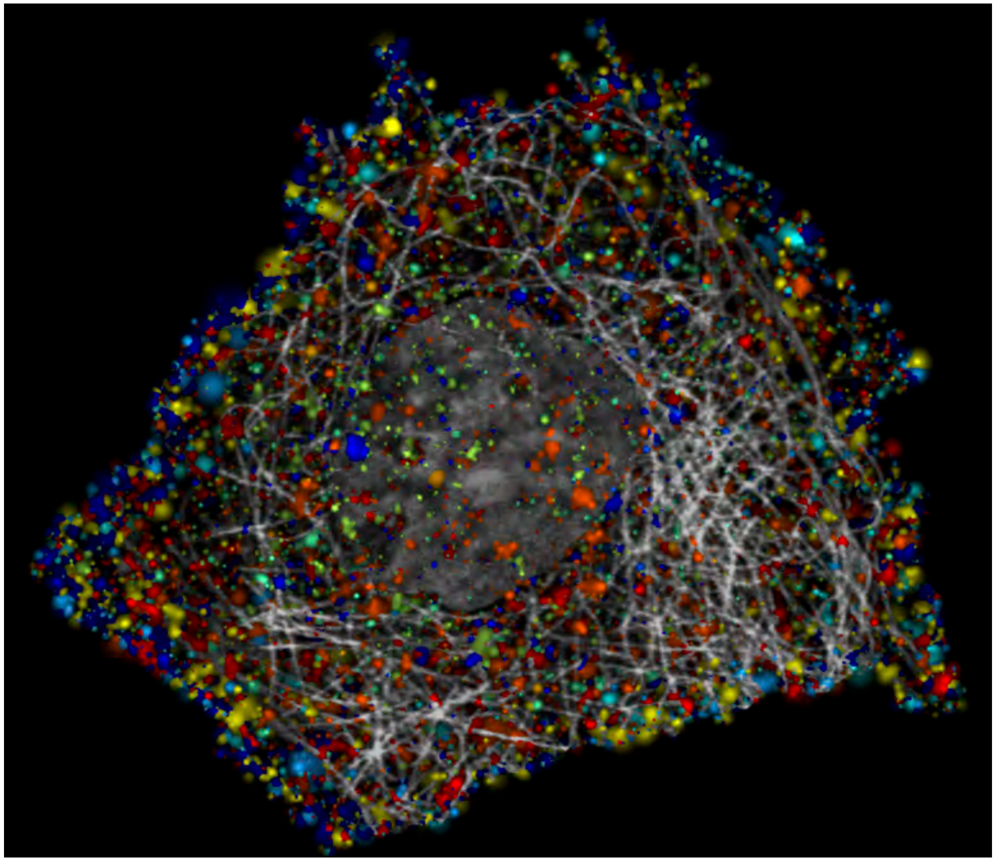
11 distinct punctate patterns using relationship to microtubules

U-251 MG	COPI	COPII	Caveolae	Coated Pits	Early Endosomes	Late Endosomes	Lysosomes	Peroxisomes	RNP bodies	Recycling Endosomes	Retromer
COPI	1	0	0	0	0	0	0	0	0	0	0
COPII	0	1	0	0	0	0	0	0	0	0	0
Caveolae	0	0	1	0	0	0	0	0	0	0	0
Coated Pits	0	0	0	0.67	0	0	0	0	0	0	0.33
Early Endosomes	0	0	0	0	1	0	0	0	0	0	0
Late Endosomes	0	0	0	0	0	1	0	0	0	0	0
Lysosomes	0	0	0	0	0	0	1	0	0	0	0
Peroxisomes	0.08	0	0	0	0	0	0	0.77	0	0.08	0.08
RNP bodies	0	0	0	0	0	0	0	0	1	0	0
Recycling Endosomes	0	0	0	0	0	0	0	0	0	1	0
Retromer	0	0	0	0	0	0	0	0	0	0	1

Overall accuracy:

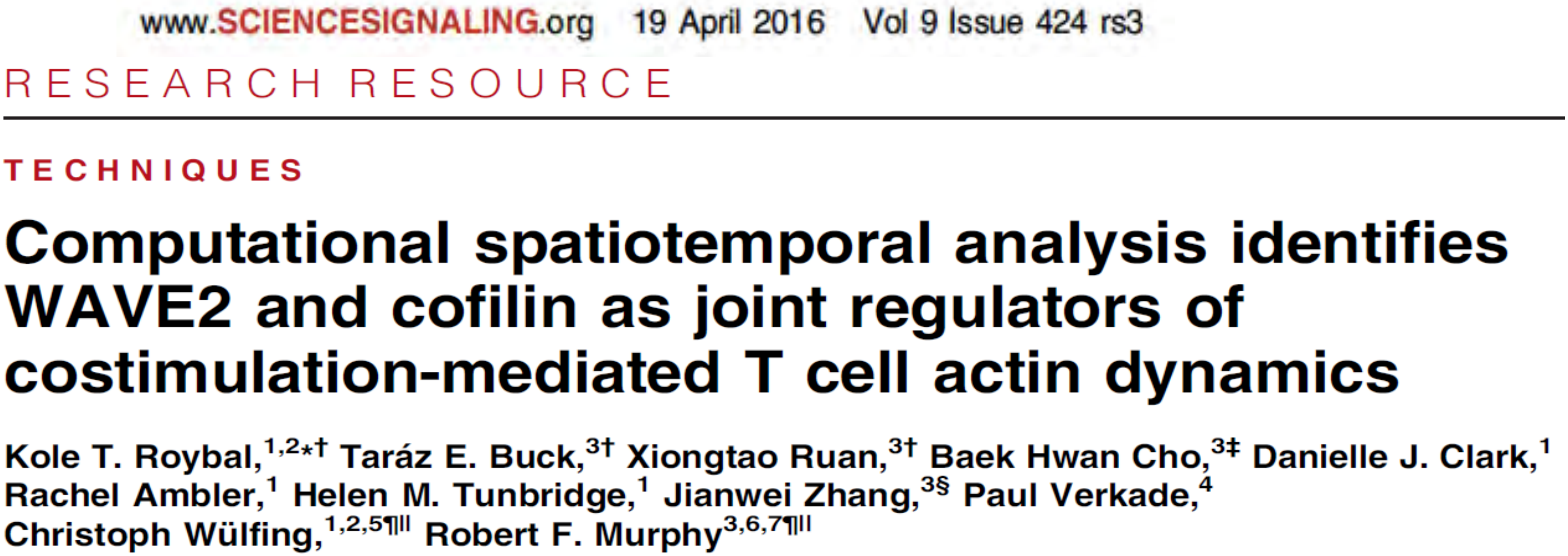
A-431	0.73
U-2 OS	0.90
U-251 MG	0.86

Example synthetic cell image with 11 punctate organelles



MODELING SUBCELLULAR DISTRIBUTION CHANGES DURING CELL SIGNALING

University of BRISTOL



www.SCIENCESIGNALING.org 19 April 2016 Vol 9 Issue 424 rs3

RESEARCH RESOURCE

TECHNIQUES

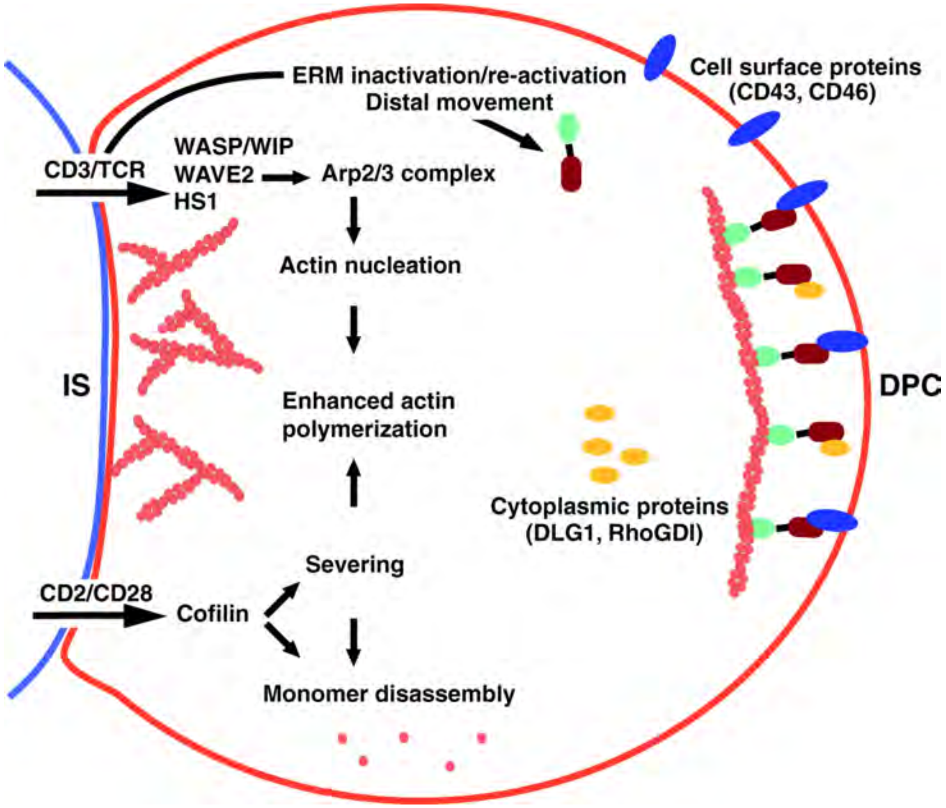
Computational spatiotemporal analysis identifies WAVE2 and cofilin as joint regulators of costimulation-mediated T cell actin dynamics

Kole T. Roybal,[1,2*†] Taráz E. Buck,[3†] Xiongtao Ruan,[3†] Baek Hwan Cho,[3‡] Danielle J. Clark,[1] Rachel Ambler,[1] Helen M. Tunbridge,[1] Jianwei Zhang,[3§] Paul Verkade,[4] Christoph Wülfing,[1,2,5¶||] Robert F. Murphy[3,6,7¶||]

Background

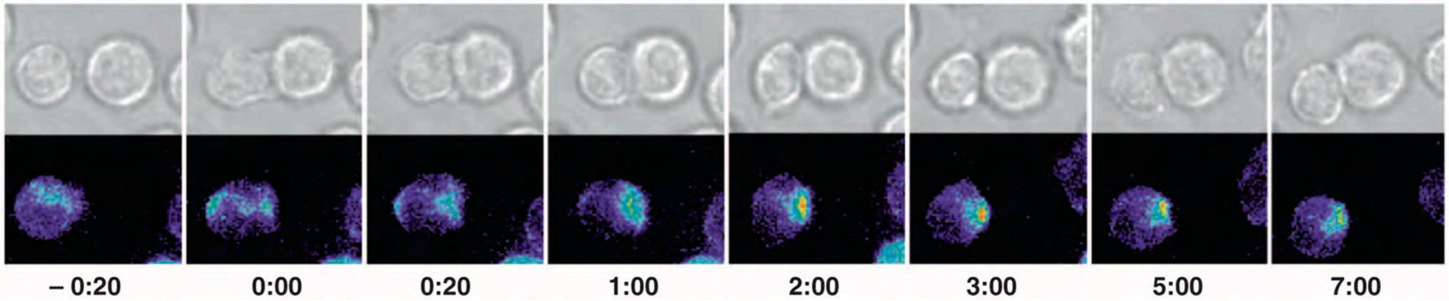
- T cells bind to APC cells triggering stimulation
- Actin and its regulators are recruited to the interacting region.

Question: how do the proteins involved in Actin dynamics regulate each other?



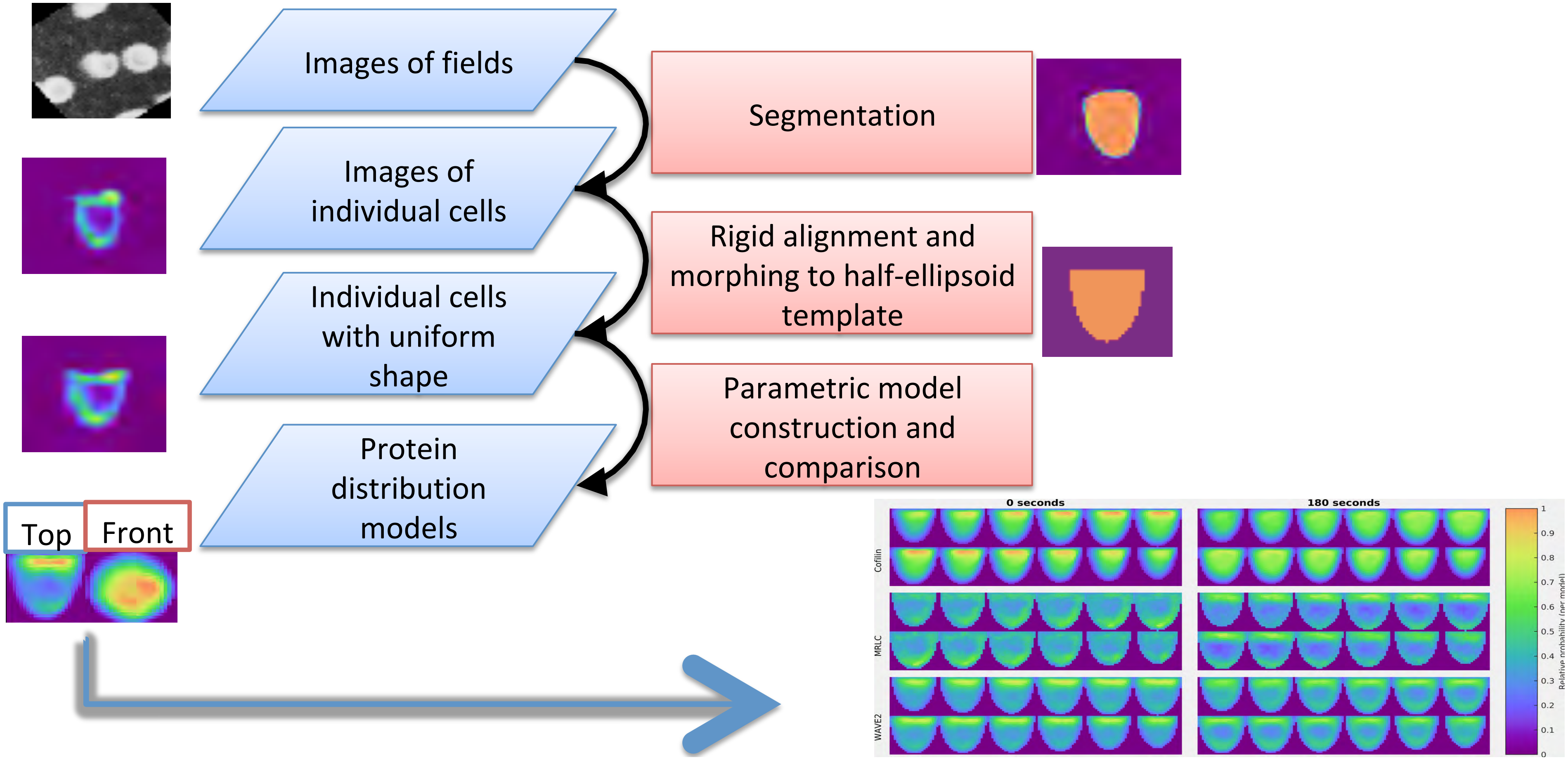
Huang & Burkhardt, 2007

Data



- We start from DIC and fluorescence movies of GFP-tagged proteins at different time points before and after immunological synapse formation (~100 cells per protein)

Image processing pipeline



Images of fields

Images of individual cells

Individual cells with uniform shape

Protein distribution models

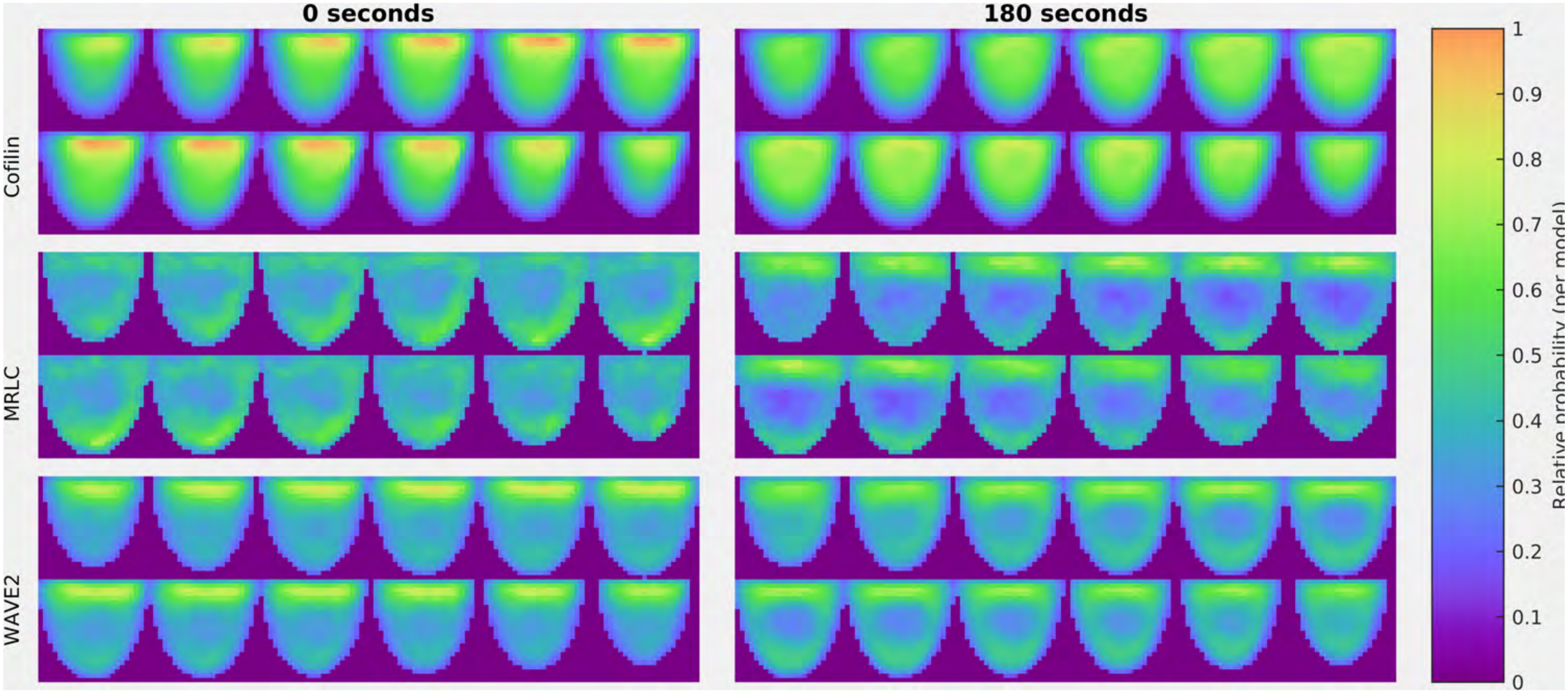
Segmentation

Rigid alignment and morphing to half-ellipsoid template

Parametric model construction and comparison

Summary of the Automatic analysis

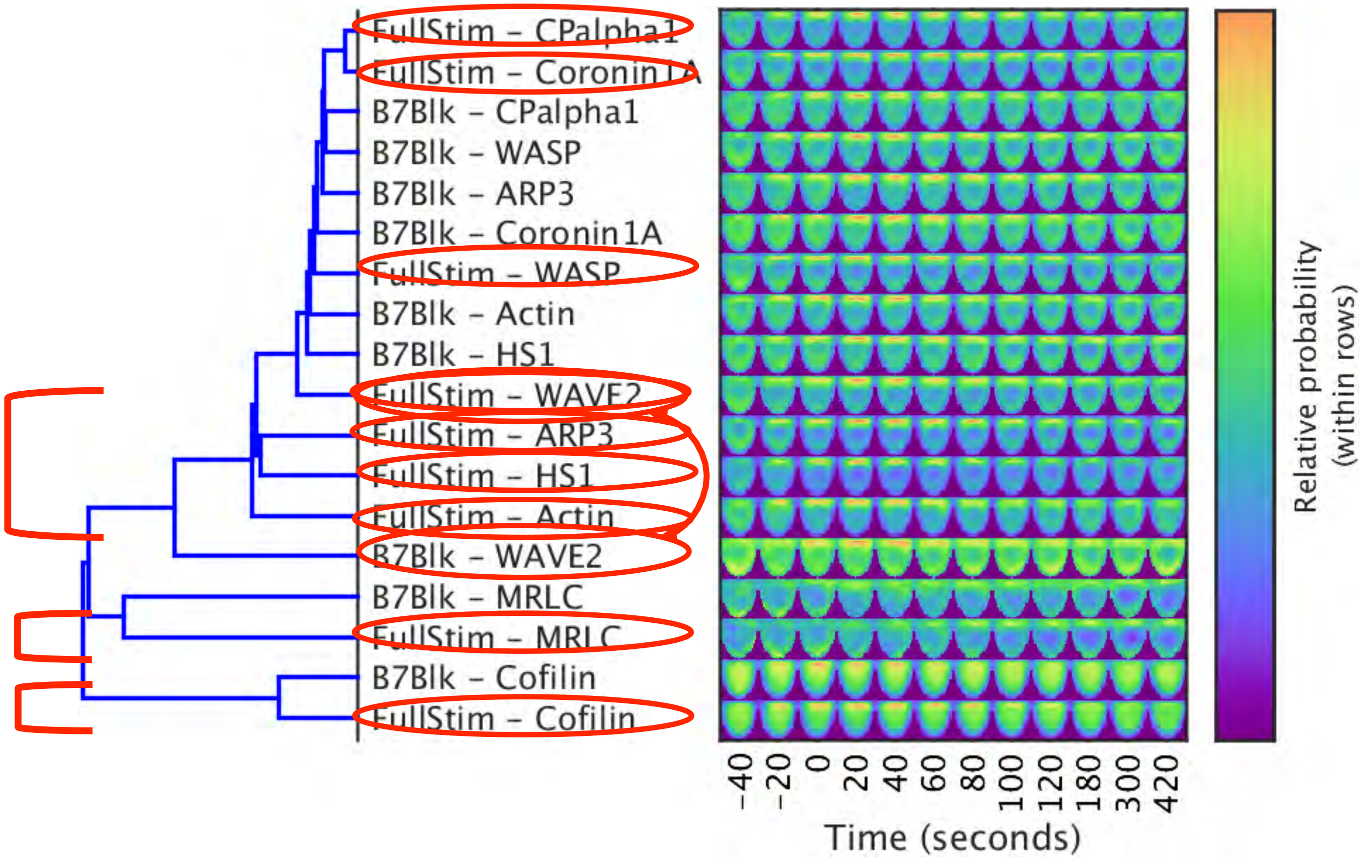
- More than 17,000 cell pairs were analyzed.
- Two conditions: Full stimulus, B7 blockade
- Ten proteins: ARP3, Actin, cofilin, Coronin1A, CPalpha1, HS1, MRLC, WASP, WAVE2, LAT


0 seconds
180 seconds
Cofilin
MRLC
WAVE2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Relative probability (per model)

Applications of the model

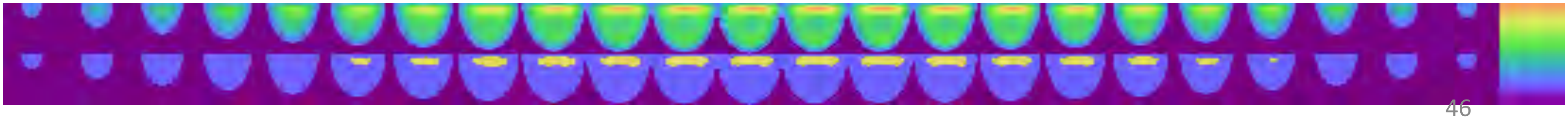
- Use voxel concentrations as features to compare different proteins across all time points across different conditions
- Enrichment analysis: see how proteins accumulate in specific locations over time
- Visualize the spatiotemporal dynamics of selected proteins

Clustering reveals differences

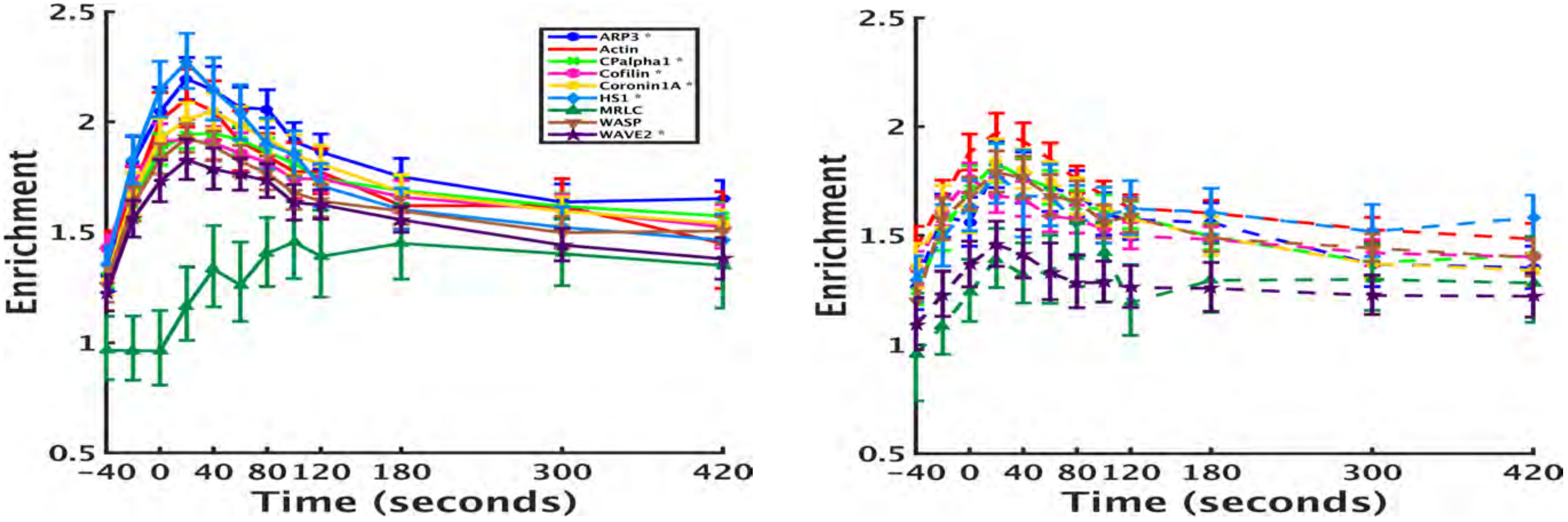


Enrichment analysis

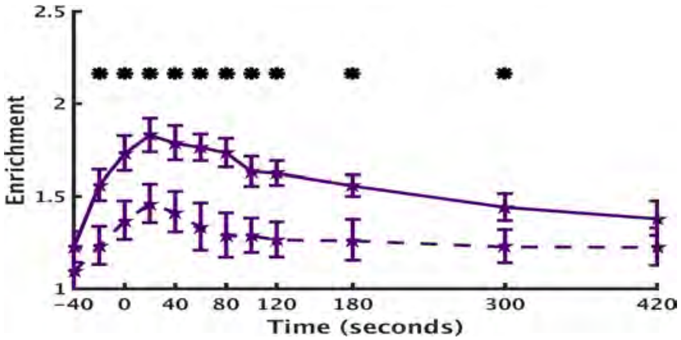
- Actin is recruited to the synapse region, we would like to measure kinetics of recruitment of other proteins
- Idea: define an enrichment region where it contains the top 90% fluorescence in the average model map for all models.
- The enrichment is defined as the ratio of the mean intensity in the region to the mean intensity out of the region.



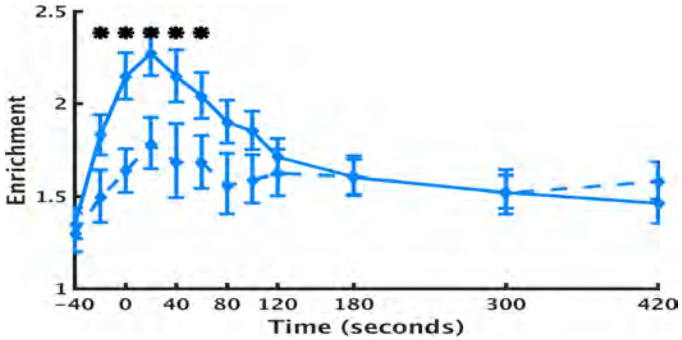
Enrichment of all proteins



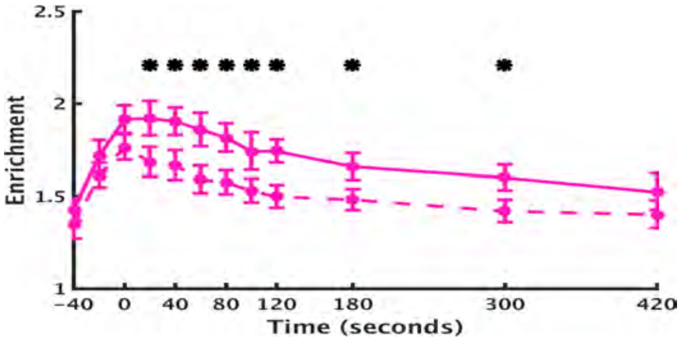
Several proteins have significant enrichment



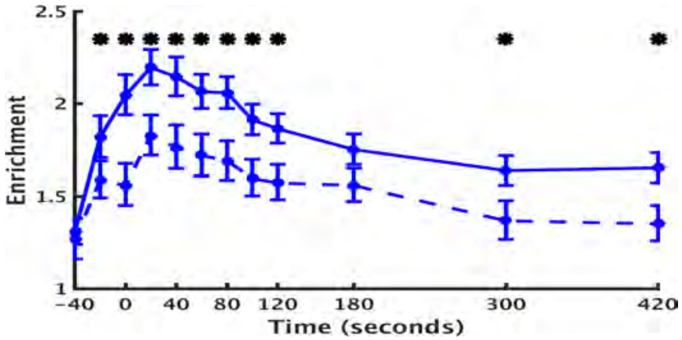
WAVE 2



HS1



Cofilin

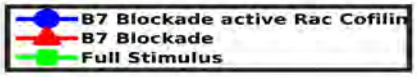


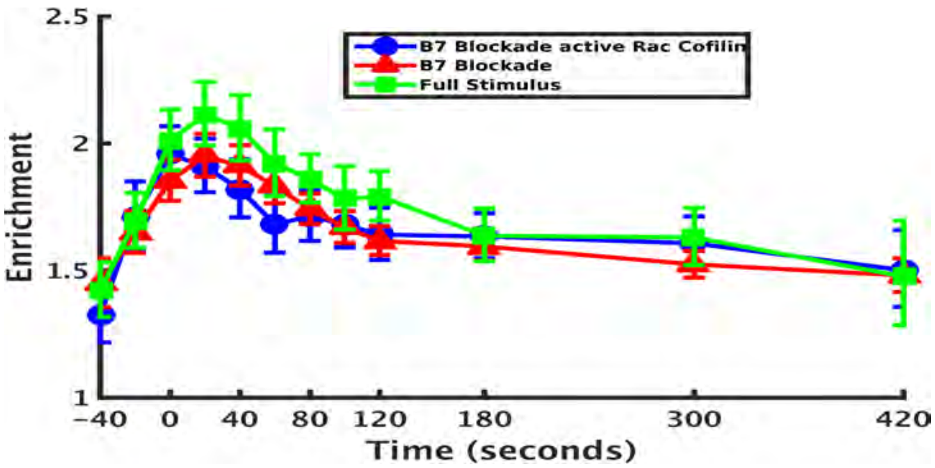
ARP3

Validation of candidate regulators

- We identified Wave2 and cofilin as candidate regulators in costimulation-mediated Actin dynamics.
- Question: could selective activation of these two regulators promote actin dynamics and synapse formation even under the B7 blockade condition?

Reconstruction

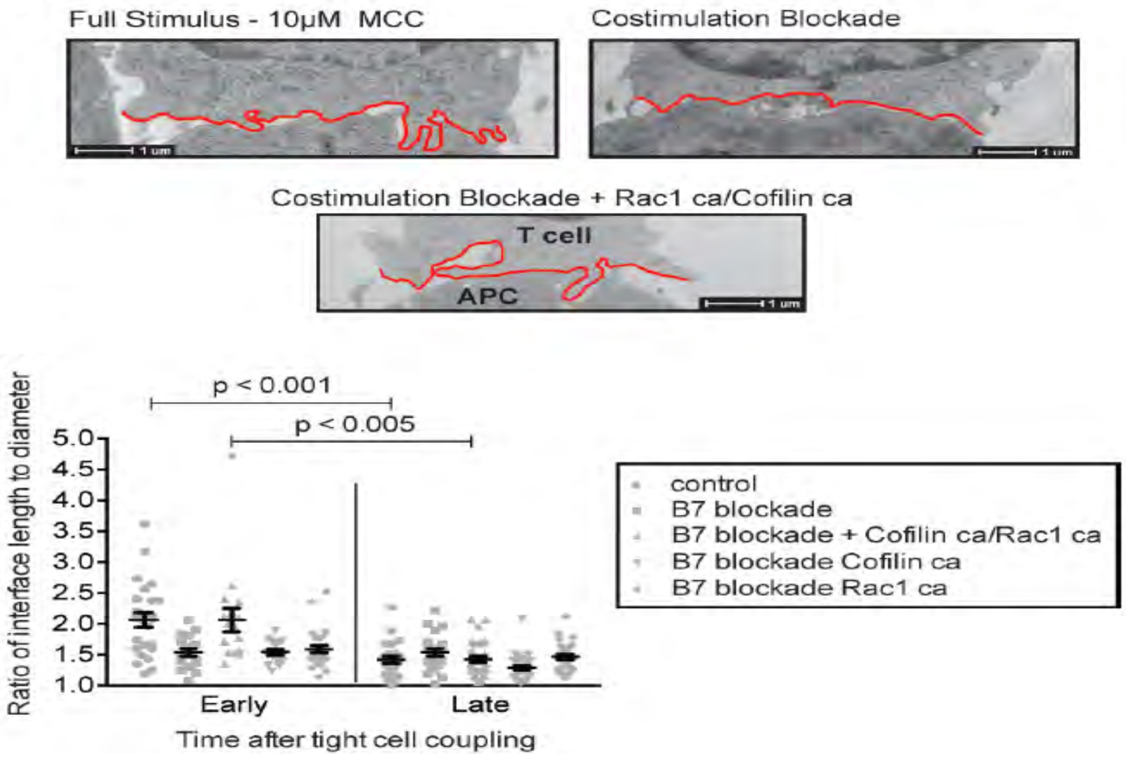




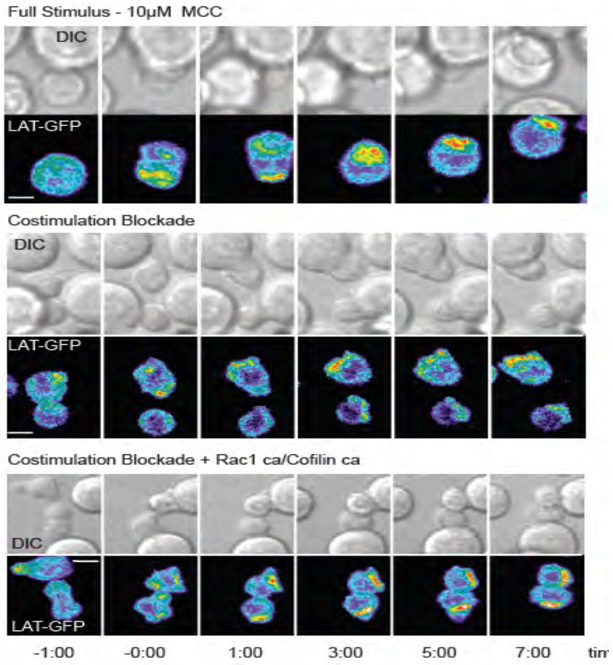
WAVE2

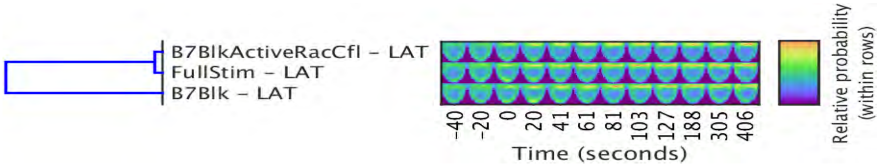
Actin

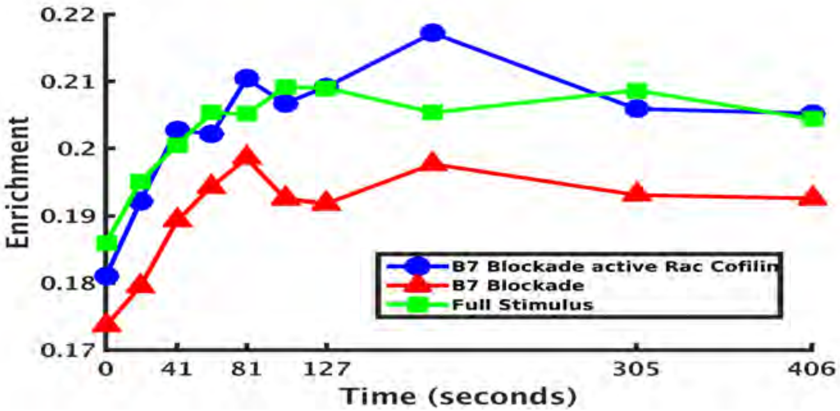
Reconstruction



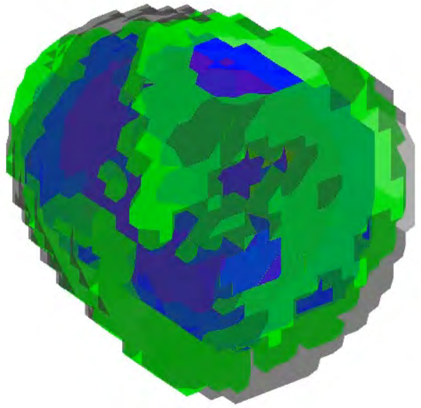
Active Rac and cofilin restore defective LAT location





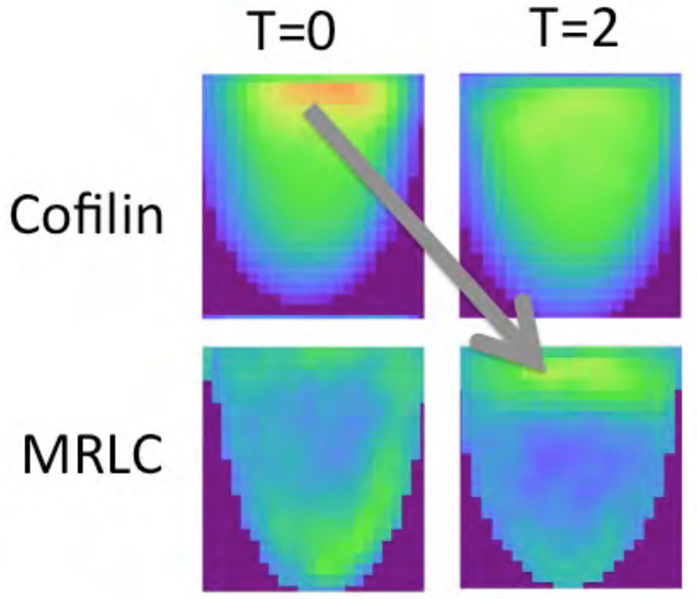


Spatiotemporal distribution of proteins



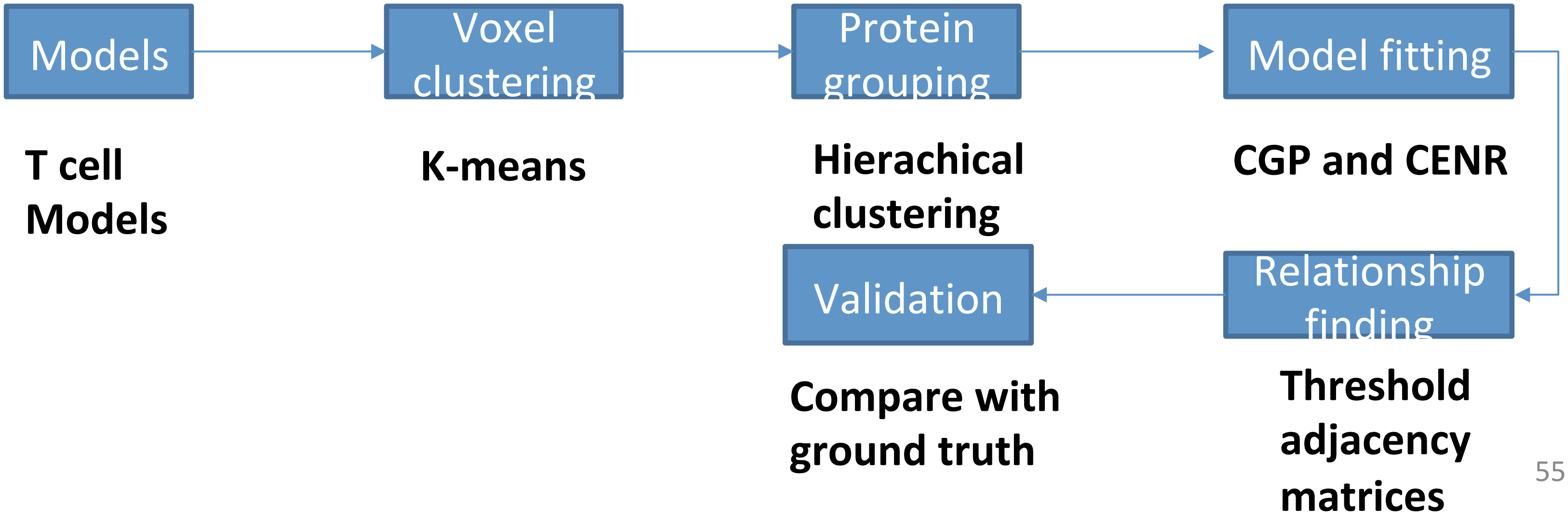
cofilin MRLC WAVE2

Using Spatiotemporal Maps to Learn Putative Regulatory Relationships



- Given spatiotemporal maps for multiple proteins, we sought to determine whether a change in one protein in one region of the cell precedes a change in another protein in another region
- For this we applied methods for learning causal graph process models
- Nodes represent a specific protein at a specific location, edges represent a possible predictive relationship between nodes

Approach



Models

T cell Models

Voxel clustering

K-means

Protein grouping

Hierachical clustering

Model fitting

CGP and CENR

Relationship finding

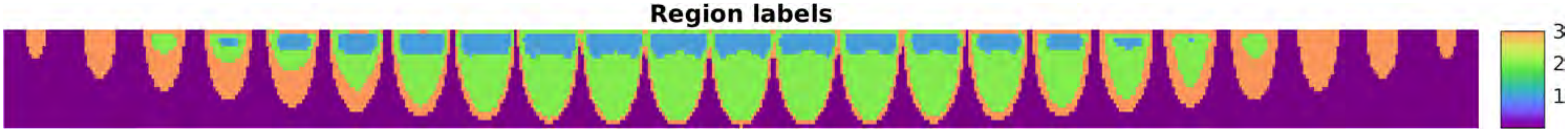
Threshold adjacency matrices

Validation

Compare with ground truth

Reducing graph size

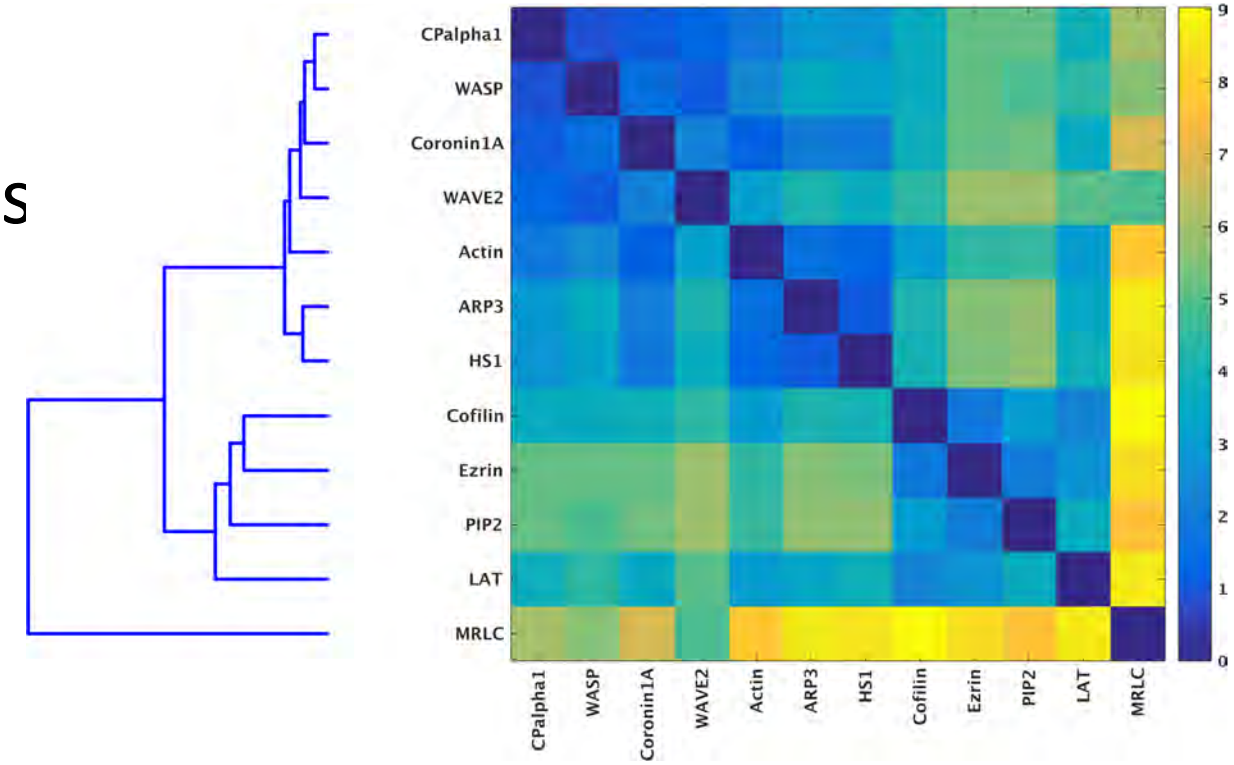
- The spatiotemporal maps have 6628 voxels per cell, and there is one map for each of 12 proteins
- The graph model requires an edge between every pair of nodes: too many edges, need to reduce the number
- Solution: represent each voxel by a vector of the intensities at all time points for all proteins
- Use K-means clustering of voxels to form regions



- Represent each region by the average intensity of all voxels in it

Removing confounding by correlations

- Highly correlated proteins/regions allow self prediction
- Collapse them into one representative



Causal graph process model

Let *x*[*t*] be concentration of all proteins in all regions

Goal is to find single model to predict all times

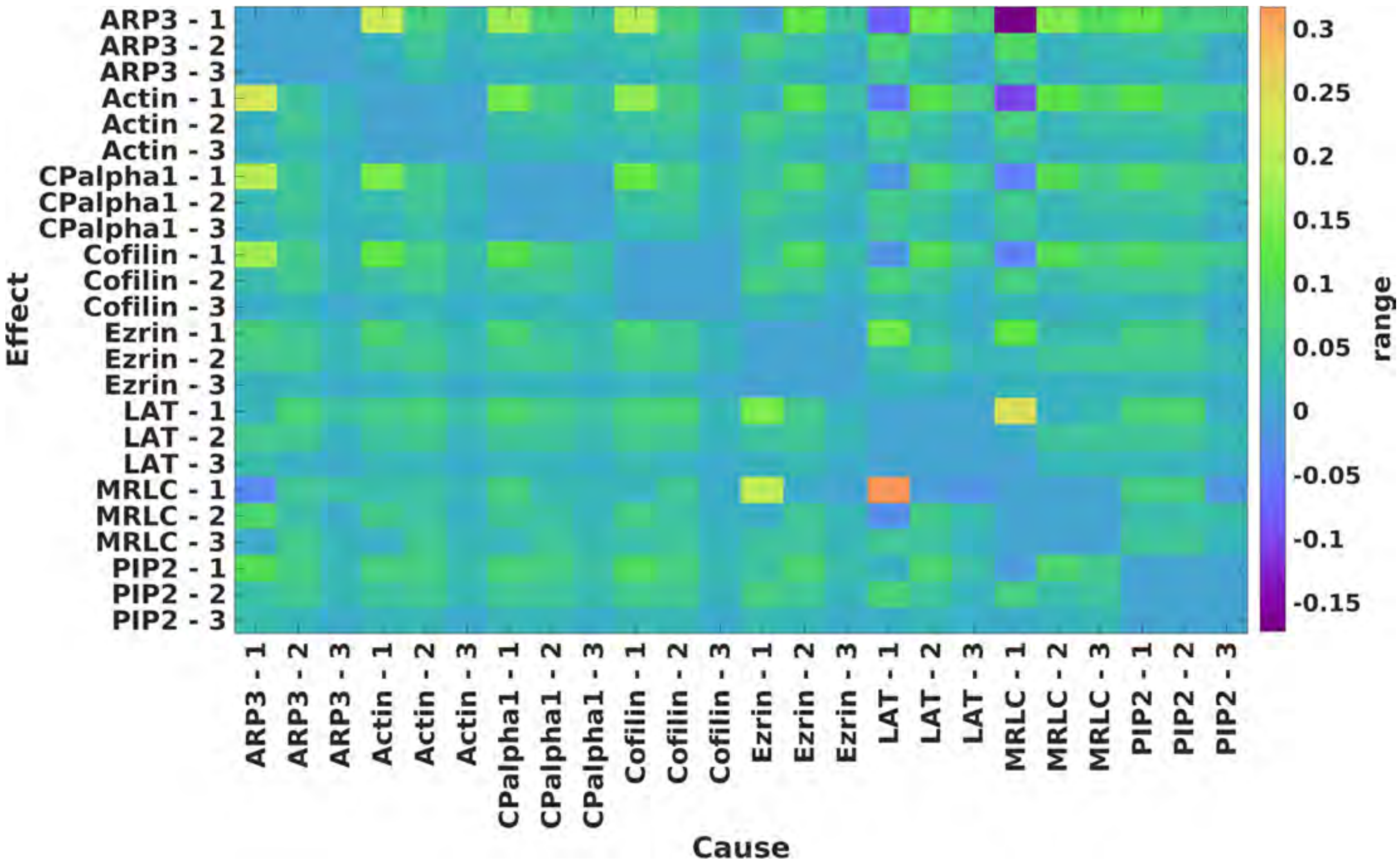
$$
\begin{aligned}
x[t] &= w[t] + \sum_{i=1} P_i(\mathbf{A})x[t-i] \\
&= w[t] + (c_{10}\mathbf{I} + c_{11}\mathbf{A})x[t-1] \\
&+ \ (c_{20}\mathbf{I} + c_{21}\mathbf{A} + c_{22}\mathbf{A}^2)x[t-2] + \cdots \\
&+ \ (c_{M0}\mathbf{I} + c_{M1}\mathbf{A} + \cdots + c_{MM}\mathbf{A}^M)x[t-M]
\end{aligned}
$$

Mei & Moura, 2015

We also used a second method we developed called CENR

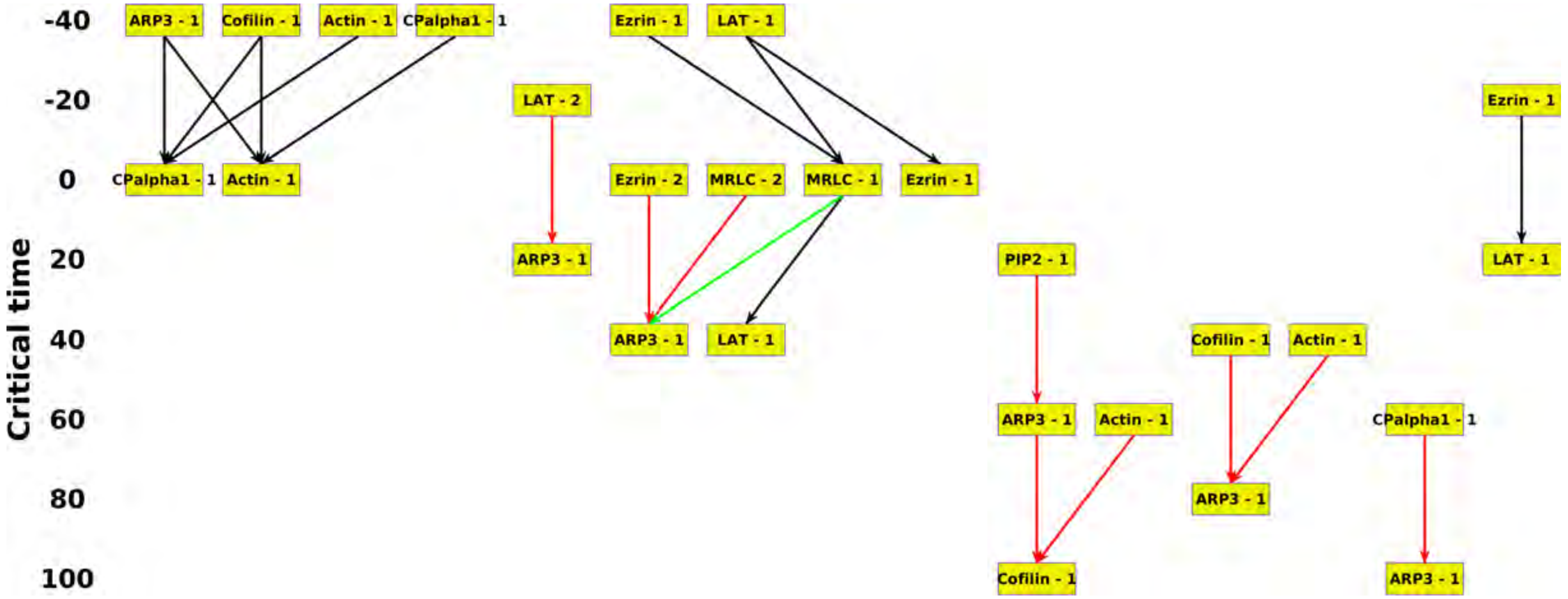
Given *x*[*t*] and M, find **A** and *c* to minimize *w*[*t*]

Adjacency matrix of CGP method



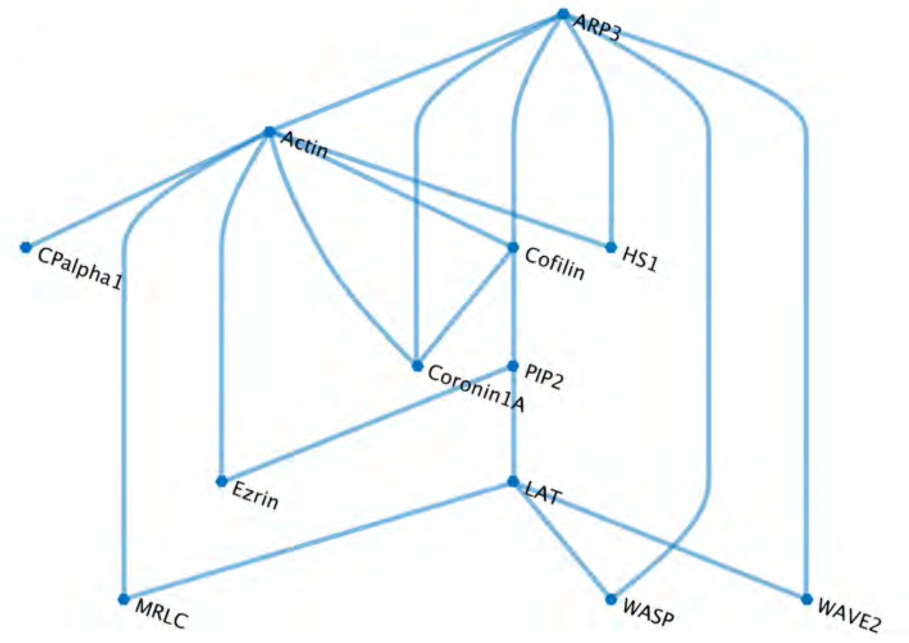
Summarizing relationships

- We threshold the strength of the relationships, (elements in the adjacency matrices), and identify the time when each is most strongly observed.



Evaluation

- Make list of known or suspected regulatory relationships from literature



Evaluation

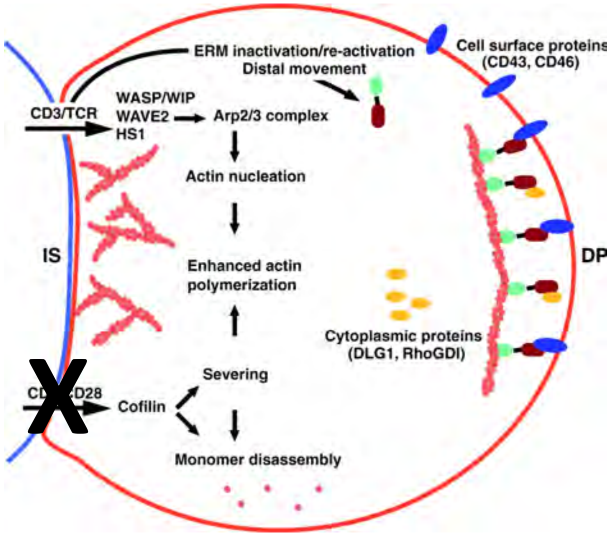
- Measure how well learned models capture these known relationships using a Receiver-Operator Curve (quantitate by “Area Under the Curve”)

	CGP	CENR
AUC	0.709	0.644

- Note some “False positives” may be real positives that are not yet known

Evaluation

- Data used so far was from costimulation conditions (stimulation through both TCR and CD28)



- Additional maps available for conditions where CD28 costimulation is blocked (“B7 blocked”)

Evaluation

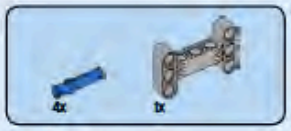
- Using the model learned from costimulation condition, make predictions from early time points for blocked condition at later time points

	CGP	CENR
Prediction error from cross-validation on training images (full stimulation)	8.6%	6.9%
Prediction error on testing images (costimulation blocked)	12.0%	8.5%

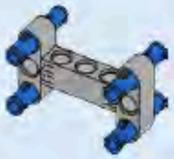
HIERARCHICAL ASSEMBLY MODELS

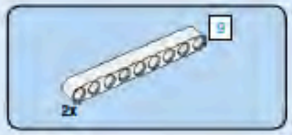
Spatial models

- Most of you probably have built models: LEGO's, K'nex, etc.
- You get different colored parts and a set of instructions
- The instructions are *hierarchical*

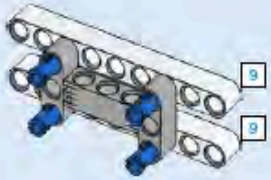


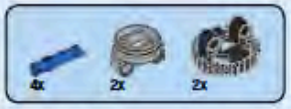
115



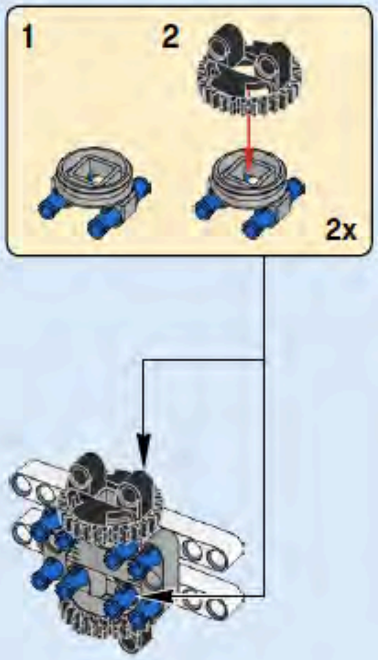


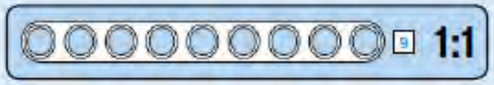
116

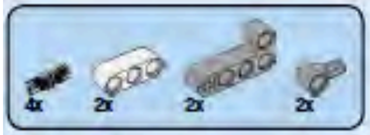




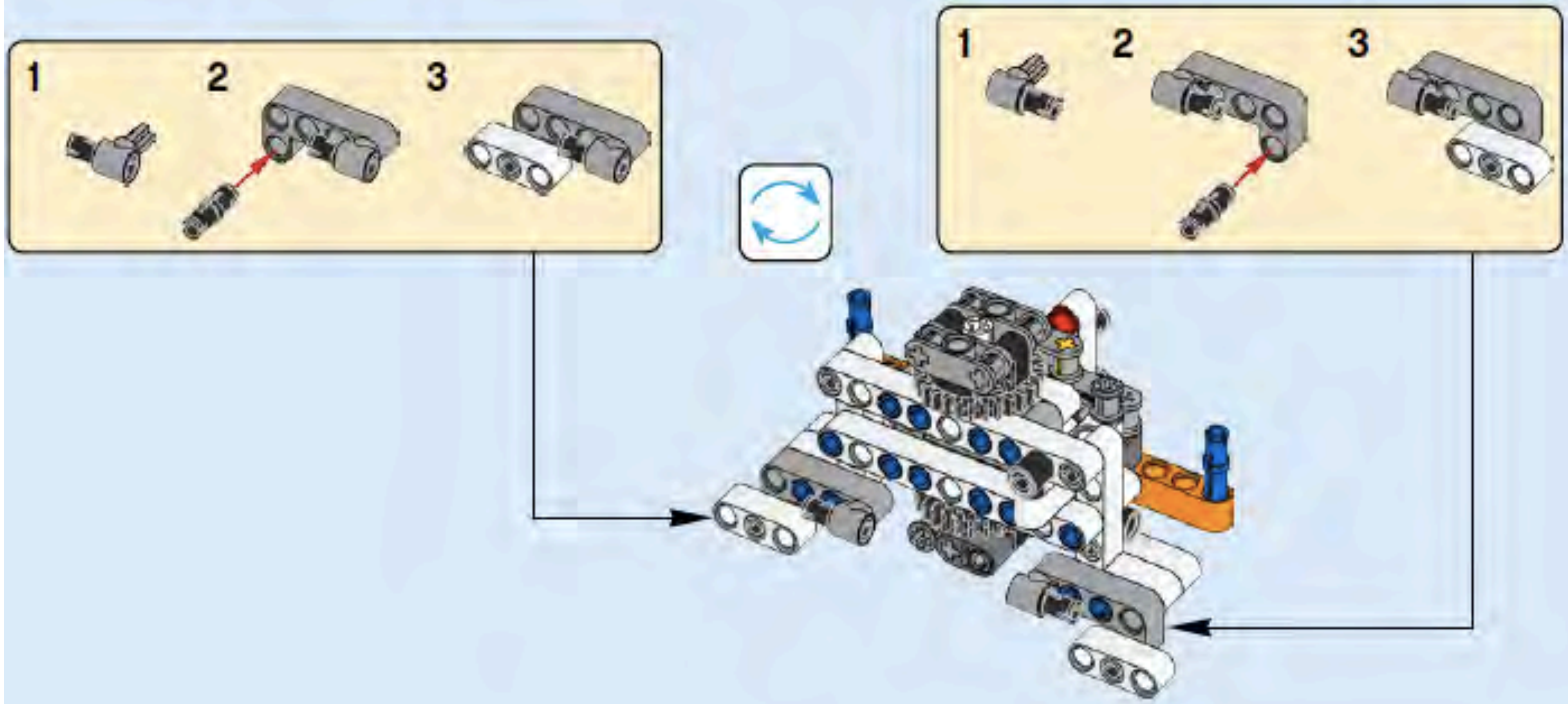
117

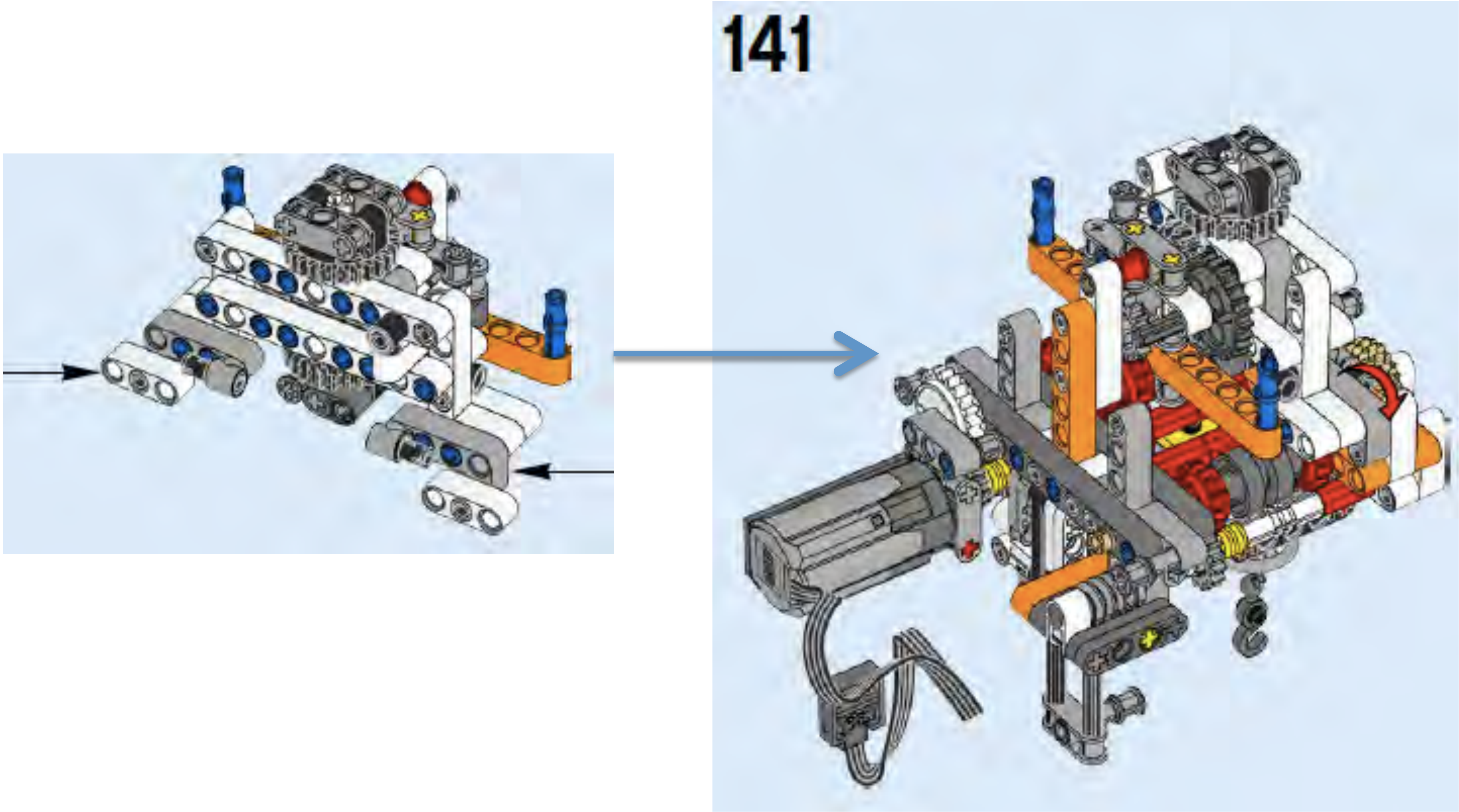






139



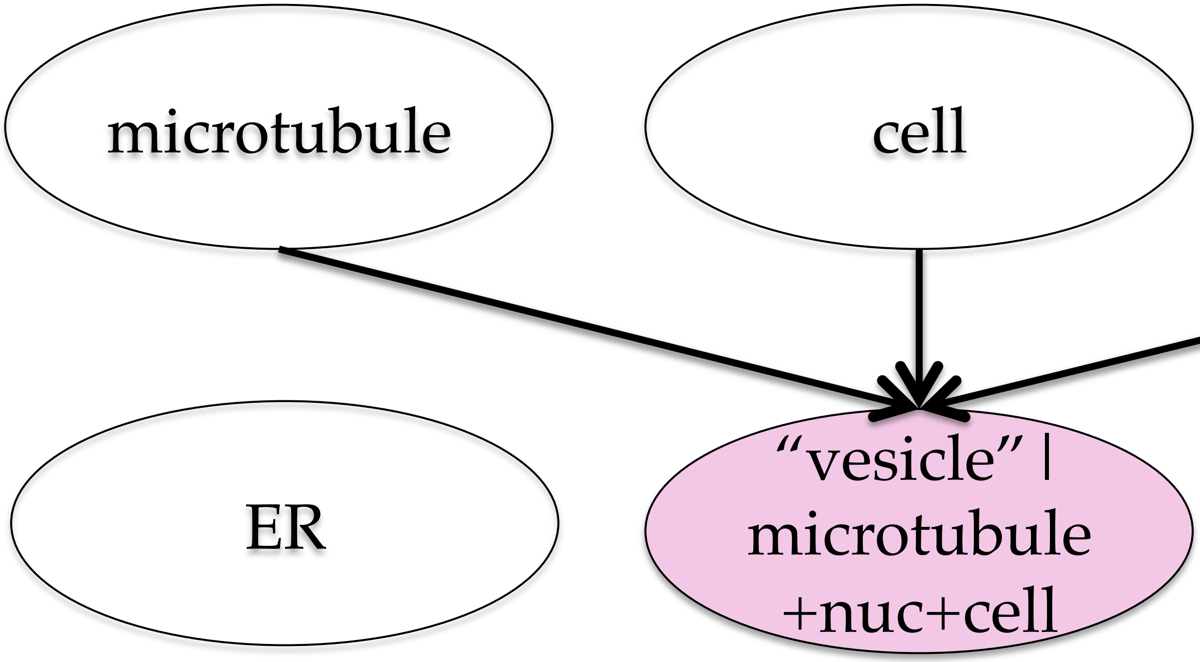

141

Bayesian network / graphical model

- One way to think of this hiearchical assembly process is as a graphical model
- Nodes correspond to parts or previous assemblies
- Edges correspond to *dependencies* − parts required to produce/localize an assembly

Li et al (2016) *Cytometry*

Bayes net for punctate organelles

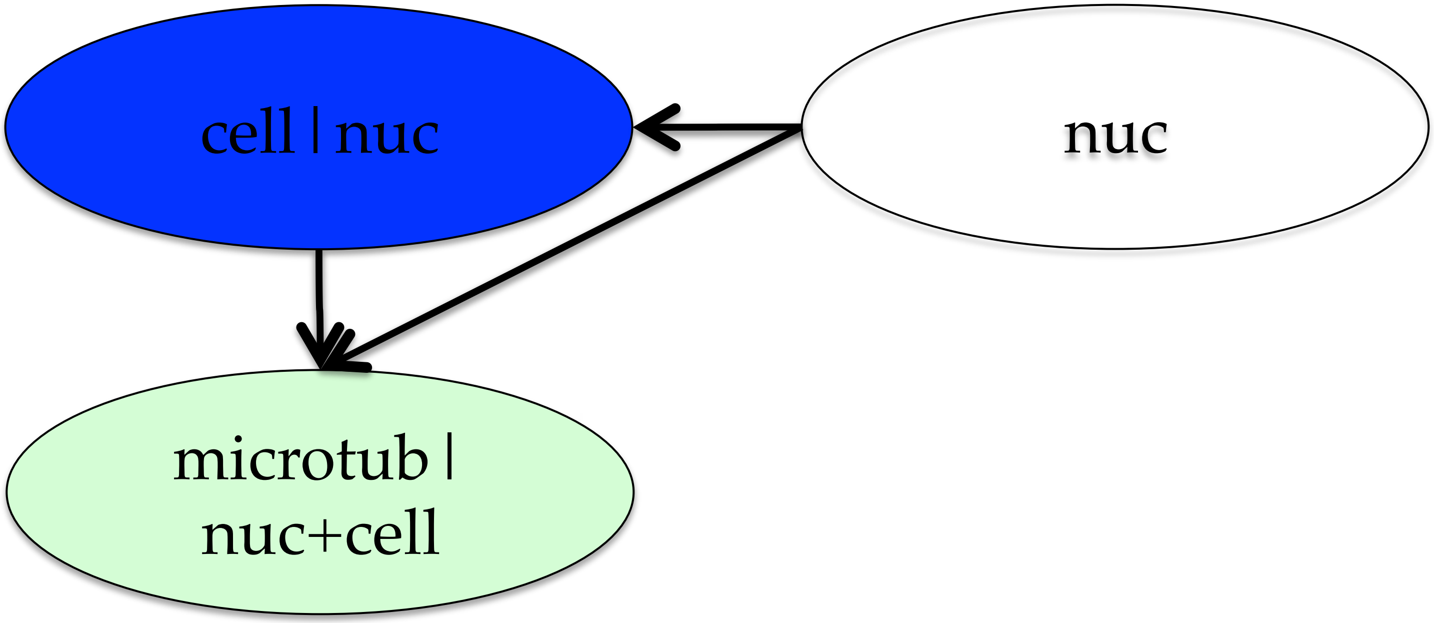


Node function=point process

But are nuc, cell and microtubule independent from each other?
They are not puncta so can't estimate with point process.

Johnson et al (2015) *Mol. Biol. Cell*

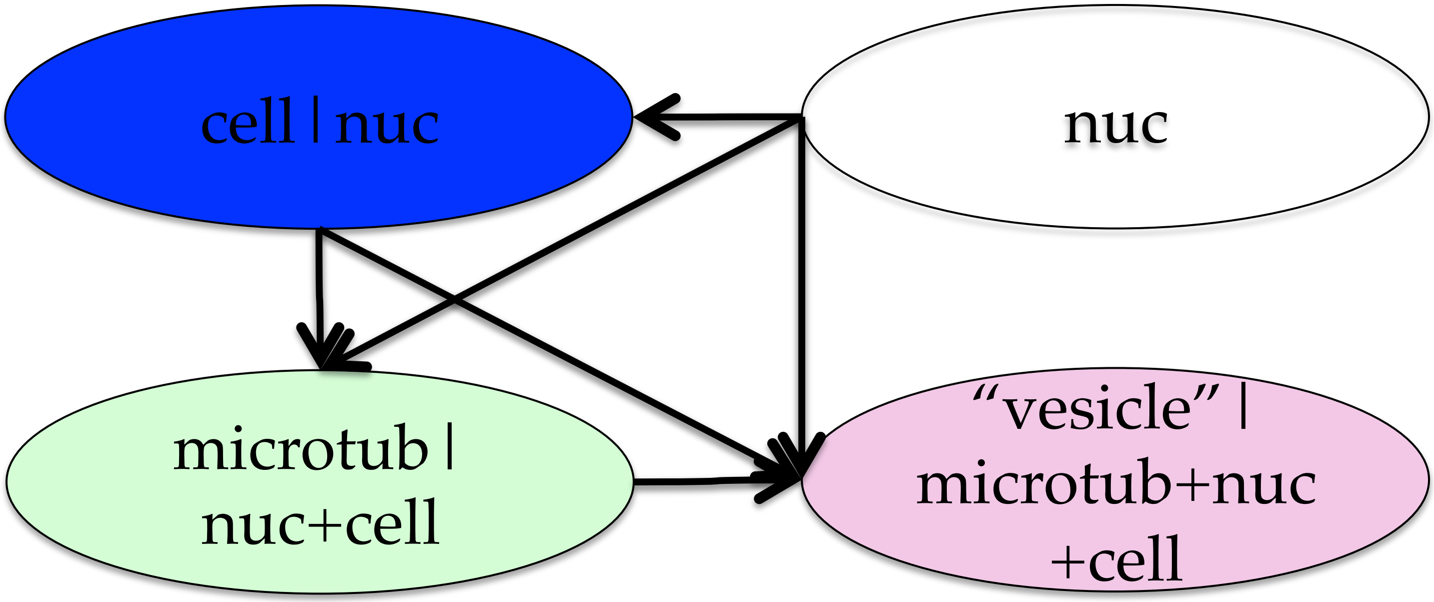
From other experiments



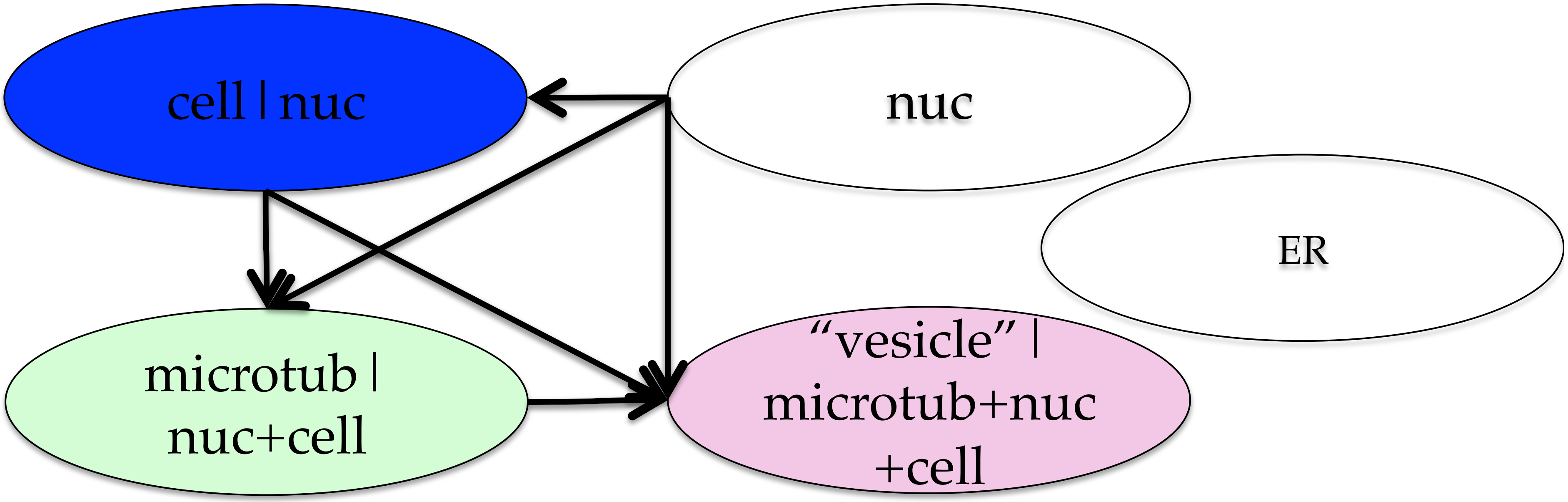
Node function=growth model

Node function=various

Merged Bayes net

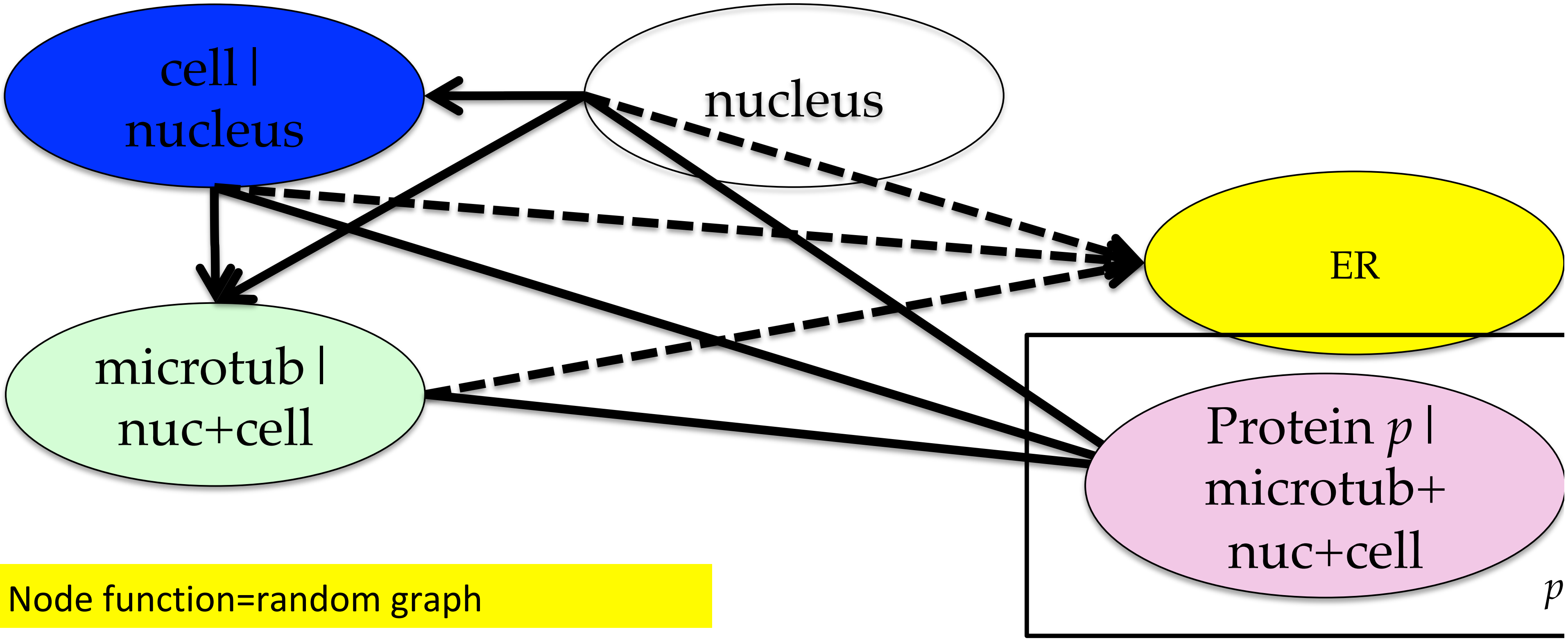


Merged Bayes net



Liu-Huang et al (2017) submitted

From other modeling…



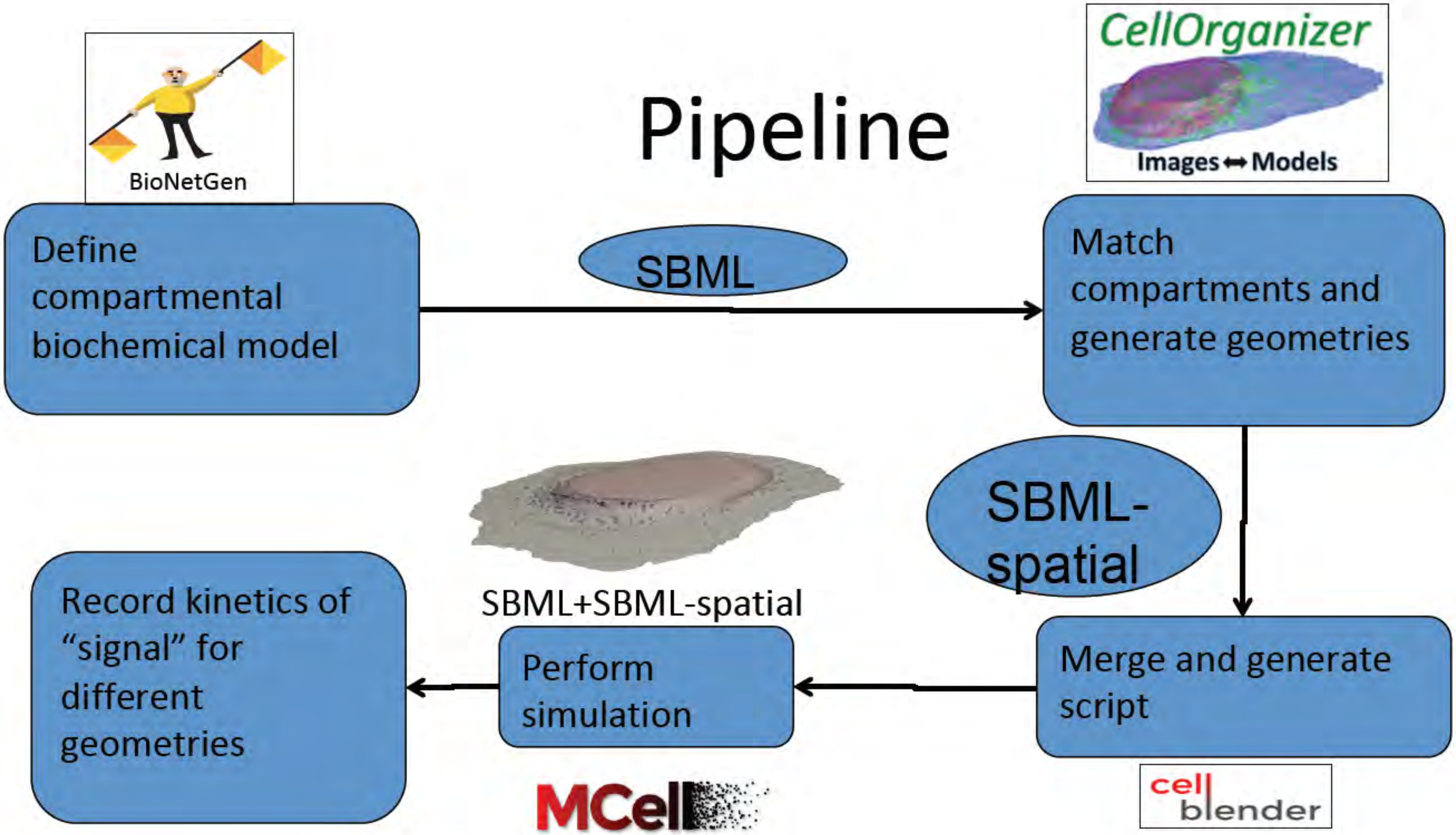
Node function=random graph

HIGH-THROUGHPUT CELL SIMULATIONS

High-throughput spatially realistic simulations

- Study the effects of spatial variance caused by
  - Cell cycle
  - Diseases
  - Drugs
  - Inherent cell variance
- Model large systems with high spatial realism
- Validate generative model accuracies

Pipeline



Define compartmental biochemical model

SBML

Match compartments and generate geometries

SBML-spatial

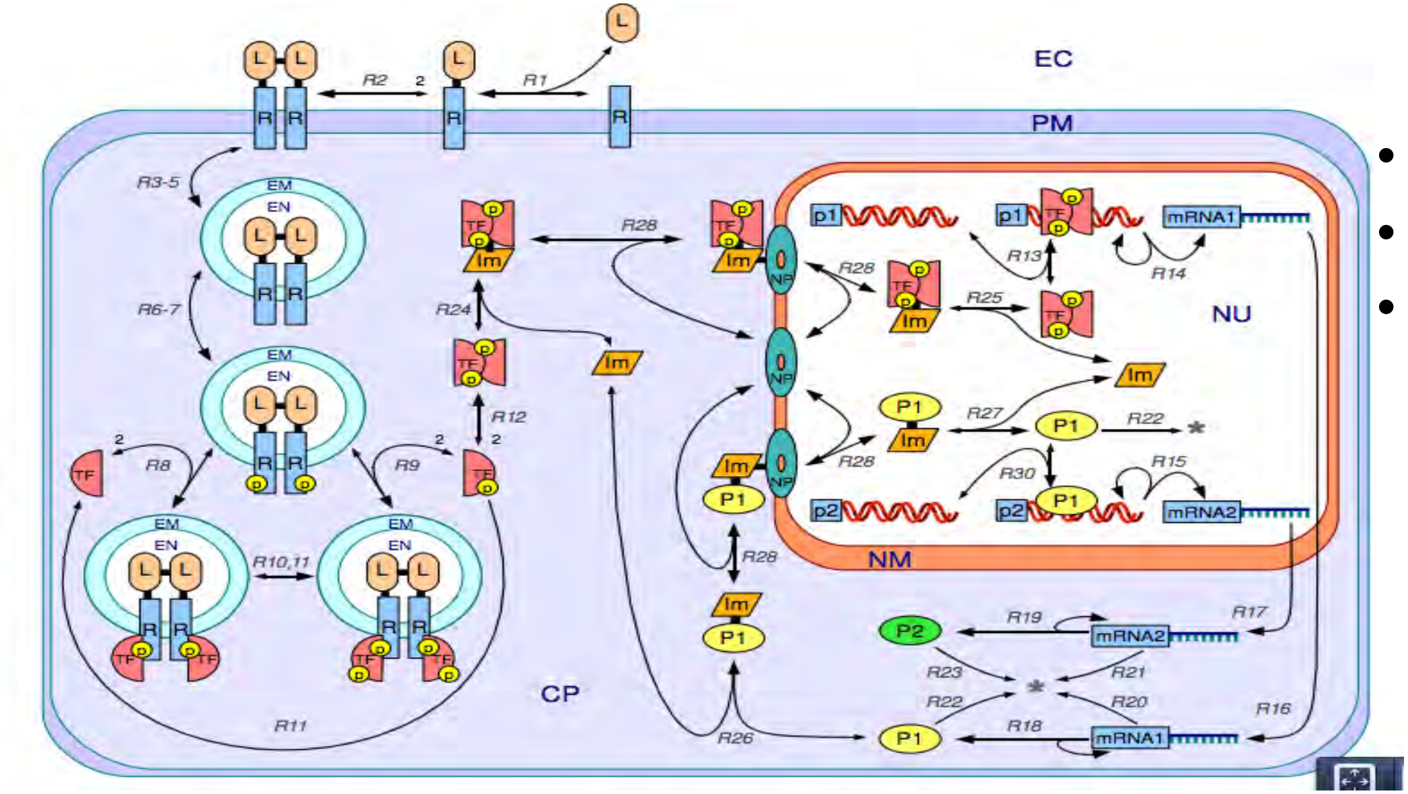
Merge and generate script

SBML+SBML-spatial

Perform simulation

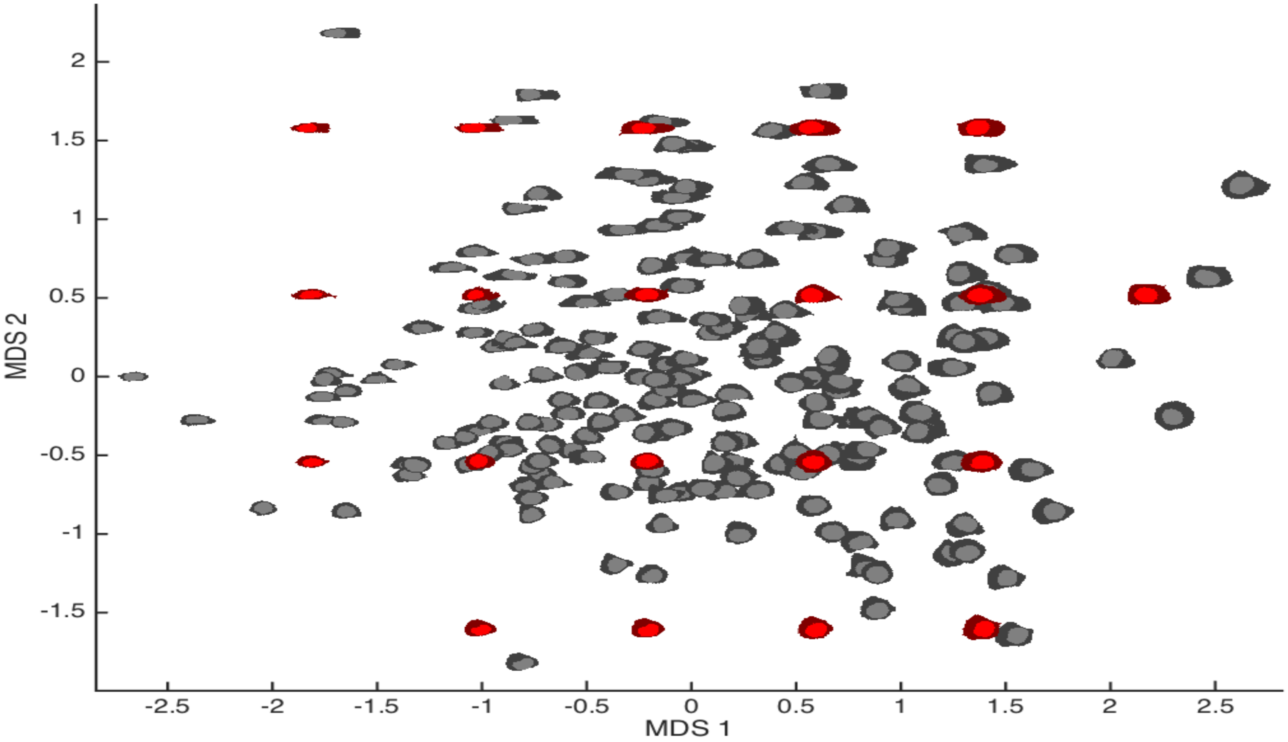
Record kinetics of “signal” for different geometries

“Simple” biochemical model

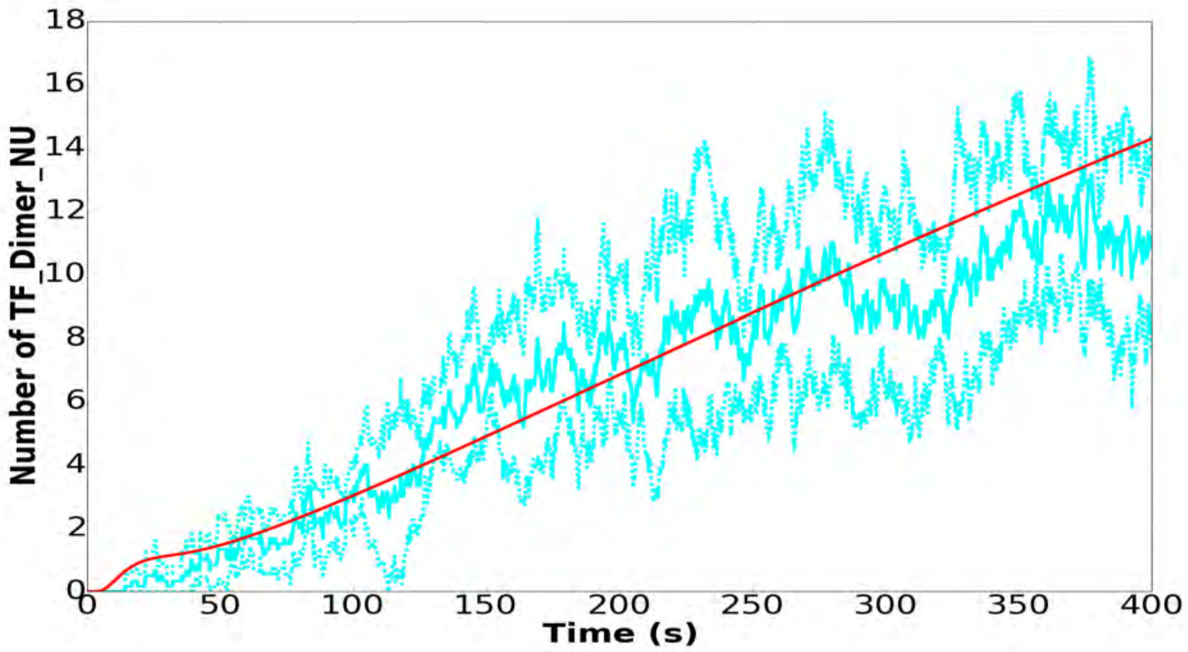


- 354 reactions
- 78 species
- 7 “compartments”

Selected geometries from shape space



Different geometries lead to variation in signaling



Mean +/- S.D.

Conclusions

- Tools becoming available to construct models of cell components directly from images
  - Better comparison across instruments/cell types
  - Provide input geometries for cell simulation
  - Provide simulated images for testing algorithms
  - Learn putative spatiotemporal causal relationships
- Need to combine images and data from various other experiments to create overall spatiotemporal models

Supported by: NSF BBSRC bioscience for the future NIH National Institute of General Medical Sciences

CellOrganizer Team

Project Leaders

Robert F. Murphy

Gustavo Rohde

Gregory R. Johnson

Current Team Members

Xiongtao Ruan

Kelvin Liu-Huang

Ivan Cao-Berg

Collaborators

Jörn Dengjel

Christoph Wülfing

Past Contributors

Ting Zhao

Tao Peng

Wei Wang

Aabid Sharif

Joshua Kangas

Jianwei Zhang

Jieyue Li

Baek Hwan Cho

Taraz Buck

Devin Sullivan

Ying Li

Tim Majarian