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
Honorary Professor, Faculty of Biology, University of Freiburg, Germany
Senior Fellow, Allen Institute for Cell Science

Carnegie Mellon University
Computational Biology Department

MMBioS
An NIH Biomedical Technology Research Center

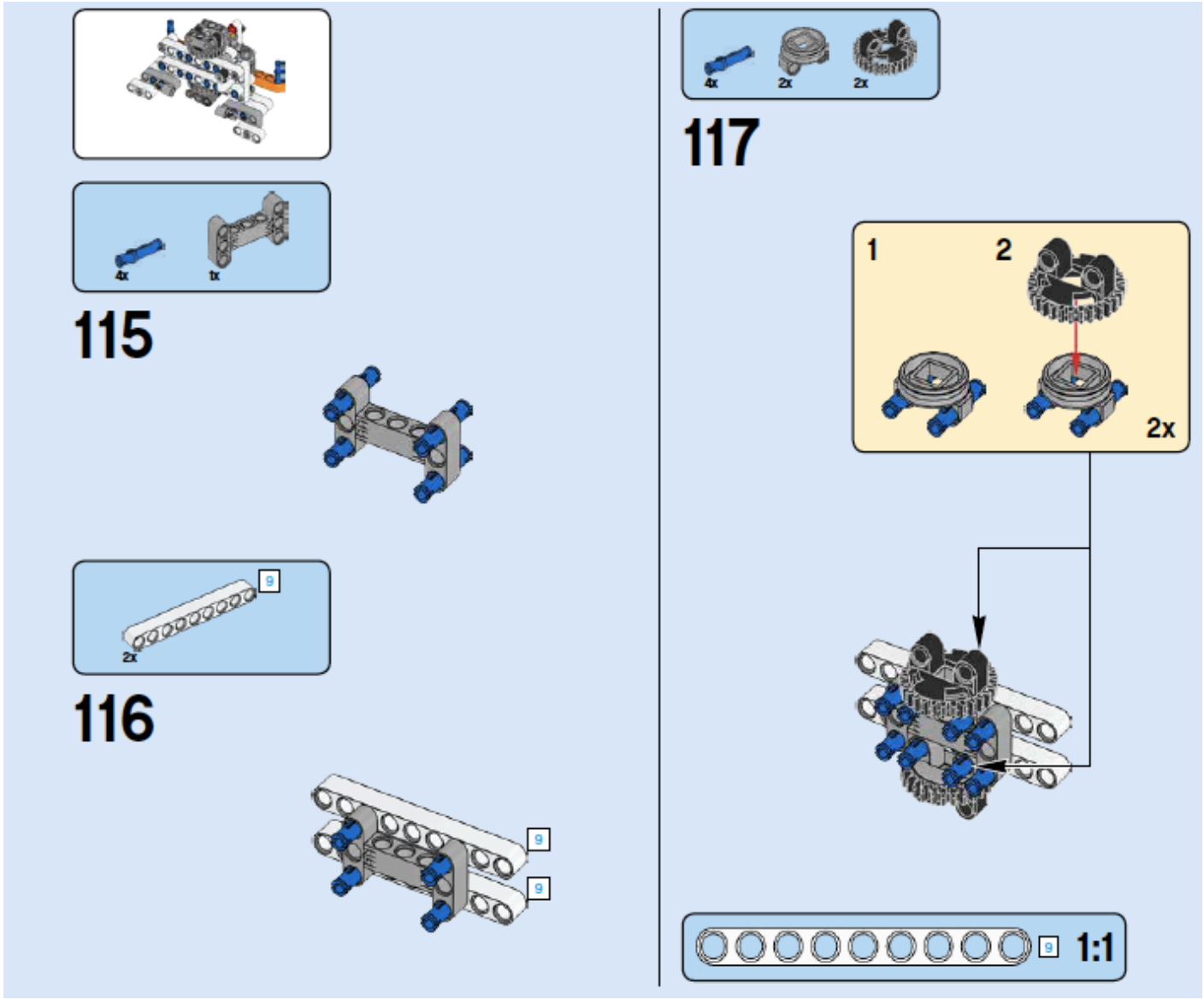
An example

- A fundamental goal in cell and molecular biology is to learn the spatial relationships between all of the components of a cell
- Can think of this as learning the assembly instructions of a cell – a ***generative model***

Spatial models

- This is like reverse engineering a model built with LEGO's, K'nex, etc.
- The forward process is given in instructions that are *hierarchical*

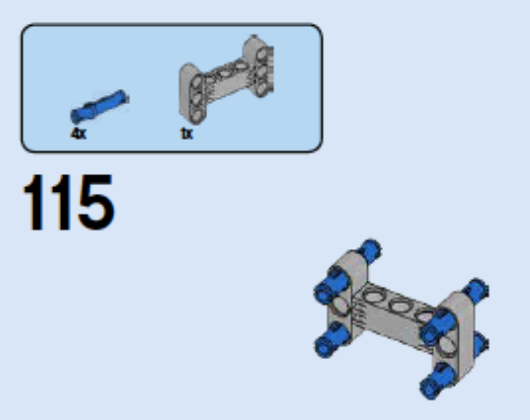
This illustrates the concept of spatial dependency: the location of the blue part depends on the location of the gray part



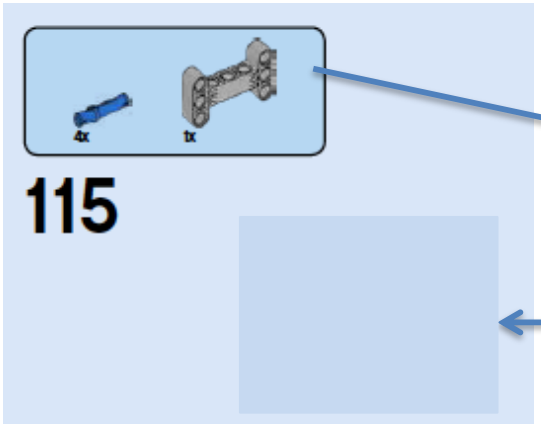
Bayesian network / graphical model

- A Lego model assembly is deterministic
- But “cell assembly” is probabilistic
- Can represent the hierarchical assembly process as a ***graphical model***
  - **Nodes** contain probabilistic spatial distributions of parts or previous assemblies
  - **Edges** correspond to *dependencies* – parts required to produce/localize an assembly
  - **Node functions** produce the spatial distributions of an assembly from the spatial distributions of the parts

Simple example



Simple example



Binary image of the location of the gray part

0	0	0	0	0	0
1	0	0	0	1	0
1	1	1	1	1	0
1	0	0	0	1	0
0	0	0	0	0	0

Node function

gray part

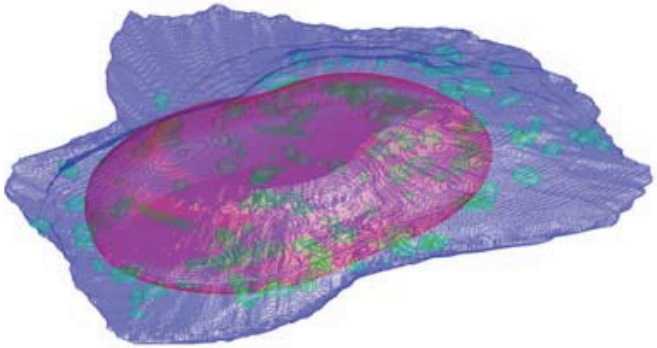
blue part

Resulting binary image of the location of the blue part

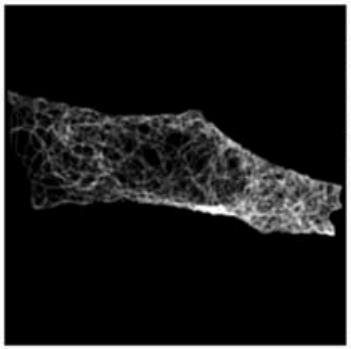
0	0	0	0	0	0
1	0	0	0	1	0
0	0	0	0	0	0
1	0	0	0	1	0
0	0	0	0	0	0

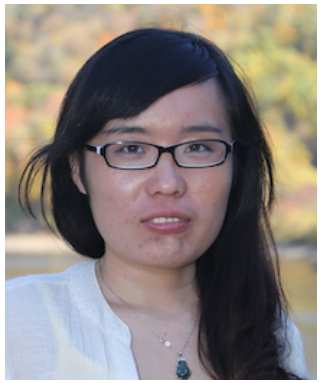
First generative models of cell organization

- Individual cell organelles dependent on cell and nuclear shape
  - 2D: Zhao & Murphy (2007)
  - 3D: Peng & Murphy (2011)



- Microtubules dependent on cell and nuclear shape
  - 3D: Shariff, Murphy & Rohde (2010)





Ying Li

Tim Majarian

Armaghan Naik

Greg Johnson

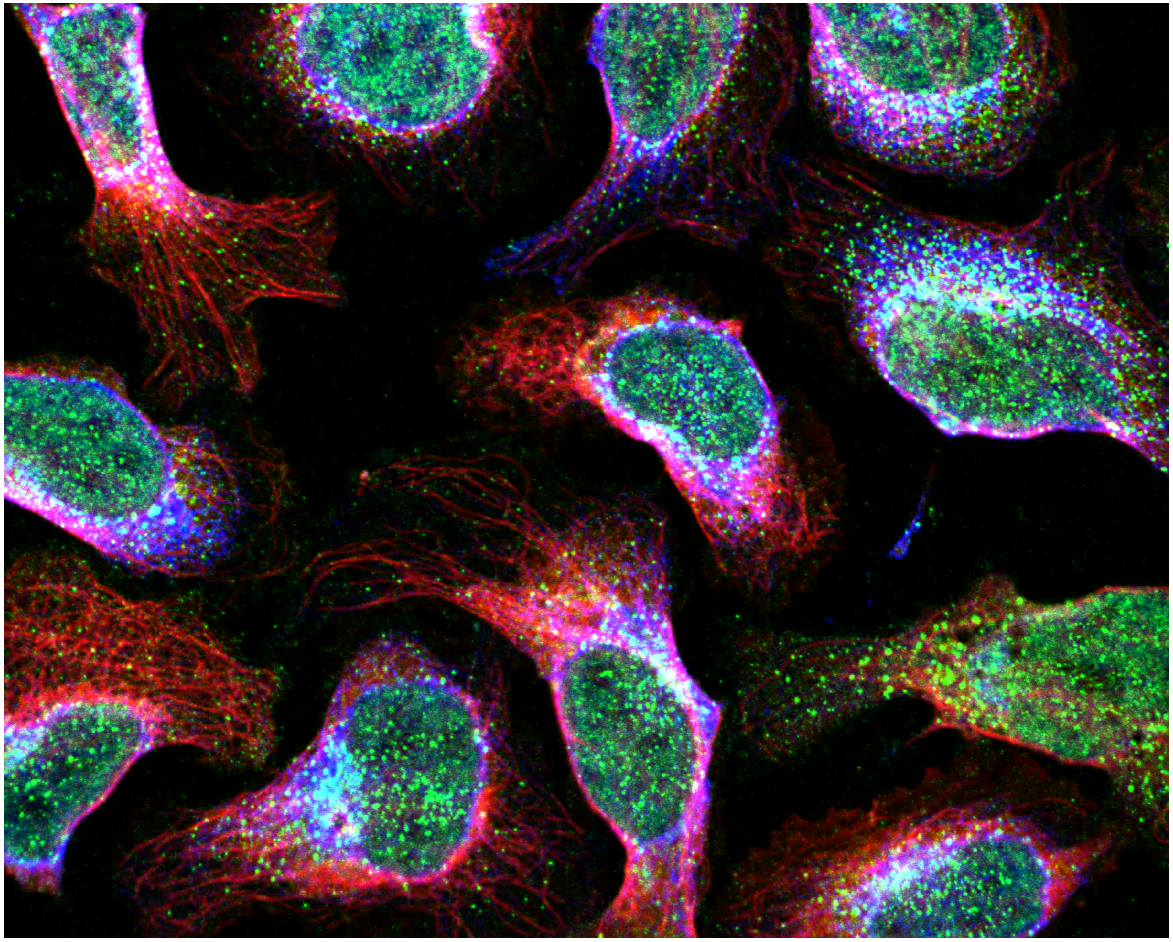
POINT PROCESS MODELS FOR ORGANELLE SPATIAL DEPENDENCIES

Learning higher levels of dependency

- Overall goal: learn which organelles the spatial distributions of other organelles depend on
- Use Human Protein Atlas images containing markers for nucleus, microtubules, endoplasmic reticulum and various punctate organelles

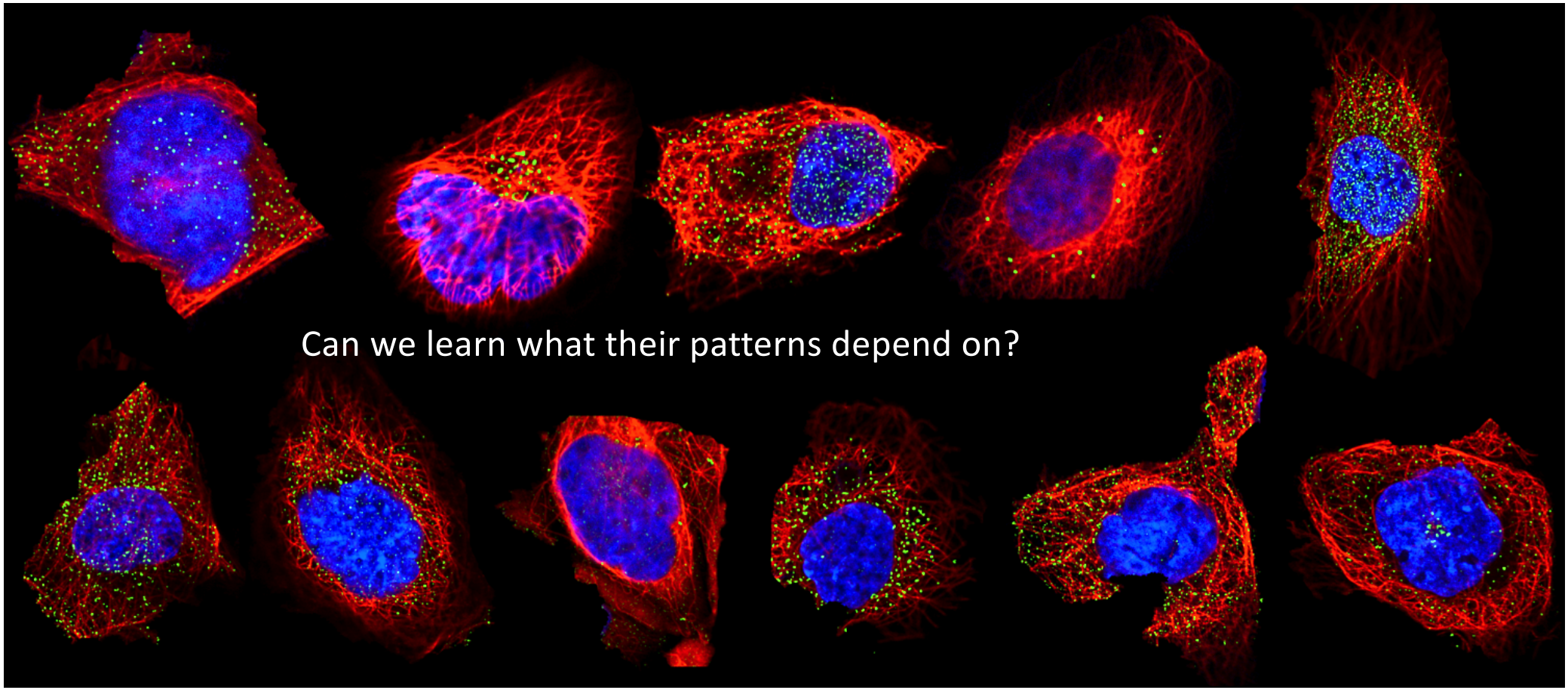
Example HPA Immunofluorescence Image

Red=tubulin
Blue=ER
Green=Sec23b



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

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



Point Process Models

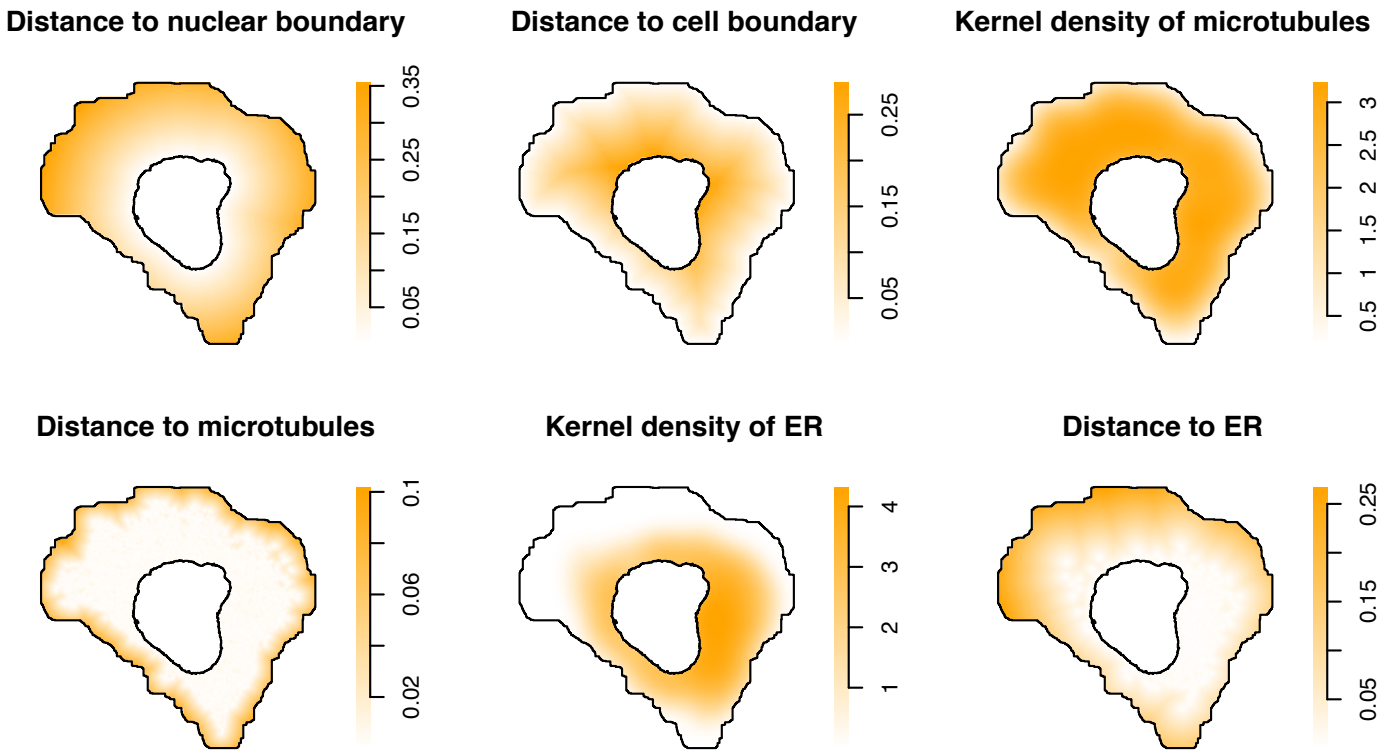
- Model the probability of a punctate organelle occurring at a position *X* in a cell using functions of the distributions of other components of that cell (called **factors**, *F*)

$$p(X) = b_{\vartheta}(X) = \vartheta_1 F_1 + \vartheta_2 F_2 \ldots + \vartheta_N F_N$$

Li et al (2016) *Cytometry*

Factors for point process models

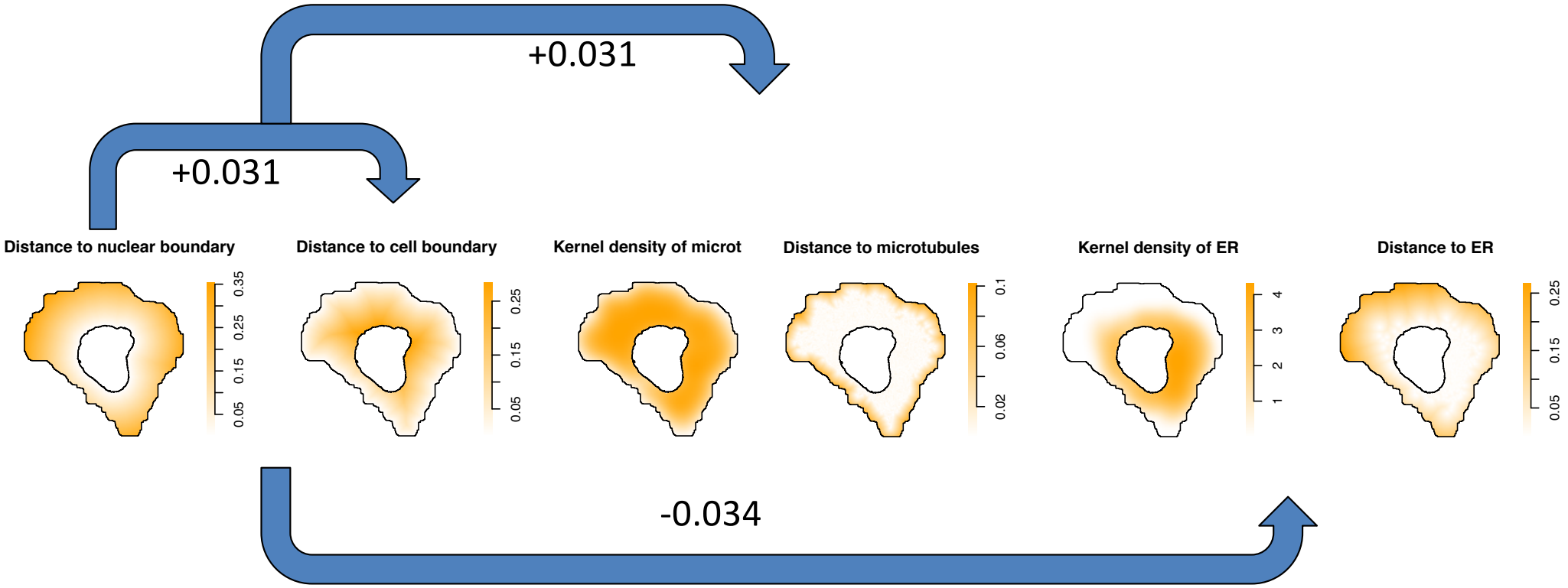
- The factors are variables for which values are known at all positions in the cell



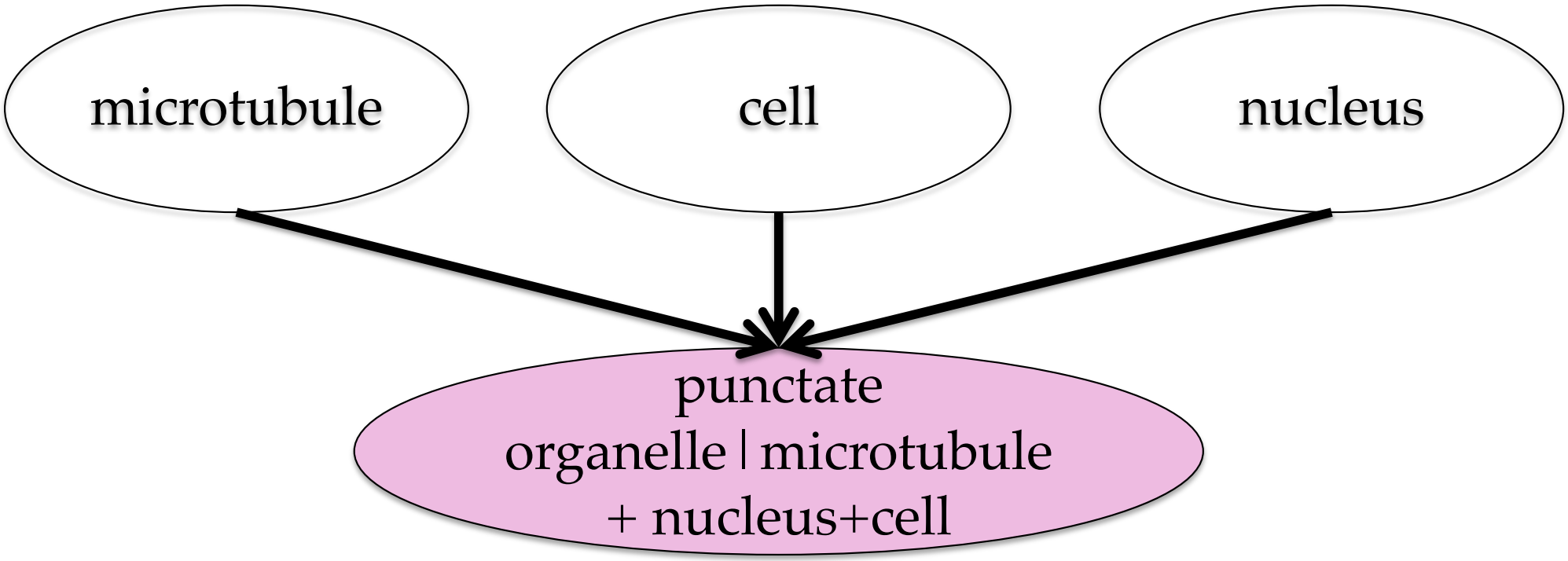
Learning dependencies on factors

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

Contributions of different factors



The resulting graph



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

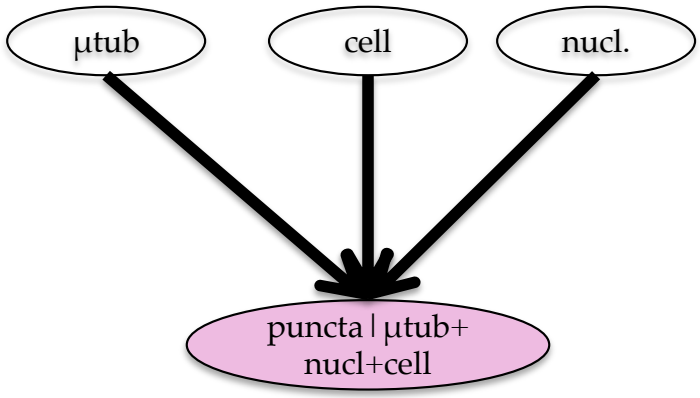
U-251 MG	COPI	COPII	Caveolae	Coated Pits	Early Endosomes	Late Endosomes	Lysosomes	Peroxisomes	RNP bodies	Recycling Endosomes	Retromer
COPI	1	0	0	0	0	0	0	0	0	0	0
COPII	0	1	0	0	0	0	0	0	0	0	0
Caveolae	0	0	1	0	0	0	0	0	0	0	0
Coated Pits	0	0	0	0.67	0	0	0	0	0	0	0.33
Early Endosomes	0	0	0	0	1	0	0	0	0	0	0
Late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

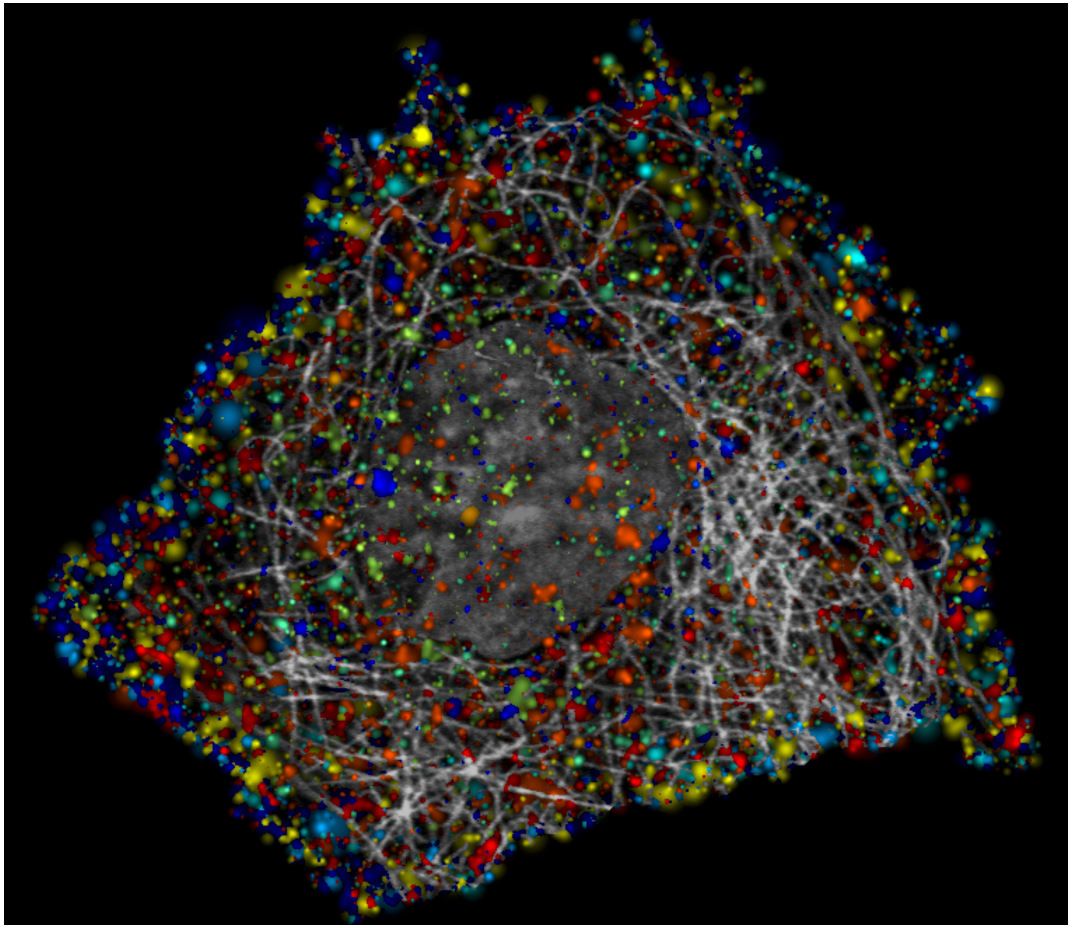
A-431	0.73
U-2 OS	0.90
U-251 MG	0.86

But…

- Need to be able to construct these graphs eventually for all cell components
- But can't do all components in a single image
- So also need to be able to infer how different graphs might be connected to each other



Example synthetic cell image with 11 punctate organelles



Tim Majarian

Seema Lakdawala
University of Pittsburgh

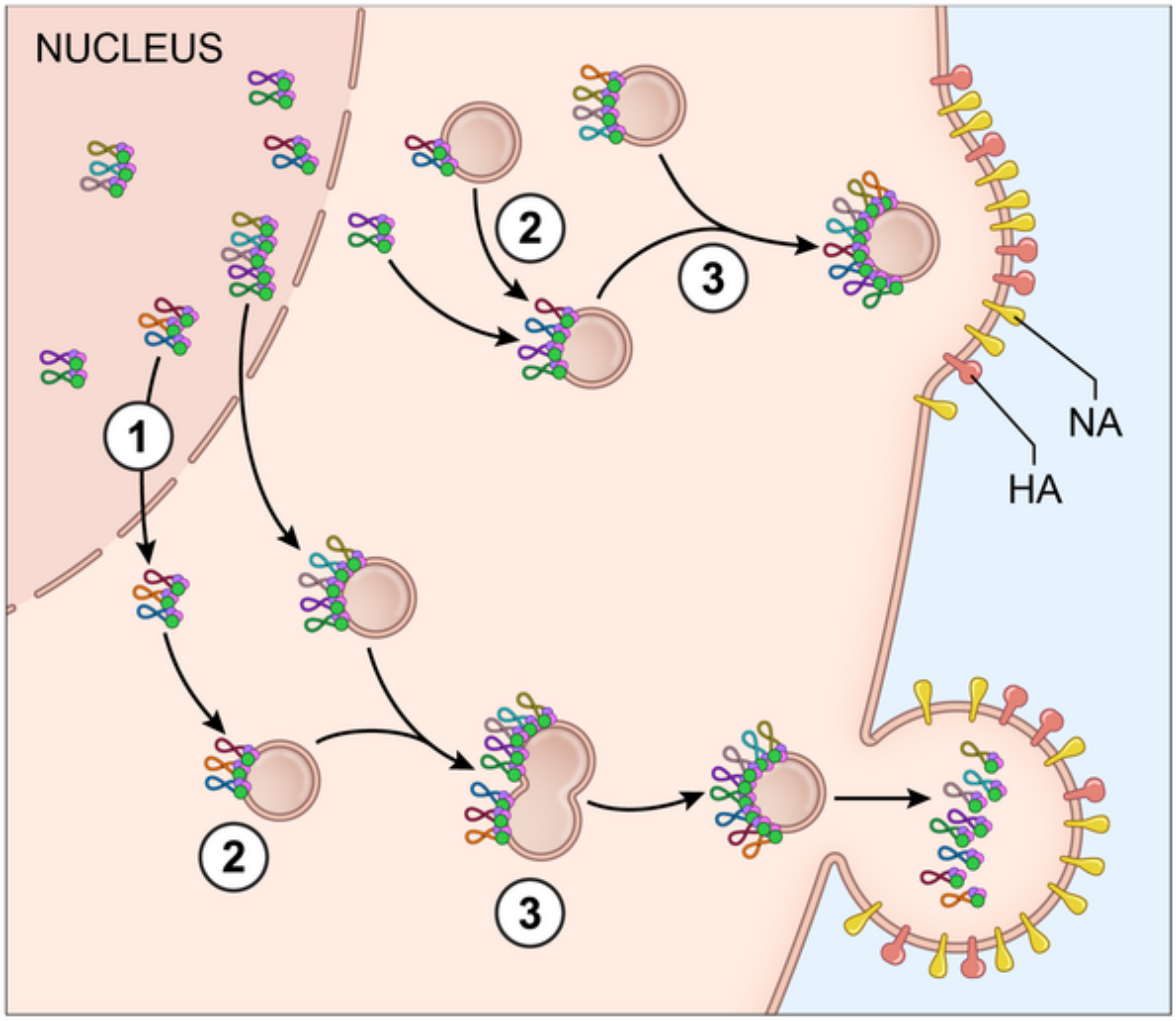
INFERRING INFLUENZA VIRUS RNA ASSEMBLY PATHWAYS

Concept

- Try to learn how to put together complex for which we can only image a subset of its components simultaneously
- Learn models of spatial dependency of different parts of a complex on each other
- Construct most likely what in which distribution of full complex could be assembled from parts

Influenza virus assembly

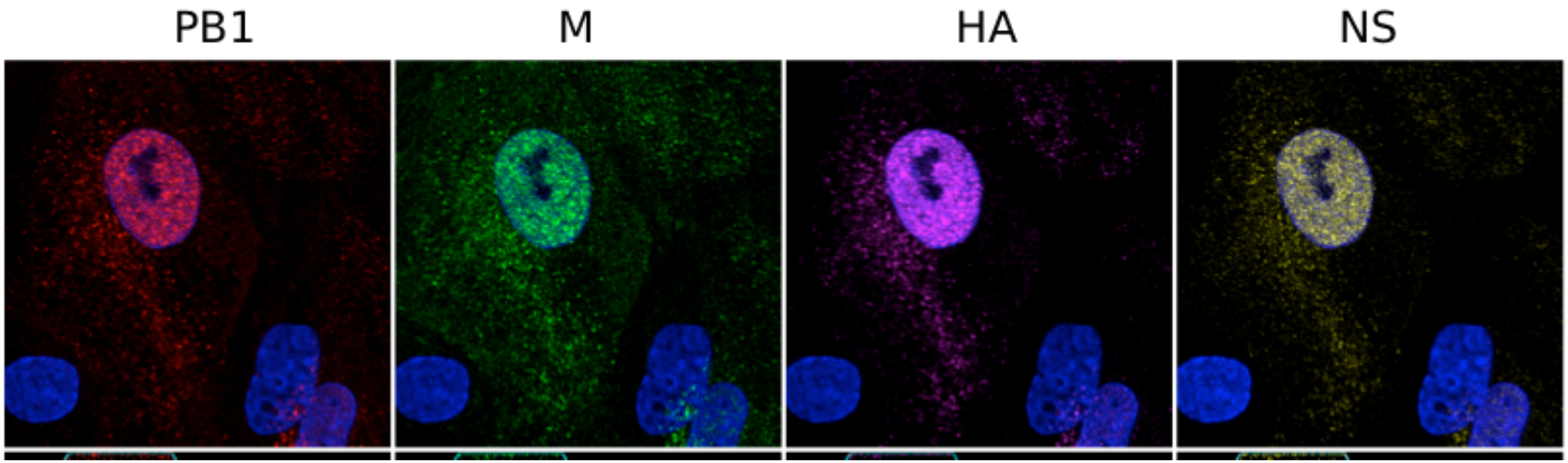
- Flu virus consists of one of each of eight ribonucleoprotein particles (RNPs)
- Mechanism of assembly unknown



Lakdawala SS, Wu Y, Wawrzusin P, Kabat J, Broadbent AJ, et al. (2014) Influenza A Virus Assembly Intermediates Fuse in the Cytoplasm. PLOS Pathogens 10(3): e1003971. https://doi.org/10.1371/journal.ppat.1003971 http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003971

Data

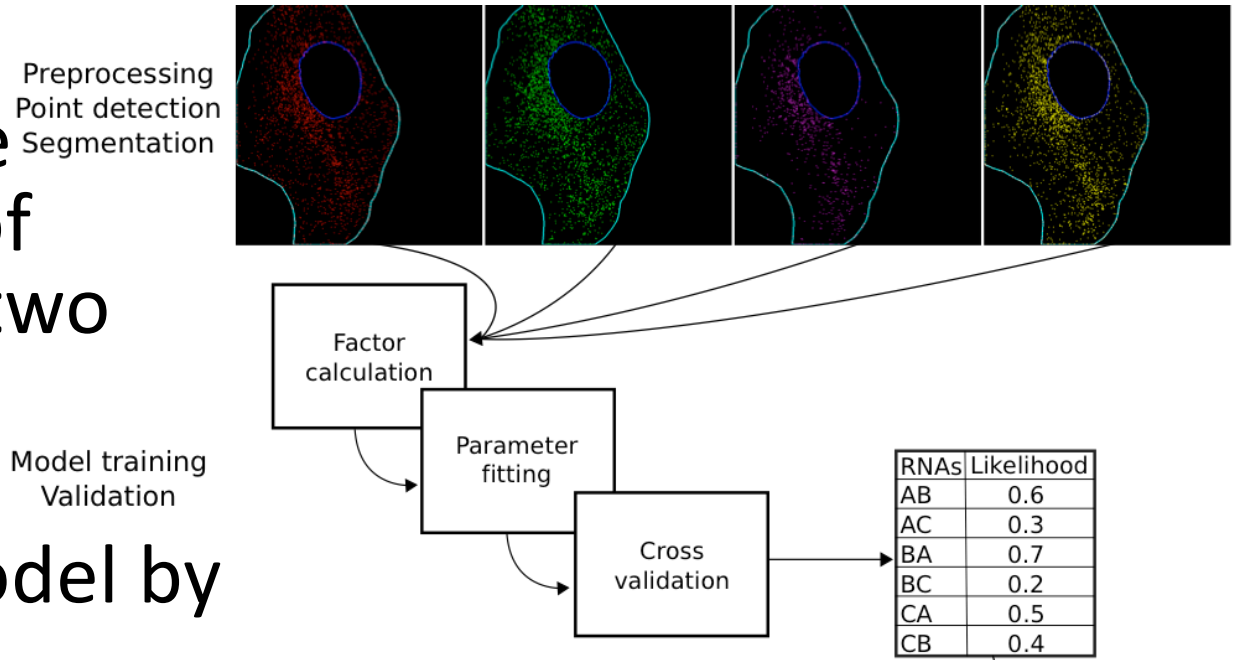
- Previous study acquired fluorescence in situ hybridization (FISH) images for many sets of four probes for different RNPs



Approach Part 1

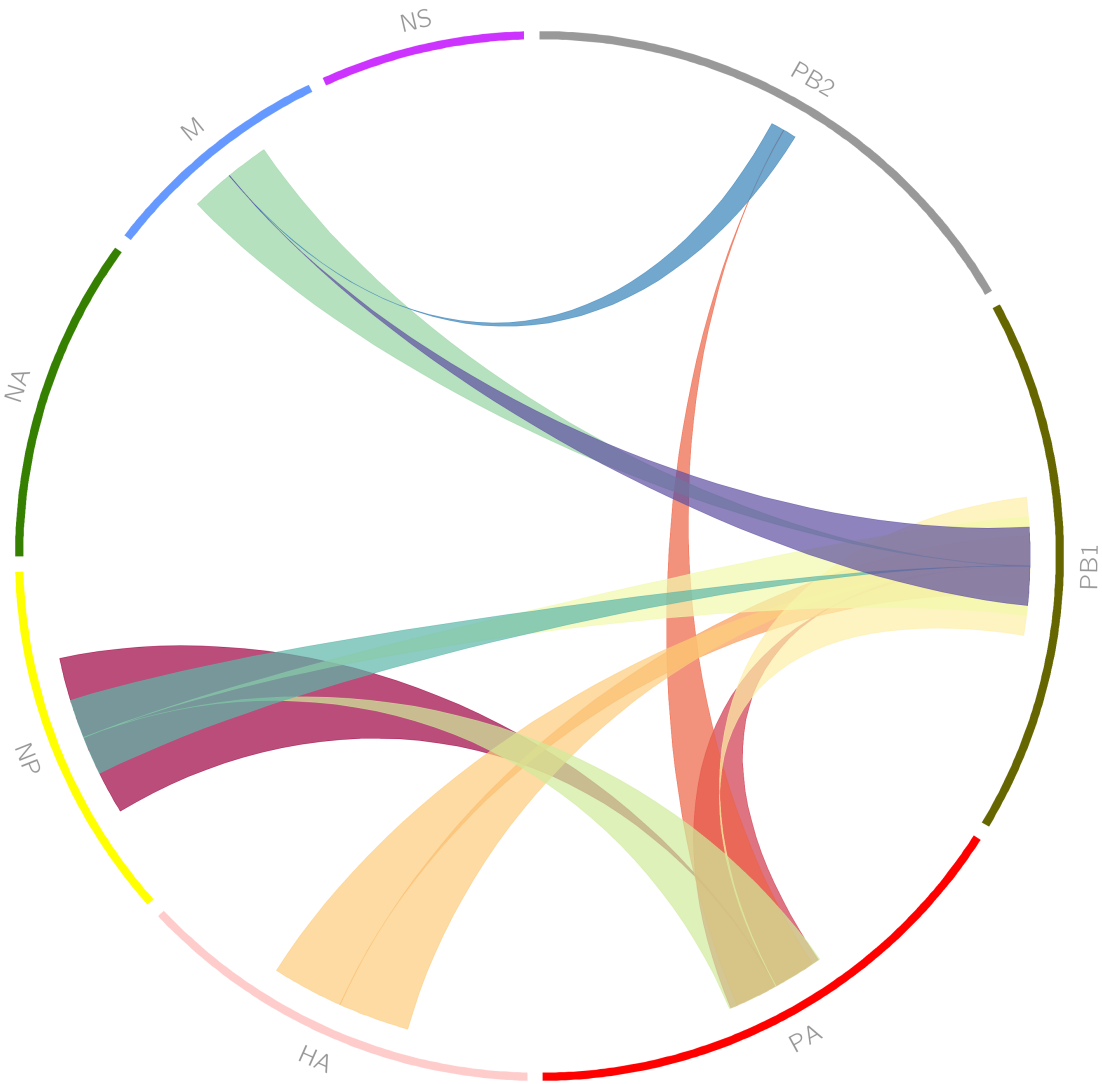
Majarian et al (2018) *PLoS Comp. Biol., in press*

- Process each channel to identify position of RNPs
- Learn point process models to predict the subcellular location of each RNP from one, two or three other RNPs
- Measure the likelihood of each model by cross-validation



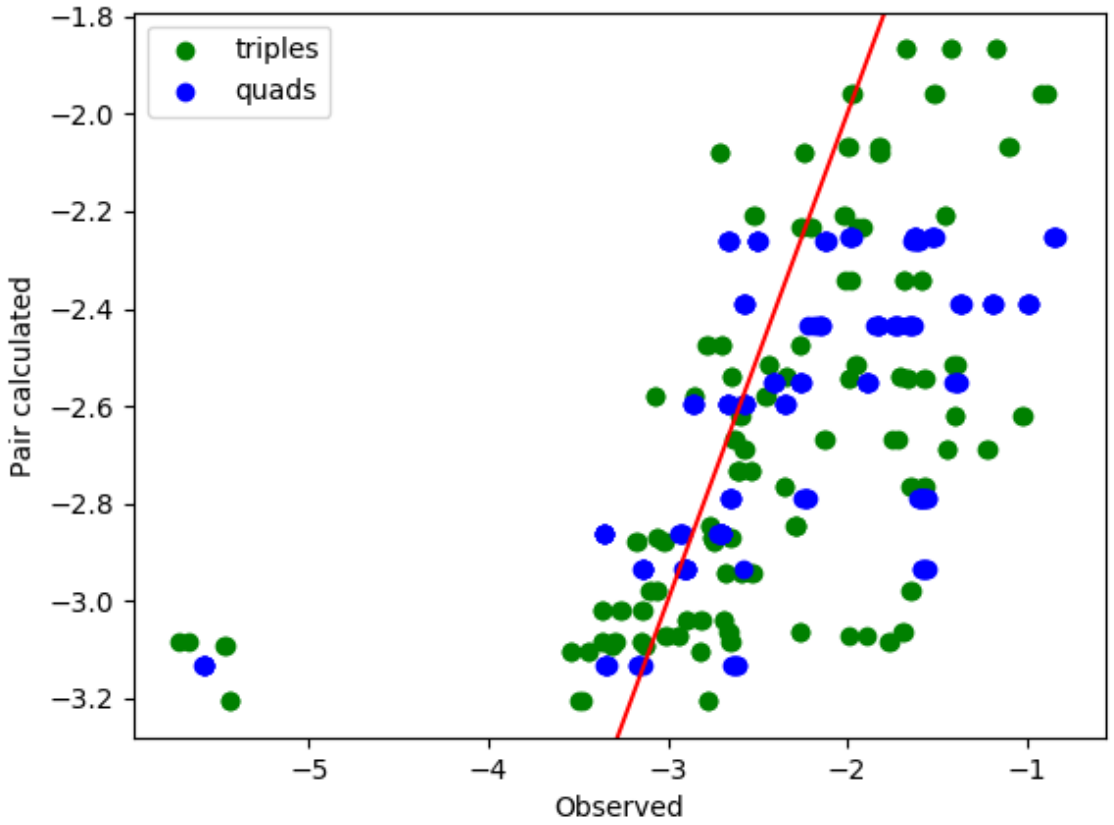
Pairwise likelihoods

- Width of base shows strength of prediction
- Can combine these to predict likelihood of triples, quads, etc.



Test predictions

- Predict likelihood for 3 or 4 RNPs from likelihoods measured from pairs
- Compare to measured likelihoods
- Decent correlation except for outlier cluster (PB2-NA + PB1 and/or NS)

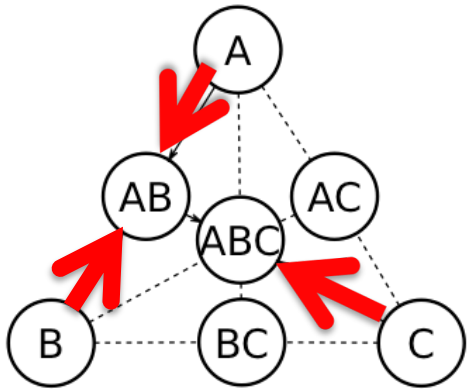


Approach Part 2

- Use dynamic programming to find the most likely path by which all eight RNPs can be assembled

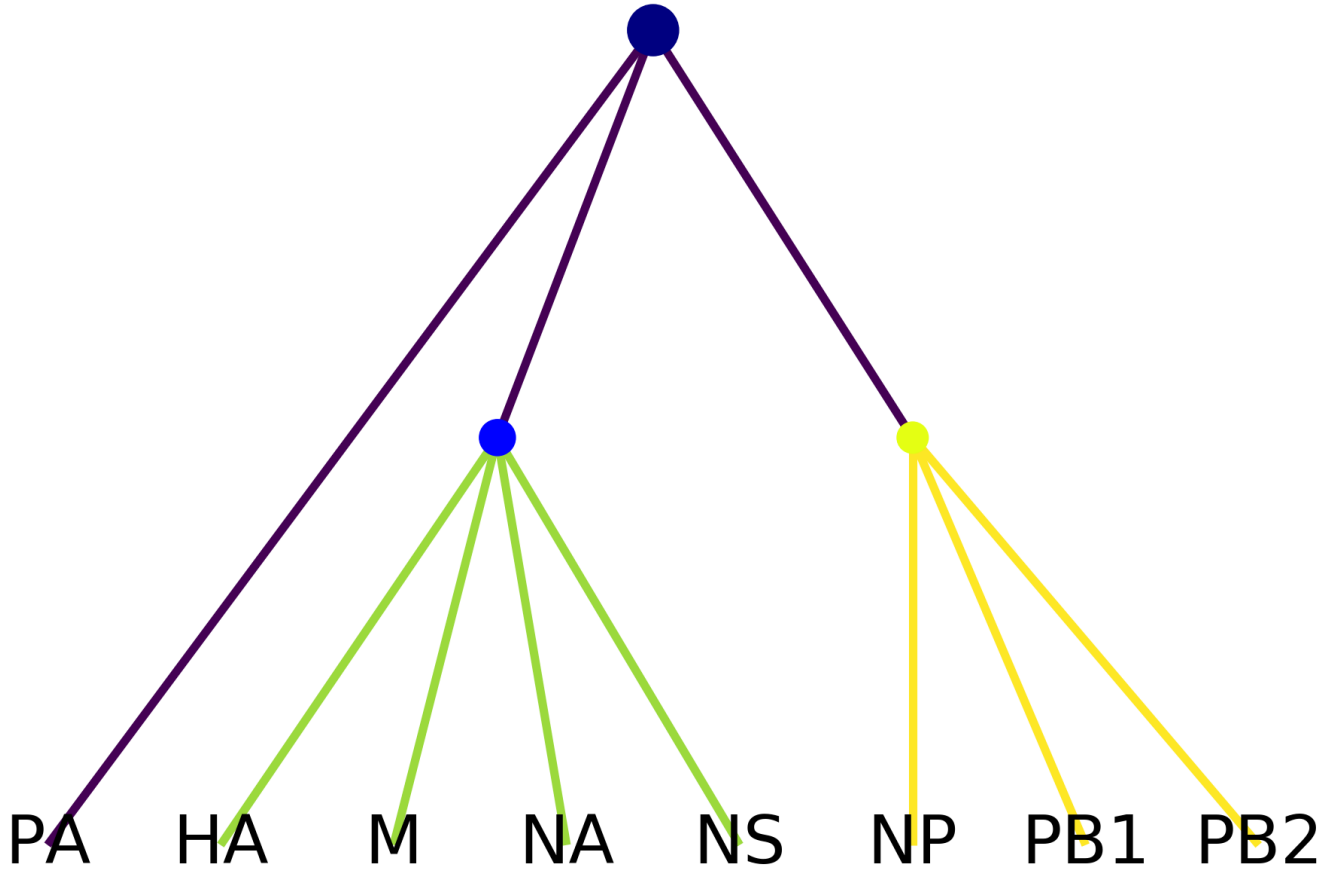
A:B	0.6
A:C	0.3
B:A	0.8
B:C	0.2
C:A	0.5
C:B	0.4
AB:C	0.3
AC:B	0.2
BC:A	0.2

[A:B+B:A]/2*AB:C	0.21
[A:C+C:A]/2*AC:B	0.08
[B:C+C:B]/2*BC:A	0.06



Majarian et al (2019) *PLoS Comp. Biol. 15:e1006199*

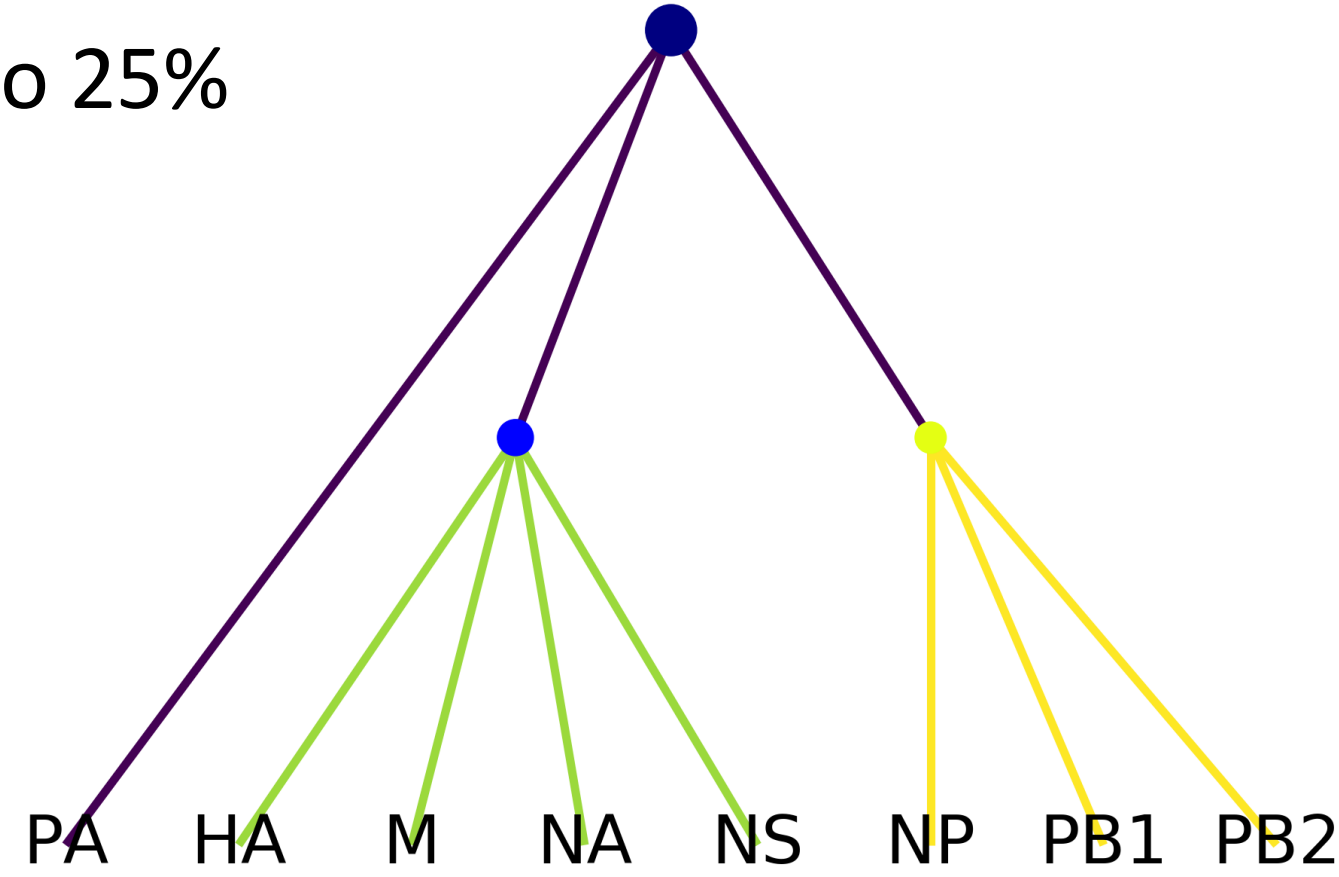
Most likely tree



Stability of solution

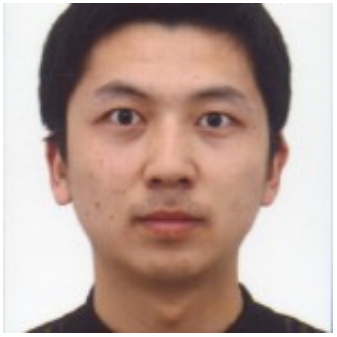
Add noise up to 25% of the average likelihood

The same tree is found!



Summary

- Very consistent complexes predicted to be major assemblies
  - (HA,M,NS) +/- NA
  - (NP,PB1,PB2) +/- PA
- Results indicate feasibility of inferring higher order complexes



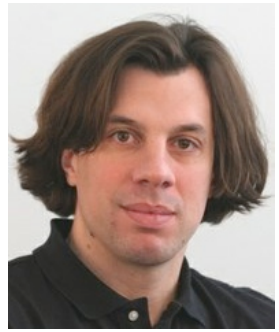
Xiongtao Ruan

Greg Johnson

Iris Bierschenk
University of Freiburg

Roland Nietzsche
University of Freiburg

Melanie Boerries
DKFZ



Hauke Busch
University of Lübeck

CELL AND NUCLEAR SHAPE MODELING FOR DYNAMICS

Concept

- Learn *kinetics* of cell organization changes from large collection of *static* images of different cells at different time points

Application to differentiation

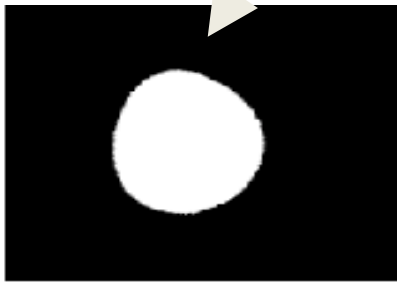
- Took 3D ***images*** of different PC12 cells stained with a mitochondrial probe at various times after addition of Nerve Growth Factor to induce differentiation
- Can't take ***movies*** due to photosensitivity of cells
- Developed approach to ***synthesize*** likely movies

Cell and nuclear shape modeling

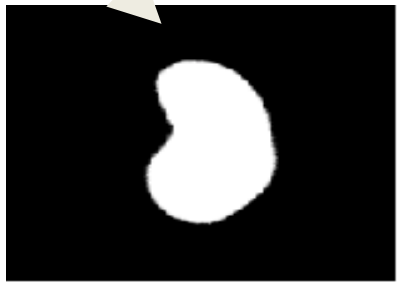
- First need methodology to construct *generative* rather than *discriminative* model of cell and nuclear shape
  - The *representation* used needs to be able to reconstruct shapes with high accuracy
  - Evaluate existing methods

Diffeomorphic modeling: Morphing one shape into another

Starting shape



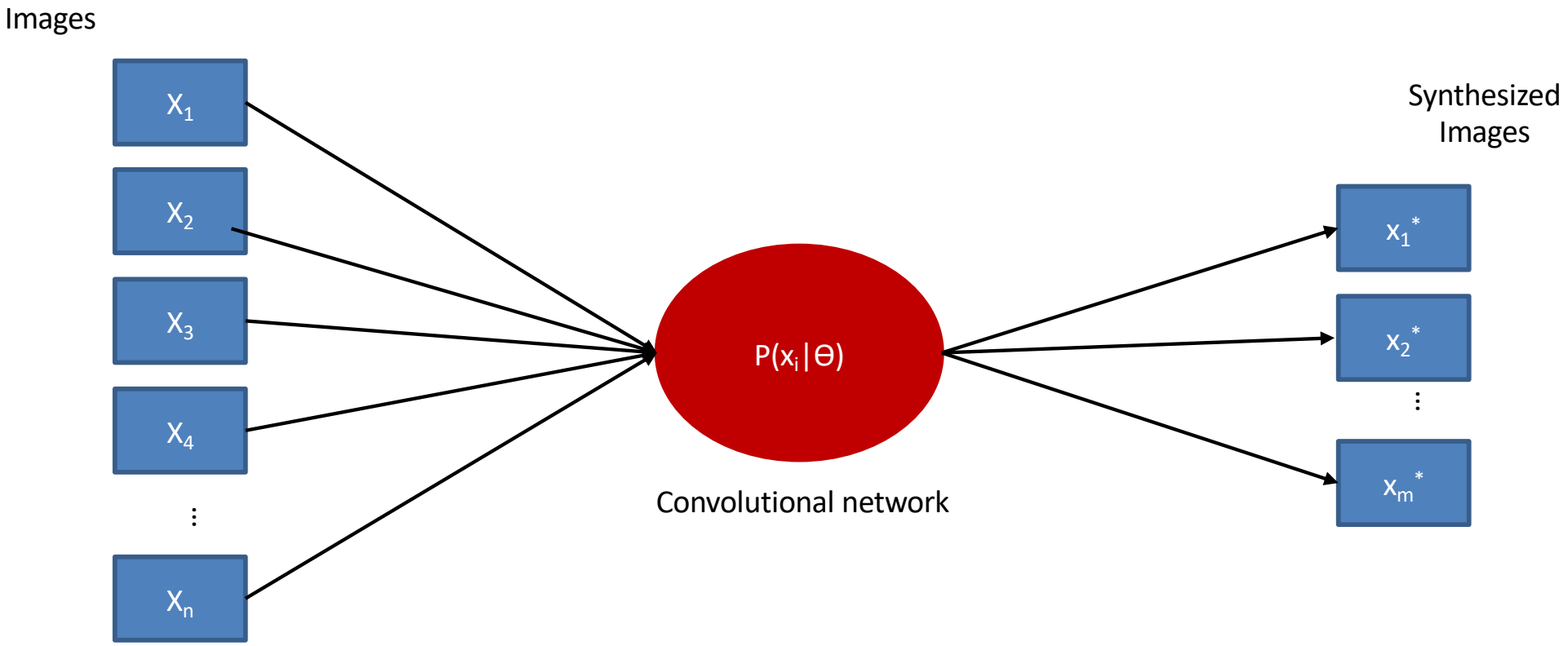
Target shape



0 0.0165 0.0191 0.0194 0.0195

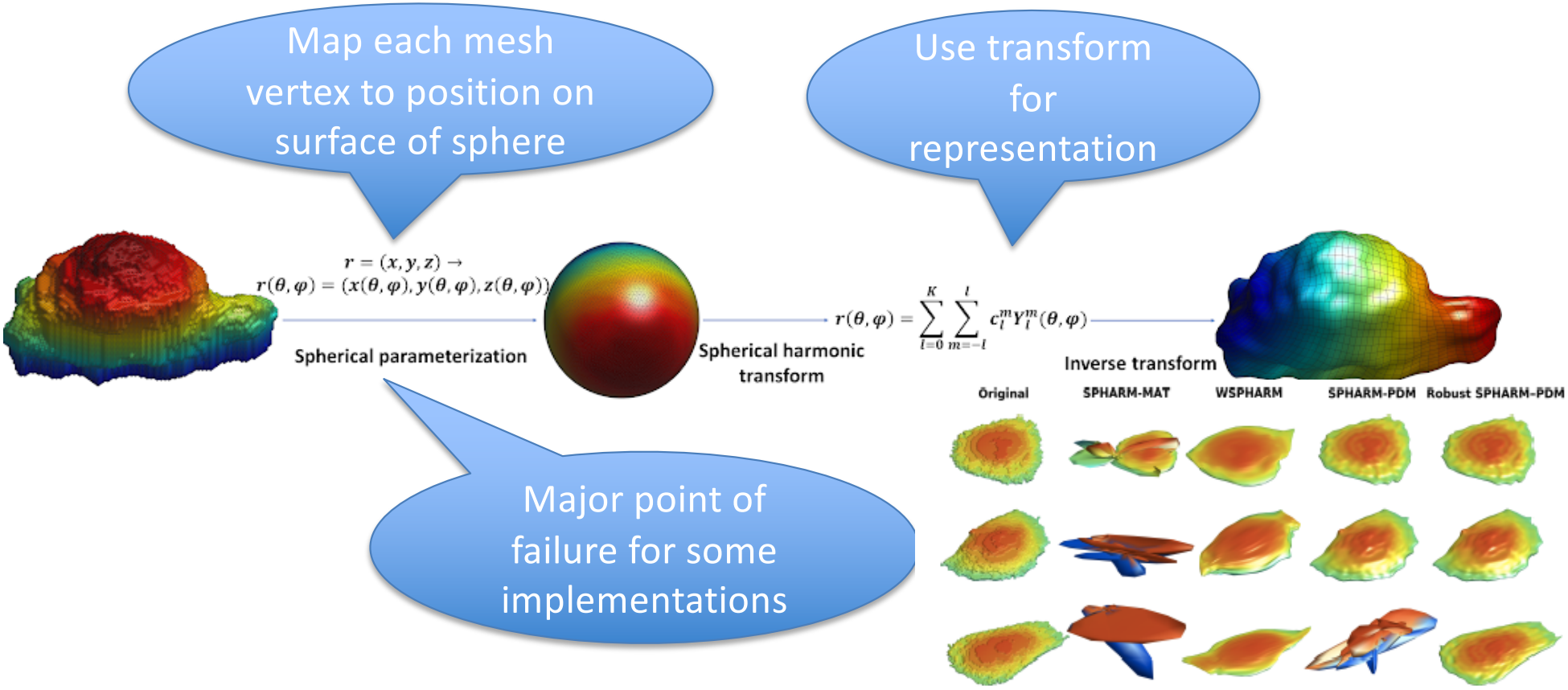
Distance

Deep learning models (e.g., autoencoders, GANs)



Ruan & Murphy (2019) *Bioinformatics., in press*

Spherical harmonic methods



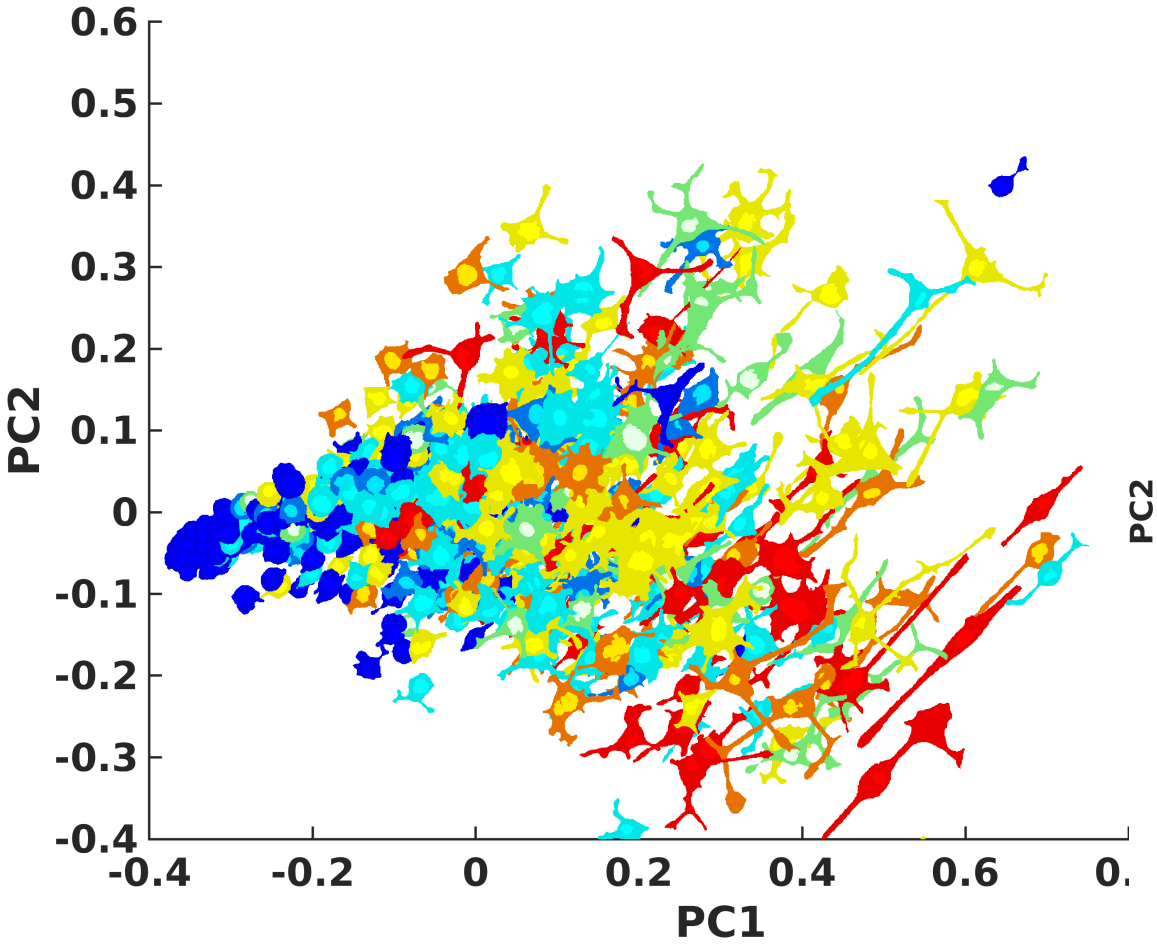
Ruan & Murphy (2019) *Bioinformatics., in press*

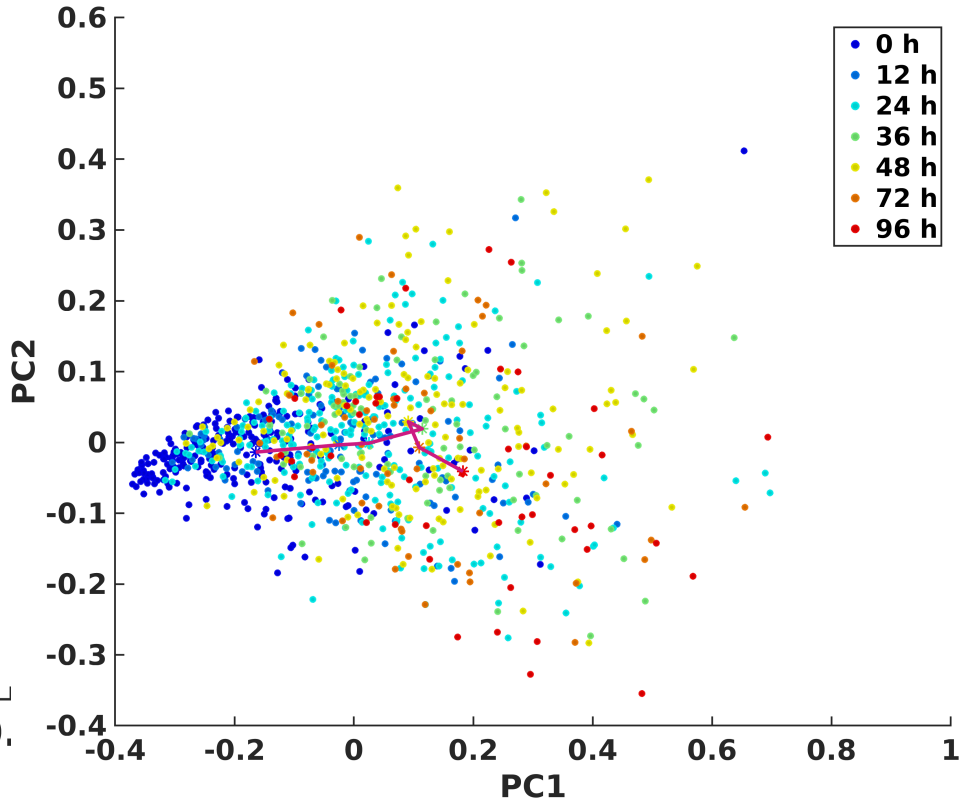
Reconstruction Errors

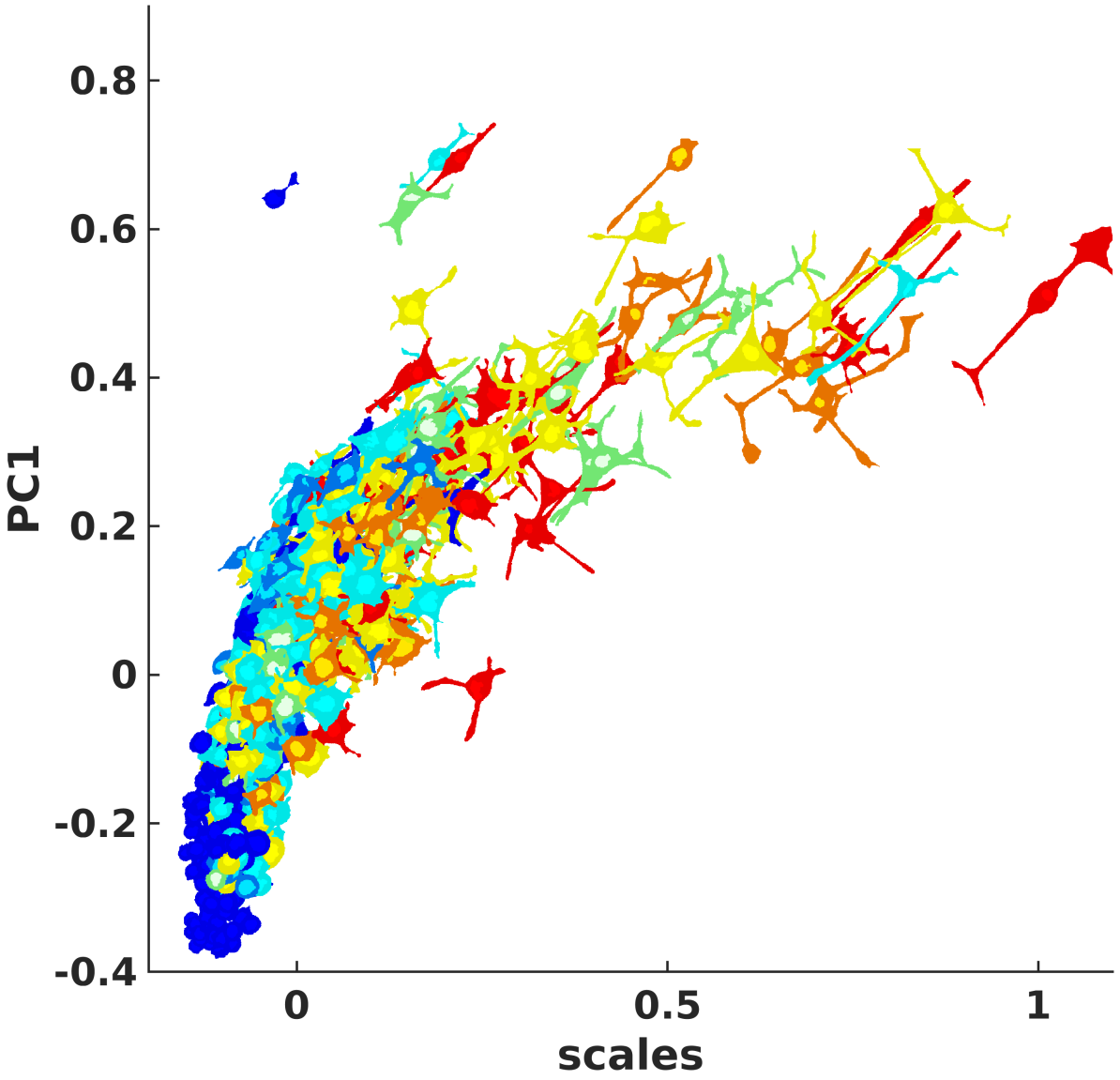
dim	datasets	spharm-rpdm	wspharm	Diffeomorphic	AE	SRAE	VAE	O-AE
7	HeLa	**8.38**	20.4	14.8	16.2	16.9	16.2	40.9
	SNL 3D	**8.64**	21.4	–	52.7	132	52.7	57.9
	SNL NR2	**12.7**	24.6	–	80.1	143.8	80.2	73.1
100	HeLa	**4.89**	20.2	–	7.93	16.9	10.4	42.4
	SNL 3D	**4.02**	16.3	–	7.45	147.3	28.4	56.5
	SNL NR2	**5.28**	21.3	–	8.19	135	80.1	81.5

Ruan et al (2019) *Mol. Biol. Cell, submitted*

PC12 cells at different times after NGF





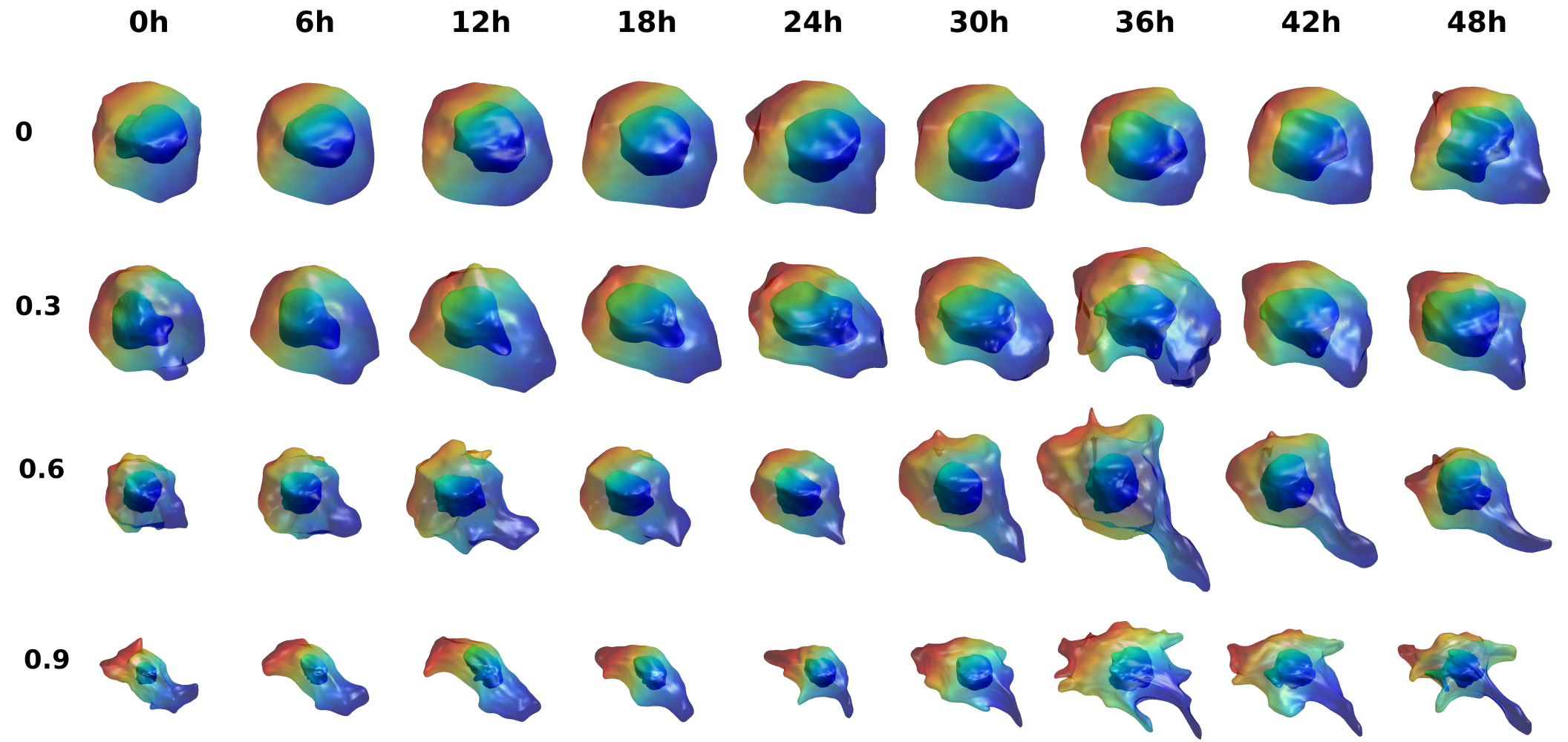


Shape vs. size

Create synthetic trajectories

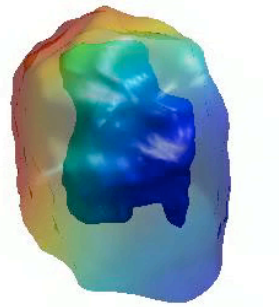
- Have over 100 cells for each time point
- Build single shape space for all cells at all time points
- Use weighted maximum bipartite matching method (Hungarian algorithm) to match cells at adjacenct time points that are the most likely to represent a single cell at two time points
- After matching, have a pseudo-trajectory for each initial cell.
- Interpolate to generate “movie”

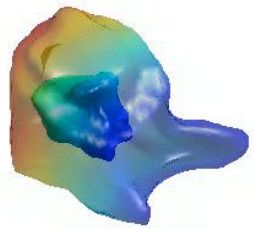
Linked “trajectories”



Ruan et al (2018) *PLoS Comp. Biol.., submitted*

Example synthetic movies





Conclusions: Cell Organization

- A number of approaches available for learning models of cell organization from images
- Important theme is inferring what we can't measure
- All tools are available in open source CellOrganizer system
- **http://CellOrganizer.org**

CellOrganizer Team

Project Leaders

Robert F. Murphy

Gustavo Rohde

Gregory R. Johnson

Collaborators

Jörn Dengjel

Hauke Busch

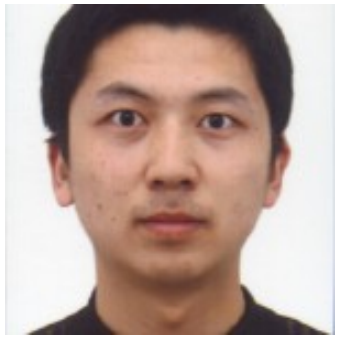
Christoph Wülfing

Melanie Boerries

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

Current Team Members



Xiongtao Ruan

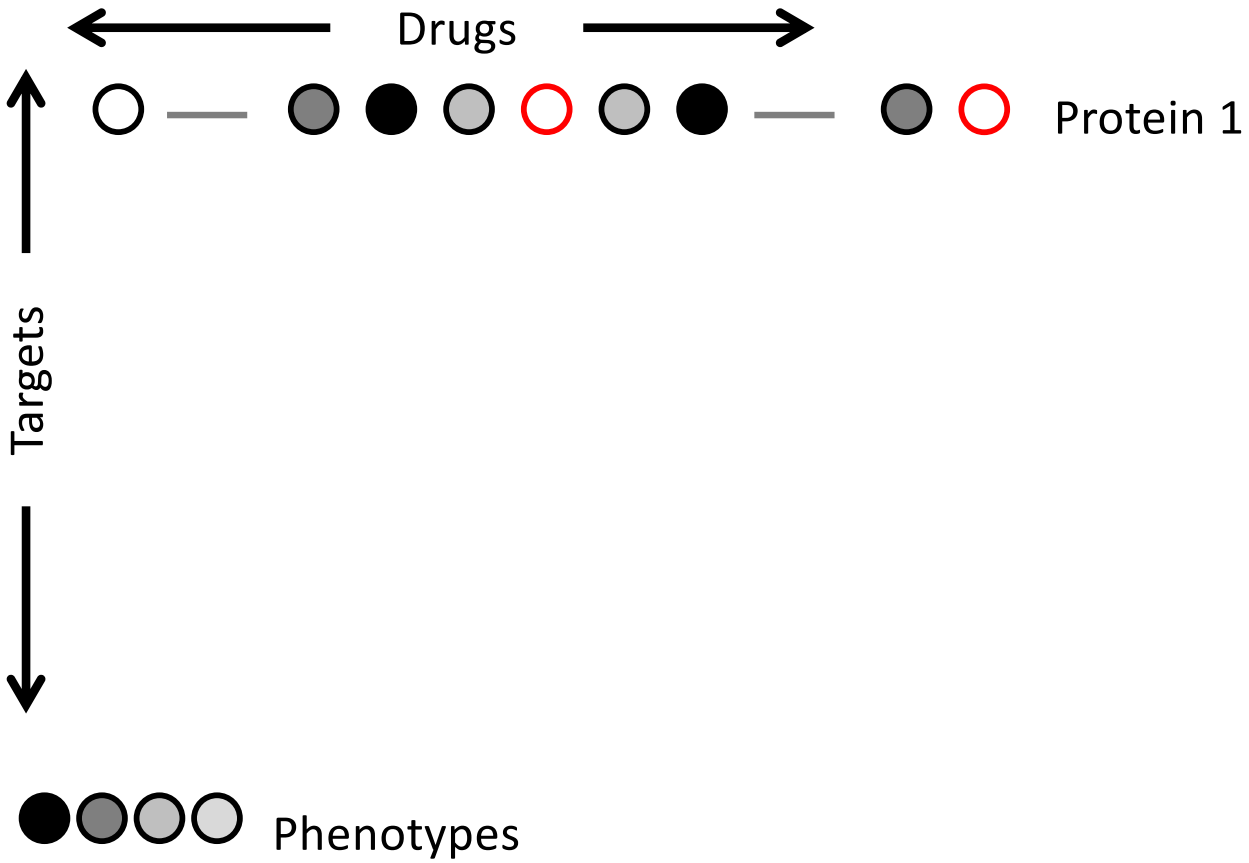
Kelvin Liu

Ivan Cao-Berg

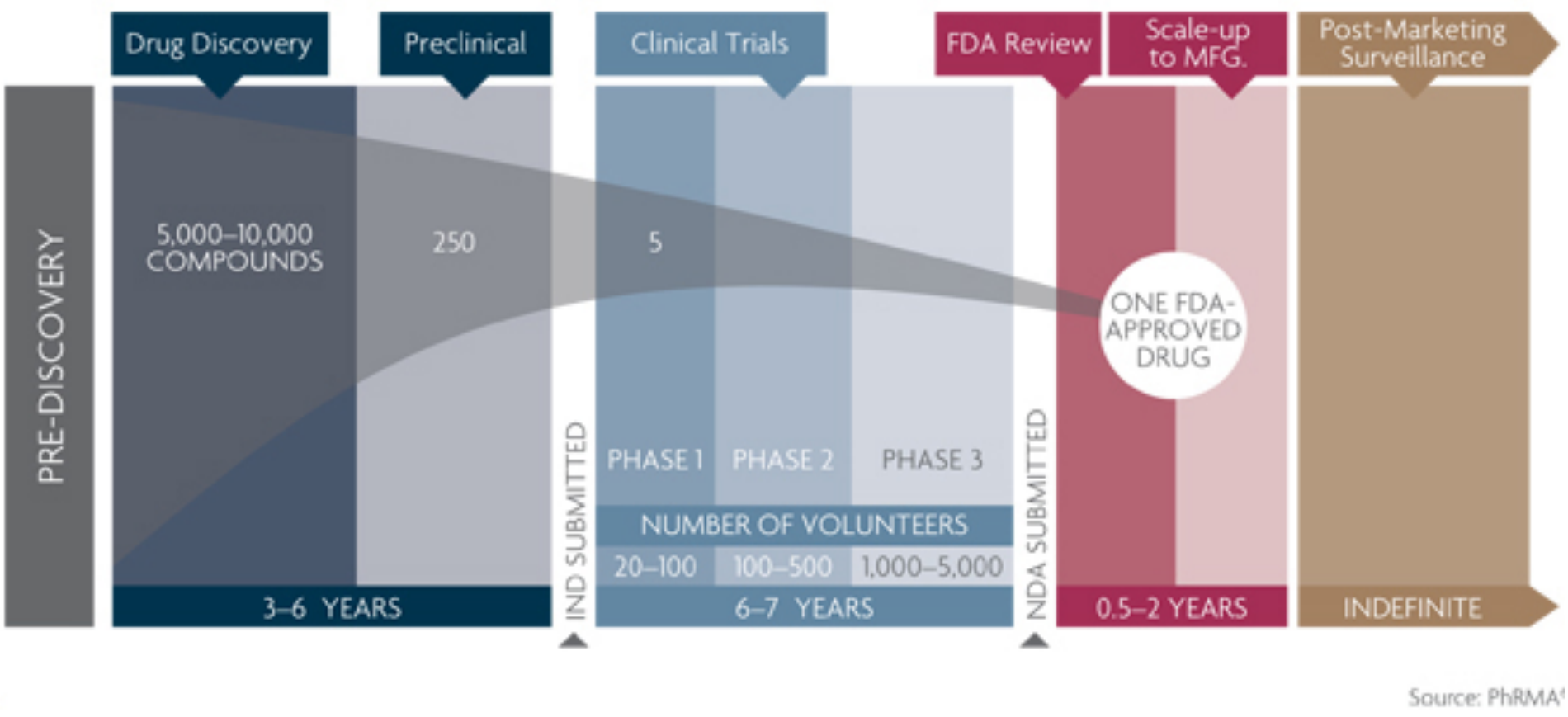
NIH GM075205, GM090033, GM103712, NSF MCB1121919

But…

- Can’t infer everything – need to choose combinations of components/proteins to image together to build the full model
- For projects like this we need *active machine learning* to decide which experiments to do and which are not needed
- For illustration, consider drug development

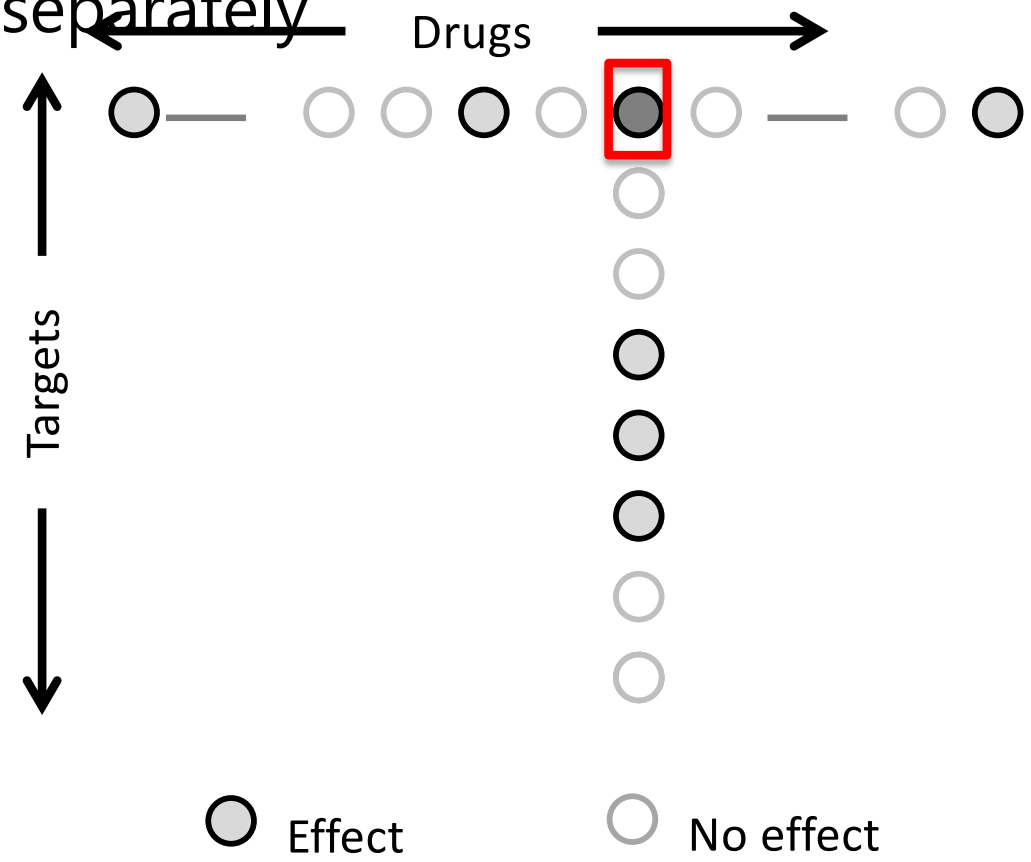

Drugs
Protein 1
Targets
Phenotypes

But the task is not just finding hits…

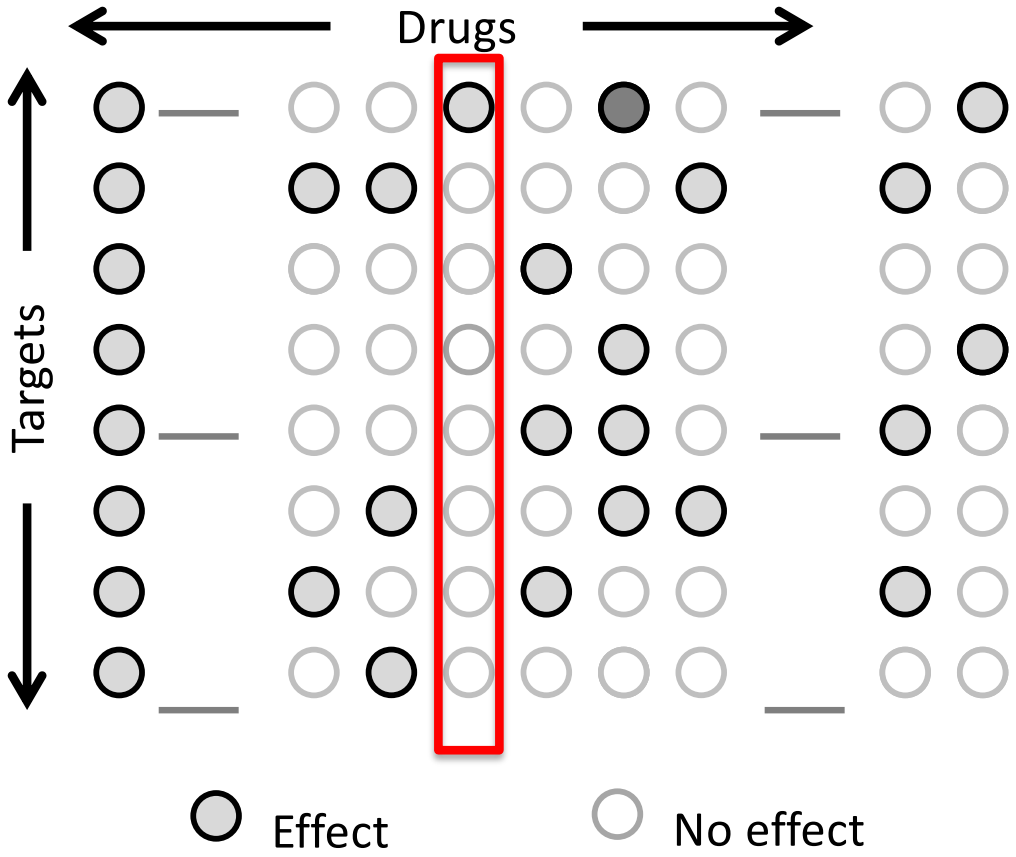


Source: PhRMA

Typical drug development: consider each target separately



Where we'd like to be: measure all drugs for all targets

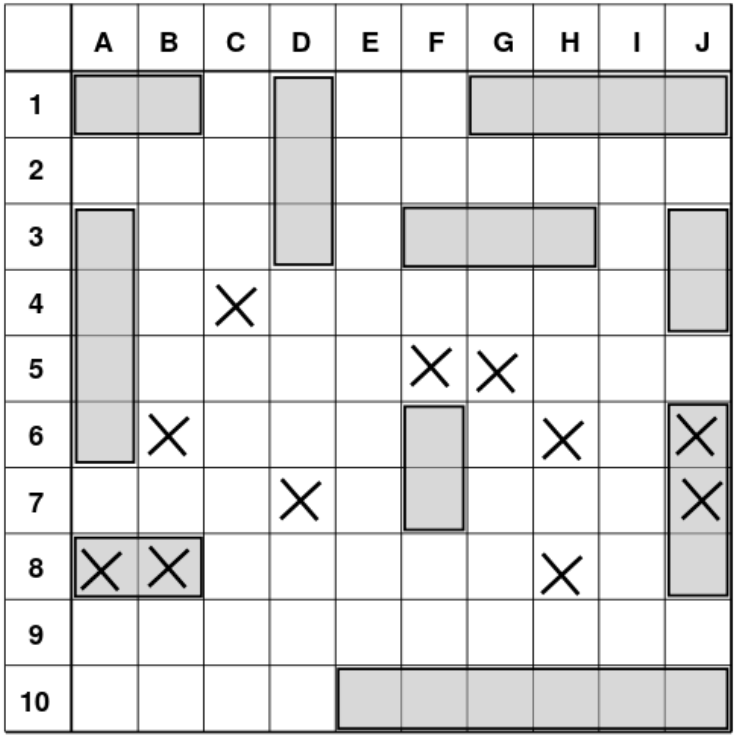


But too many combinations

- Approximately 10,000 targets
- Approximately 1,000,000 potential drugs
- How would active learning help?

Playing Battleship with Drugs and Cells

Source: Wikipedia



Testing retrospectively (with existing data)

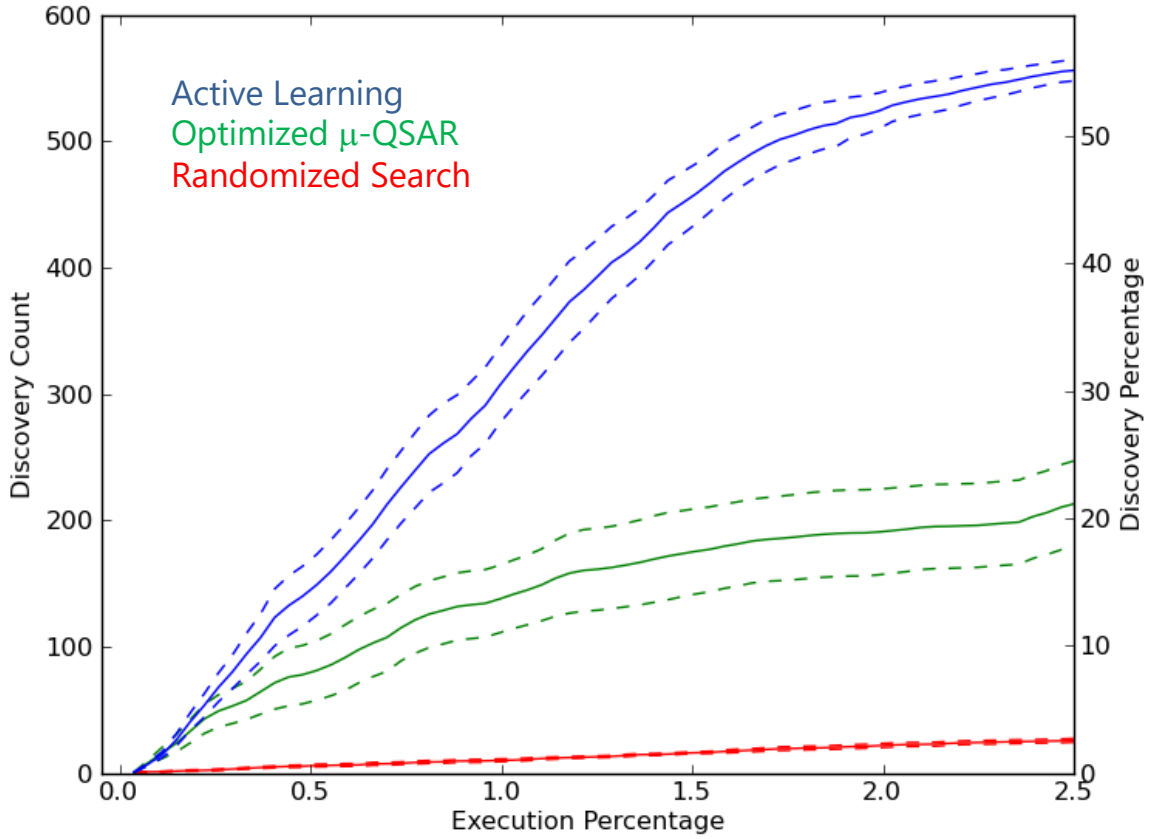
- Large database on effects of drugs on targets
- Very expensive to generate
- Would active learning have been able to save time and money?

Testing retrospectively (with existing data)

- “Hide” the PubChem data (like in Battleship) and only reveal the results when asked
  - as if we were doing that experiment for the first time
- Use different methods to choose what experiments to do

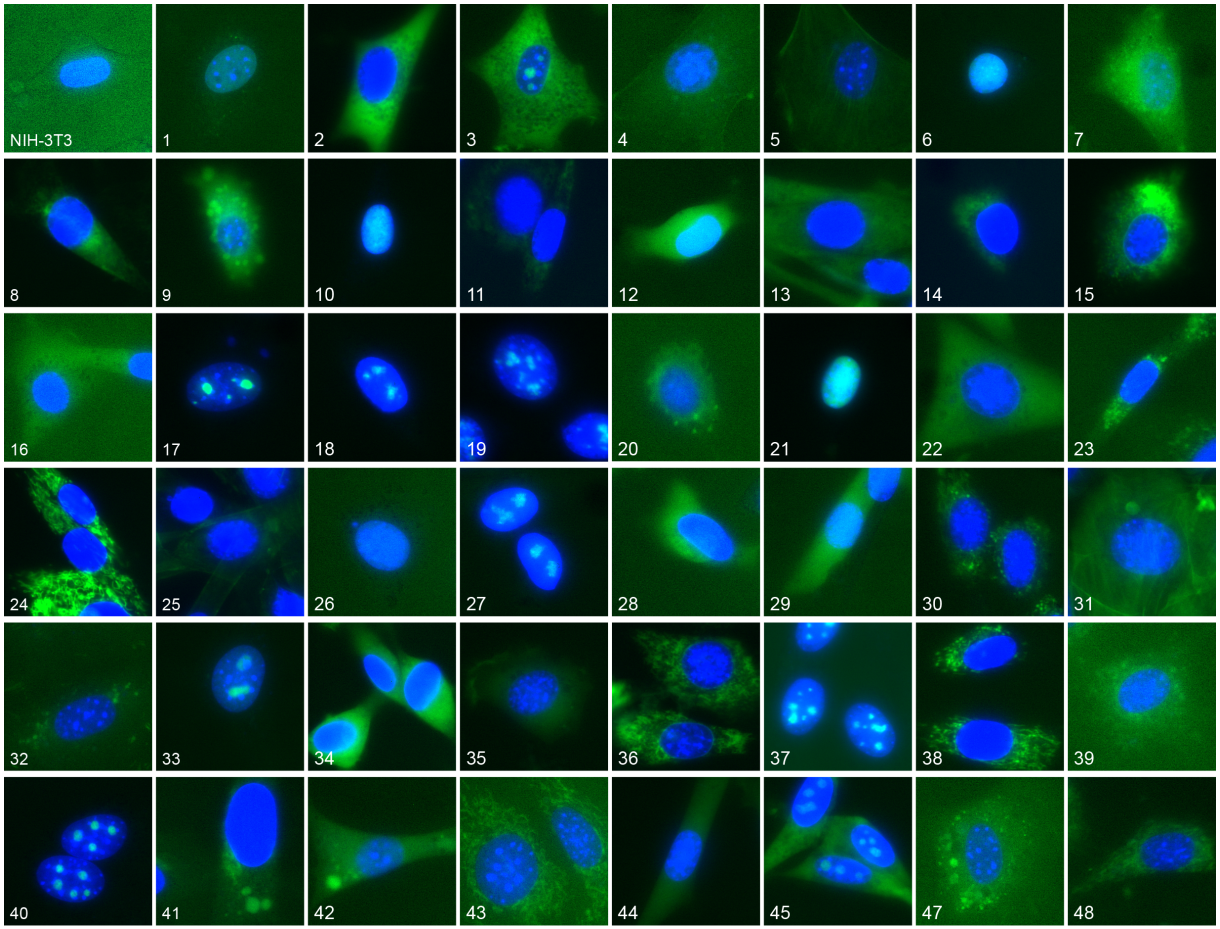
With only **2.5%** of the matrix covered, we can identify **57%** of the active compounds!

Kangas, Naik, Murphy, BMC Bioinformatics 2014



Now try this ***prospectively*** for an even harder problem

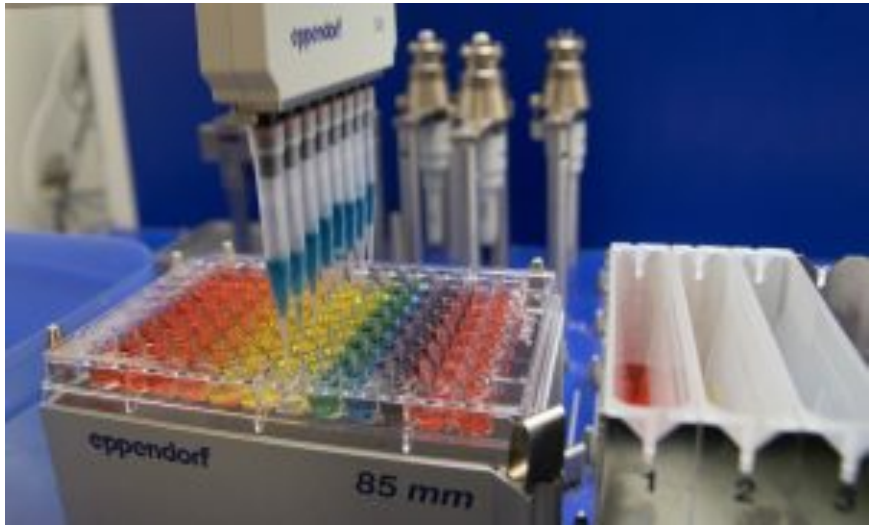
Try to learn the effects of 96 drugs upon 96 GFP-tagged proteins, *without doing experiments for all drugs and proteins,*
and where the *kinds* of effects drugs might have are unknown



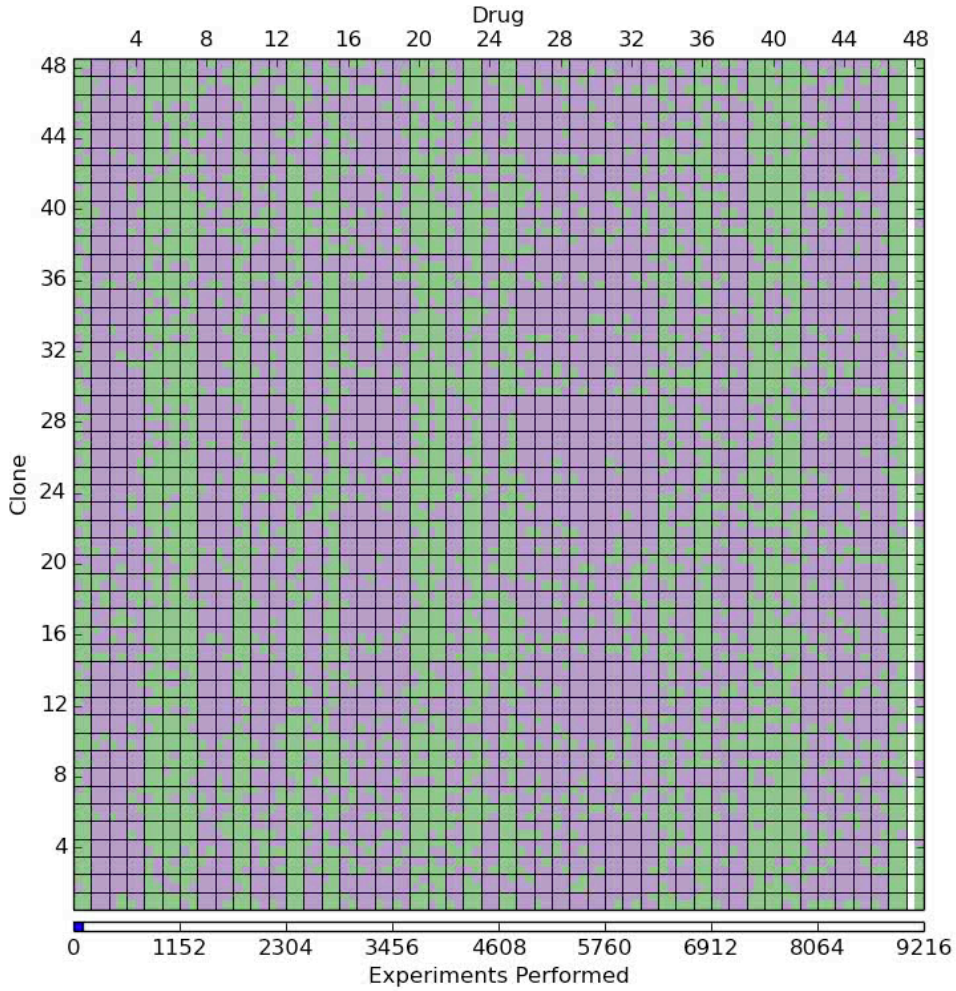
Automated Execution

Use liquid handling robots and automated microscope to execute experiments chosen by an active learner

Source: Beckman Coulter

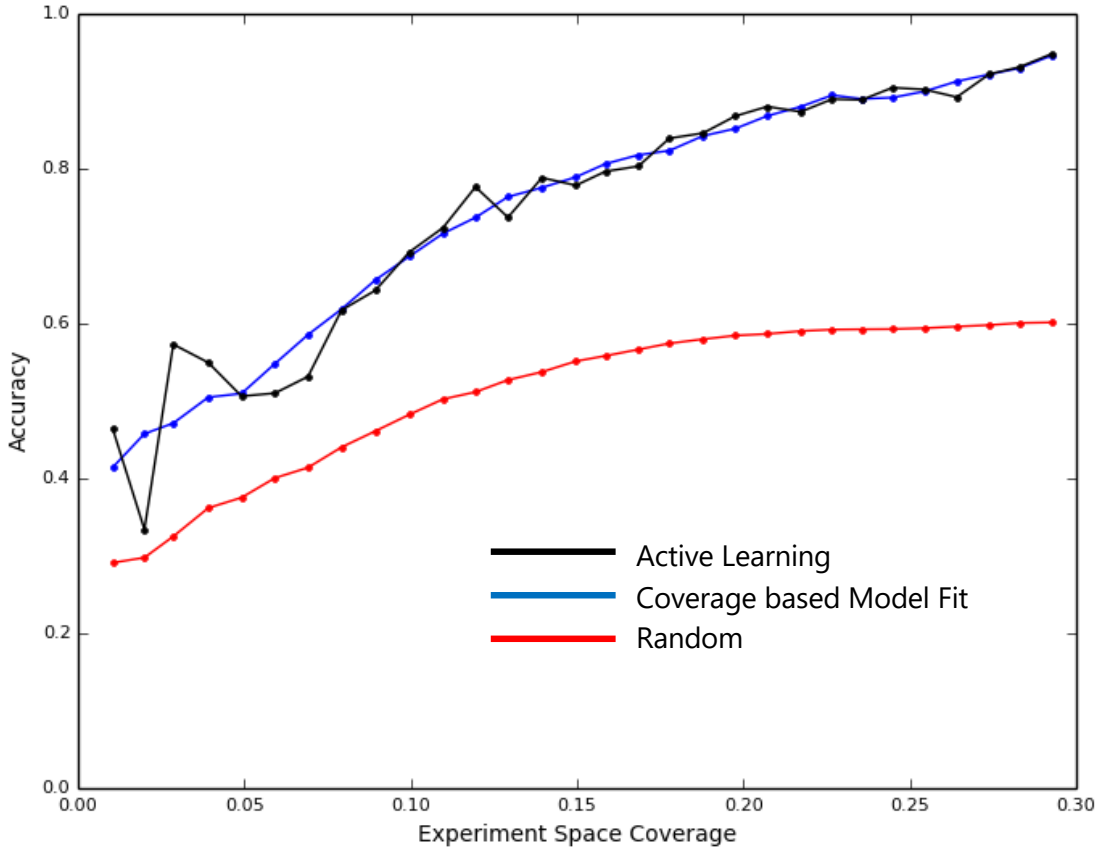


- Each small box is one drug and one target
- Green shows accurate prediction, purple is inaccurate, white shows experiments done



After doing 28% of possible experiments, model is 92% accurate and 40% more accurate than would have been obtained by random choice of experiments

Naik, Kangas, Sullivan, Murphy, eLife 2016



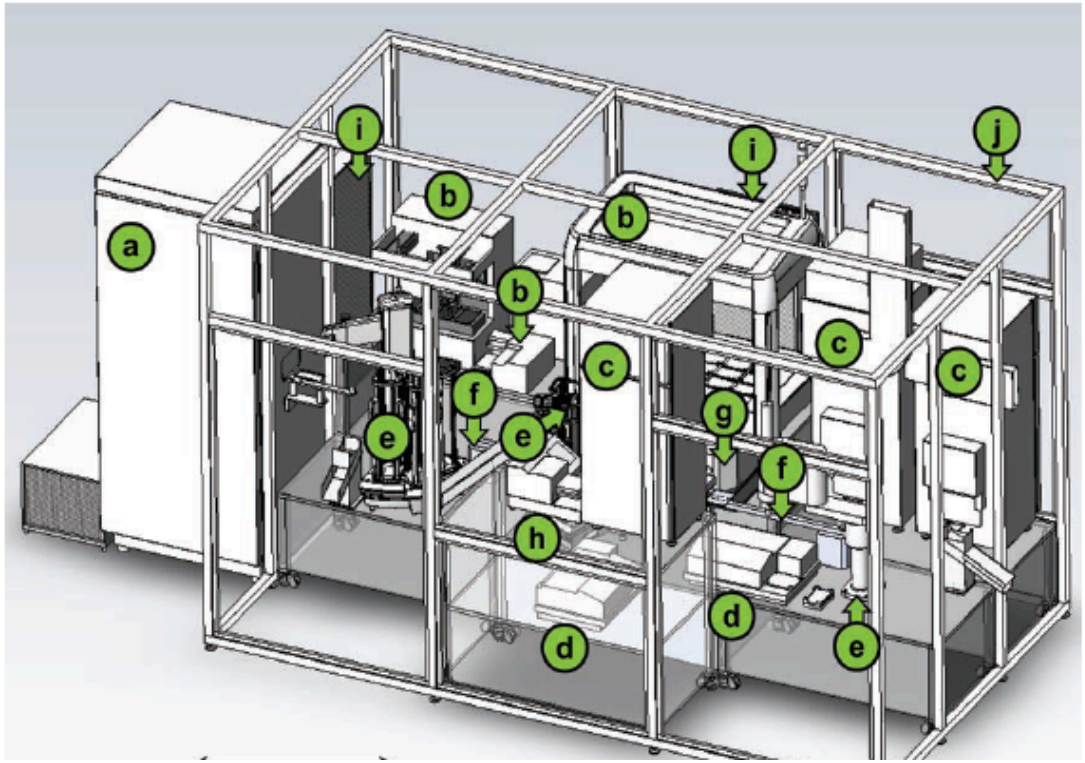
Automated science

- Additional precedent in the work of Ross King

Functional genomic hypothesis generation and experimentation by a robot scientist

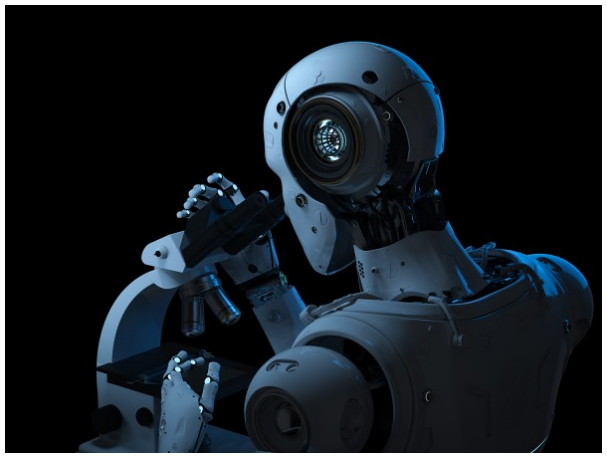
Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1], Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4] & Stephen G. Oliver[5]

[1]*Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK*
[2]*School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK*
[3]*Department of Computing, Imperial College, London SW7 2AZ, UK*
[4]*Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK*
[5]*School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK*



Automated science

- These results provide strong support for the idea of doing “Automated Science” in which not only the ***execution*** of experiments is done robotically but the ***choice*** of experiments is done robotically



- **“Self-driving instruments!”**