Data Science Provides Actionable Insights for Biologists

Brandon Gregory Horan

August 8, 2019

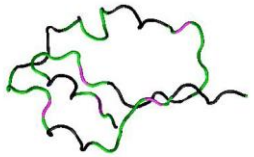
Ph.D. Funding

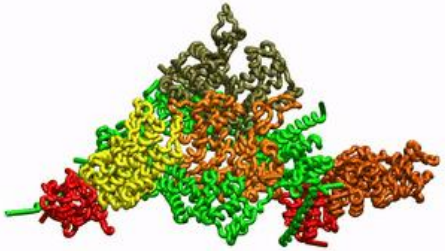
For project

NIH

For simulation time, undergraduate research

XSEDE
Extreme Science and Engineering Discovery Environment



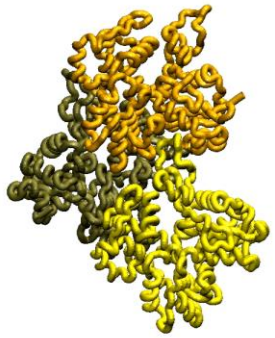


Poster Awards:

BPS PA Regional Meeting 2017

ABRCMS meeting 2018

Lehigh University



Additional Acknowledgements:

Advisors

Lehigh University Springboard

Advisees

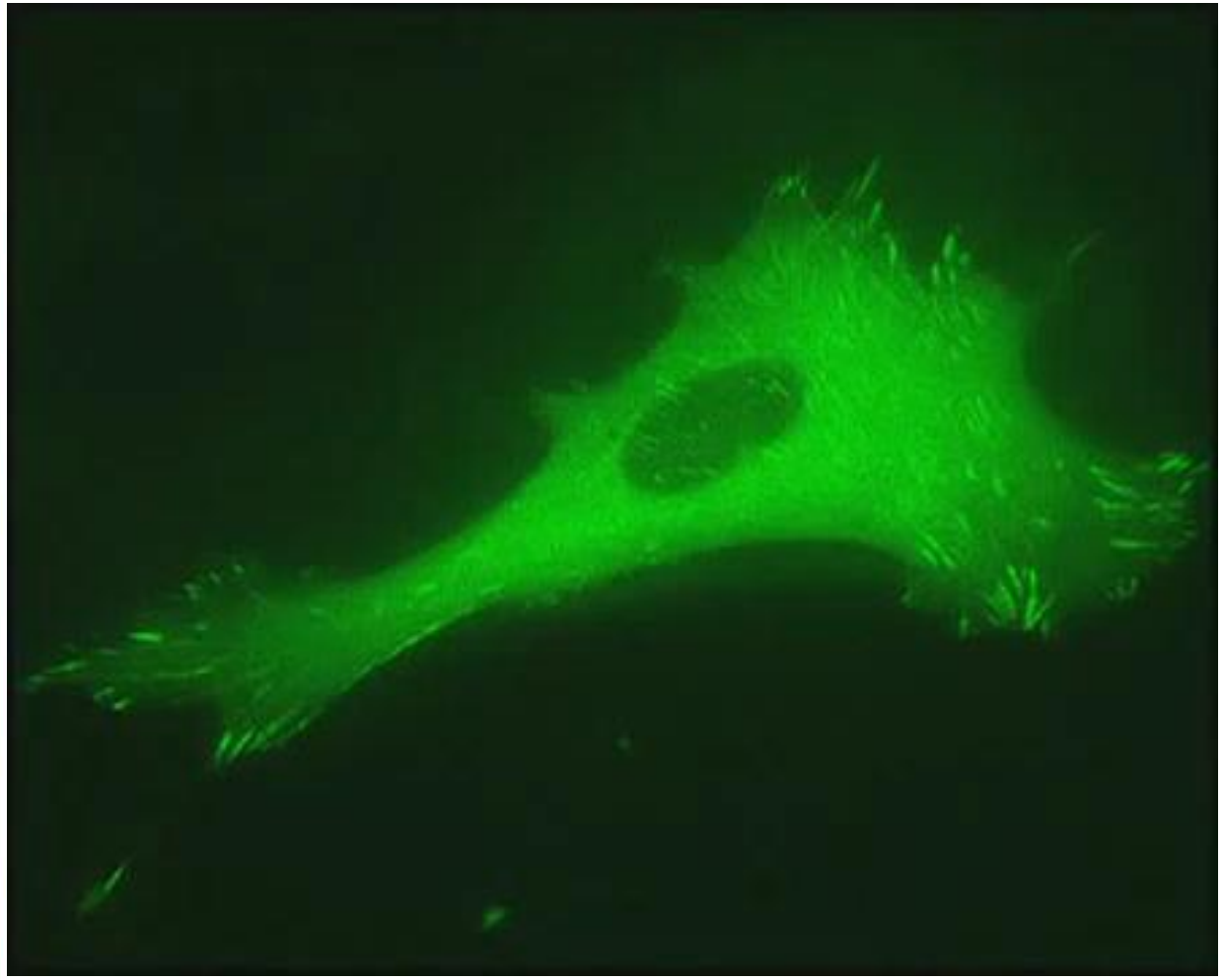
Lehigh University

Publications:

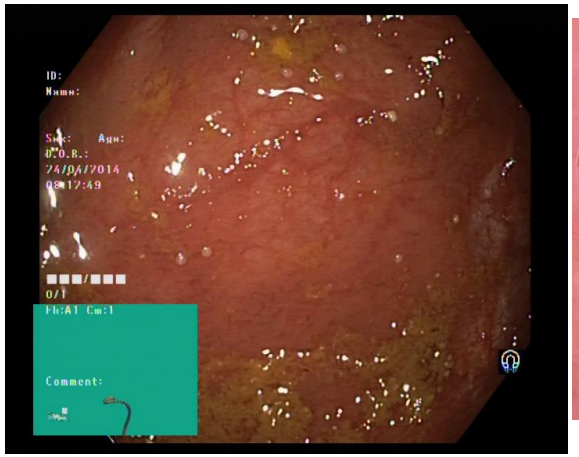
1) Vavylonis, D. & Horan, B.G. Cell Biology: Capturing Formin's Mechano-inhibition, Curr. Bio. 2017
2) Horan, B. G. and Zerze, G. and Kim, Y. C. and Vavylonis, D. and Mittal, J. Computational modeling highlights disordered Formin Homology 1 domain's role in profilin-actin transfer. FEBS Letters 2018. ***FEATURED RESEARCH ARTICLE***
3) Horan, B. G. and Vavylonis D. Insights into actin polymerization and nucleation using a coarse grained model. *In review* (also on bioRxiv)
4) Quintana, F. M. and Kodera, A. and Horan, B. G. and Yamashiro, S. and Mittal, J. and Watanabe, N. and Vavylonis, D. Mechanism of formin-mediated actin polymerization: alternate delivery of profilin-actin to the barbed end. *Manuscript in preparation*
5) Poddhar et al. Cytokinesis triggered frequent fluctuations of intracellular calcium concentration in fision yeast cells. 2019. *Manuscript in preparation*

Overview

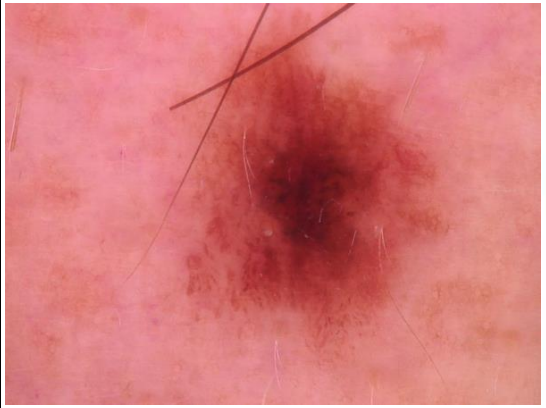
- Cellular processes such as cell migration, division, shape regulation driven by actin cytoskeleton.
- Presence/proper functionality of actin & cofactors important for avoiding health issues such as hearing loss, neurodegenerative diseases, & cancer. *Horan 2019*



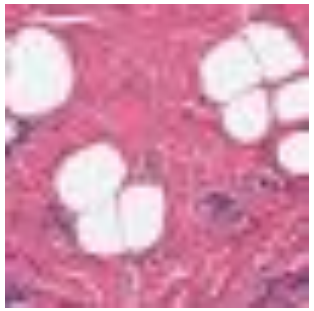
- Automated detection of disease is a critical step in treating disease before they become deadly. Modern modeling approaches can provide success in this process. *Rajaraman et al 2018 PeerJ*



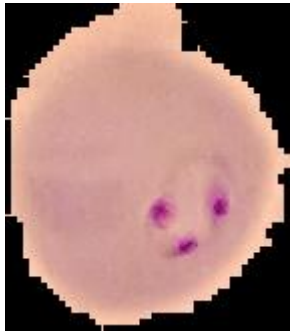
Ulcerative colitis



Skin Cancer



Breast Histopathology

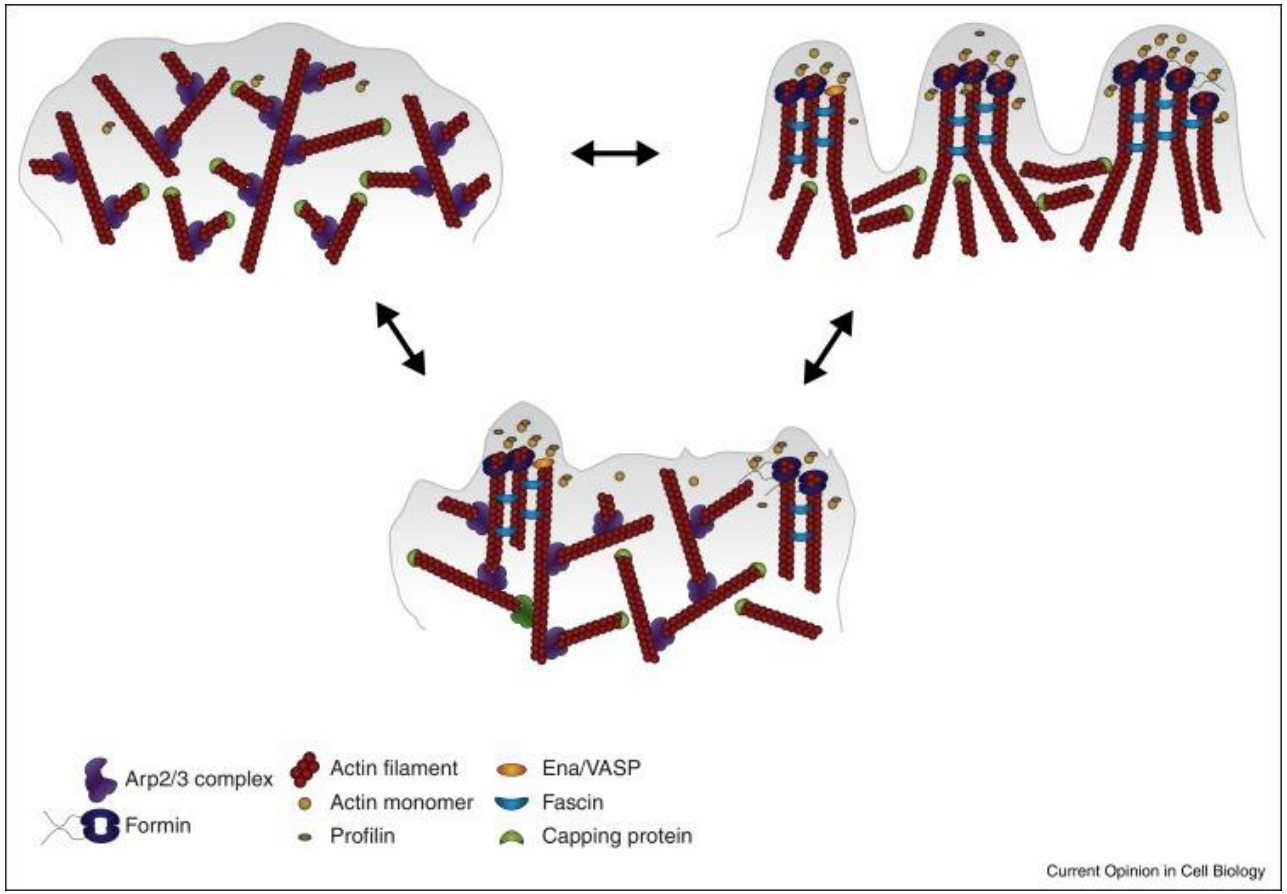


Malaria

The Actin Cytoskeleton Structure

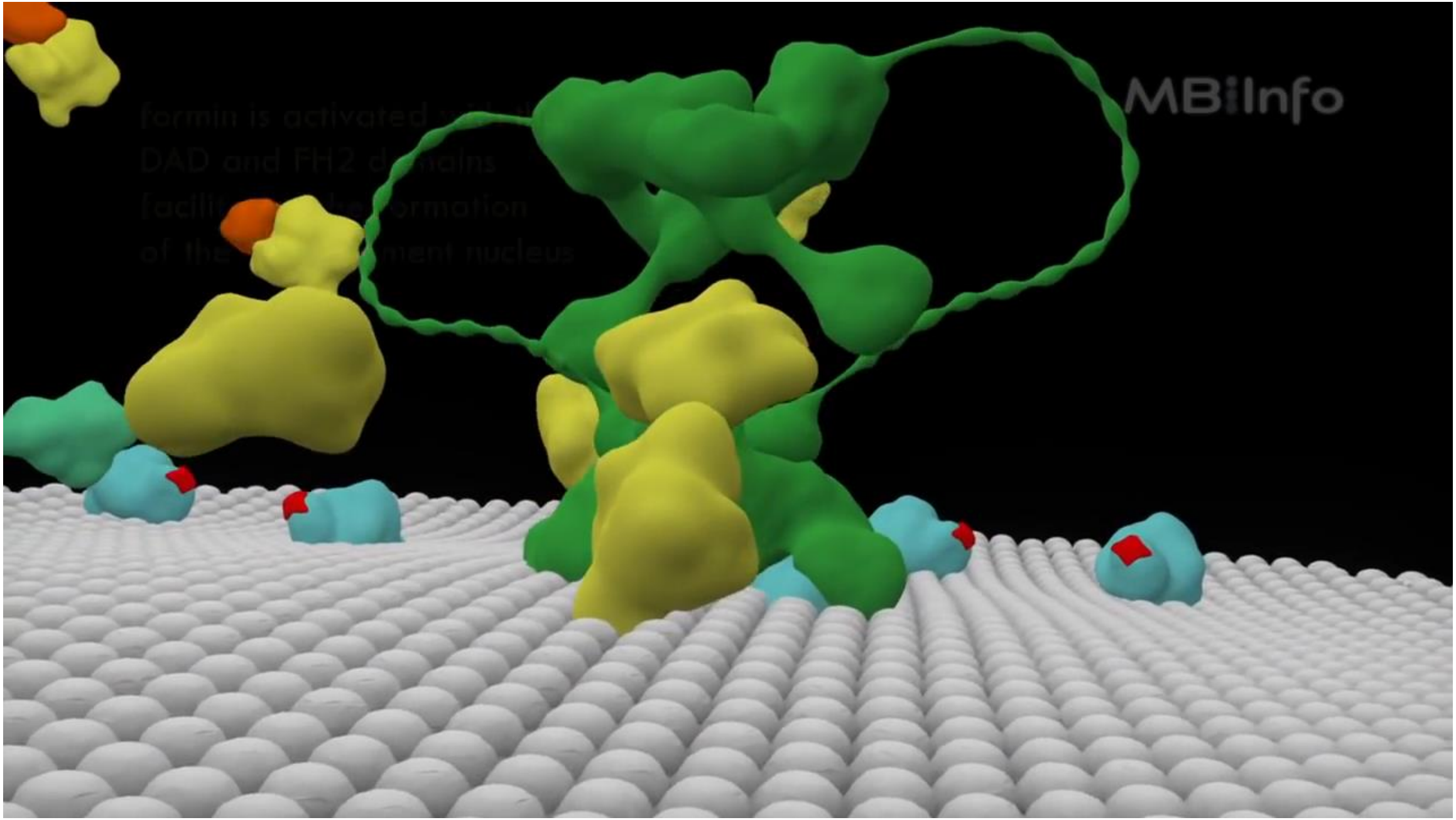
- Goal: understand molecular mechanisms behind assembly of actin cytoskeleton
- Purpose: Can lead to design of drugs targeted at specific aspects of actin system for treating disease.

- The actin cytoskeleton is assembled primarily with the help of two regulators: formin (responsible for linear protrusions) and Arp2/3 complex (branched network).



Swaney & Li Curr. Op. Cell Bio. 2016

Formin-Mediated Polymerization

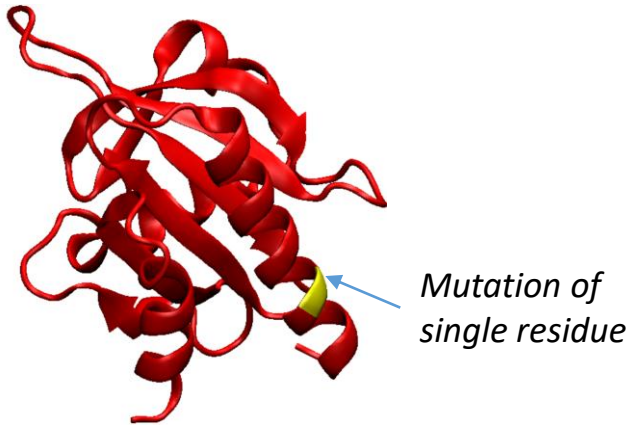


Multiscale Actin Regulation is Biologically Critical

- Regulation of actin polymerization begins at the level as small as the nanoscale, and has affects at scales as large as the cell.

Nanoscale Difference

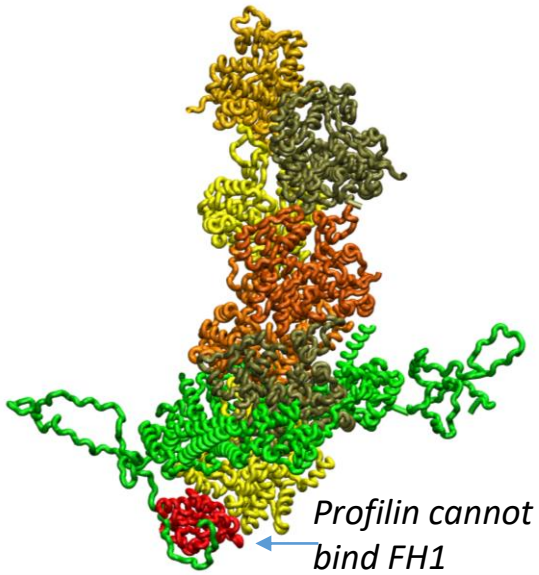
Point mutation in profilin inhibits binding of formin FH1 domain to profilin (Kursula et al. 2008 J. Mol. Bio).



All-atom/coarse grained molecular dynamics

Microscale Affect

Formin FH1 is unable to accelerate actin polymerization.

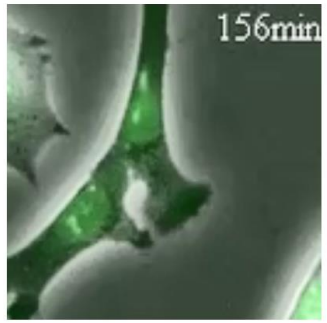


Coarse-grained/ultra-coarse-grained molecular dynamics

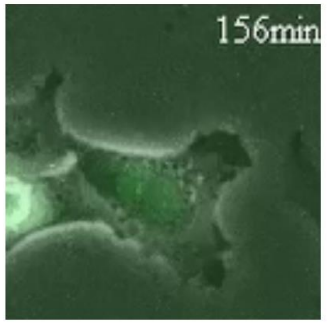
Cell Scale Affect

Cytokinesis defects/failure, lack of stress fiber formation, etc.

With Formin



Without Formin

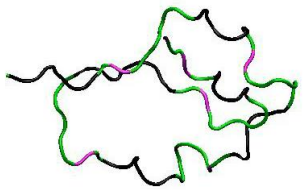


Watanabe et al. Molecular Biology of the Cell (MBoC) 2008.

Continuum-level differential equation/kinetic Monte-Carlo

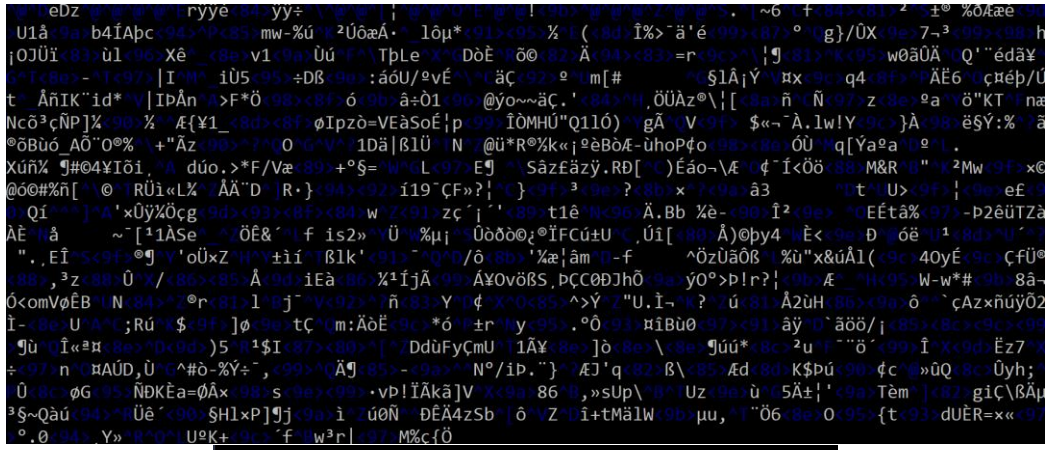
Molecular Dynamics Simulations

- Molecular Dynamics is a commonly used modeling technique which employs classical mechanics to study a wide variety of systems, including biomolecules.
- The process is straight-forward:
  1) Define system of particles to simulate
  2) Define interactions between particles
  3) Numerically integrate classical equations of motion



Horan et al FEBS Letters 2018

- Base output is a binary file which can be converted to coordinates in a timeseries.
- Features must be extracted for data analysis.



```
_1125
Atoms. Timestep 0.0
 CA     46.596   13.745   -2.786
 CA     50.428   13.839   -2.625
 CA     51.977   13.331   -6.210
 CA     48.580   12.829   -7.888
 CA     45.109   14.450   -7.967
 CA     43.242   11.517   -6.231
 CA     39.714   11.911   -4.527
 CA     37.213   10.250   -2.100
 CA     34.757    7.571   -3.259
```

- Examples of analysis approaches:
- Use standard metrics or identify/define alternate metrics & examine their timeseries/distributions
- Apply machine learning to extract patterns in data

FH1: IDP With High Propensity PPII Helices (1 of 2)

Flexible: a reasonable representation for FH1?

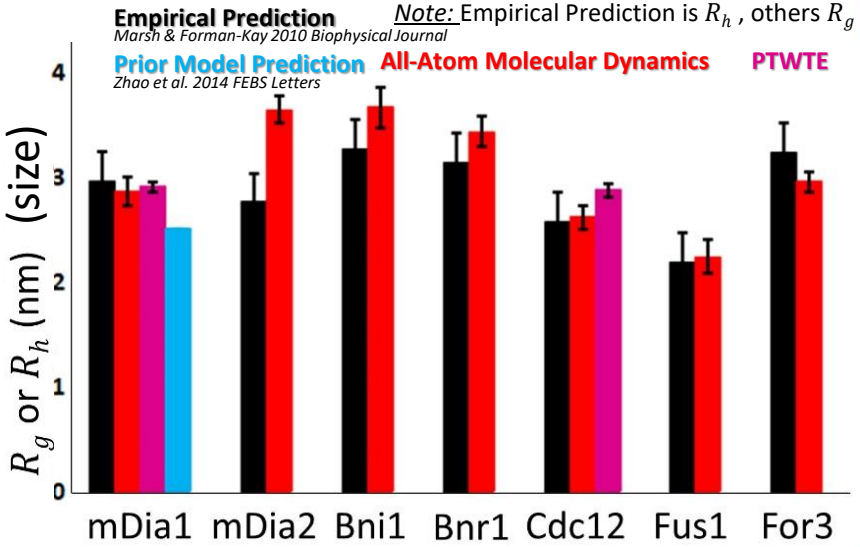
- Well-studied representative sequences of formin FH1 domains selected for this exploratory analysis.
- Abundance of proline in FH1 domains (>>5%) *(Morgan & Rubenstein 2013 PLOS One).*
- FH1 believed to be disordered (*Vavylonis et al. 2006 Mol. Cell*).

mDia1 (6PRM)	IPPPPPLPGV GPG	ASIPPPPPLP	GATAIPPPPP	LPGATAIPPP	PPLPGGIGIP	PPPPPLPGSV	GVPPPPPLPG
mDia2	AELQAFKSQF PPPPPLGFLG	GALPPGIKIP GQSSIPLNLP	LQPSVEGEAG F	PSALPPAPPA	LSGGVPPPPP	PPPPPPPPLP	GMPMPFGGPV
Bni1	LSTQSSVLSS PKGEIPPPPP	QPPPPPPPPP LPSVLSSSTD	PVPAKLFGES GVIPPAPPMM	LEKEKKSEDD PASQIKSAVT	TVKQETTGDS SPLLPQSPSL	PAPPPPPPPP	PPPPMALFGK
Bnr1	QLVPEVVKLP LPESLSMNKG	QLPPPPPPPP PSNHDLVIPP	PPPLPQSLLT APPLPNGLLS	EAEAKPDGVS SSSVSINPTT	CIAAPAPPPL T	PDLFKTKICG	AVPPPPPPPP
Cdc12	NNSKITNFDI PPLVSAAGGK	PNDATSLPTI FVSPAVSNNI	ITHPTIPPPPP SK	PLPVKTSLNT	FSHPDSVNIV	ANDTSVAGVM	PAFPPPPPPP
For3	SDTVEEQQKL GGSRYYAPAP	LLKSPPPPPP QAEPEPKIDE	AVIVPTPAPA TSLTEEQKIQ	PIPVPPPAPI LEEARKQRKA	MGGPPPPPPP ADDAARAAIE	PGVAGAGPPP	PPPPPPAVSA
Fus1	MRIKEVIDGN VYASKIP	PFKAPPPAPL	PPPAPPLPTA	MSSLQKFEKN	DSQIFRKTII	IPENISIDDI	FKFCSGSESE

At least 3 successive proline residues **Other proline residues**

Horan et al. FEBS Letters 2018

- Radius of Gyration (R_g) and Hydrodynamic Radius (R_h) are similar measures of size.
- R_g is average rms distance between an object and its COM. Easily calculated by simulation, difficult to calculate experimentally.
- R_h represents average conformation of a spherical approximation of polymer; difficult to calculate by simulation, more suited for experiment.



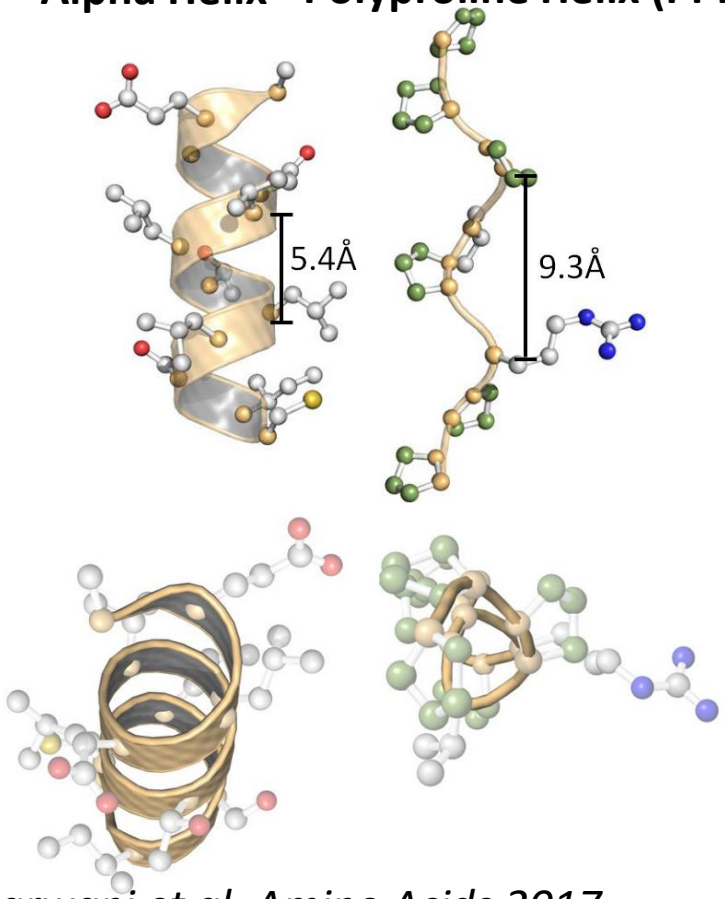
- Simulation radius of gyration agrees well with empirical prediction for size (hydrodynamic radius) of IDP.

FH1: IDP With High Propensity PPII Helices (2 of 2)

Flexible: a reasonable representation for FH1?

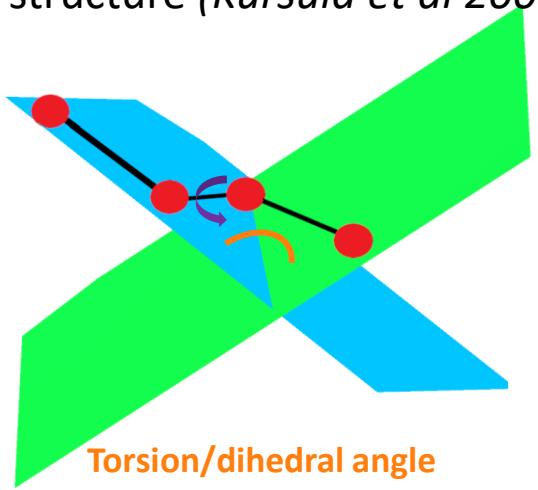
- All-atom molecular dynamics simulations reveal that FH1 domain is indeed disordered, with the PRMs having high polyproline helix propensity.
- The polyproline helix is an extended structure seldom identified in proteins.

Alpha Helix **Polyproline Helix (PPII)**

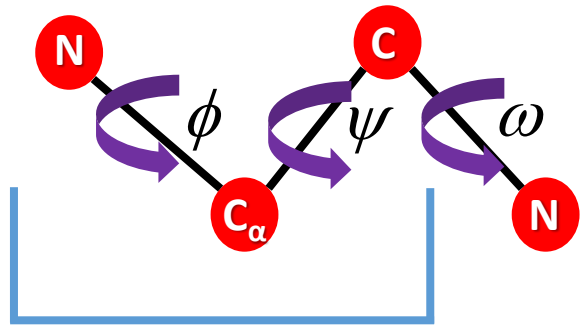


Narwani et al. Amino Acids 2017

- Polyproline is in ppii conformation in profilin-bound FH1 crystal structure *(Kursula et al 2008 J. Mol. Bio.)*



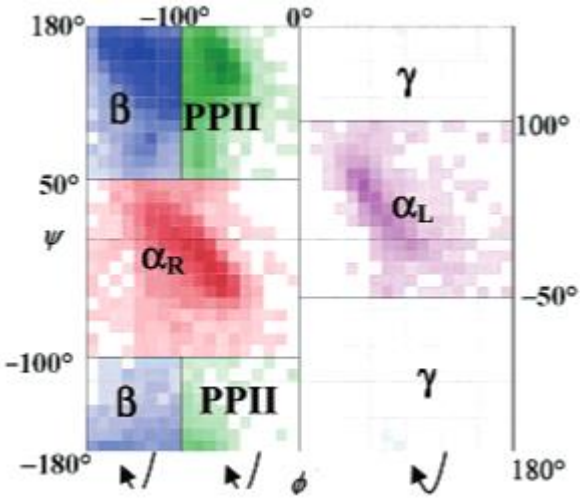
Torsion/dihedral angle



Single Amino Acid Backbone

Polyproline Backbone Dihedral Angles in Crystal Structure (deg)

ψ	ω	ϕ
141	182	–76
158	176	–71
146	182	–80
163	172	–71
157	184	–65
149	183	

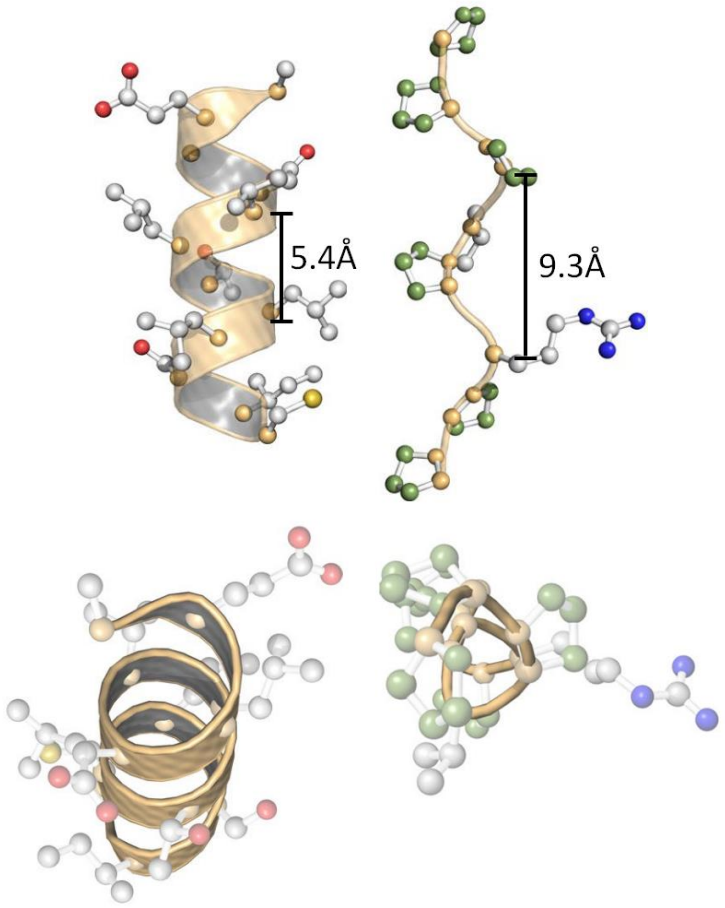


Jha et al. Biochemistry 2005.

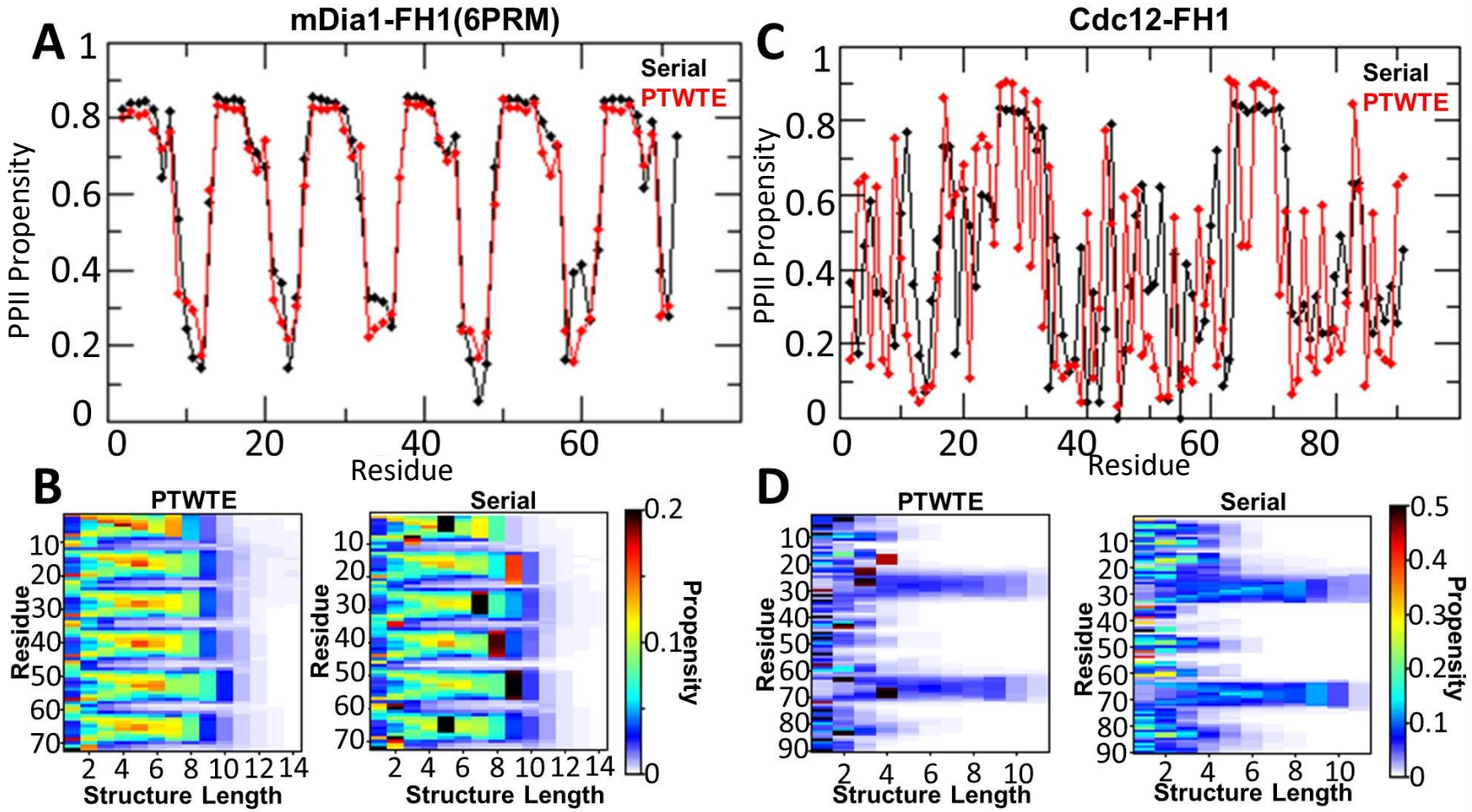
FH1: IDP With High Propensity PPII Helices (2 of 2)

Flexible: a reasonable representation for FH1?

- All-atom molecular dynamics simulations reveal that FH1 domain is indeed disordered, with the PRMs having high polyproline helix propensity.
- The polyproline helix is an extended structure seldom identified in proteins.



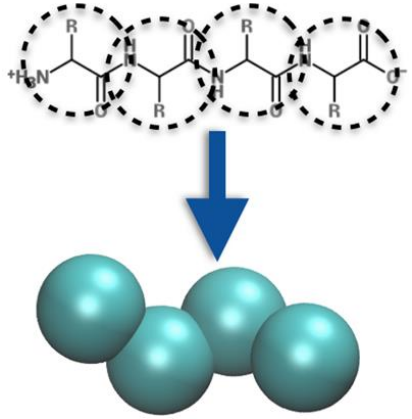
Narwani et al. Amino Acids 2017



Horan et al. FEBS Letters 2018

Coarse-Grained Model Retains FH1 Size

Effect of profilin(-actin) occupancy on FH1?



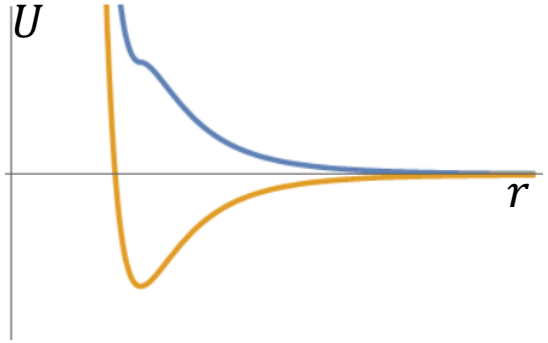
Bead Definition

- Alpha-carbon based model, retains residue's charge, mass, interaction radius.

Bead Interactions

$$U = U_{bonds} + U_{nonbonded\ pairwise} + U_{electrostatic}$$

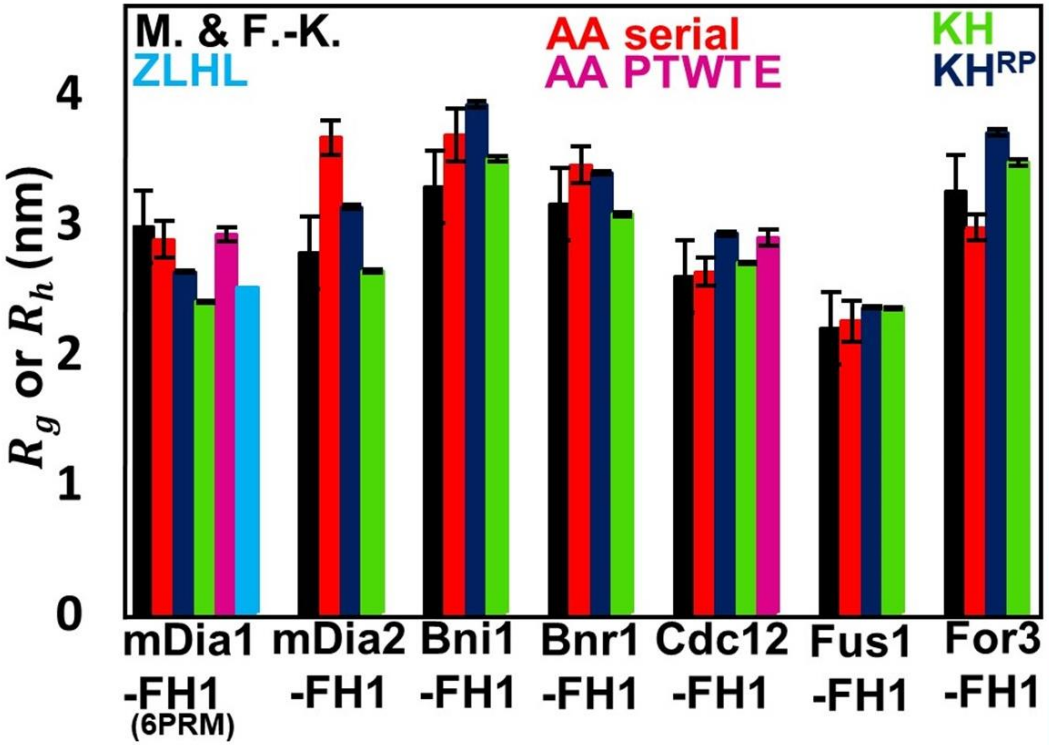
- Harmonic bonds.
- Debye-Hückel electrostatics
- Miyazawa-Jernigan based attractive or repulsive Lennard-Jones like nonbonded pairwise interactions.



- Force on a given bead is given by the interaction potential, along with two additional terms to account for the water implicitly (much faster than explicit water) for friction and thermal fluctuations.

$$\vec{F} = -\nabla U + \vec{F}_{friction} + \vec{F}_{random}$$

- Cα model of FH1 is consistent with size of FH1. Polyproline helices are maintained explicitly by initializing FH1 in PPII and keeping the PRMs as rigid bodies.



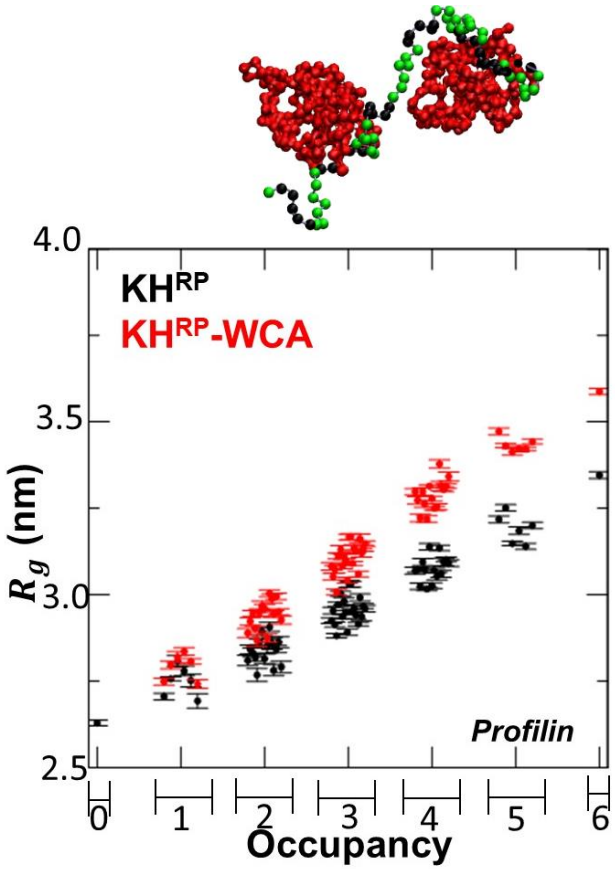
Model based on: Kim & Hummer 2008 J. Mol. Bio.

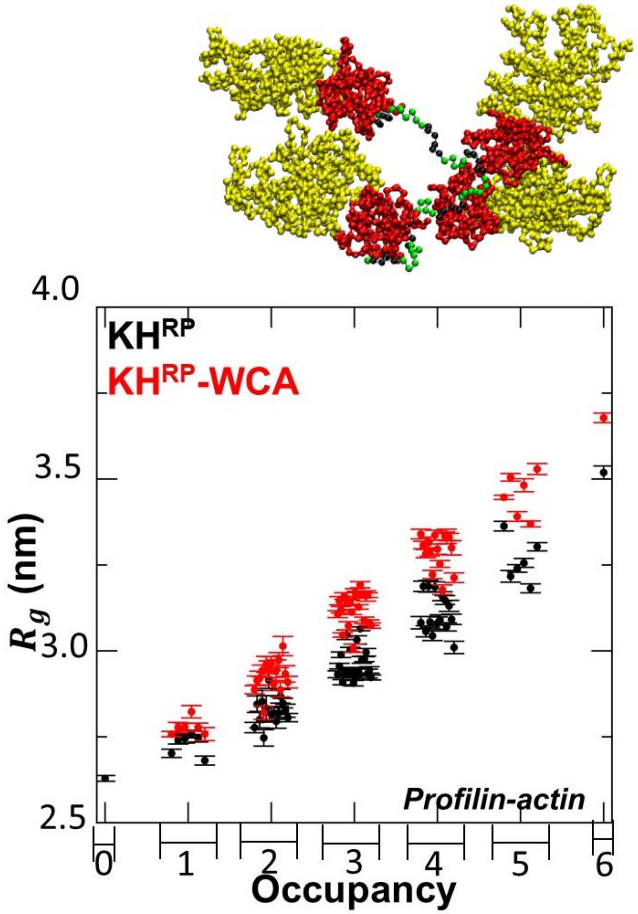
Horan et al. FEBS Letters 2018

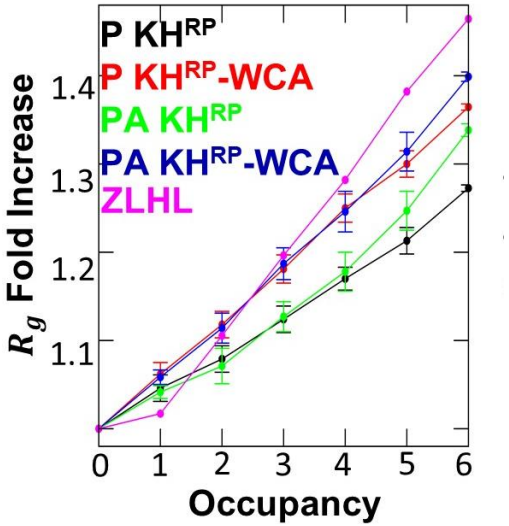
Profilin(-actin) Occupancy Weakly Expands FH1

Effect of profilin(-actin) occupancy on FH1?

- FH1 believed to expand upon profilin or profilin-actin binding *(Zhao et al. 2014 FEBS Letters, Byant et al. 2017 Cytoskeleton)*.
- Exploratory simulations reveal that all PRMs on FH1 may be simultaneously occupied by profilin or profilin-actin, weakly expanding FH1.





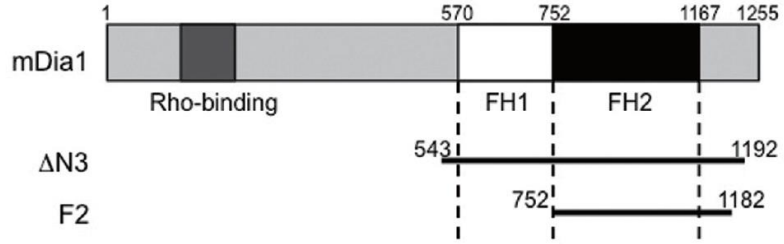


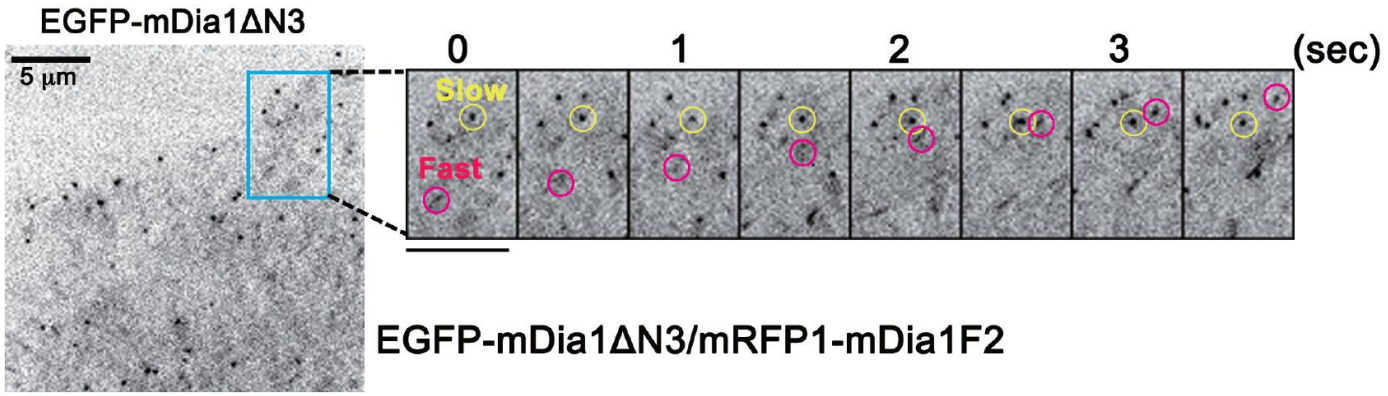
Horan et al. FEBS Letters 2018

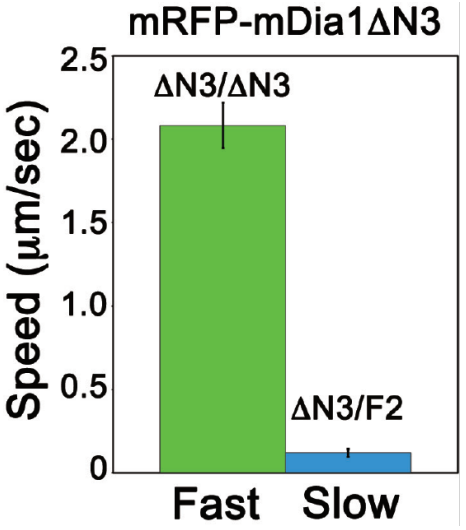
Experiment Suggests Possible Transfer Mechanisms

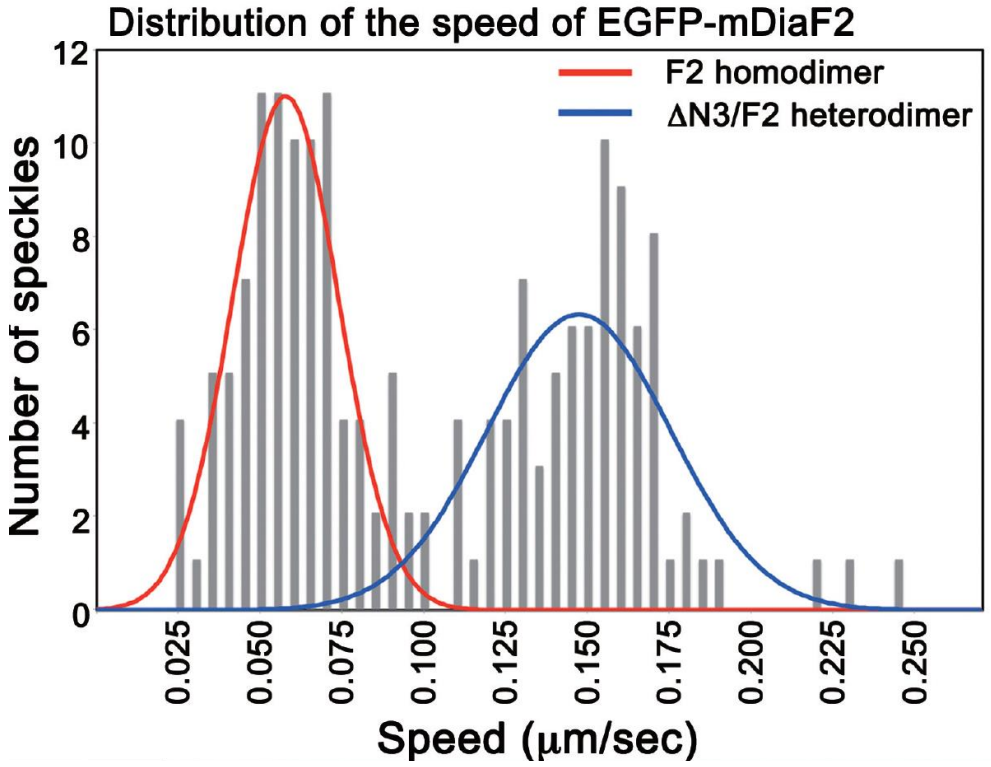
What is the FH1-mediated profilin-actin delivery mechanism?

- Collaborators performed single-molecule speckle microscopy experiments with mDia1 constructs in XTC cells *(unpublished, Watanabe group, Kyoto, Japan)*.









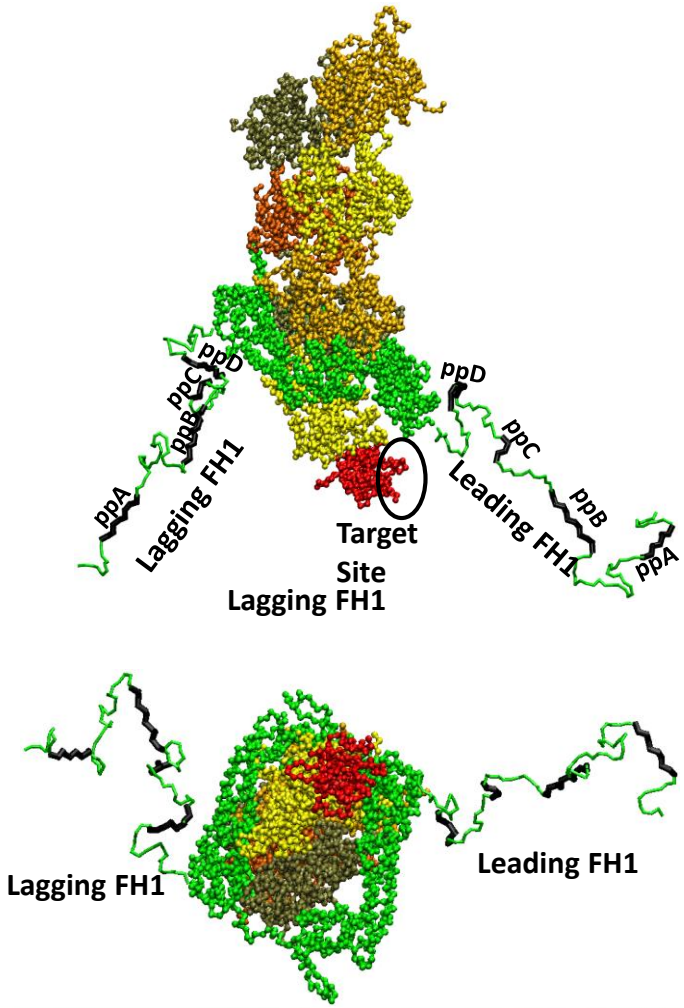
- Three populations of speckle speeds were observed, corresponding to homodimers with and without FH1, as well as heterodimer with only one FH1.

- Difference in speed suggests possible mechanism:
  - *Alternating delivery from each FH1*
  - *Simultaneous delivery from both FH1s*

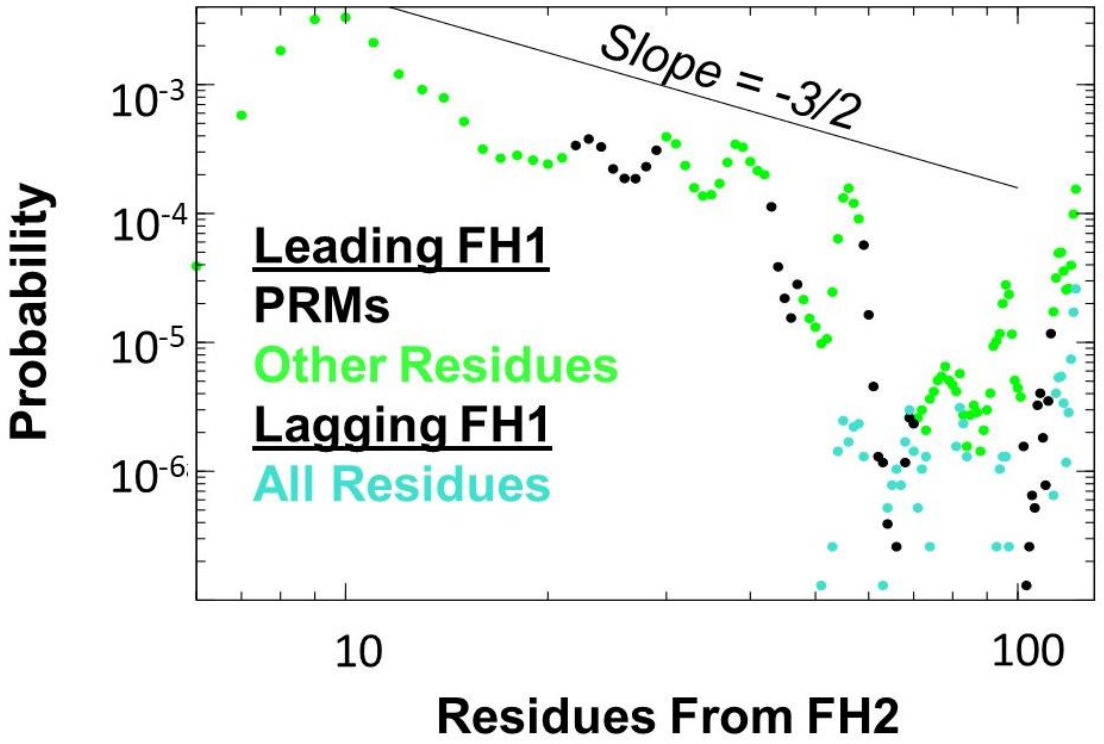
Profilin-actin Delivery to Barbed End is FH1-specific

What is the FH1-mediated profilin-actin delivery mechanism?

- Alignment of profilin to Bni1-FH2 barbed end (*structure from Baker et al. 2015 Structure*) and attachment of FH1 domains allows us to determine closure probability (proportional to closure rate, *Vavylonis et al. 2006 Mol. Cell*).



- Leading FH1 in better position for delivery.
- PRMs closer to FH2 more likely to contribute to polymerization.



Horan et al. FEBS Letters 2018

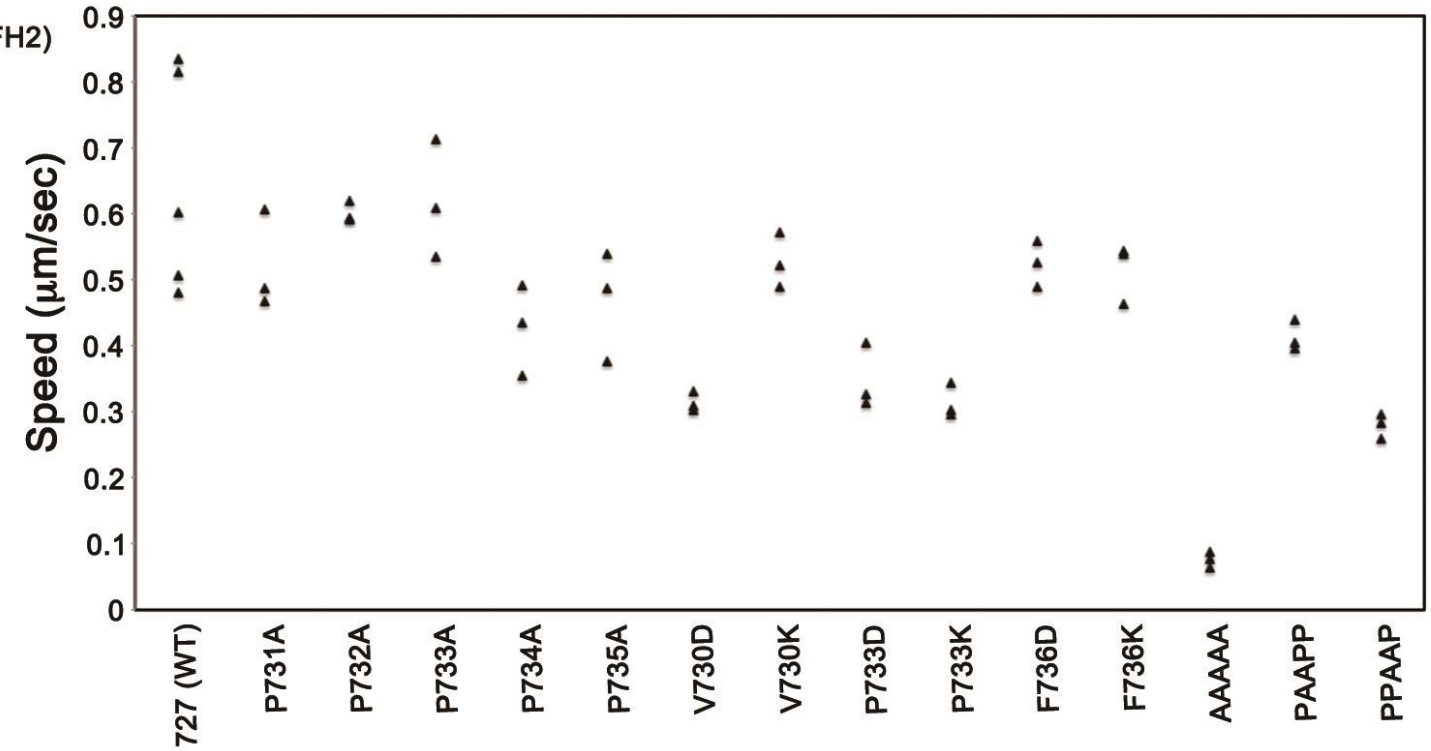
Specific Interactions Important in Profilin-FH1 Binding

What is the FH1-mediated profilin-actin delivery mechanism?

- Experiment confirms the importance of specific interactions in profilin-FH1 binding.

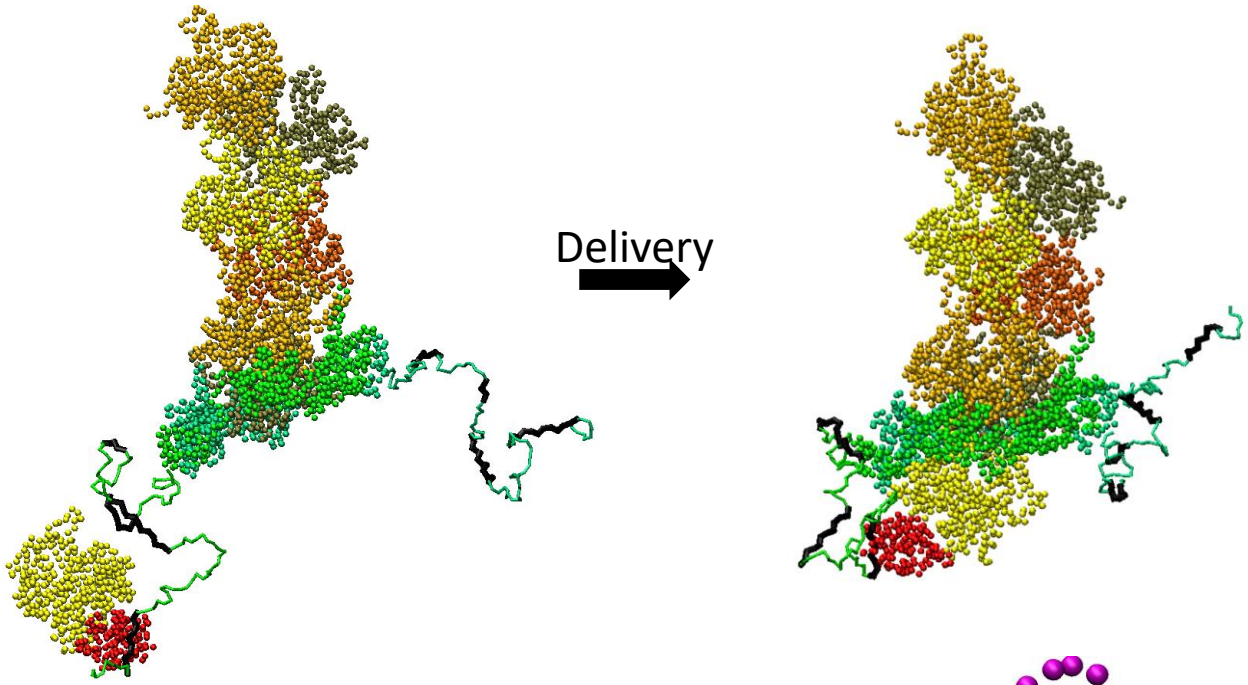
- Mutation of single proline residue or two proline residues to an alanine (less likely to be in a polyproline helix) or of the surrounding valine or phenylalanine or the central proline to a charged (positive or negative) residue have unique affects on polymerization speed.

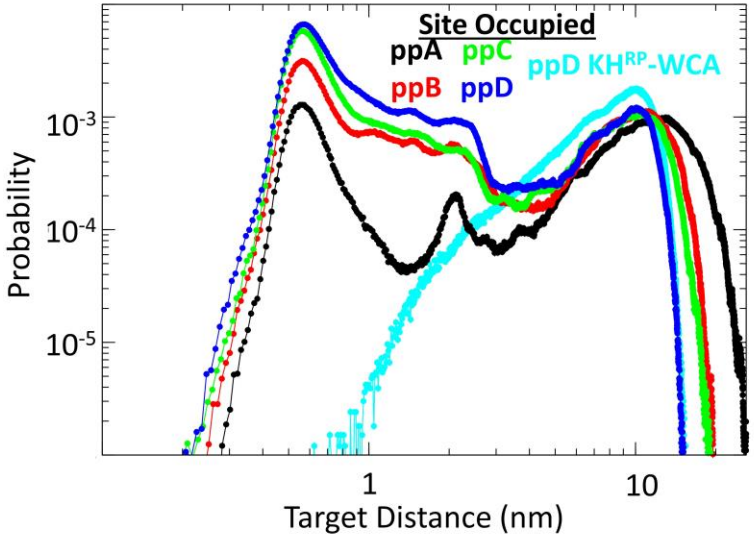
```
                    727
                730                 751
727(WT)  : AAAAAGMGVPPPPPFGFGVPAAPVLPFGLT - (FH2)
P731A    : AAAAAGMGVAPPPPFGFGVPAAPVLPFGLT
P732A    : AAAAAGMGVPAPPPFGFGVPAAPVLPFGLT
P733A    : AAAAAGMGVPPAPPFGFGVPAAPVLPFGLT
P734A    : AAAAAGMGVPPPAPFGFGVPAAPVLPFGLT
P735A    : AAAAAGMGVPPPPAFGFGVPAAPVLPFGLT
V730D    : AAAAAGMGDPPPPPFGFGVPAAPVLPFGLT
V730K    : AAAAAGMGKPPPPPFGFGVPAAPVLPFGLT
P733D    : AAAAAGMGVPPDPPFGFGVPAAPVLPFGLT
P733K    : AAAAAGMGVPPKPPFGFGVPAAPVLPFGLT
F736D    : AAAAAGMGVPPPPPDGFGVPAAPVLPFGLT
F736K    : AAAAAGMGVPPPPPKGFGVPAAPVLPFGLT
AAAAA    : AAAAAGMGVAAAAAFGFGVPAAPVLPFGLT
PAAPP    : AAAAAGMGVPAAPPFGFGVPAAPVLPFGLT
PPAAP    : AAAAAGMGVPPAAPFGFGVPAAPVLPFGLT
```

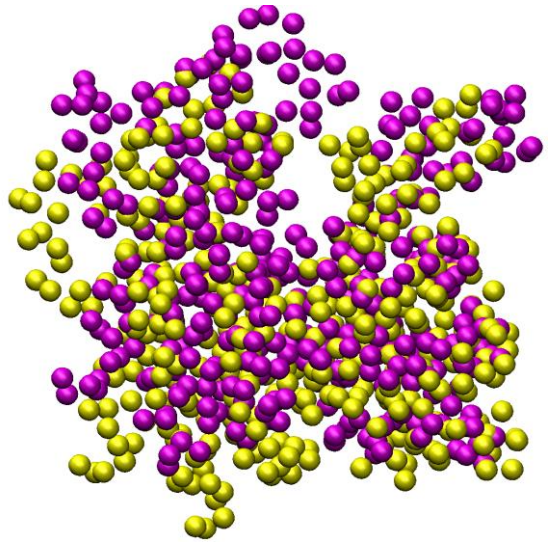


Model Captures Profilin-actin Delivery to Barbed End

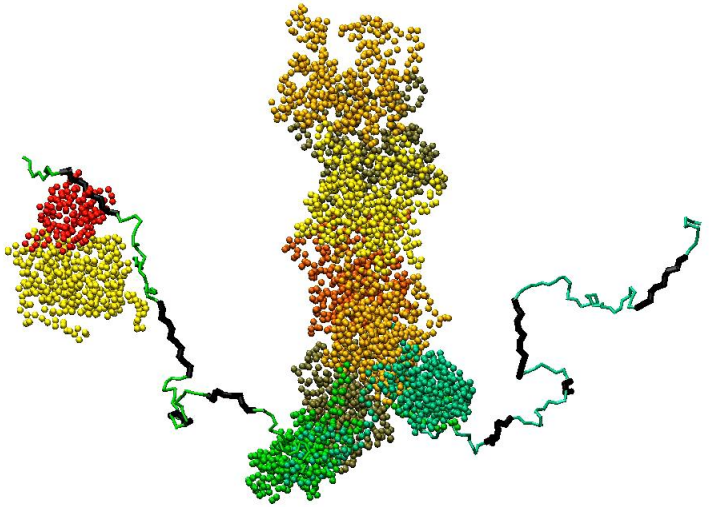
What is the FH1-mediated profilin-actin delivery mechanism?







- Removal of the terminal profilin-actin and placing it on FH1 supports direct transfer mechanism (Vavylonis et al. 2006 Mol. Cell).



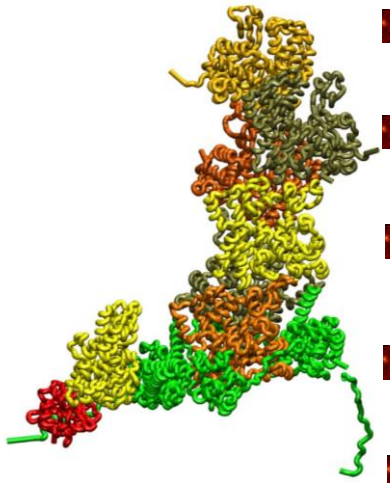
- Incoming monomer can find near-correct orientation for polymerization.

Horan et al. FEBS Letters 2018

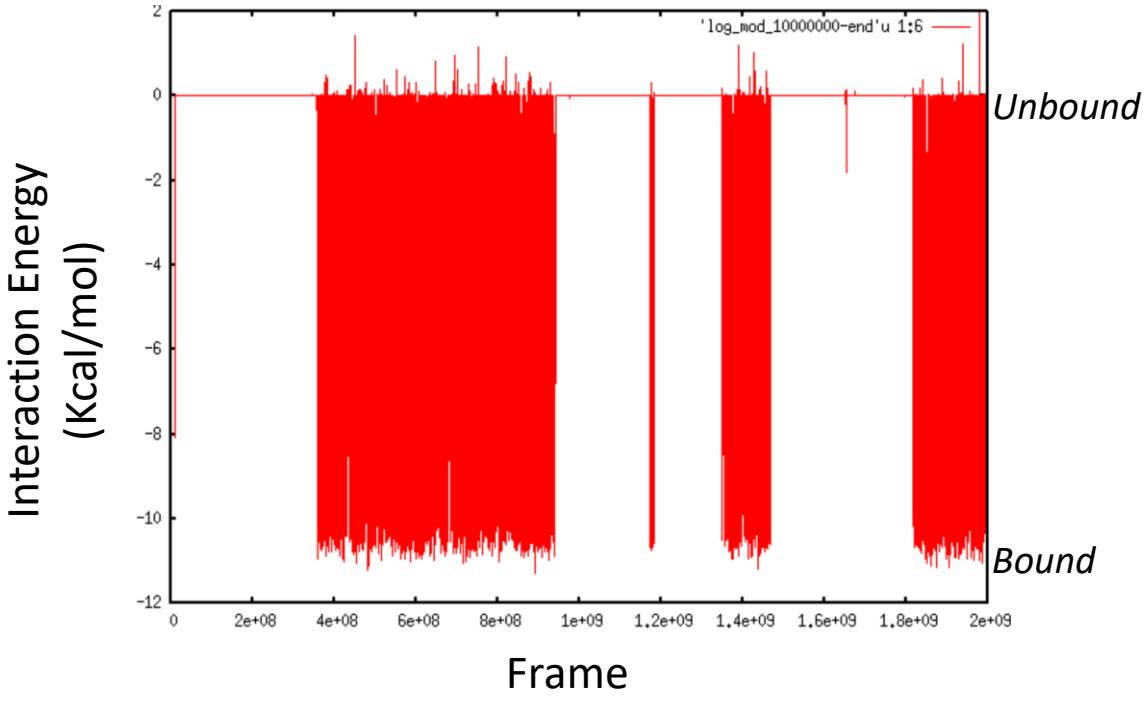
Enhanced Sampling (REMD)

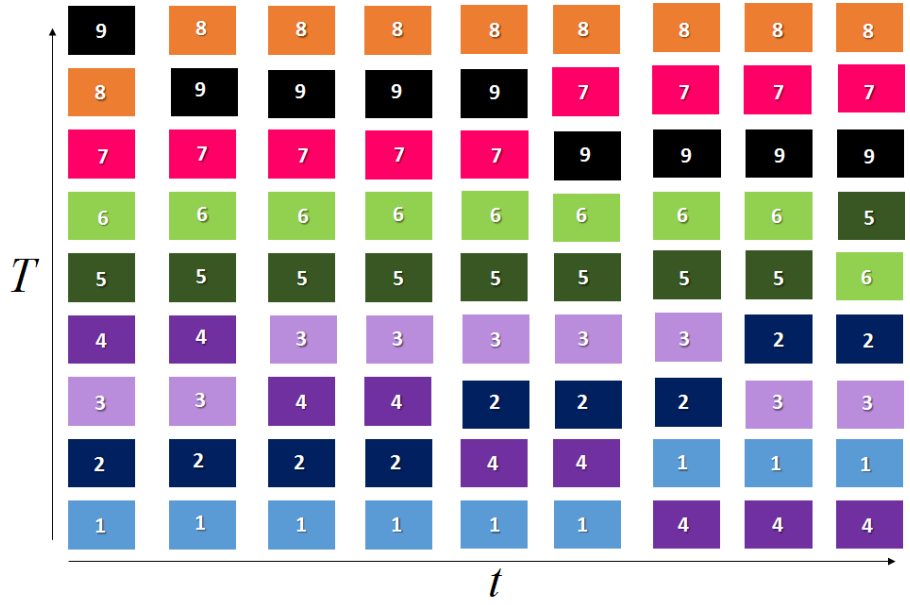
What is the FH1-mediated profilin-actin delivery mechanism?

- Using standard MD simulations, it would be difficult to have good sampling of delivery (binding) events.



- Sample data (small system) took **~1 day** to obtain.
- Appx **7 binding events** (binding=1 event, unbinding=1event) in sample data.
- Sufficient sampling will probably have **100s to 1000s binding events**.
- Appx time to obtain this is **140 days** (for 1000 events).
- This is **too long** to wait for data.



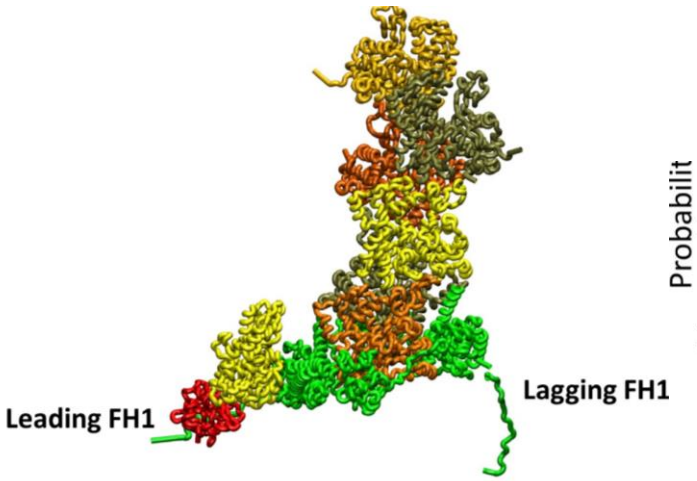


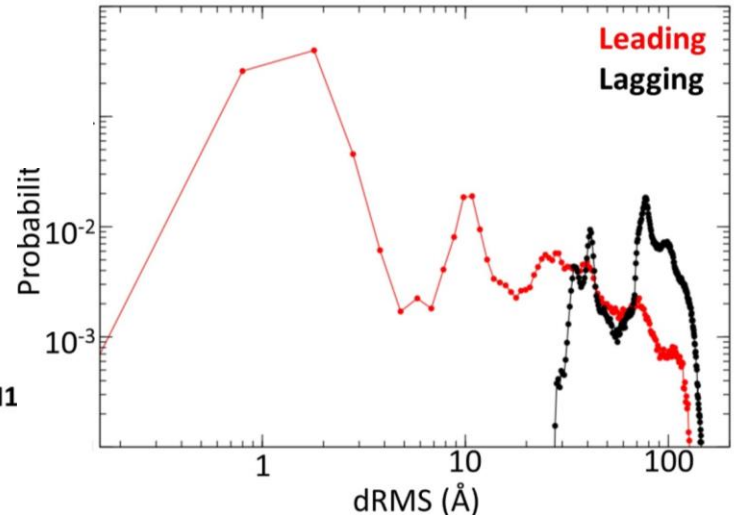
- We can overcome this limitation by using **enhanced sampling**. Here, we employ **replica exchange** (parallel tempering).
- Multiple copies (replicas) of the simulation run at different temperatures.
- Replicas at high temperatures easily overcome potential barriers.
- Periodically exchange replicas in neighboring temperatures so all replicas sample whole potential energy landscape.

Delivery Simulations Support Alternating Mechanism

Possibility of Alternating Delivery?

- Using enhanced sampling, we examine delivery from mDia1-FH1 in more detail.



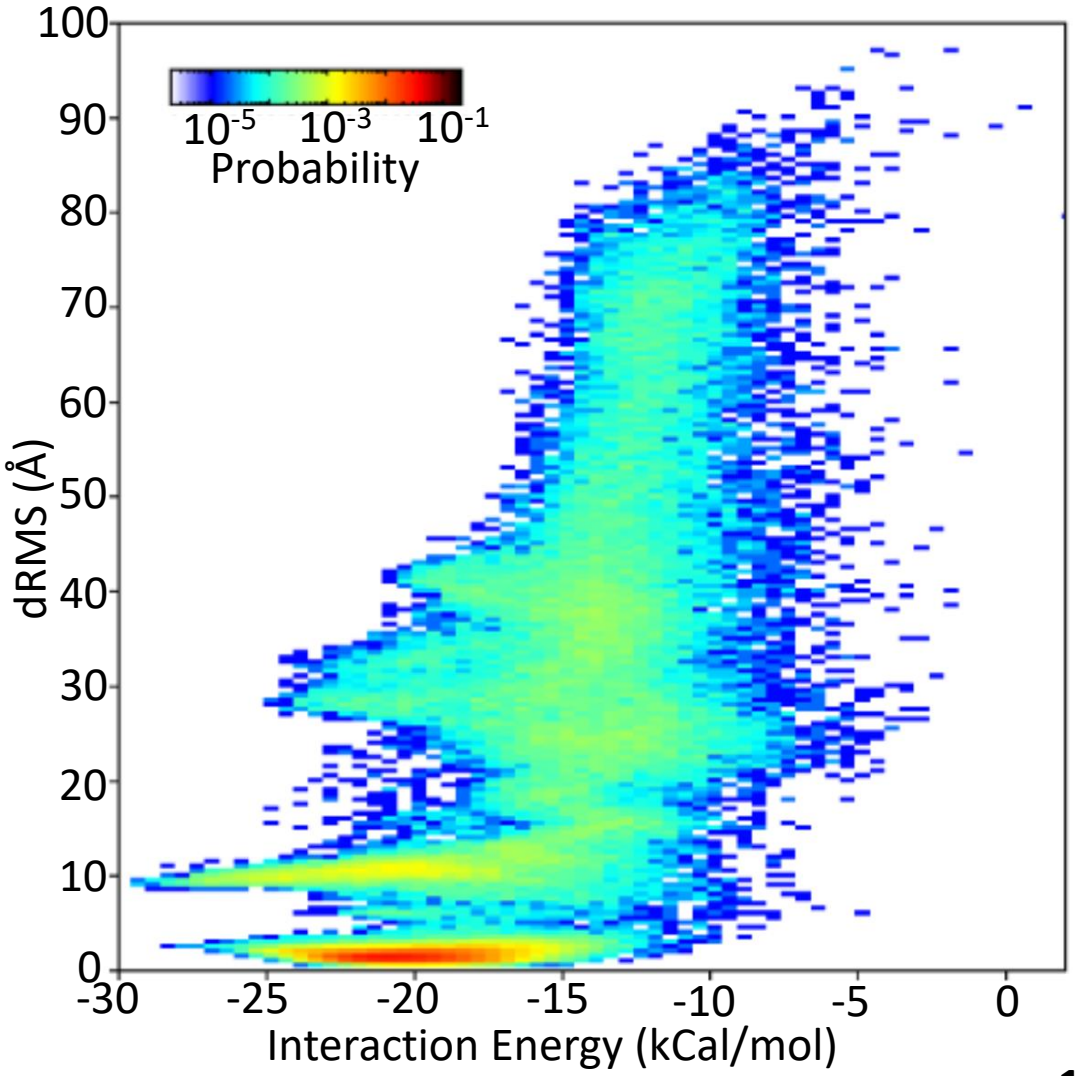


- dRMS measures the average absolute deviation of distances between two objects in a reference structure (here crystal/model) as compared to a test structure (here simulation).

$$dRMS = \frac{\sum_{i,j} \left| d_{ij}^{\,test} - d_{ij}^{\,ref} \right|}{N}$$

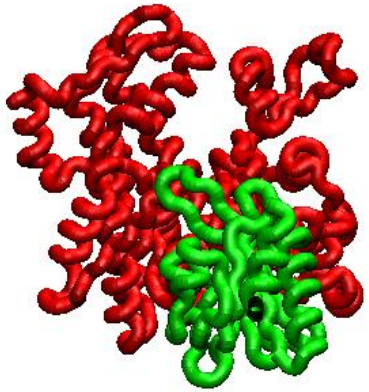
- dRMS <~ 5-10 Angstroms indicates high degree of similarity between reference and test structure.
- If reference structure is experimentally validated structure, low dRMS indicates native-like binding.

- Visually, there are two main bound clusters plus some noise.
- Utilize clustering to identify common structures.

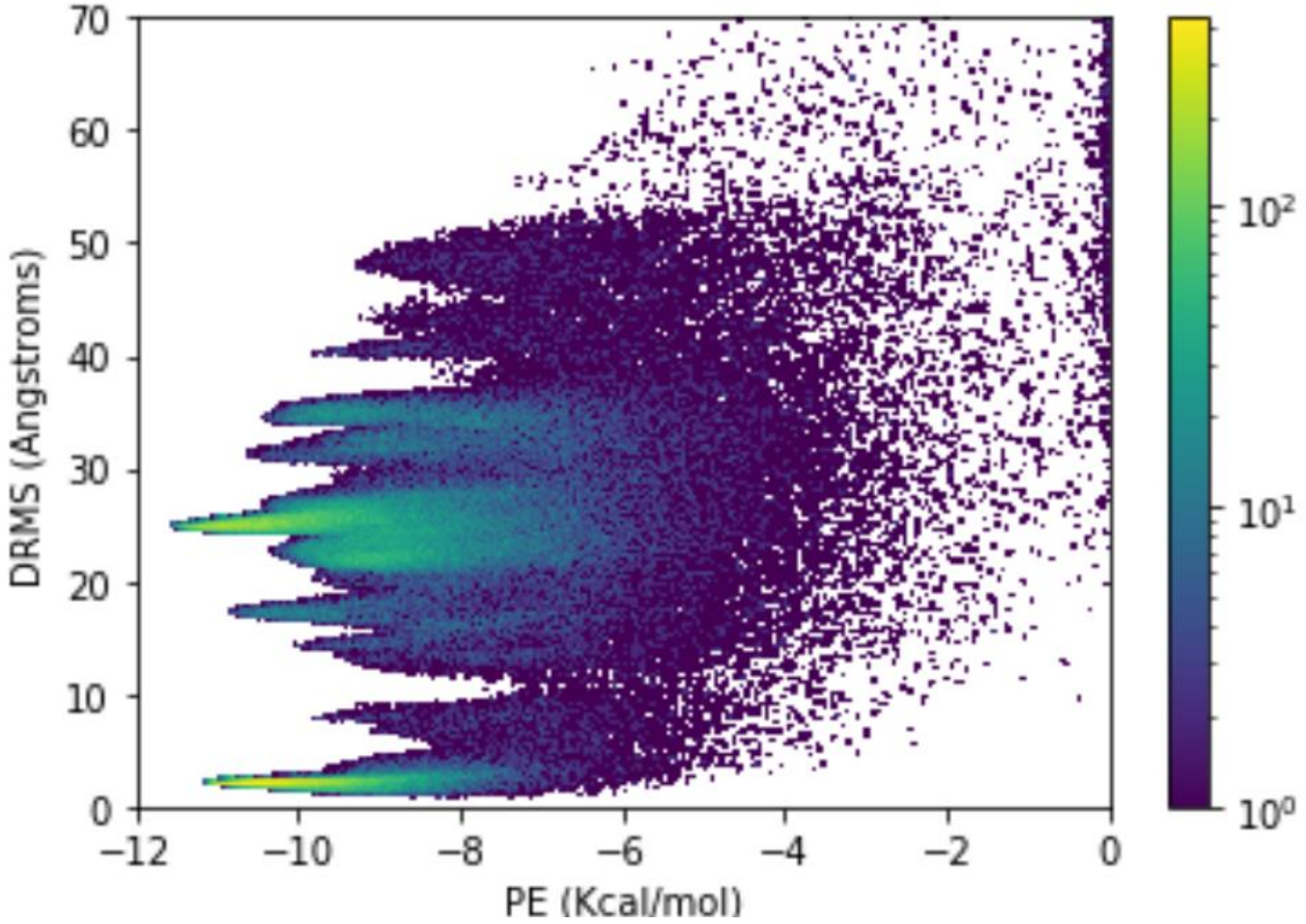


Clustering Approaches

- I discuss a clustering approaches primarily in the context of a similar system: profilin-actin binding

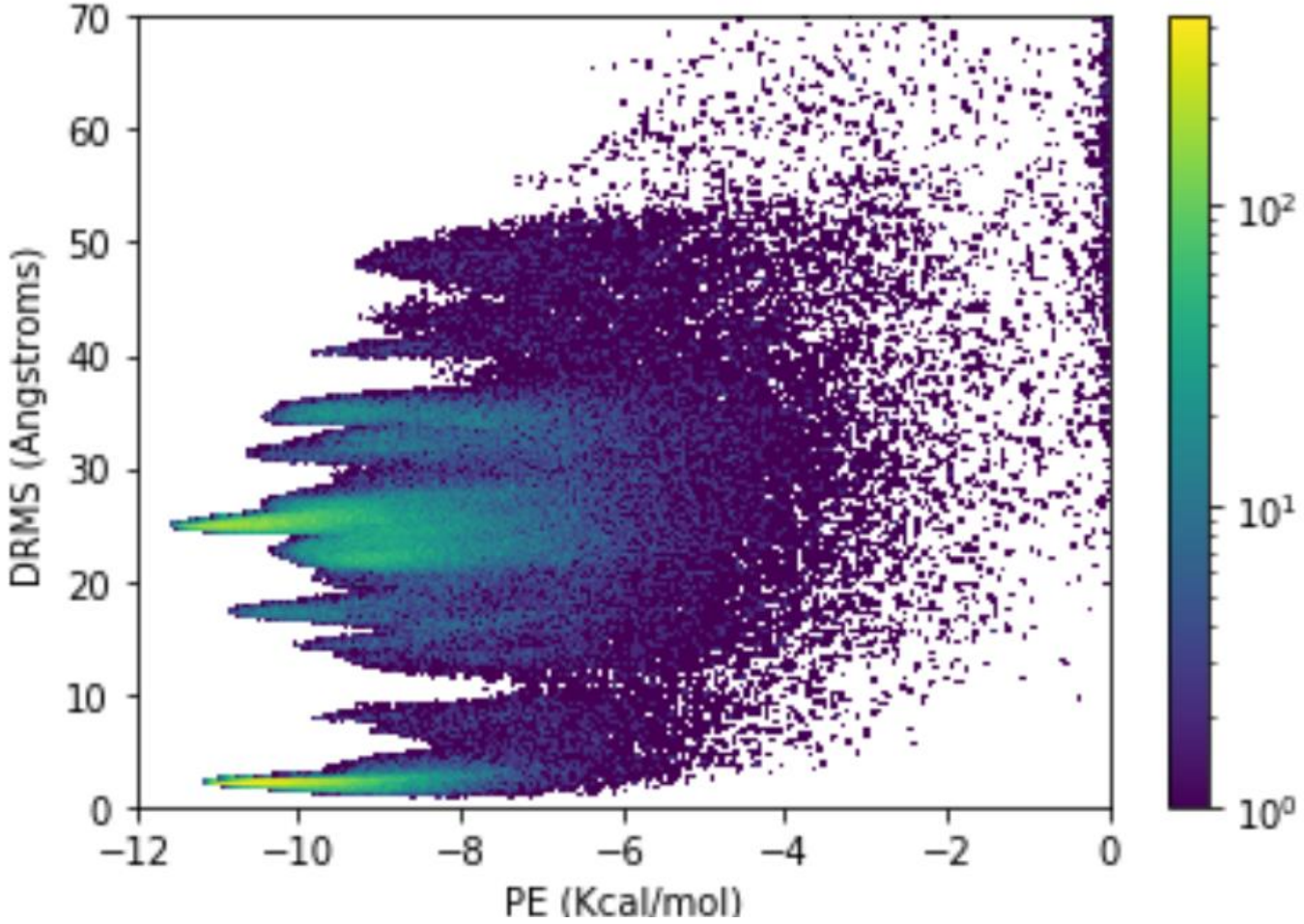


- dRMS and energy are common metrics in this field (*Kim & Hummer 2008 J. Mol. Bio)*

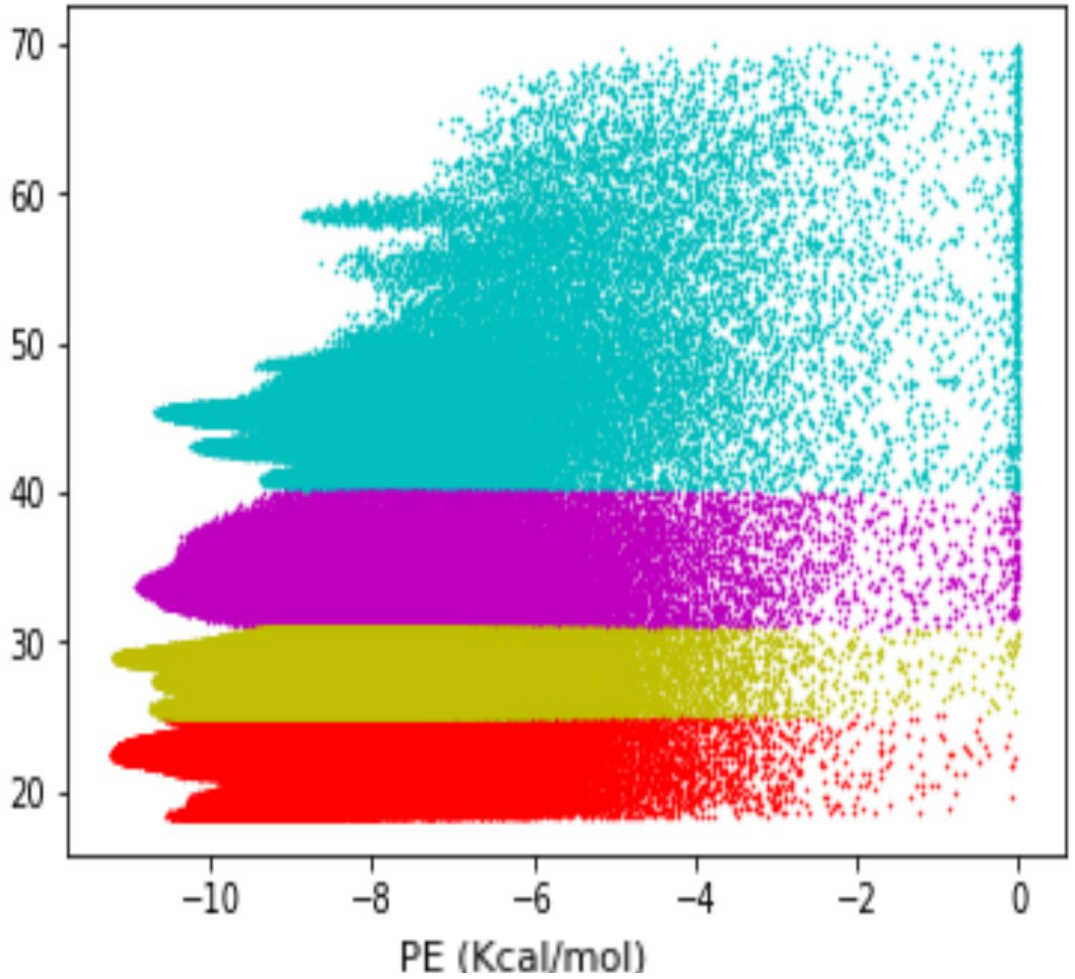


Clustering Approaches

- K-Means doesn't give good clusters on this data.



Example of clusters coming from K-Means

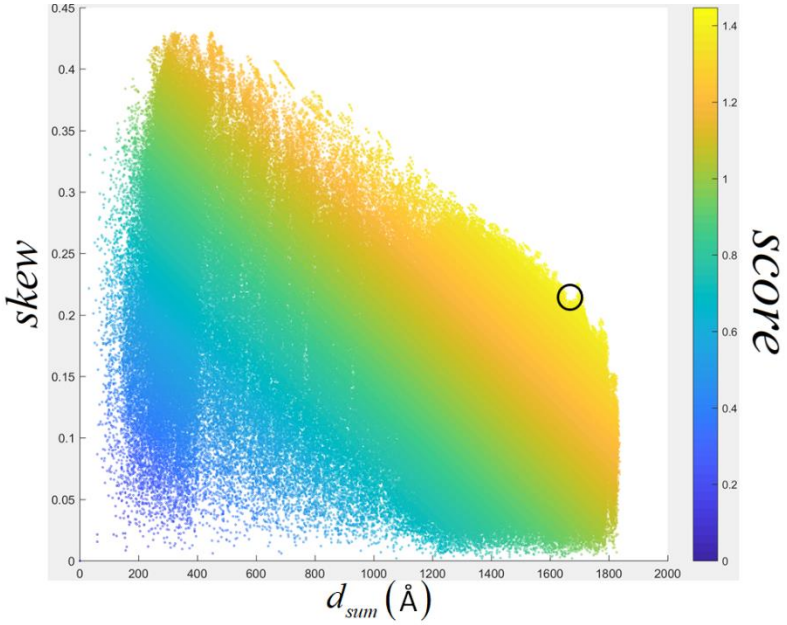


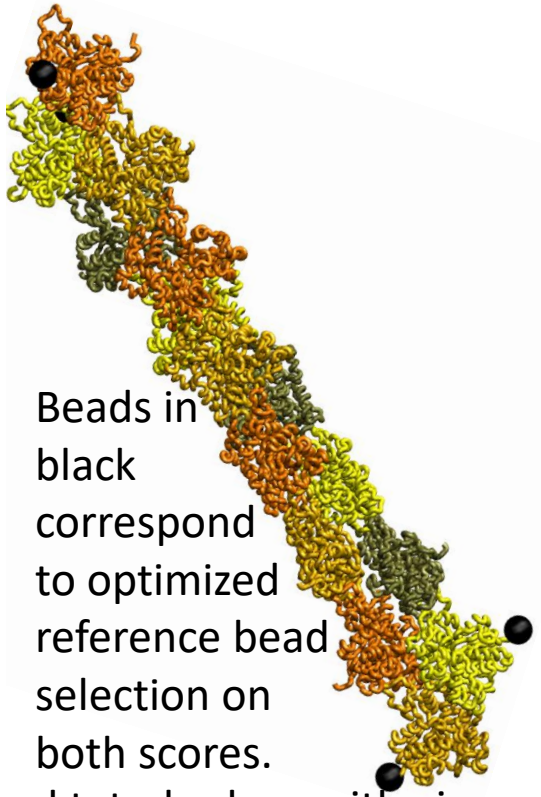
Reference Bead Selection

- Using a lossless space doesn't give improved results with K-Means, but MeanShift is better.

- Rigid bodies can be described by 4 reference points.
- Selection of 4 points must be optimized to reduce numerical error.

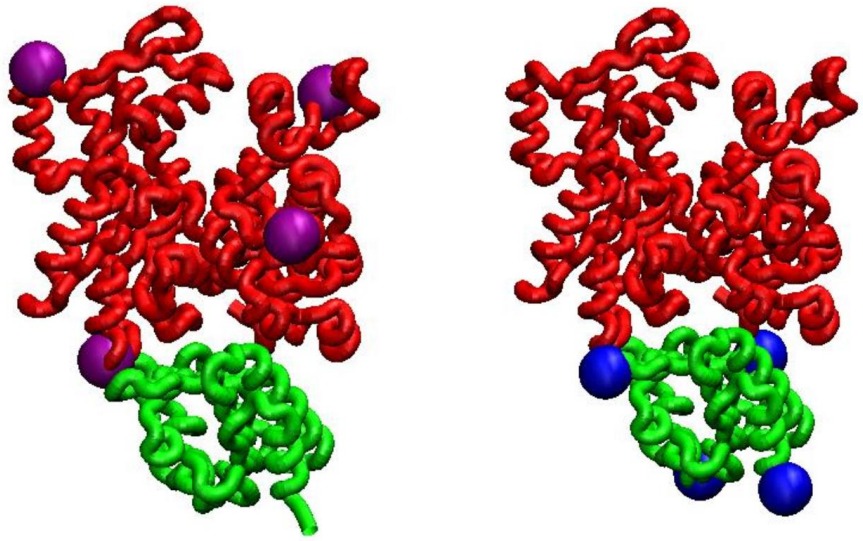
Example of reference point selection





Beads in black correspond to optimized reference bead selection on both scores.

- Reference beads selected from algorithm for actin and profilin shown in images.



- Since volume is optimized on the convex hull, only the convex hull is searched for reference point selection (speeds up algorithm from days to seconds for large systems such as in example on left).

- The set of distances between reference points in one protein and in the other protein defines a 16-dimensional lossless space to cluster on.

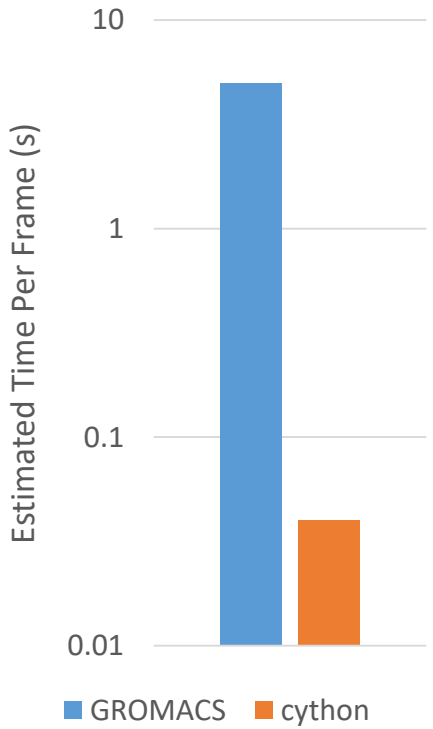
- Algorithm optimizes skew relative to ideal tetrahedron with given distances between reference points (volume) and sum of distances between reference points.

Cython

- Using a lossless space doesn't give improved results with K-Means, but MeanShift is better.
- Without Cython, this task would also take unfeasibly long to solve.
- Example of Cython usage: calculating distances between points

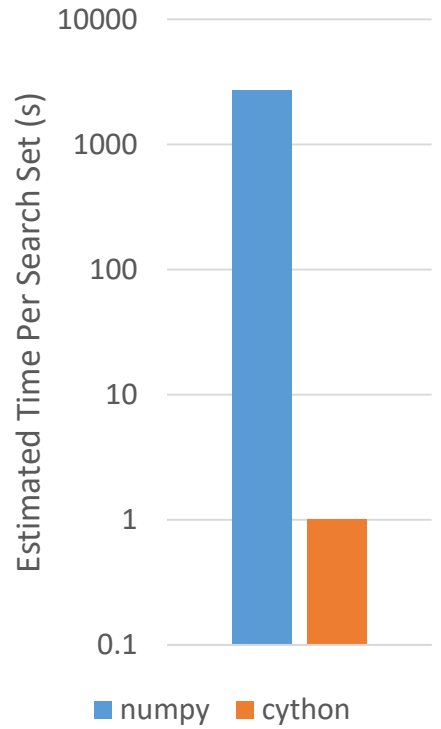
```
def run(d_, length_,hb_,lx_,ly_):
    cdef float [:,:] d = d_
    cdef double dx = -1.0
    cdef double dy = -1.0
    cdef double dz = -1.0
    cdef int length = length_
    cdef double hb = hb_
    cdef double b = hb*2.0
    cdef int lx = lx_
    cdef int ly = ly_

    cdef double [:] ds = np.zeros(length)

    cdef Py_ssize_t i = 0
    cdef Py_ssize_t j = 0

    cdef int c = 0
```

```
for i in range(0, lx):
    for j in range(i+lx, lx+ly):
        dx = d[i][0] - d[j][0]
        dy = d[i][1] - d[j][1]
        dz = d[i][2] - d[j][2]

        applyPBC()

        ds[c] = (dx**2.0+dy**2.0+dz**2.0)**0.5
        c += 1

return ds
```



- Typing variables with cdef declarations and labeling functions cdef that don't need python objects yields significant performance improvement.

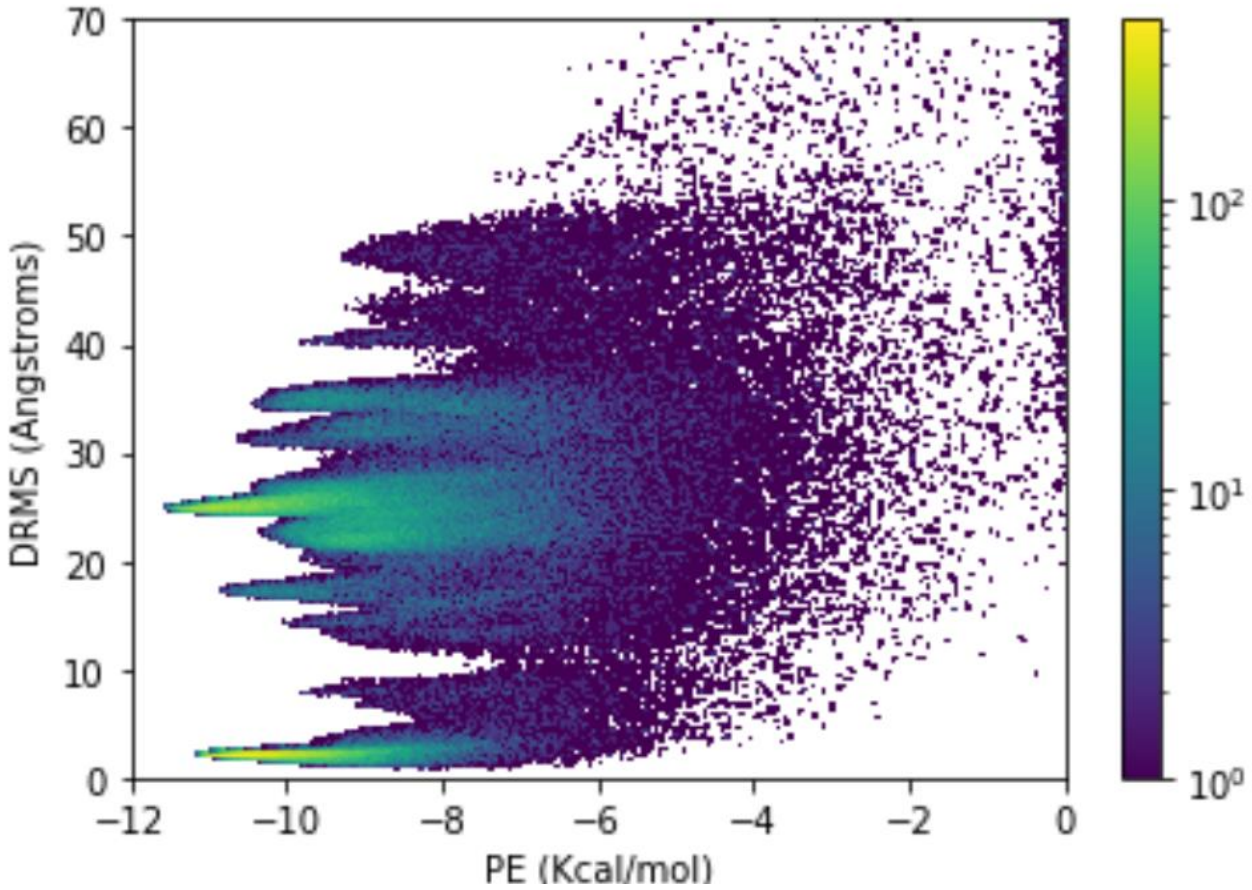
- Additional caution: some numpy functions are too slow and better performance can be achieved with Cython.
- I was performing an index of operation on an array (for the malaria detection project I will discuss later).
- Writing an optimized index of (constant time lookup) yields significant performance improvement.

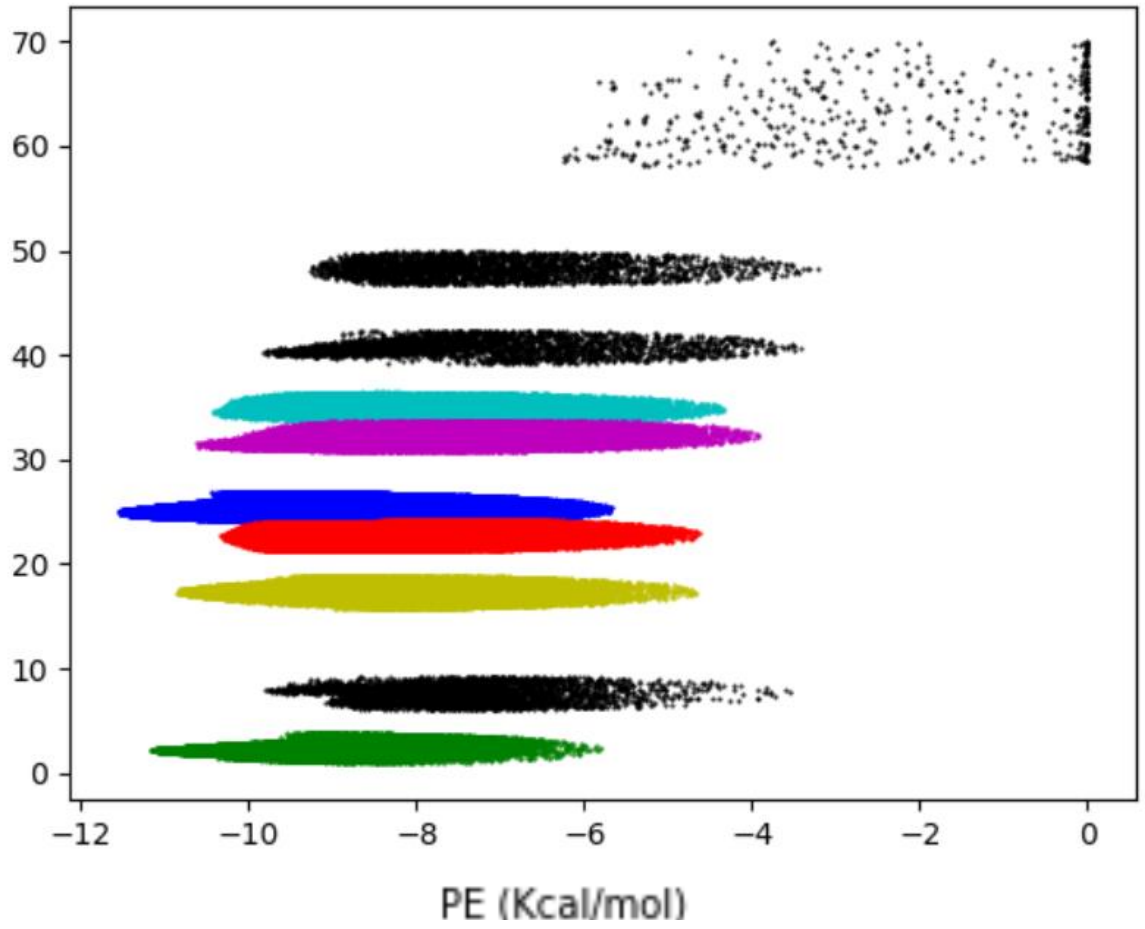


Clustering Approaches

- Using a lossless space doesn't give improved results with K-Means, but MeanShift is better.

Example of output from MeanShift

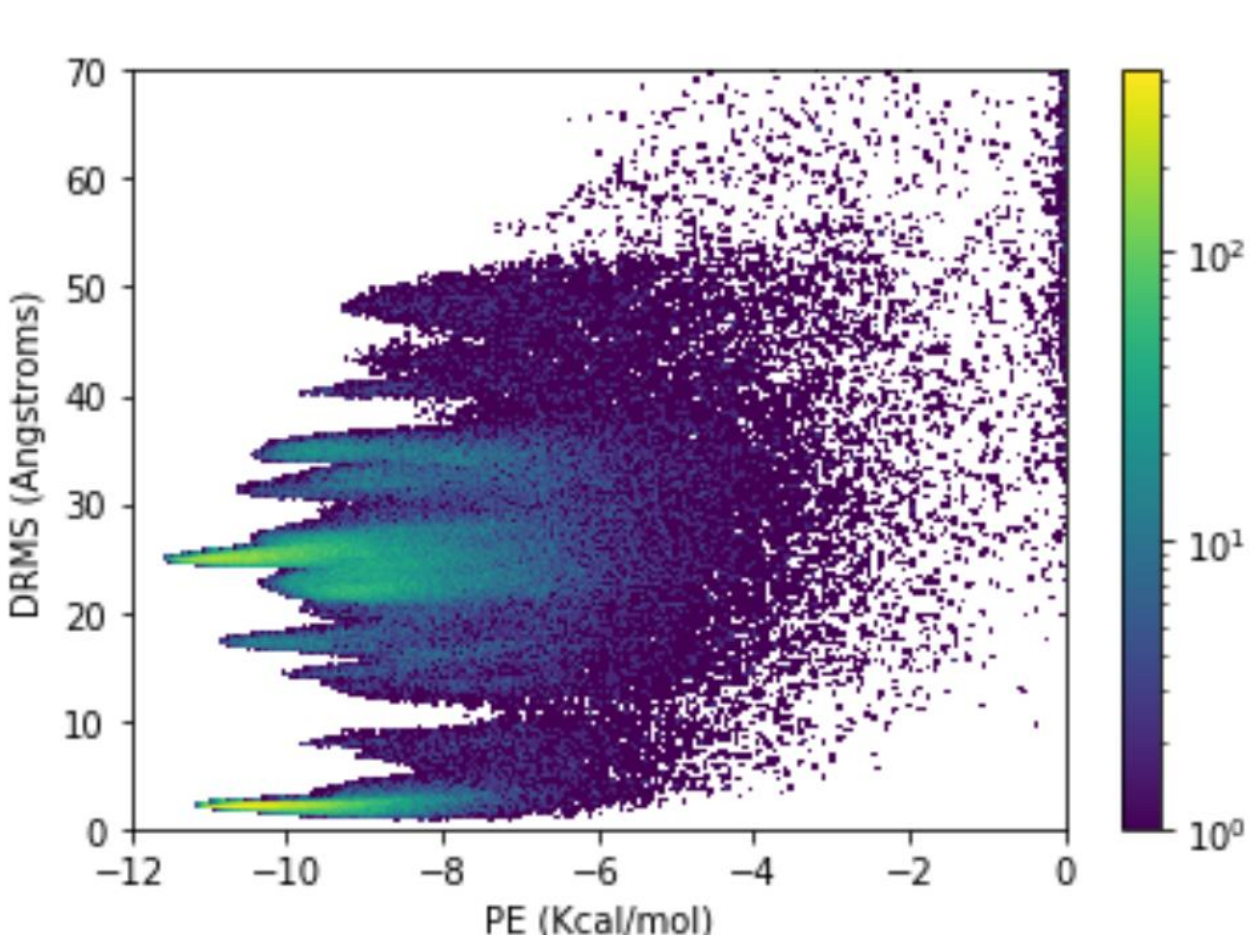




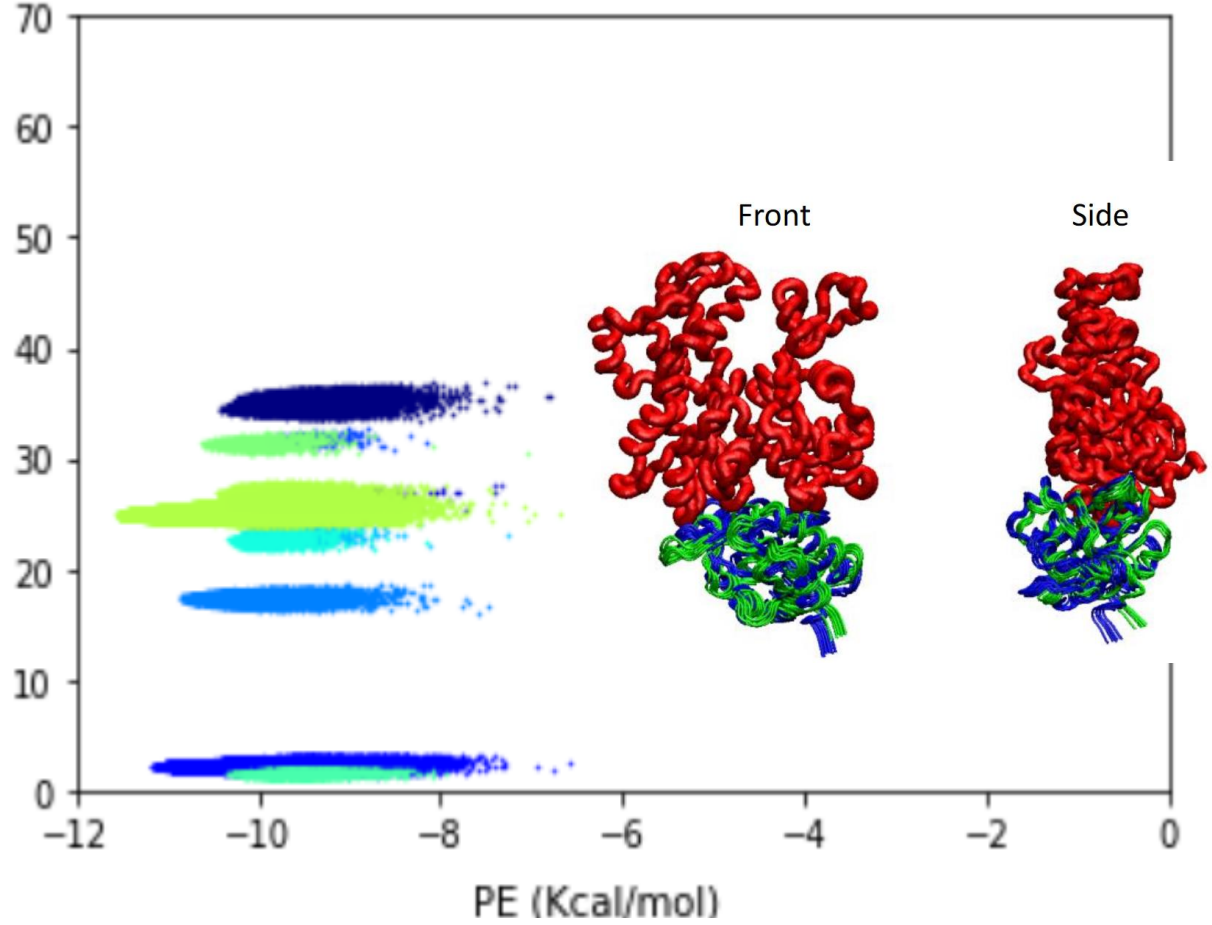
- MeanShift discriminates clusters in the right locations, but they are still relatively wide in their geometry, not truly capturing bound modes distinctly.

Clustering Approaches

- The best results with standard clustering algorithms is achieved using DBSCAN.



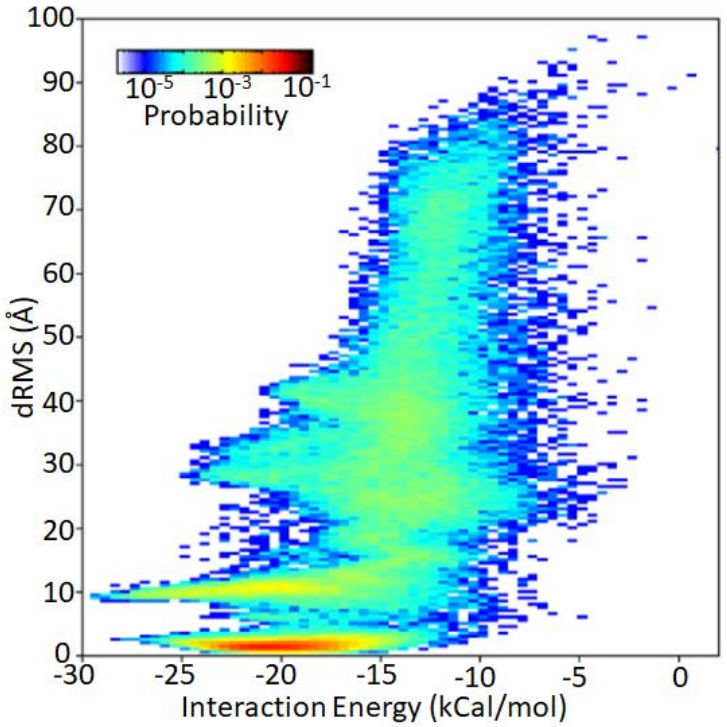
Example of output from DBSCAN



- DBSCAN can even discriminate between very similar modes of binding (so similar I did not detect them by eye until the algorithm revealed this information).
- The only downside is that optimizing parameters for DBSCAN can be a slow process.

Clustering Approaches

- Returning to original data to cluster:



- Using the same spirit as DBSCAN (density-based, includes outliers, has minimum number of points to include in cluster), I developed a clustering algorithm which clusters the data very quickly.

Clustering Algorithm*:*

Input: raw data, bin size, cutoff on maxima for searching for cluster.
n: Number of data points

-Create probability map of size s << n, reducing size by order(s) of magnitude: O(n)
- Define region around each maxima (using statistical cutoff) (recursively), i.e.

-Identify local maxima greater than cutoff O(s^2)
*- **For** each maximum*
check if neighbors within cluster cutoff
***For** each neighbor within cluster cutoff*
check for additional neighbors within cluster cutoff
*-**For** each overlapping cluster*
If not already, cut off cluster at minimum probability boundary between clusters and divide bin probability between clusters sharing these bins.

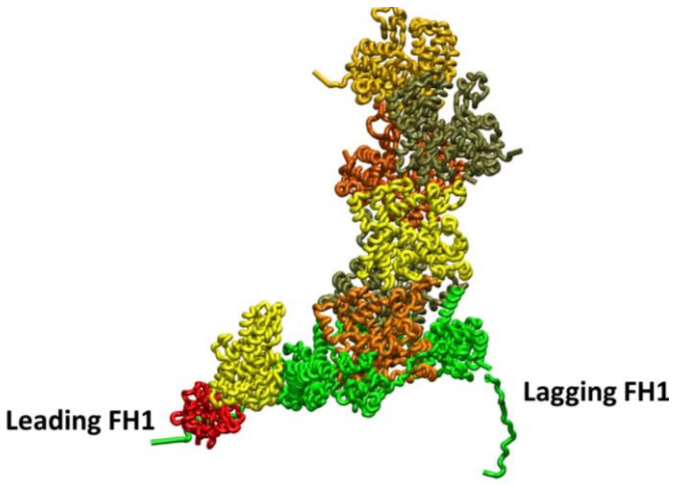
- While the algorithm is still quadratic, it is quadratic in the density space, so given large enough bin size, it clusters much faster than DBSCAN.

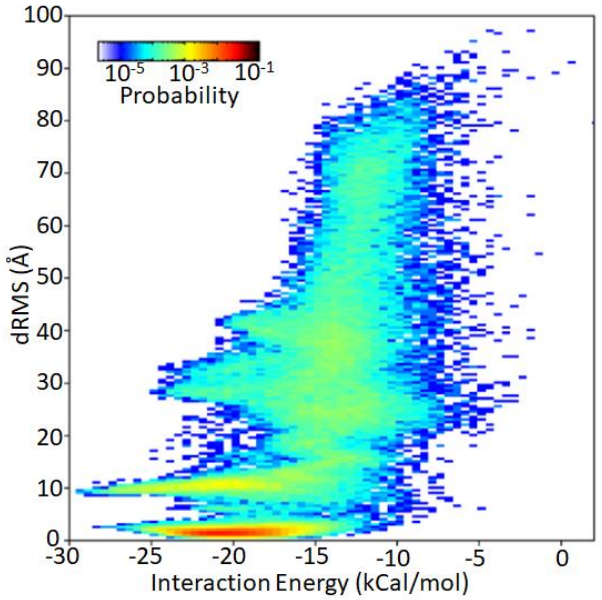
- Currently, points in neighboring clusters support shared cluster assignment rather than unique cluster assignment.

Delivery Simulations Support Alternating Mechanism

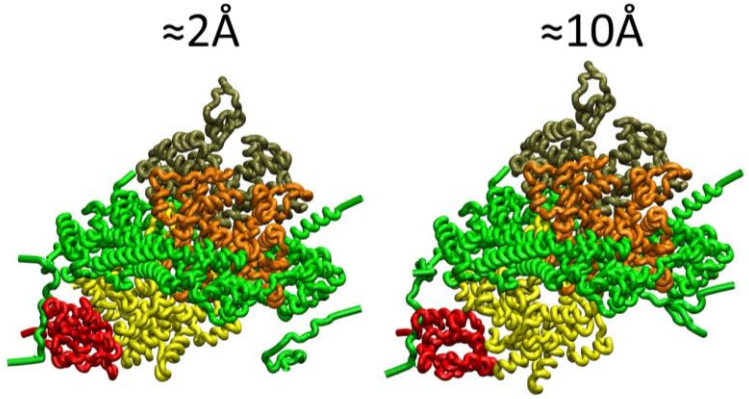
Possibility of Alternating Delivery?

- Using enhanced sampling, we examine delivery from mDia1-FH1 in more detail.





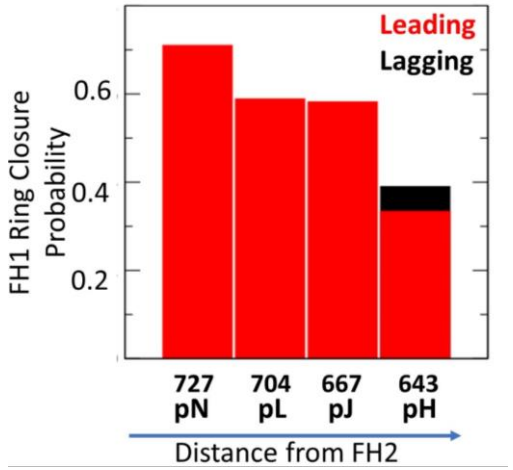
- Delivery simulations yield two delivery type complexes.



- dRMS measures the average absolute deviation of distances between two objects in a reference structure (here crystal/model) as compared to a test structure (here simulation).

$$dRMS = \frac{\sum_{i,j} \left| d_{ij}^{\,test} - d_{ij}^{\,ref} \right|}{N}$$

- dRMS <~ 5-10 Angstroms indicates high degree of similarity between reference and test structure.
- If reference structure is experimentally validated structure, low dRMS indicates native-like binding.

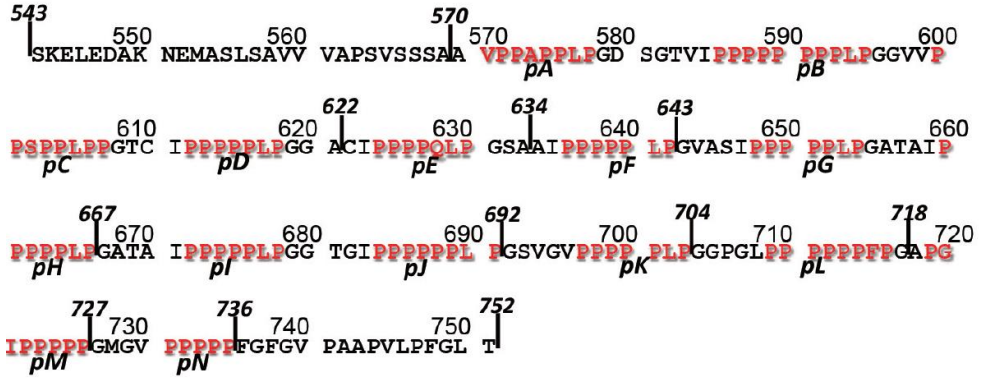


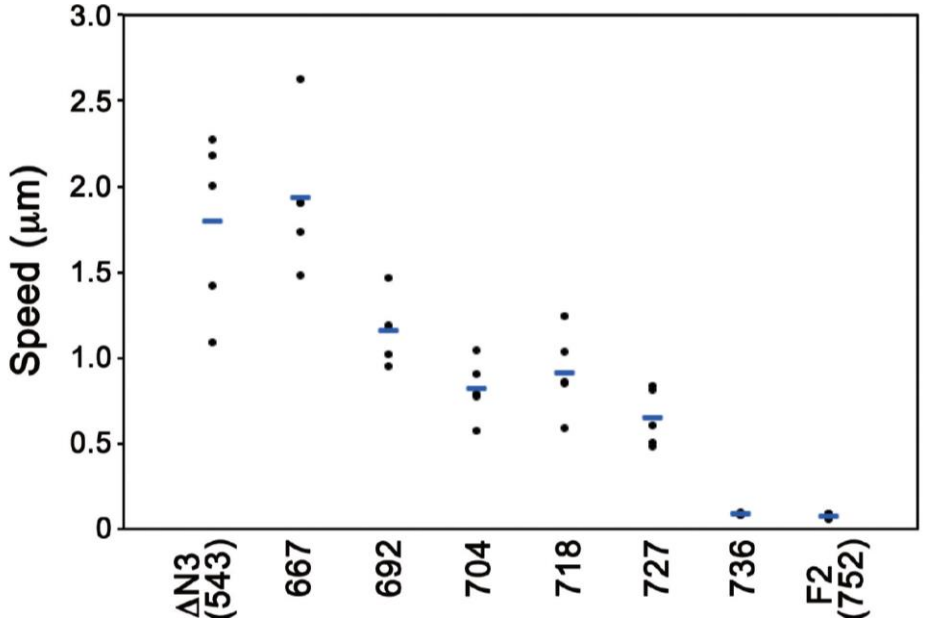
- Delivery from lagging only occurs from distal PRMs and with low probability, consistent with FH1 ring closure simulations.
- Supports alternating delivery mechanism.

Experiments Support Alternating Mechanism

What is the FH1-mediated profilin-actin delivery mechanism?

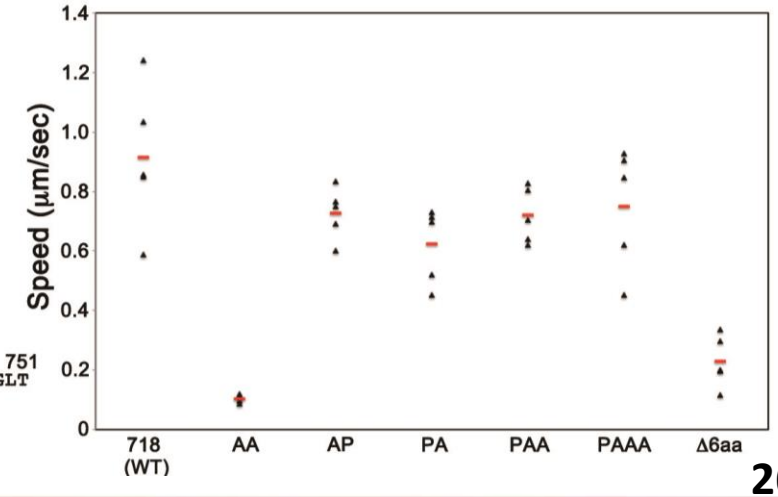
- Mutant constructs with FH1 of various length.
- The more PRMs, the faster actin polymerizes.
- Distal PRMs only minimally contribute.
- Distance required to get lagging FH1 delivery in simulation suggests only leading FH1 delivers profilin-actin.





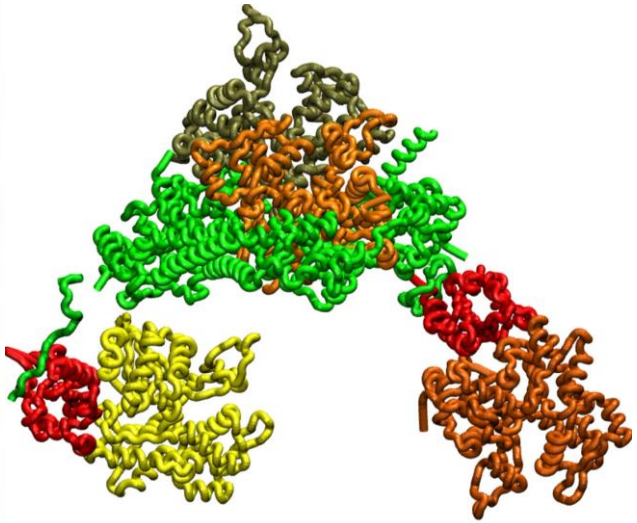
- Unless simultaneous delivery is likely, this supports alternating delivery mechanism.

- Single PRM constructs confirm nearby PRMs have similar ability of delivery.
- Bringing closest PRM 6 amino acids closer to the FH2 also confirms the minimum distance from FH2 requirement for profilin-actin delivery.

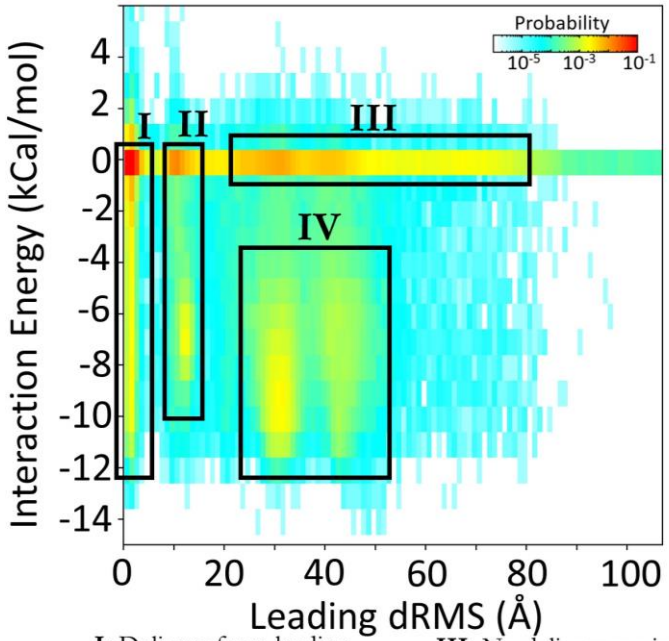


Simultaneous Delivery Not Excluded, Yet Unlikely

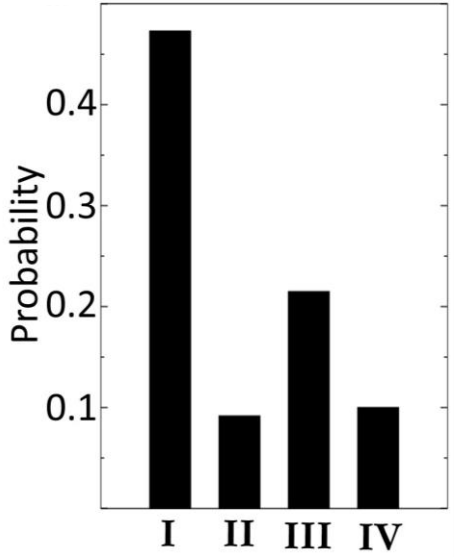
Possibility of Simultaneous Delivery?

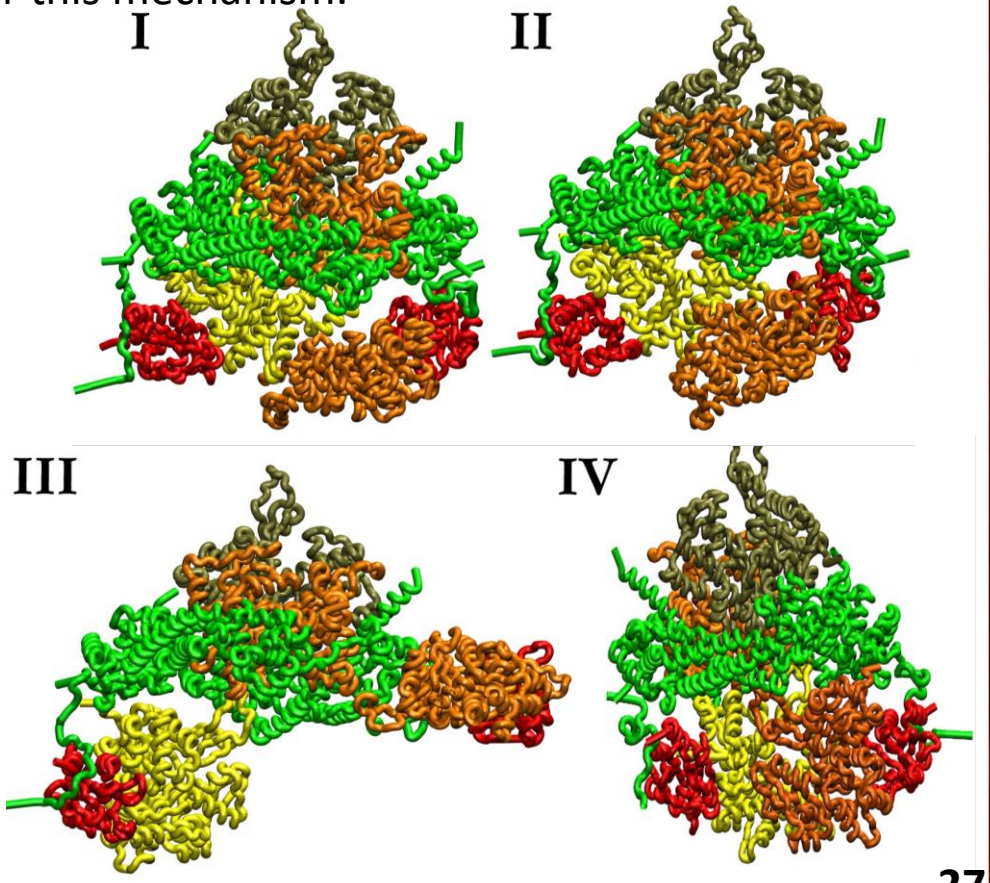


- Profilin sterically occludes possibility of simultaneous delivery. If simultaneous delivery possible, simulation should pick up some intermediate in this process.
- Classifying clusters of data points reveals four types of complexes observed (delivery, near-delivery, non-specific association, novel actin dimer).
- The dimer (IV) may be intermediate in possible simultaneous delivery where delivery pulls FH2 down and profilin also releases in this highly-coordinated mechanism.
- Simulation predicts low probability for this mechanism.



I: Delivery from leading
II: Near delivery from leading
III: No delivery, no interaction
IV: Interaction but no delivery

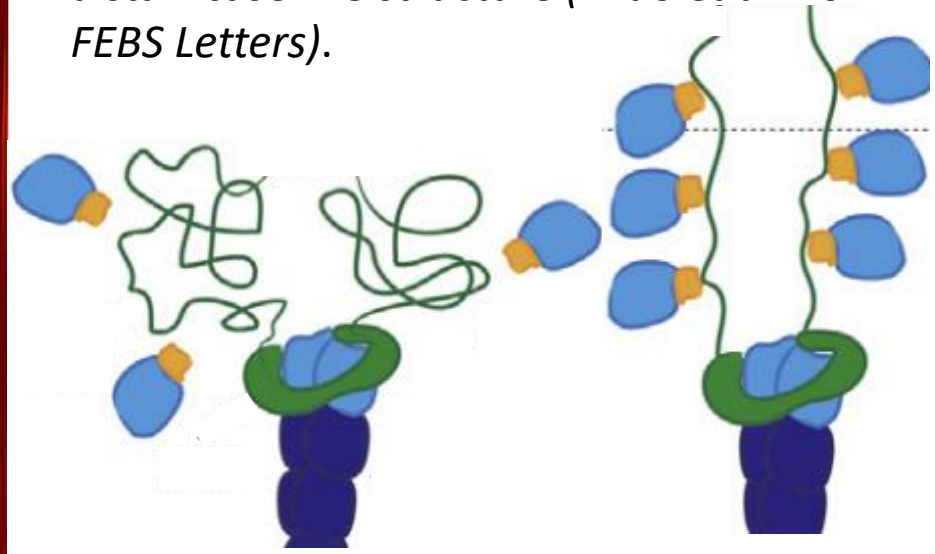




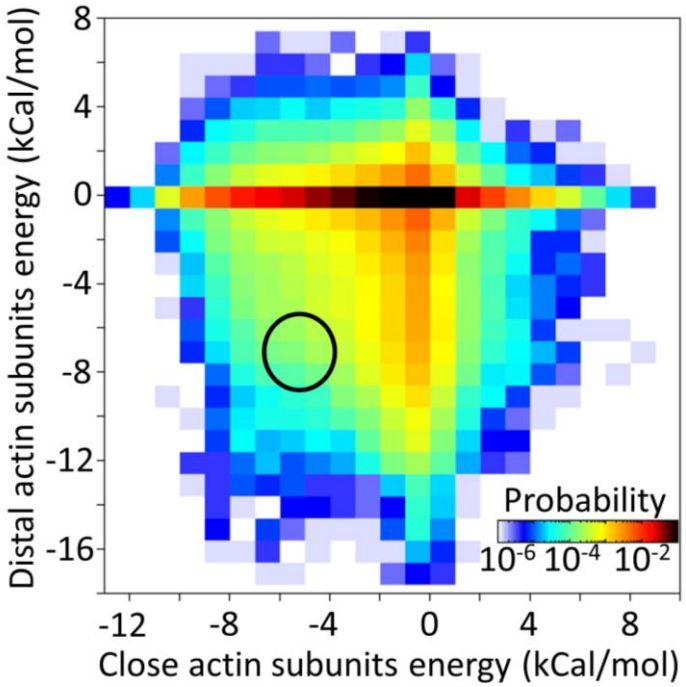
Profilin-actin Delivery to Barbed End is FH1-specific

Possibility of "stair-case" coil-to-elongation?

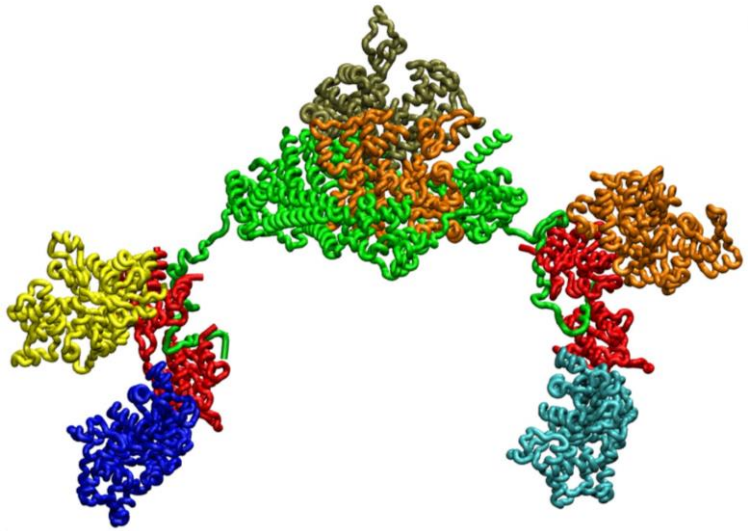
- It was proposed that FH1 expansion upon profilin-actin binding may lead to formation of a stair-case like structure *(Zhao et al. 2014 FEBS Letters)*.



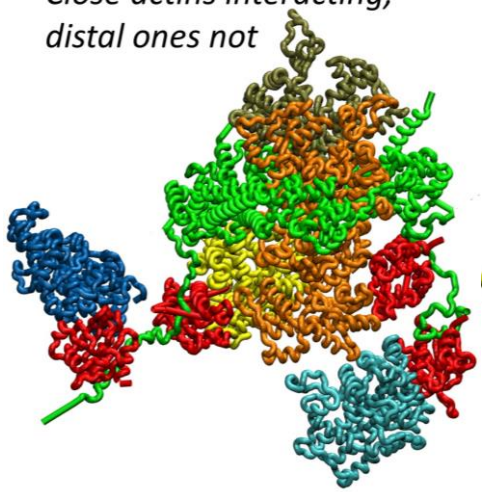
- Simulation finds low probability and stability of any stair-case type complex.



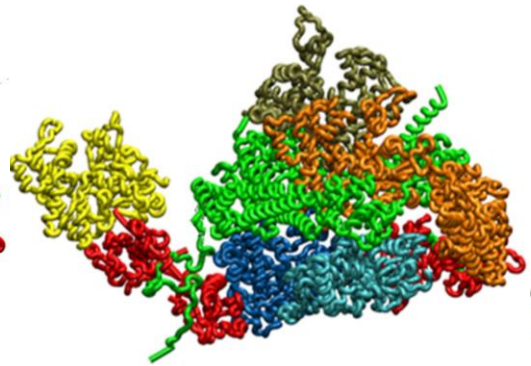
Structure destabilizes in <1ns simulation time



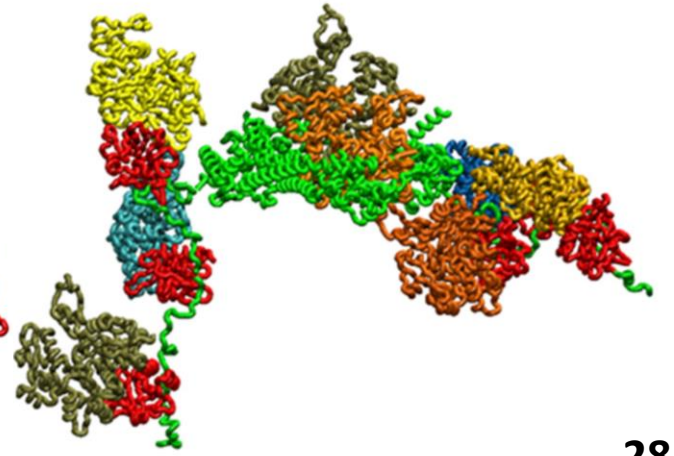
Close actins interacting, distal ones not



distal actins interacting, close ones not



No stair-case like structures observed

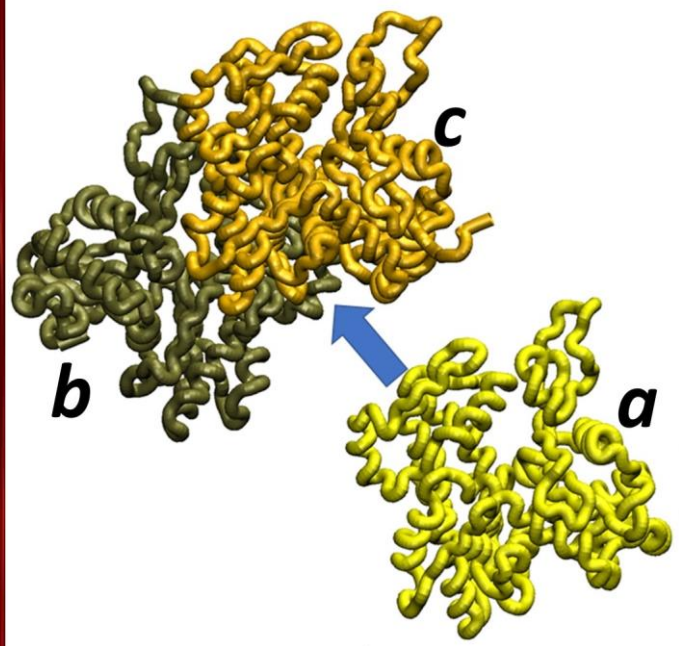


Model Captures Barbed End & Pointed End Binding

- Calculate barbed end binding affinity with minimal system: short pitch dimer and free monomer.

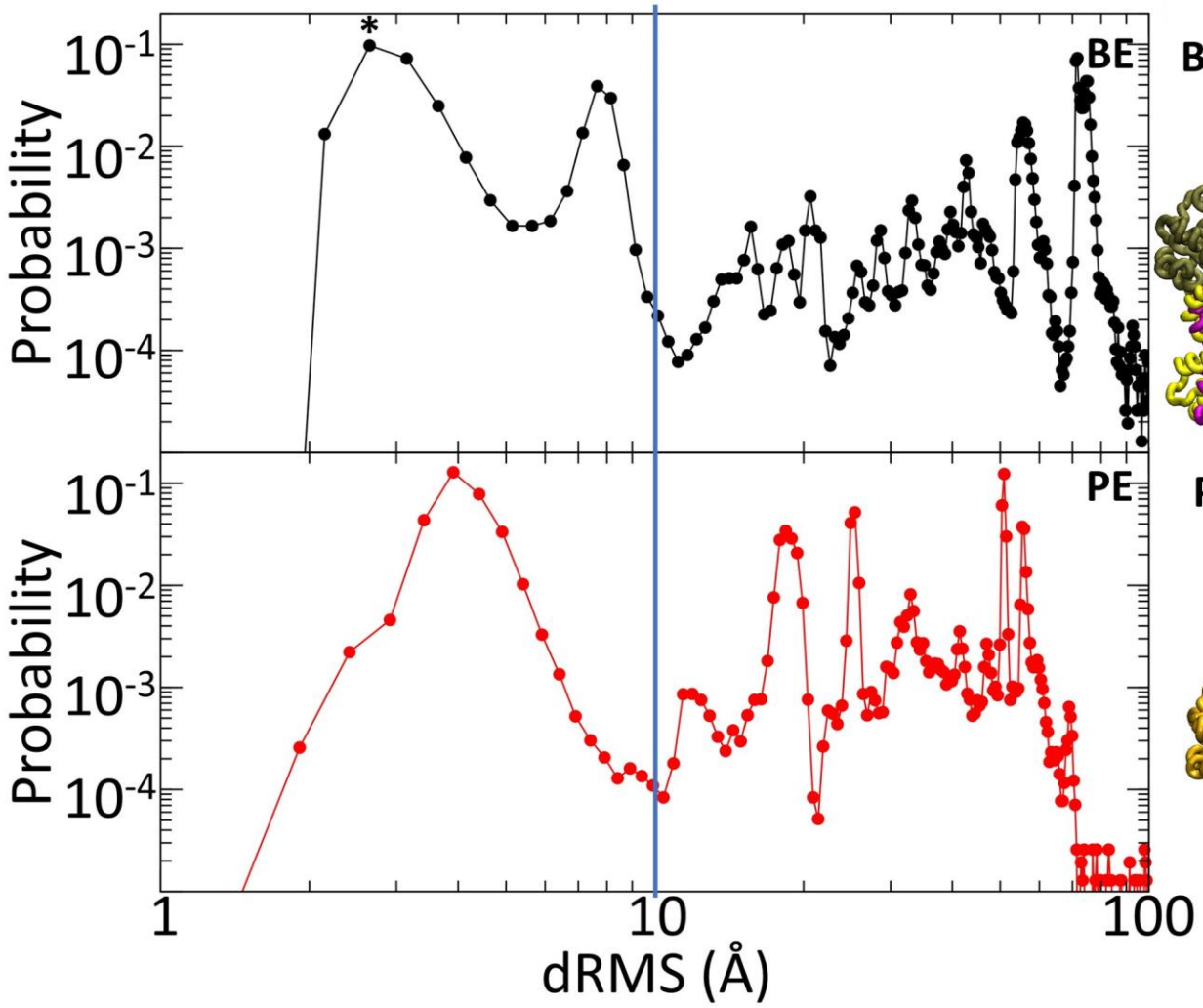
(Oda et al. 2009 Nature)

F-Oda interstrand dimer

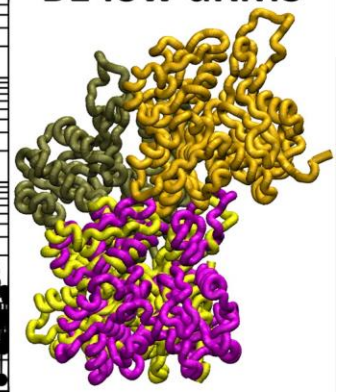


F-Oda monomer

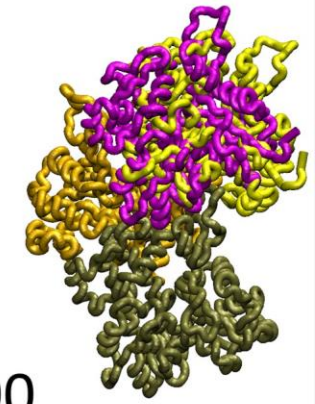
- The dRMS with respect to the Oda model reveals native-like complexes at each end of the filament.
- Binding affinity of ends similar to each other and expected affinity.

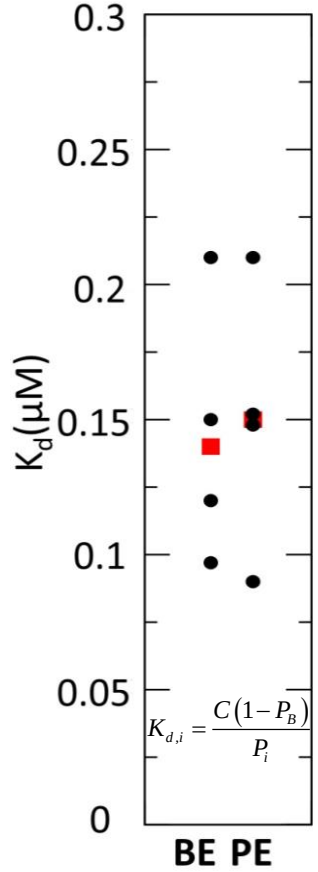


BE low dRMS*



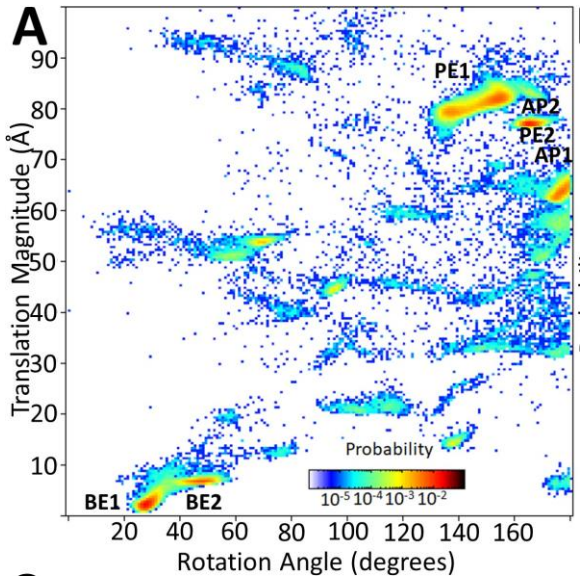
PE low dRMS



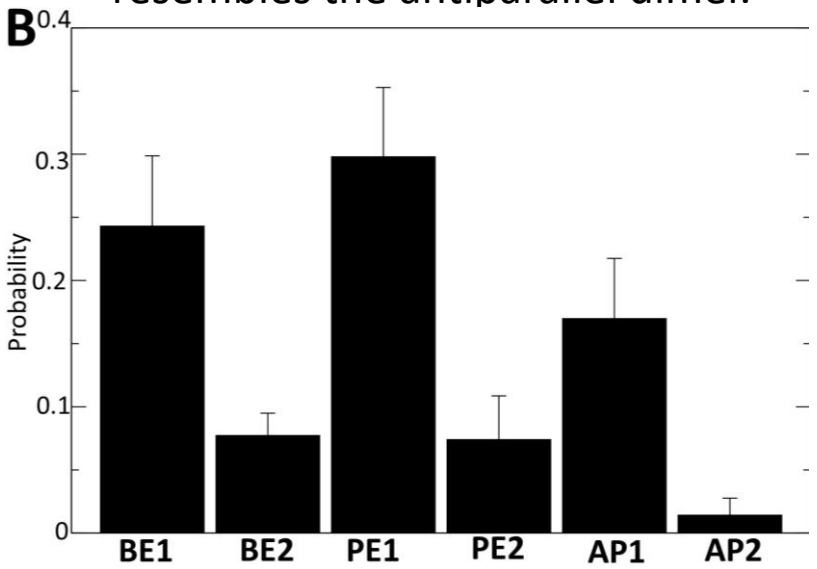


Structural Ensemble of F-Oda Dimer + Monomer

- Rigid body translation and rotation used to describe structural ensemble.

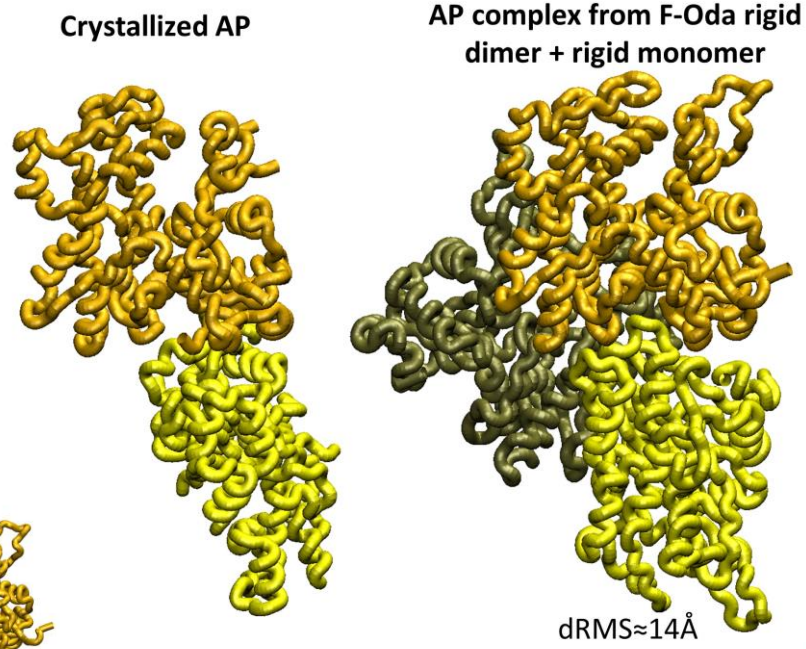


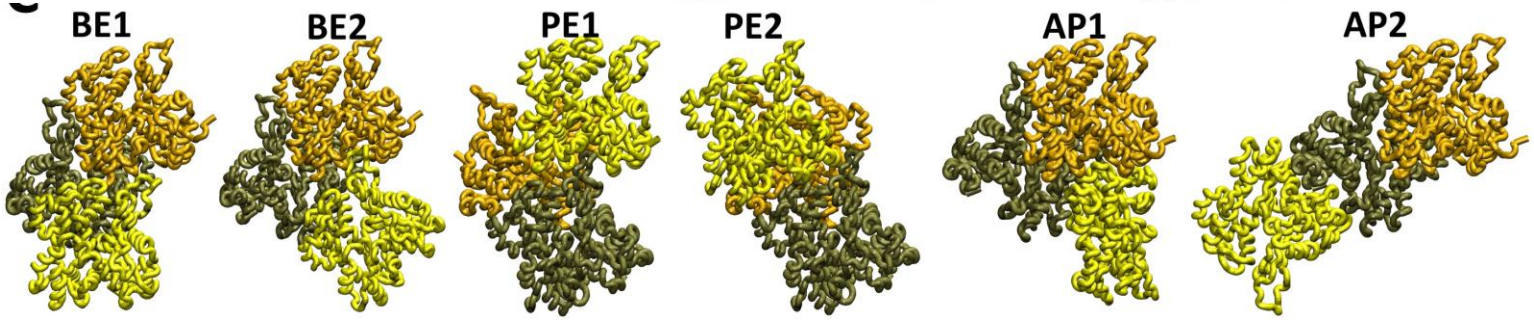
- Binding occurs at each end in filament-like and tilted conformations.
- We also detect a complex which resembles the antiparallel dimer.



- Antiparallel dimer complex bears some difference to crosslinked crystallized structures.
- Some difference to be expected.

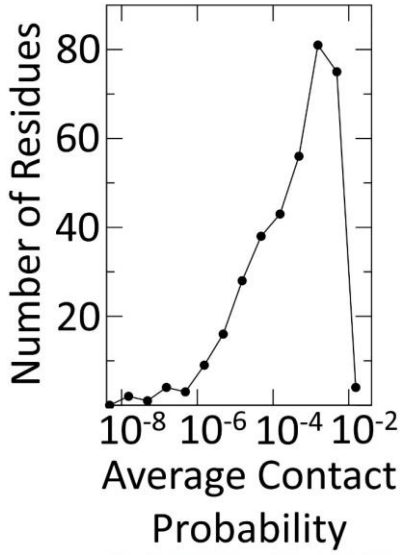
(Reutzel et al. 2004 J. Struct. Bio.)

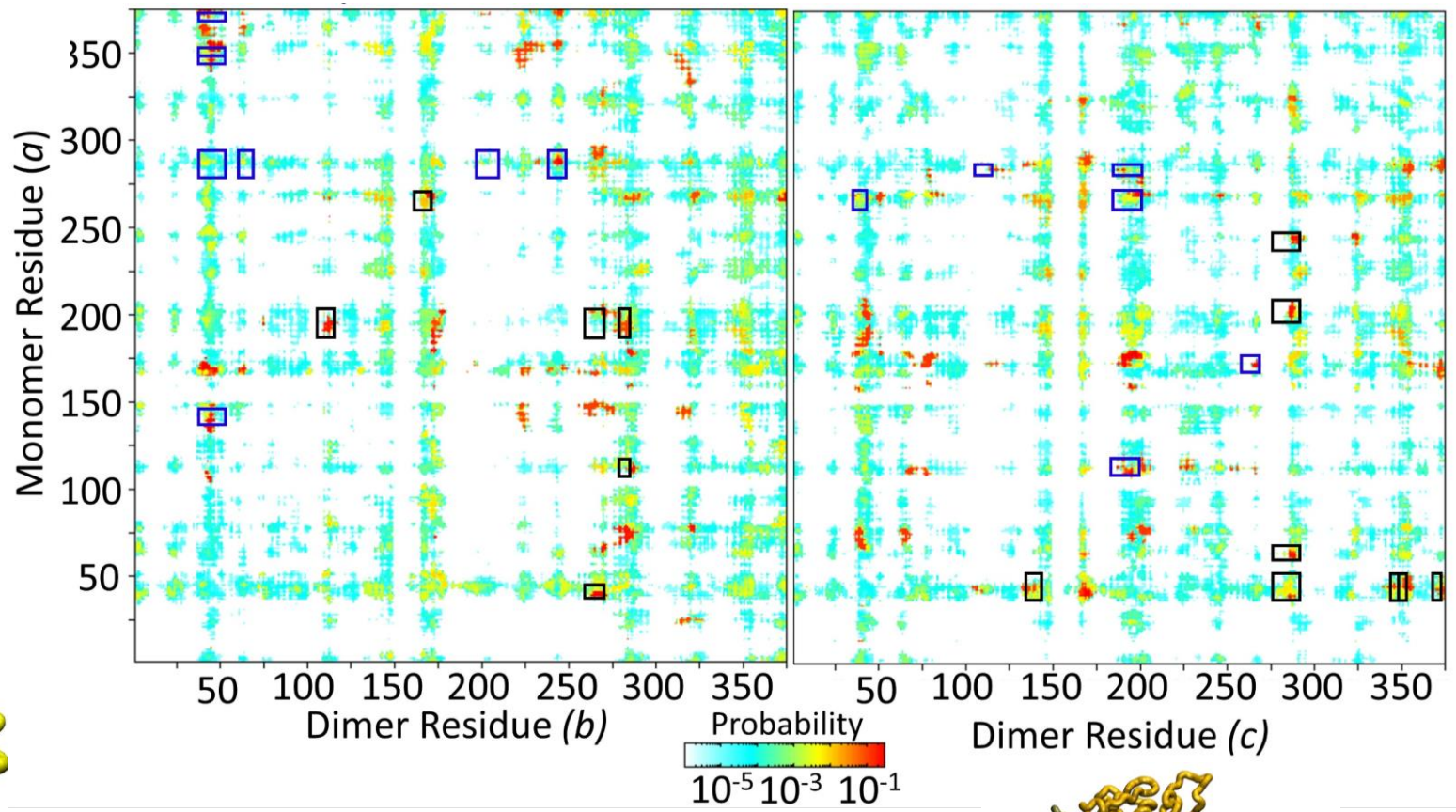


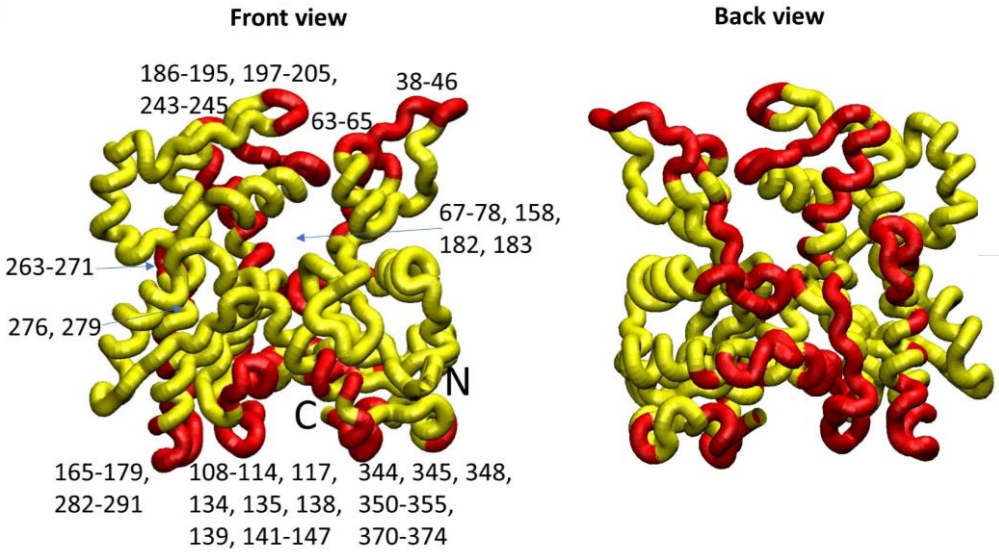


Contacts Involved in Complex Stability

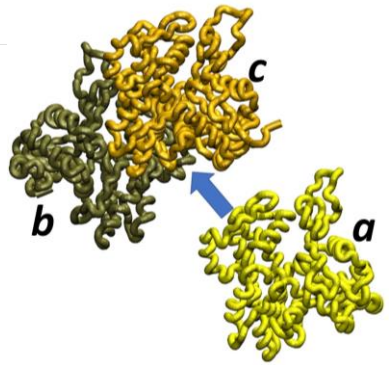
- I calculate the average probability of contact between residues; two residues are defined as “in contact” if they are within twice their interaction radius of each other.







- Contacts similar to Oda model (boxes) with additional regions of high probability contacts found.



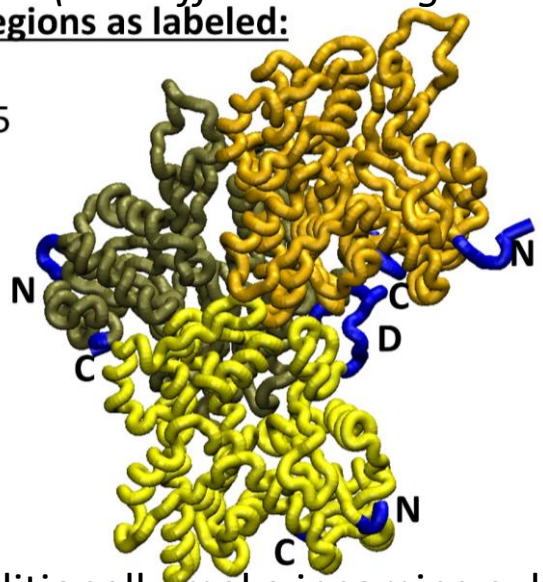
Effect of Flexibility & Monomer Conformation

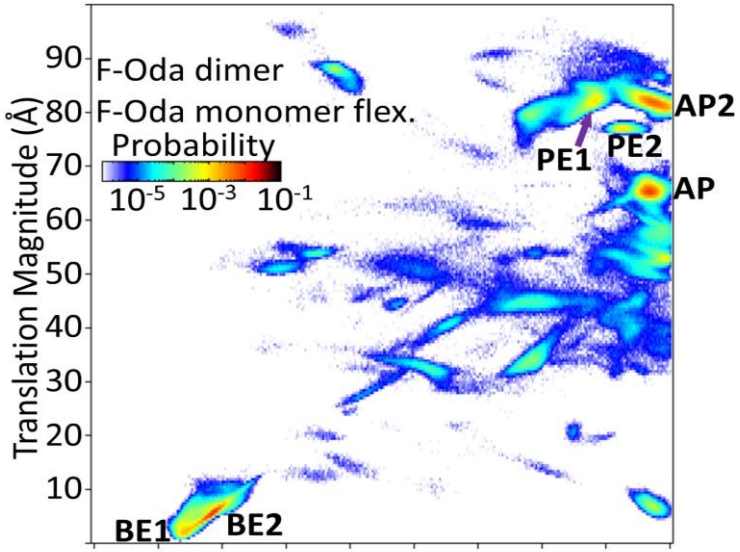
More Reasonable Actin Models

- Flexibility added where residues commonly missing from experimental atomistic structures of actin in barbed end binding interface *(Graceffa & Dominguez 2003 J. Bio. Chem.)*.

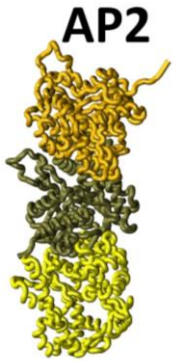
Flexible regions as labeled:

N: 1-5
C: 372-375
D: 40-51

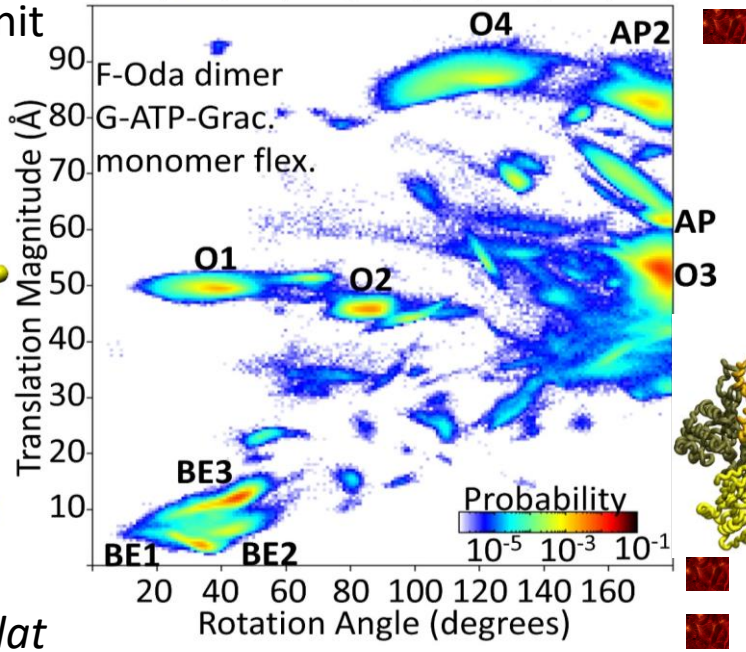




- F-Oda monomer with flexibility yields same complexes as rigid case with slightly different second AP complex.

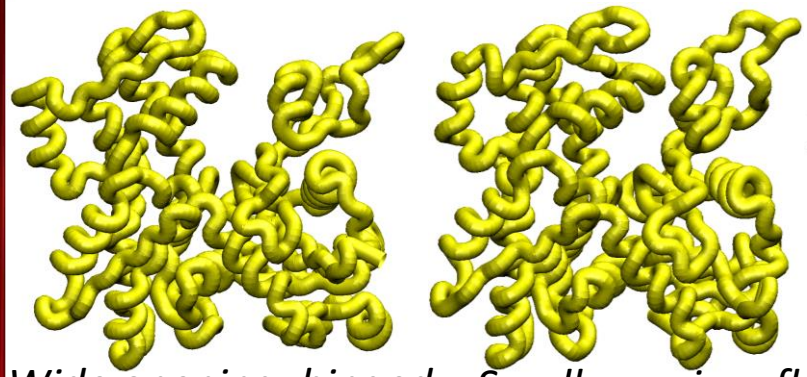


- We additionally make incoming subunit in the, more likely for new subunit addition, G-actin conformation.



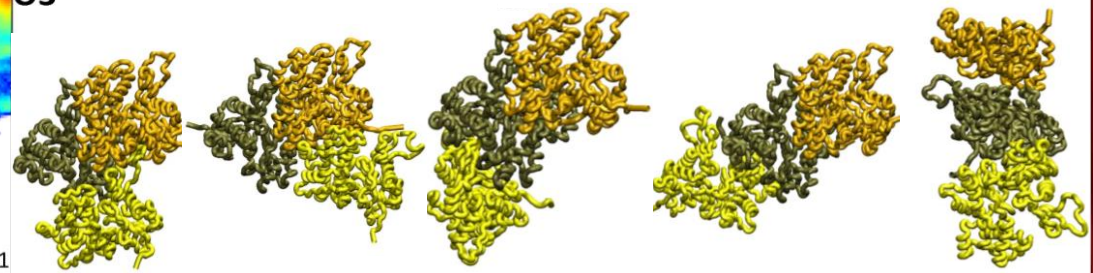
- G-ATP-Grac. monomer with flexibility yields same complexes as F-Oda flex. case with addition of other complexes, 3rd BE complex, but no PE binding.

G **F**



Wide opening, hinged *Small opening, flat*

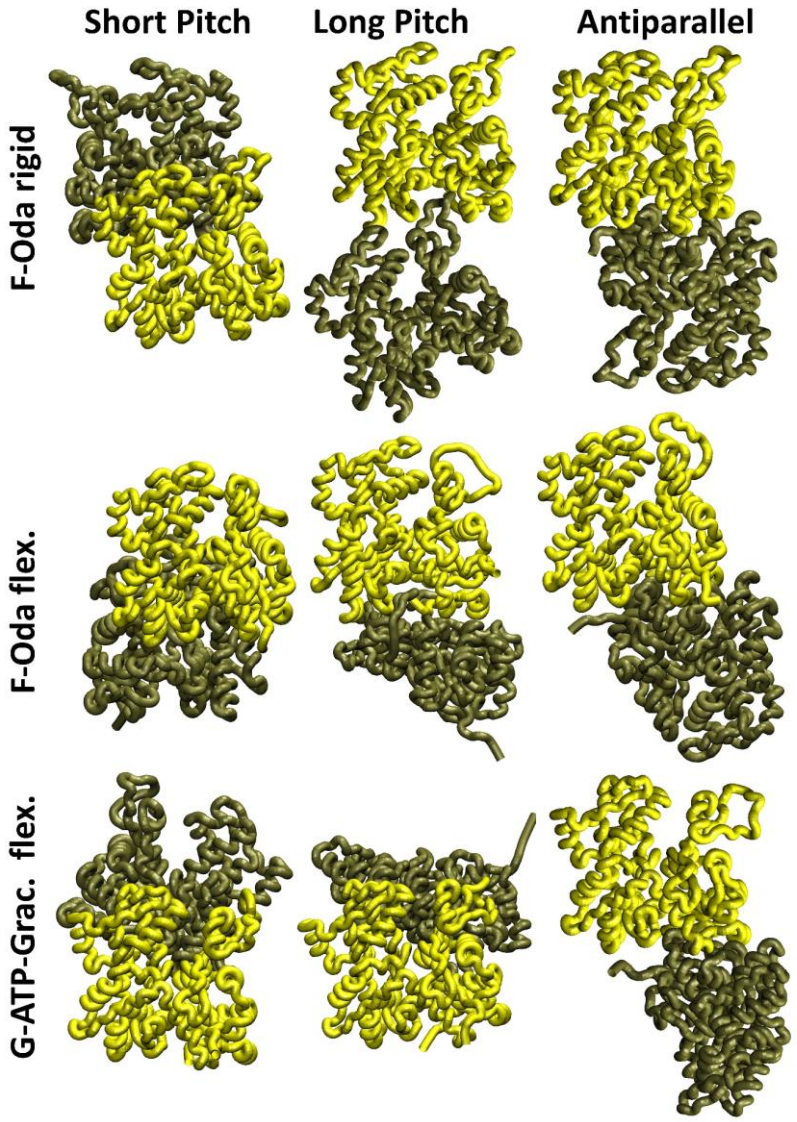
BE3 **O1** **O2** **O3** **O4**



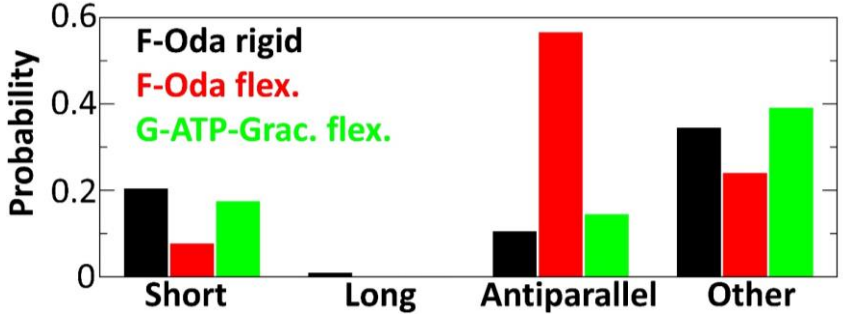
- Flexibility residues involved in stronger contacts
- Flexible D-loop may reach into binding pocket.

Dimerization Favors Short-Pitch and Antiparallel Dimer

- Simulations of each case of monomers as before performed for two free monomers.

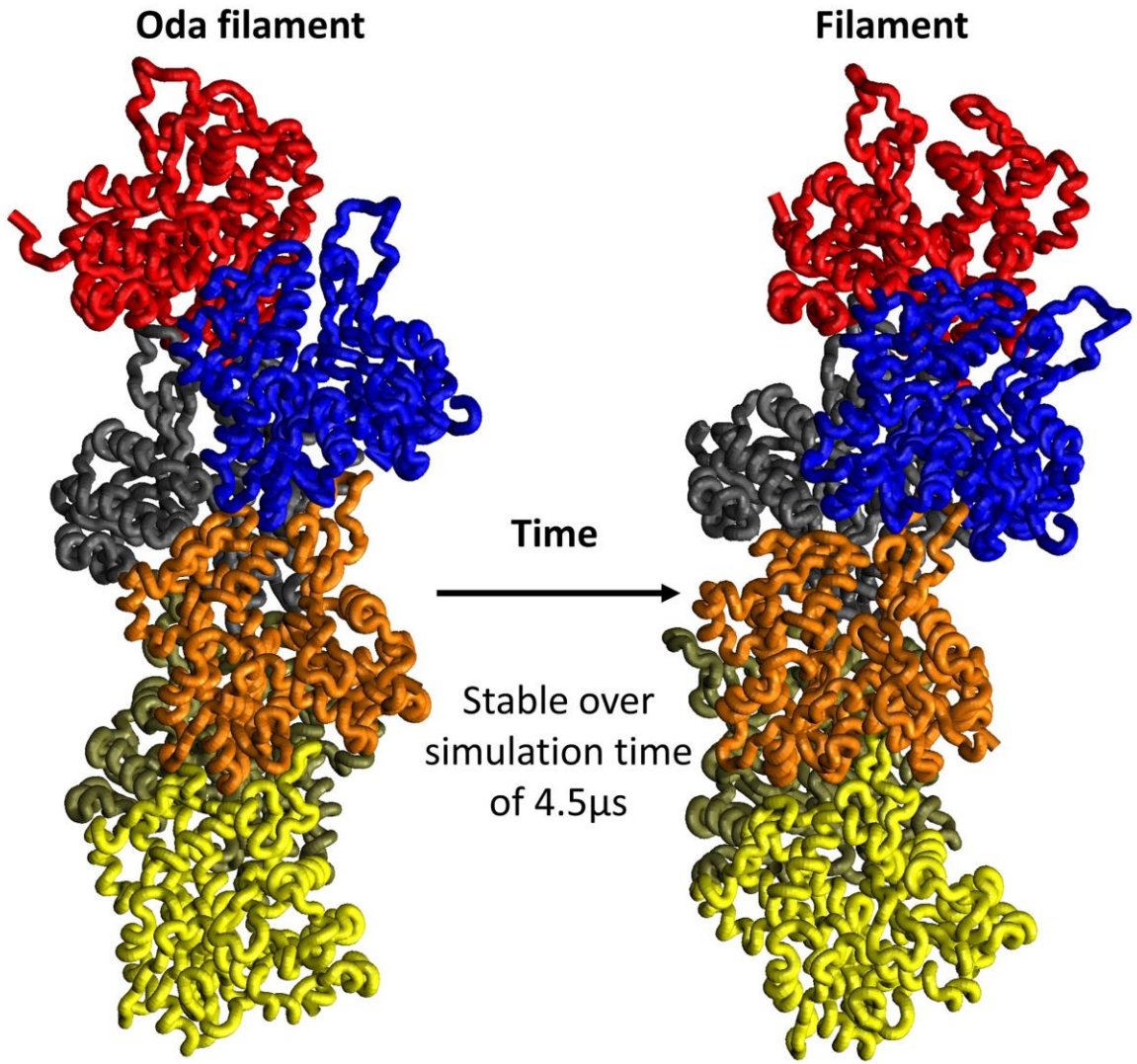


- Lowest dRMS complexes with respect to short pitch, long pitch, and antiparallel shown.
- Short pitch and antiparallel found in each case (decreasing dRMS going down the table for antiparallel).
- Long pitch only found in rigid Oda monomer case.
- Short and antiparallel may be favored because of larger binding interface.



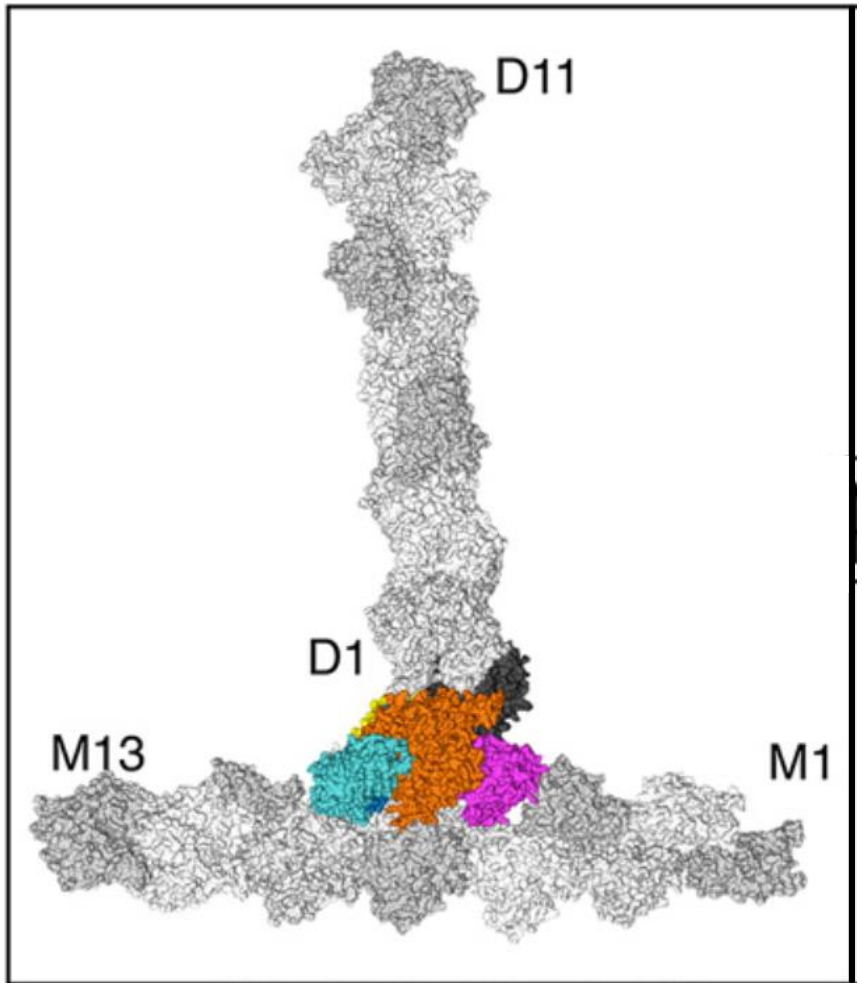
Actin Self-Assembly in Simulation

- Simulations are able to hold an actin filament stable as well as to form an actin filament from a pool of monomers.



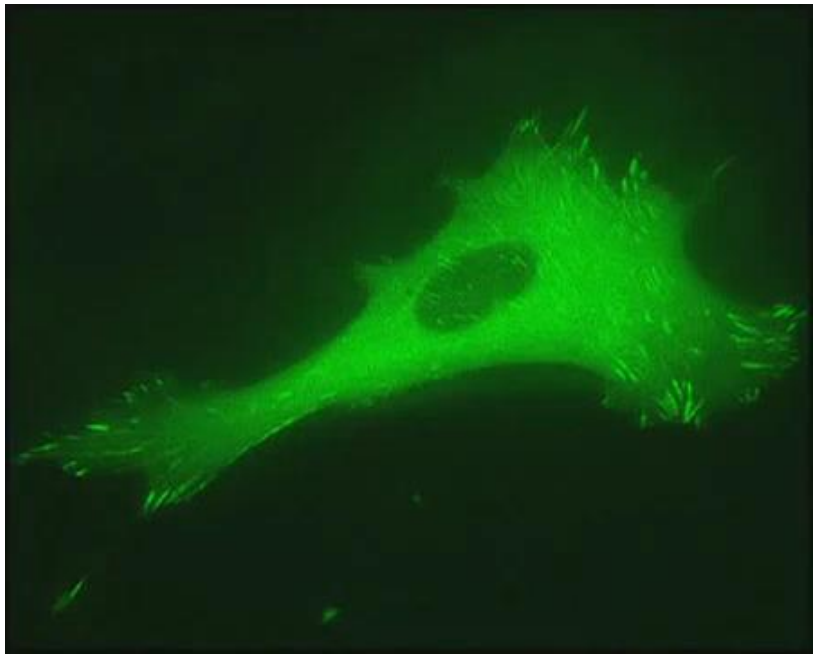
Other Actin Regulators

- Since model well-captures interactions involved in bare actin polymerization as well as in FH1- mediated profilin-actin delivery, we could try to generalize this model to other actin regulators.
- Arp2/3 Complex is other common actin regulator.
- No atomic structures are available; however, an atomic model of the Arp2/3 at an actin branch is available.



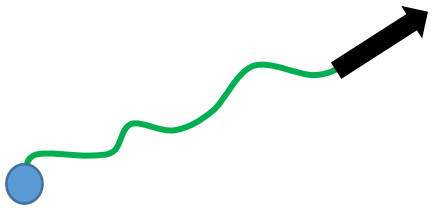
Pfaendtner et al. 2012 J. Mol. Bio.

- Can we describe binding between actin and its other cofactors (Arp2/3, profilin, formin)?
- Can we simulate actin branch formation?
- Can we bridge the gap between molecular modeling and cellular scales?



Formin Mechanobiology

- Application of force on FH1 gives various responses of actin polymerization rates.



Can give	Formin
Increase	mDia1[1-4], Bni1[5]
Decrease	Cdc12[6]
No effect	mDia2[6]

- Mechanosensitive response has been mapped to FH1 domain for mDia2 & Cdc12 [6].

- Existing models of force-dependent behavior of FH1 ignore FH1 primary sequence [6, 7].

- First measurement of an FH1 force-extension curve (for human homolog of mDia1) for forces up to 100pN. Critical force presumably for polyproline helix unfolding is ~40pN. Decent fit to WLC model ($l_p = 0.8nm$) [3].

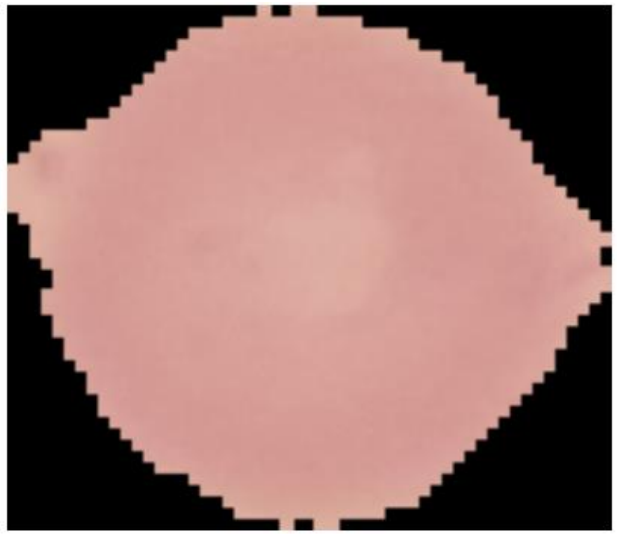
- A good starting point for studying formin mechanobiology is to quantify FH1 extension curves, and see how force affects FH1 occupancy by profilin & profilin-actin and delivery of profilin-actin to the barbed end.

1. Jégou et al. Nat. Comm. 2013
2. Yu et al. Nat Comm. 2017
3. Yu et al. Nano Letters 2018
4. Kubota et al. Biophysical Journal 2017
5. Courtemanche et al. PNAS 2013
6. Zimmerman et al. Nat. Comm. 2017
7. Bryant et al. Cytoskeleton 2017

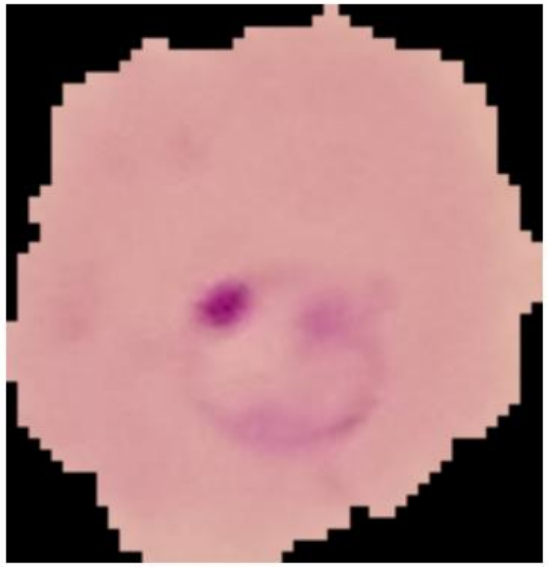
Vavylonis & Horan 2017 Curr. Bio.

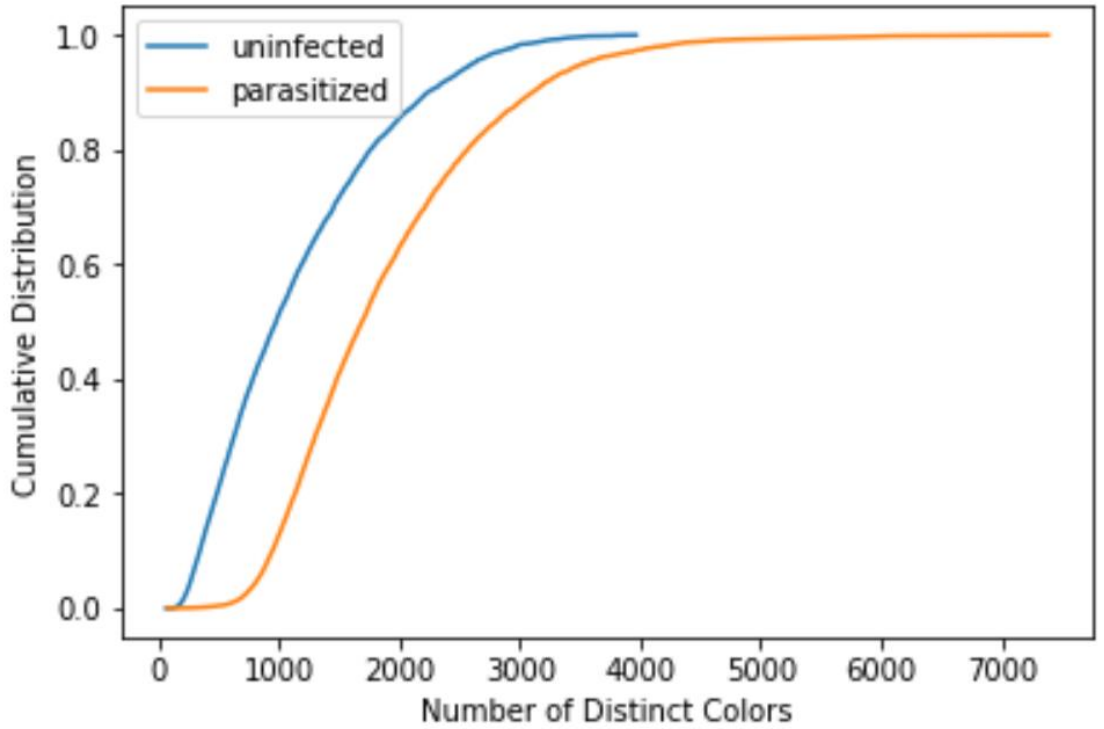
- I use the Malaria Images dataset for this project.

Uninfected



Parasitized

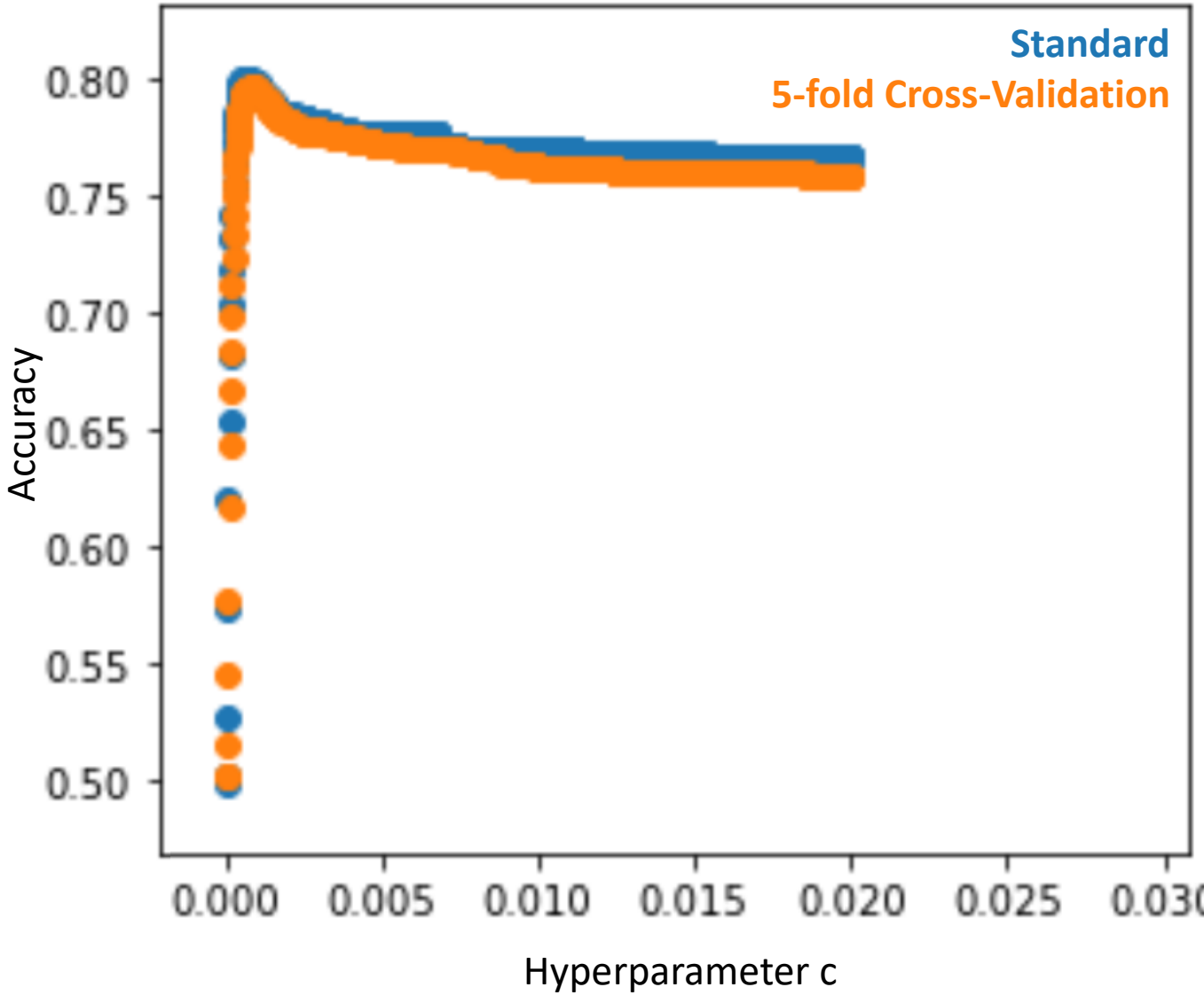




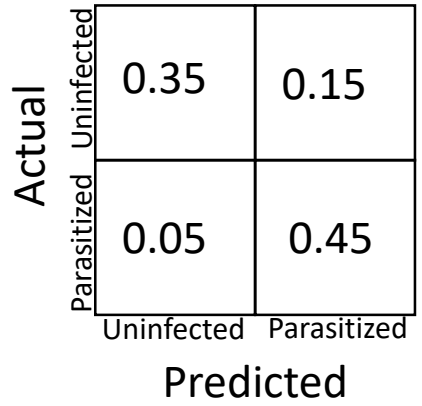

```
uninfected # colors per image: 715.6453298497714 +- 557.7370951226687
parasitized # colors per image: 1525.9140721387619 +- 740.2401933051165
```

Using Logistic Regression on Number Colors Model

- Logistic Regression yields about 80% accuracy on this feature.

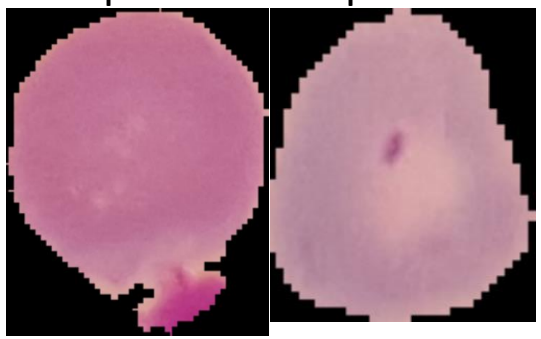


- Using hyperparameter c which maximized accuracy on model which was cross-validated.

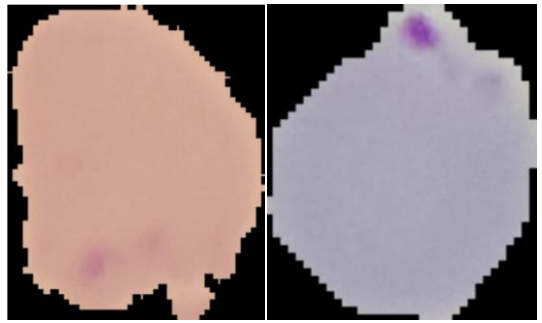


- There are more false positives than false negatives.

Examples of false positives

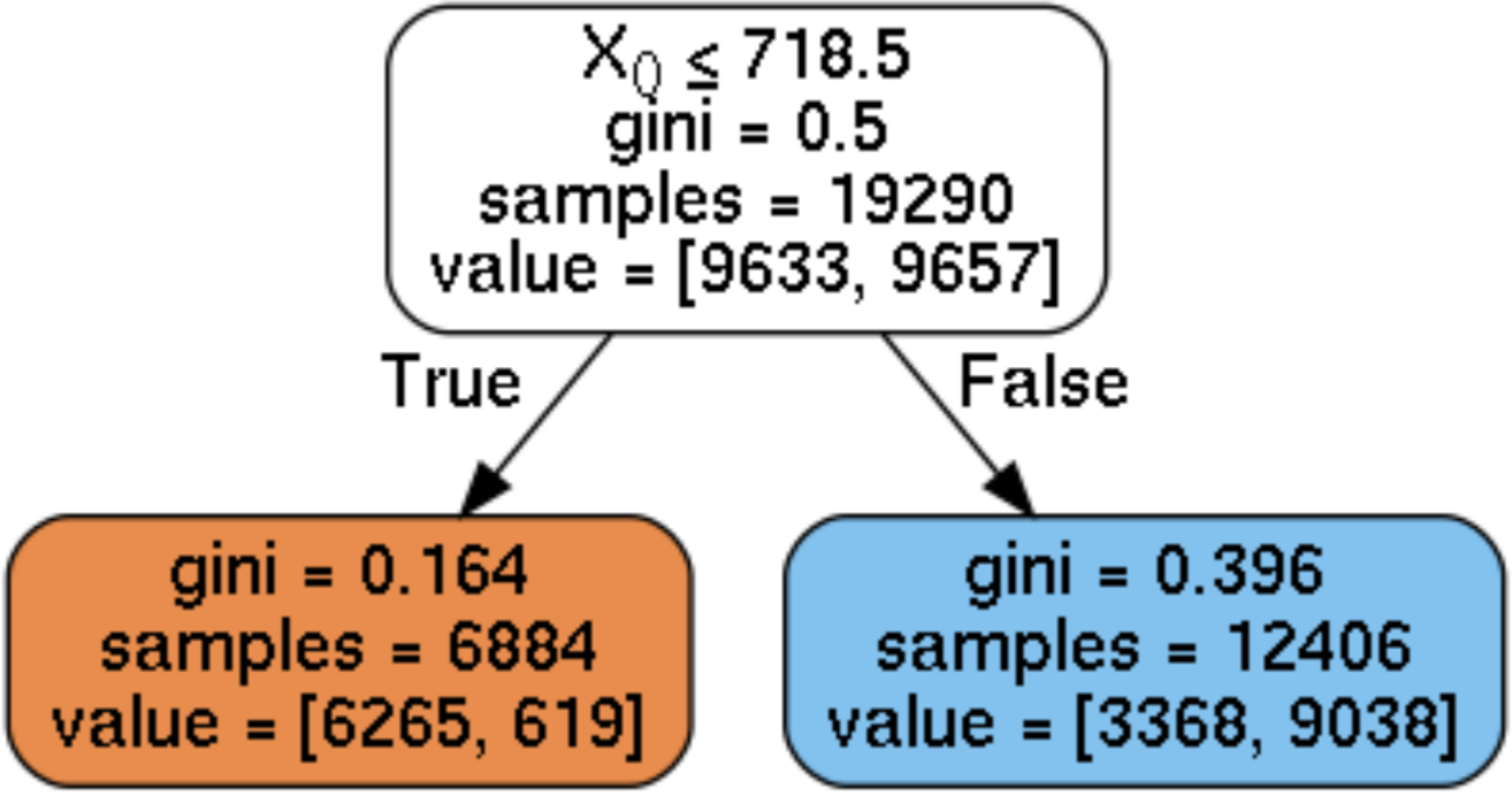


Examples of false negatives



Using Tree Models

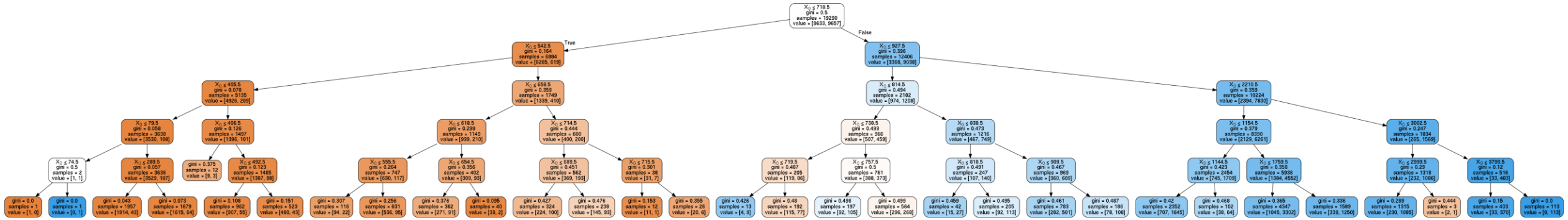
- A simple decision tree yields similar accuracy to the logistic regression model and explains the data quite easily.



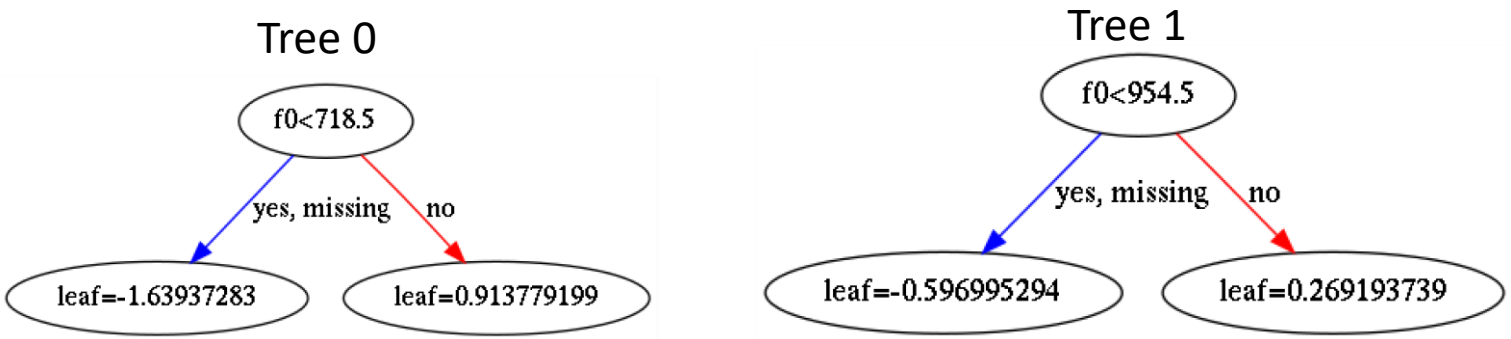
Using Tree Models & AutoML

- Decision Trees. Larger depth can pick up on subpopulations in the images.

- Could be overfitting, but using random forest with same max depth gives same accuracy as single tree → overfitting is likely not happening here.



- I also compared the simple decision tree model to a boosted tree model (XGBoost)



- Giving additional depth to the tree here also does not yield overfitting, but still doesn't give improved performance.

- Using both autosklearn and h2o.ai give models without improved accuracy than I achieved by designing my own models, suggesting I have likely capped my performance models using only this feature.

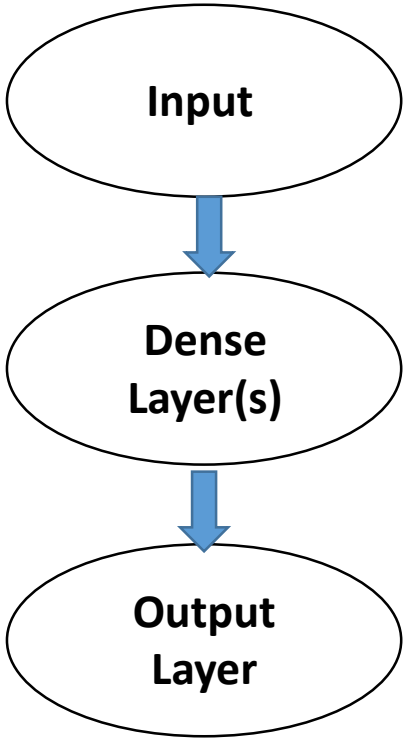
Using Neural Networks

- I designed simple (deep) neural networks with a variety of architectures and trained & tested each network on the same data.

- Each network contained the following elements:

  Sequential model
  Fully connected layers (relu activation)
  Output layer (either as regression with sigmoid or binary classification with softmax activation)
  Loss: binary cross-entropy
  Optimizer: Adam

- Despite architecture variability, accuracy did not exceed 74%.

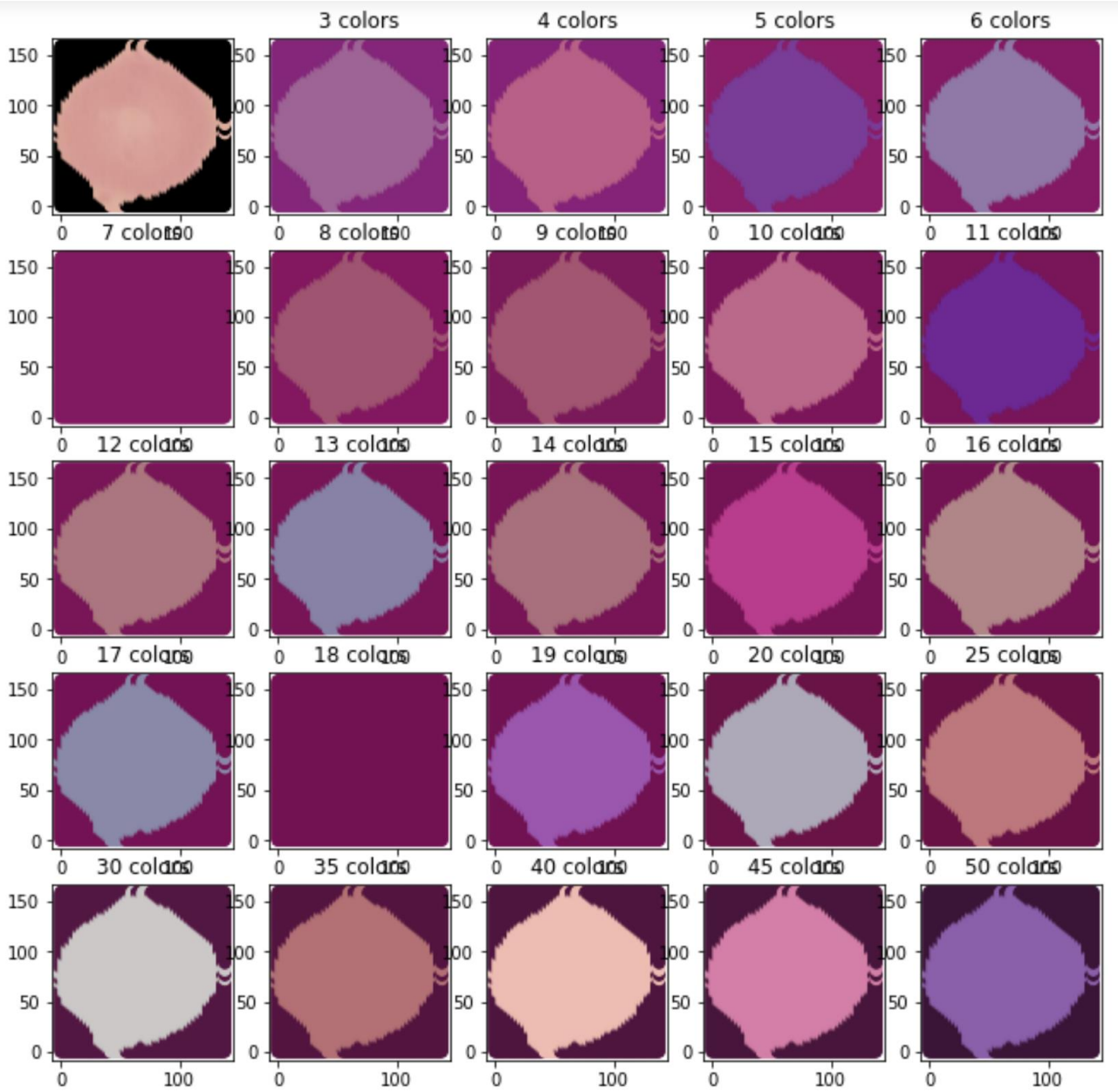


- I also plan to use autokeras to perform architecture search (currently must be installed on separate python distribution than autosklearn, as they require different versions of sklearn).

- While I did get autokeras to work with CNNs (on the MNIST digits data), I have not yet had luck to work on a standard neural network.

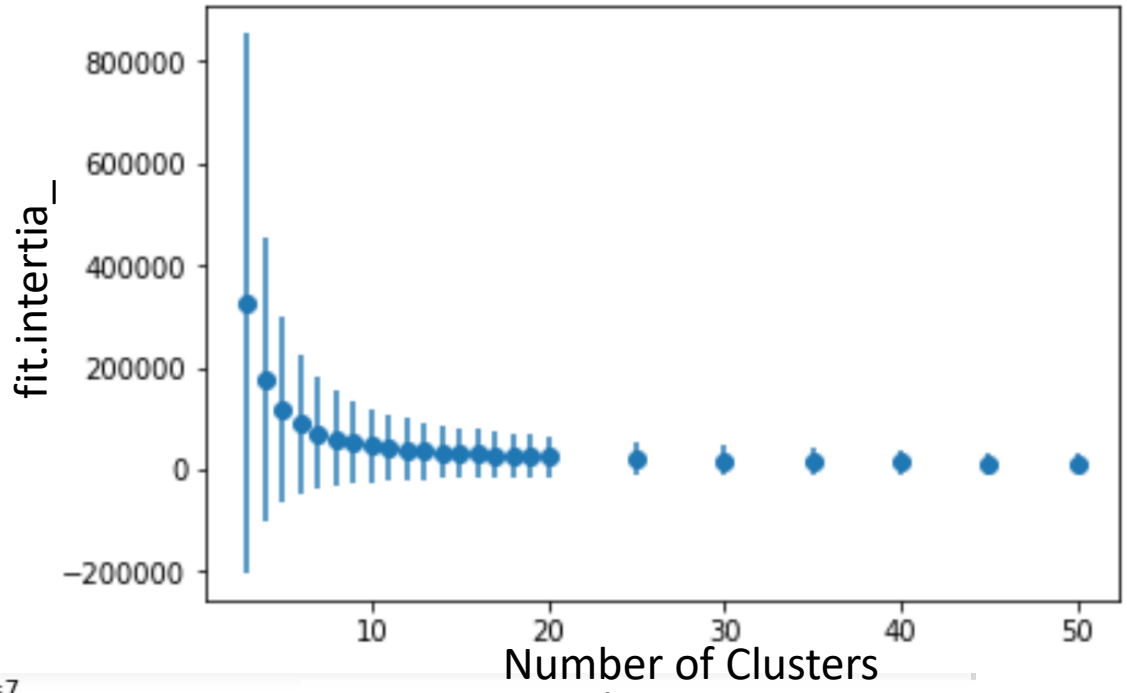
Additional Feature Extraction

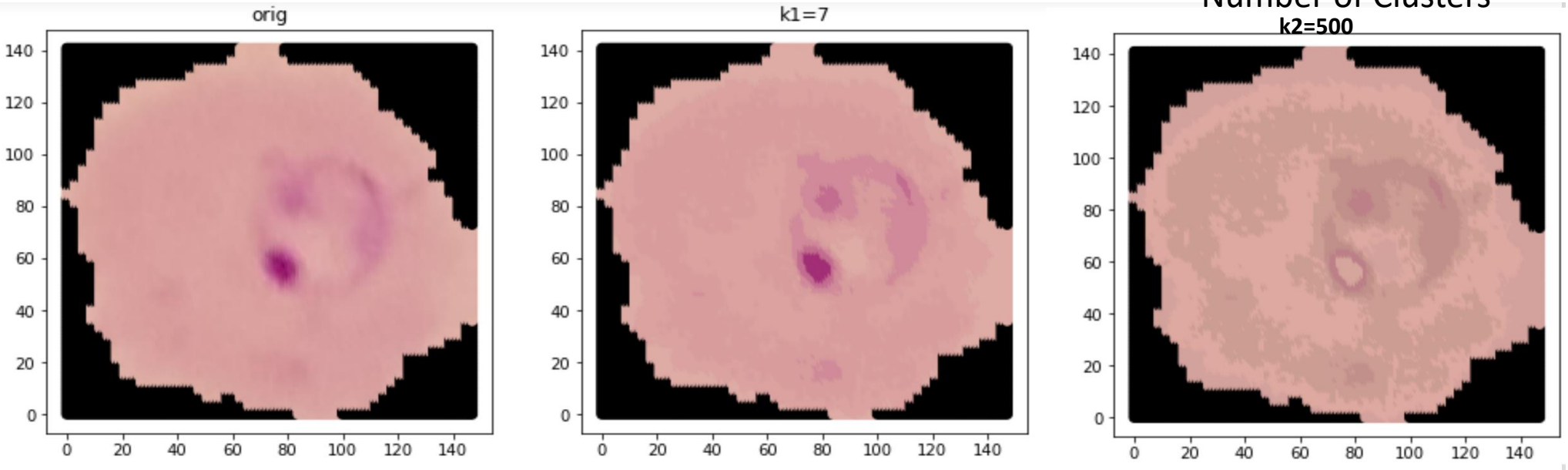
- In the number of colors models, most algorithms give the same result because there is only 1 feature.
- An alternative method to featurize the data is to identify the fraction of each image which is a given color.
- Counting all the colors gives millions of distinct colors.
- Using KMeans to quantize colors across all images does not seem successful.



Additional Feature Extraction

- First clustering individual images' colors seems to help, but doesn't quite seem to solve the problem.
- Only about 7 colors in a given image are needed to roughly capture the image.

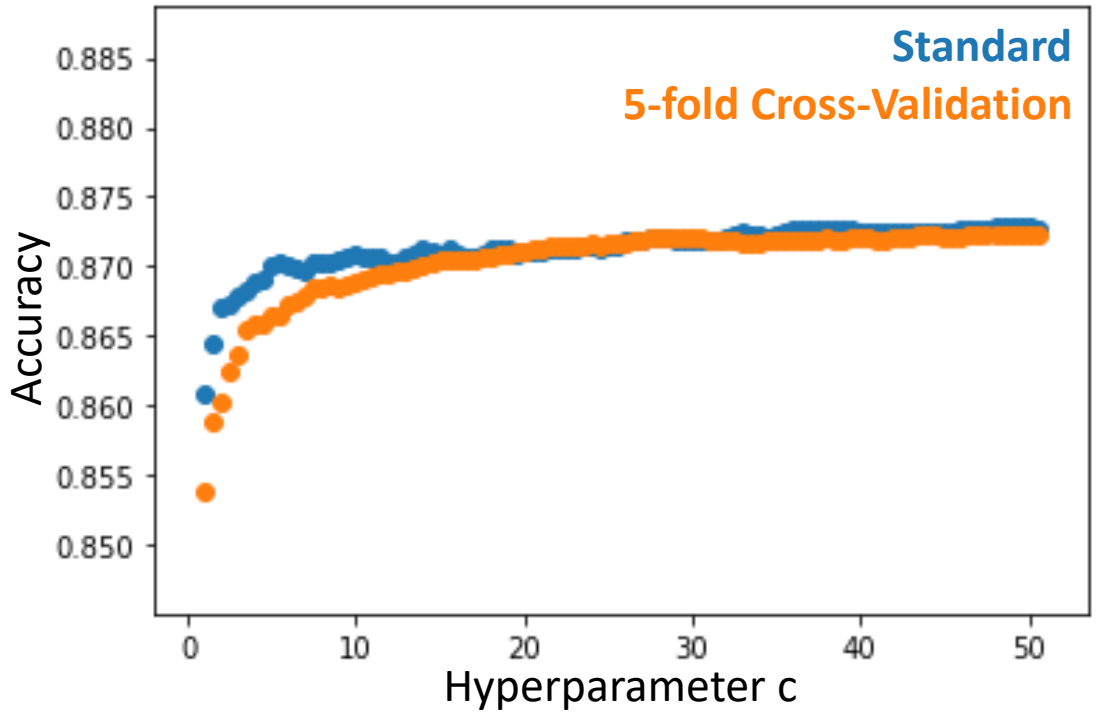




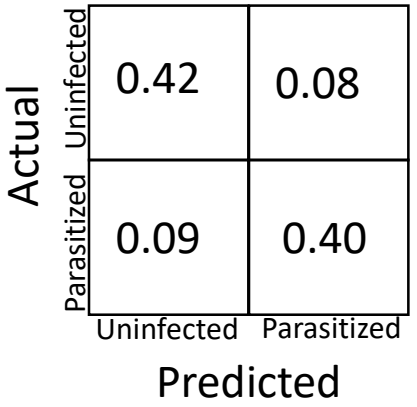
Using Logistic Regression

- An approximate model simply sorts the fraction of each of the clustered colors (but not always the same color!) and treats those as the features of another model.
- This approach is actually pretty good.

- With very large hyperparameter c, Logistic Regression model caps accuracy out at about 87%.

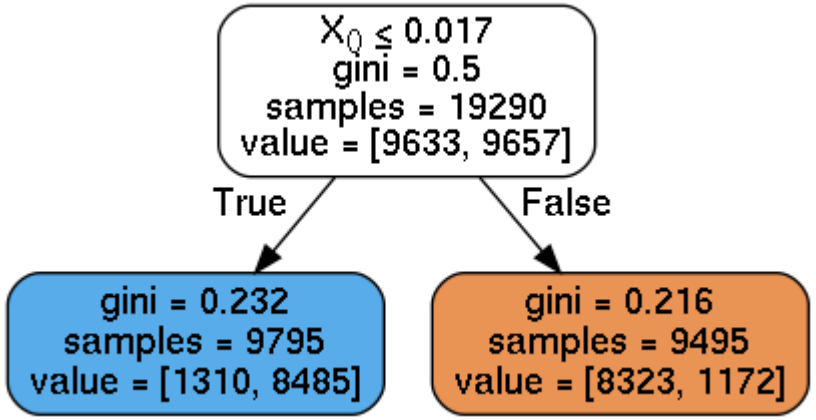


- Not only does this model perform with higher accuracy than the simple number of colors model, but it also has a more even ratio of false positives to false negatives.

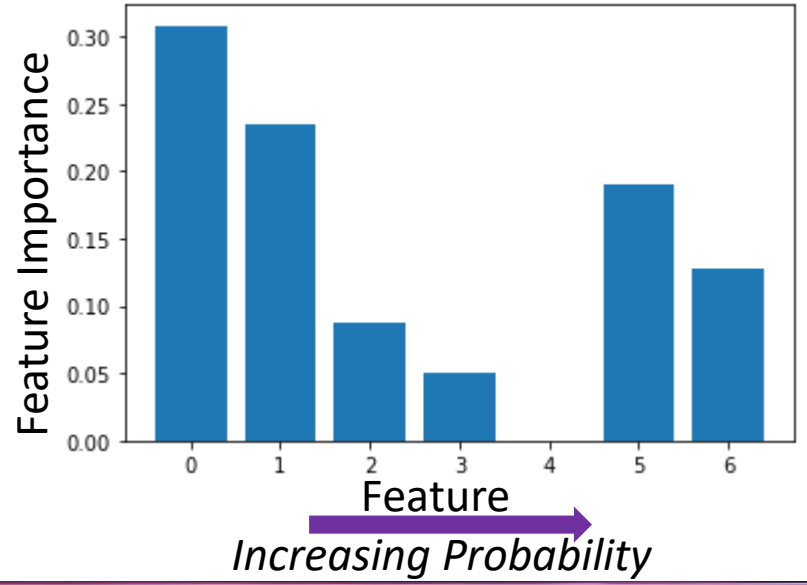


Using Tree Algorithms

- Tree depth of 1 gives near optimal model on these features.
- The least likely color to appear should have > 1.7% to be parasitized (acc: 87%).

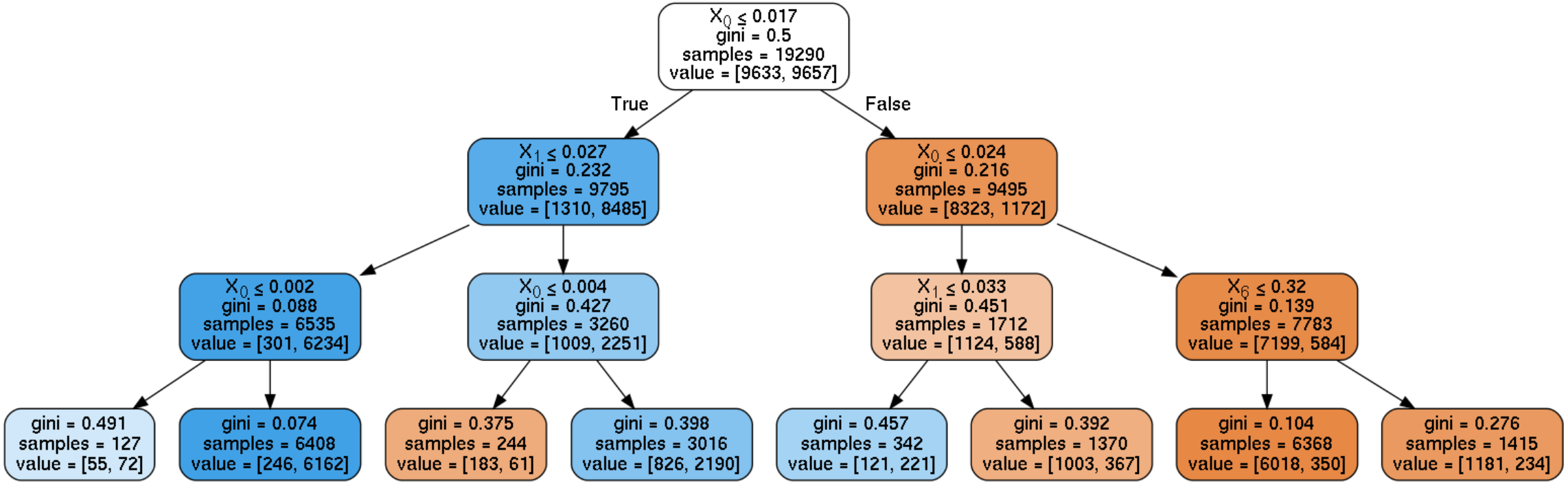


- Random forest with 1000 estimators and tree depth of 1 suggests that the lowest and highest probability colors are most important (too uniform will give uninfected, but asymmetrical color fraction distribution gives parasitized) (acc: 88%).



Using Tree Algorithms

- Additional tree depth nominally helps accuracy (up to 89%), by picking up on importances of other features.
- Because of simplicity of model, becomes difficult to explain.



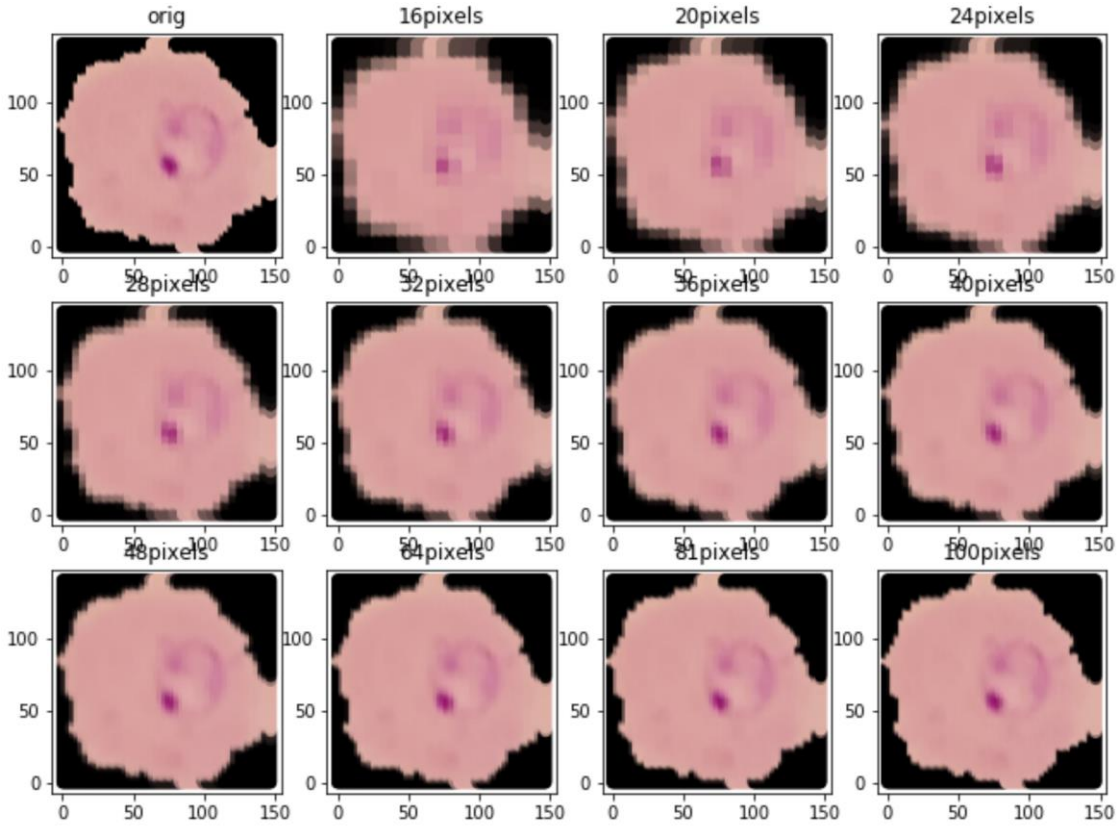
- With these features, tree algorithms perform better than logistic regression because decisions can be made on more than one feature in more than one way.

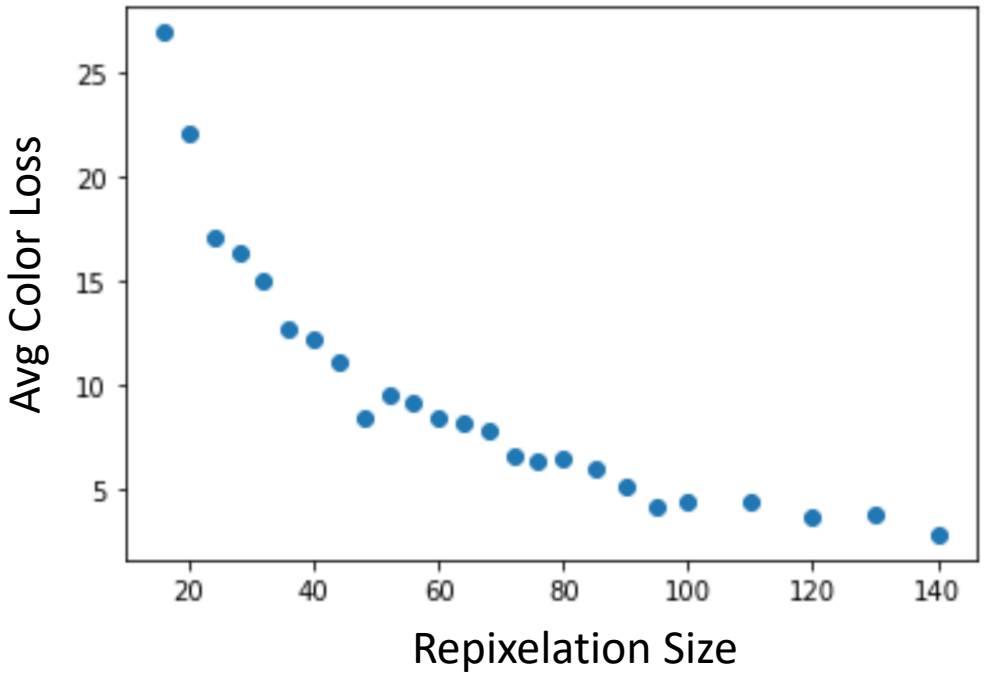
Using Convolutional Neural Networks

- To prepare images for CNNs, they must be repixelated so all images are the same size (unless I use a fully-convolutional network, not done here).

- These plots quantitatively confirm that a repixelation size of ~100 square pixels is appropriate for data preparation for this image.

- Average image size is ~130+-20 pixels, so I don't want to repixelate larger than that. Images are not square, but I just use square images.

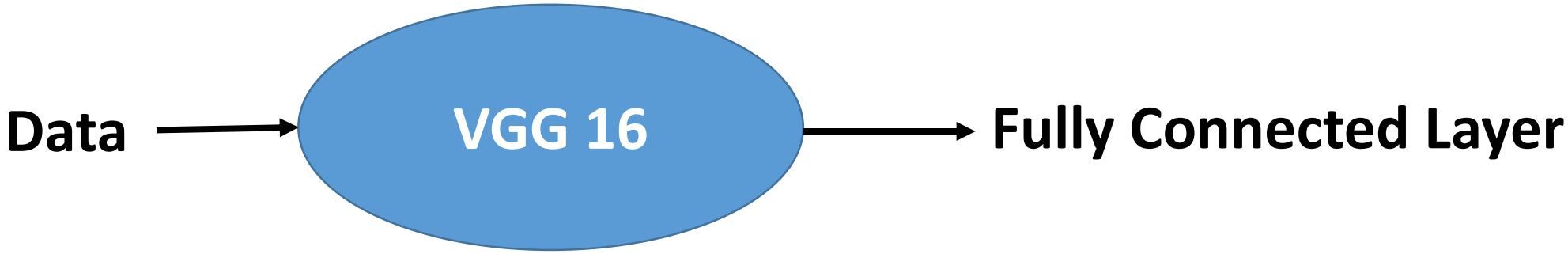




- Using simple architectures, I did not exceed 70% accuracy.
- Using autokeras takes order of weeks, so not used here.

Using Transfer Learning

- I used the VGG 16 model (trained on the ImageNet dataset)
- I added a single fully-connected layer at the end of the model

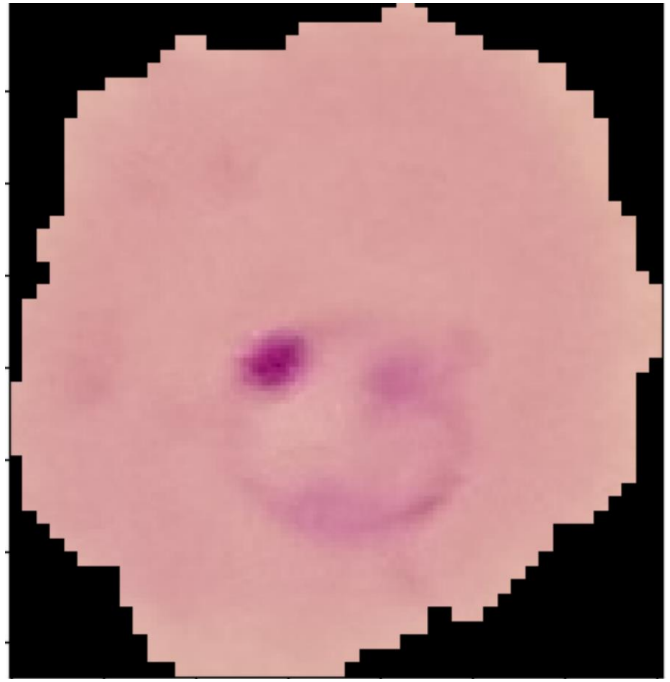


- This fast to train approach gives increased accuracy (93% on the validation set) on images repixelated to 100x100.
- Next steps for possible improvement is to repixelate images to 224x224 (original size for images training VGG 16)

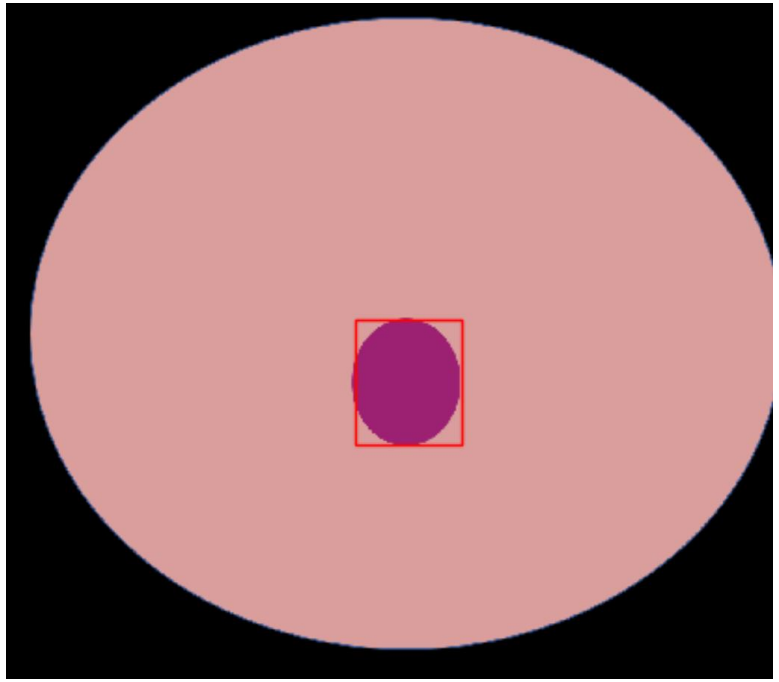
Using Computer Vision

- I applied the Mask RCNN model to identify the parasite in images.
- I found that it struggled to find the parasite in the original image.
- Simulated image parasite found with simplified features.
- Motivated using color quantized version of image where parasite able to be found.

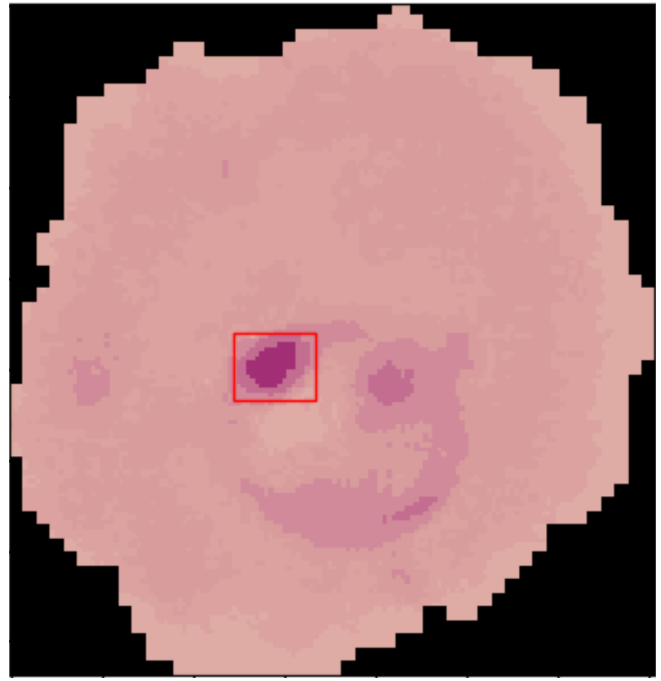
Original Image



Test Image



7 Color Clusters



Questions?

Navigation

Jump to:

Jump to Slide: