Image-derived generative modeling of complex cellular organization in both space and time

Devin P. Sullivan

CMU-CB-15-102

February 2015

Pittsburgh, Pennsylvania

Committee: Robert F. Murphy (Chair)
James R. Faeder
Gustavo K. Rhode
Ivo Sbalzarini, Max Planck Institute of Molecular Cell Biology and Genetics

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Carnegie Mellon University

This research was supported in part by the National Institutes of Health grants P41 GM103712, GM090033 and T32 EB009403. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: generative modeling, machine learning, computational biology, cell and systems modeling, bioimage informatics

To my family and friends for the good times, laughs and support

Abstract

Understanding cellular organization is a major goal of systems biology. Cellular organization affects the behavior of cells and many diseases and disorders impact the spatial organization of cells and their morphologies in turn. There are many current means of studying these systems and their effects. High-content imaging is one high-resolution way in which to study the location of proteins within cells. Advances in imaging technologies have allowed for high quality data to be acquired from live cells in three dimensions over time. Historically, imaging data have been analyzed using image-feature based approaches to create models predicting cell state using classification or regression based machine learning. Generative modeling tools such as CellOrganizer offer an alternative approach to modeling cells and their subcellular structures. The added benefit of this class of approaches is that they describe the statistical distributions of cells and can be sampled from to create realistic *in silico* instances of cells and their subcellular organization. Despite our ability to model static subcellular organization, modeling the dynamic restructuring of cells and their components remains a major challenge in systems biology. These subcellular dynamics are strongly correlated with cell cycle and disease progression and understanding them will aid in the development of treatments. Towards this goal we trained generative models describing cellular morphology dynamics by using both time series and static-time cell image datasets. At a more granular level, cell function is dependent on the proteins within it and their interactions. Not only is the organization of cells correlated with cell response, but it may also be a driving force. To study the impact of cell shape and organization on these biochemical interactions we developed a computational pipeline to perform high-throughput spatially resolved simulations using realistic cellular geometries generated with CellOrganizer. In addition to exhibiting complex responses over time, some cells such as neurons are highly morphologically complex. As such, traditional generative modeling methods are ineffective or fail completely. We addressed this issue by expanding the capabilities of CellOrganizer to include models for neuronal shape. Together these works allow for the study of cellular and subcellular structure for realistic and complex cellular morphologies and their dynamic responses over time in high-throughput.

Acknowledgements

I would like to sincerely thank my advisor Dr. Robert F. Murphy (Bob) whose advice always challenged me to become a better scientist and person. Whether his wise words included "I don't know what you're doing right now, but it's not going to be useful!" or "we can either meat in person or vegetable over the phone", his guidance has made me into the man I am today. I would also like to express my sincere gratitude to my committee, Dr. James Faeder, Dr. Gustavo Rohde and Dr. Ivo Sbalzarini for their support on various projects we've collaborated on and insight into the work included in this thesis. I am also grateful to my former lab members Dr. Armaghan Naik and Dr. Joshua D. Kangas for their support and guidance on a multitude of projects and general advice throughout my graduate school tenure. I would further like to acknowledge my undergraduate research advisor, Dr. Peter T. Cummings and collaborator Dr. Peter Pivonka for sparking my interest in the field of computational biology and nurturing my nascent interest. I would also like to acknowledge all my CellOrganizer team members, especially Gregory R. Johnson, and Ivan Cao-Berg for all the work and help they put into making the CellOrganizer system that was heavily utilized and incorporated in each of the chapters of this thesis.

For the development of Chapter 2 I would like to thank Gregory R. Johnson, Drs. Taraz Buck, Gustavo Rohde and Robert F. Murphy whose contributions to the development of the LDDMM method and its applications to cell morphologies were vital to the development of the models presented in this chapter.

For the work presented in Chapter 3 I would like to express my sincerest gratitude to Jose-Juan Tapia whose collaborative efforts were instrumental in the development of high-throughput spatial modeling pipeline. I would also like to recognize the contributions of Rohan Arepally, an undergraduate advisee who made significant contributions to the early development of the pipeline.

For the models in Chaper 4 I would like to thank two outstanding undergraduate advisees, Xuexia Jiang and Rebecca Elyanow for their efforts in the development of the stick breaking process and neurite placement models respectively. I would also like to acknowledge Dr. Armaghan Naik who initially developed the stick breaking process model with Xuexia Jiang.

I would like to thank the past and current members of Murphy lab for their support and friendship throughout my research here: Dr. Luis Pedro Coelho, Dr. Jieyue Li, Dr. Joshua D. Kangas, Dr. Armaghan Naik, Dr. Taraz Buck, Dr. Seung-Il Huh, Aparna Kumar, Gregory Johnson, Ying Li, Kelvin Liu, Xiongtau Ruan and Ivan Cao-Berg.

I would also like to thank the faculty and students of the Carnegie Mellon-University of Pittsburgh PhD program in Computational Biology for the supportive, vigorous and collaborative environment they have demonstrated during my research here. I would especially like to thank the Dr. Shannon Quinn for his friendship and collaboration in the development of the Short Term Innovative Research competition (STIR) this past year. I would like to thank all the past and current members of the CPCBGSA for stepping up and working to constantly improve the program and the directors Dr. Russell Schwartz, and Dr. Dan Zuckerman for working hand-in-hand with the CPCBGSA in this endeavor. I would also like to express my sincere thanks to Drs. Karen Thickman, Joseph Ayoob and Carl Kingsford for their friendship and guidance and the opportunity to work as their teaching assistant.

I would like thank Nichole Merritt and Thom Gulish for their friendship and support each day and for making my life and the lives of everyone in the program easier.

I would also like to express my gratitude to my friends throughout Pittsburgh. Particularly the ChemE, the Bio and BME students, Dr. Reza Rock and all the team members of any Stuck In Mitosis or Suds team I've ever been a part of for making my time in graduate school truly enjoyable and grounding me when I felt overwhelmed.

Lastly, I would like to thank those closest to me. First, members of my class of CPCB 2010 for the constant emails, lolcats, advice and support you have shown me over the years. My roommates Pat, whose friendship and wit throughout undergrad and graduate school has never failed to make me smile, and Brad who is always up for any adventure. My wonderful girlfriend Emily Lotkowictz for her care, love and support and of course her care packages that helped get me through the final push. My loving godmother Dr. Joan Sullivan for all her messages of love and support. My amazing siblings Dr. Brenna Sullivan and Shane Sullivan for their support throughout my life, putting up with me at my worst and always ready to listen, or help me in any way. Finally, my parents, Dr. Thomas J. Sullivan and Mary Katherine O'Sullivan who have loved, supported and nurtured me over the past 27 years to make me the man I am today. Thank you all I could not have accomplished this without each and every one of you.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

Cellular shape and organization play a key role in cellular behaviors including cellular division, and cellular motility[1-4]. Diseases such as cancer often impact these processes and cellular shape is known to correlate strongly with disease progression including metastasis[5]. This spatial organization is not only necessary for proper cellular function, but often includes dynamic restructuring of the proteins within the cell to perform key cellular functions [6,7].

1.1 Modeling cellular and sub-cellular structures

The field of location proteomics has grown substantially in the last 20 years and there are currently a number of methods available capable of identifying patterns in cellular organization from imaging data. Many of these methods build models discriminating cell types, protein localizations, disease progression states and cell fates [8-16]. These methods use image-feature vectors to predict labels in the case of classification or continuous values in the case of regression. In recent years, these automated machine-learning based methods have progressed to the point of consistently outperforming manual visual inspection [17]. The labels for classifying images typically use Gene Ontology (GO) terms where proteins are assigned a label from a short list of organelles describing the location of the protein [18]. Unfortunately, these labels are binary and do not describe the mixture of proteins between compartments or how their organization changes over time and labels are frequently incorrectly assigned altogether. As with many of these discriminative methods GO terms cannot be used to describe novel patterns.

One drawback of the image-feature based methods is that features are sensitive to imaging and experimental settings making it difficult or impossible to compare models built using different datasets. Additionally, these approaches struggle to identify novel phenotypes or fractional distributions of proteins between cellular localizations. Recently, statistical generative models have been used to address some of these issues by directly modeling the statistical relationship between compartments and the distribution of markers within them [19-24]. These methods use imaging data to learn models that describe the statistics of cellular organization and build models based on the underlying population variances within cell populations.

One such system for generative modeling is CellOrganizer, an open source tool currently developed by Murphy Lab [25]. CellOrganizer consists of two main classes of methods, parametric models and nonparametric models. The parametric models in CellOrganizer can be used to describe cell and nuclear shape, vesicular shape, frequency and location, and microtubule number, length and linearity. These models are somewhat limited by their parameterizations of the cell. For example, the B-spline and ratio models of nuclear and cell shape respectively are restricted to modeling star polygons [19,20,22]. In contrast, the nonparametric large deformation diffeomorphic metric mapping (LDDMM) [26] approach in CellOrganizer can be used to model arbitrary polygons or sets of polygons jointly assuming shapes can be properly registered using non-rigid image registration [24,25]. This approach measures the pair wise differences in cell shape through non-rigid image registration and generates a “shape space” by embedding the resulting distance matrix in a lower number of dimensions.

The models within CellOrganizer not only capture the statistical distributions of compartments and their conditional relationships allowing for comparison of models, but are also capable of synthesizing *in silico* instances sampled from these statistical distributions to create novel hypothetical cellular instances. A major advantage of this generative approach is the use of consistent “global” parameters.

This allows for the combination of models from a variety of experiments, and the direct (and humanly interpretable) comparison across experiments, cell types, and conditions.

One current limitation of both generative and discriminative cellular modeling is the inability to describe the dynamic spatiotemporal changes in cellular organization. A major reason for this is the phototoxicity and photobleaching of current live cell fluorescent imaging technologies [27]. Phototoxicity occurs due to the excited fluorescent molecules interactions with oxygen producing free radicals. Photobleaching is caused by the repeated excitation of fluorescent molecules that breaks covalent bonds necessary to produce fluorescence. Both of these technological limitations increase over time as samples are exposed to excitation photons. This leads to a limited "photon budget" that describes the amount of excitation photons that a sample can be exposed to, limiting the number of times a single cell can be observed.

Recently, Hu et al. [28] addressed the lack of temporal analysis by demonstrating that the use of temporal texture features significantly improved the ability of a model to distinguish protein subcellular localization patterns for 2D time series images though only slight improvements were observed for 3D time series images. Images in this study consisted of NIH 3T3 cells acquired every 45 seconds for 6 minutes, a relatively short time span over which no dramatic subcellular restructuring events were expected to occur. Efficacy of temporal feature extraction to capture relevant changes may drastically improve when analyzing systems for which dynamics of proteins are expected to play a larger role such as over longer times and in response to perturbations. As such, one remaining challenge in temporal imaging is designing experiments that maximally capture temporal dynamics while using a small number of samples to conserve the photon budget.

One of the most noticeable mechanisms of cellular dynamics is the restructuring of the actin network, leading to changes in cell morphology [29]. One approach to modeling these changes in cellular morphology is to extract image features from fluorescent microscopy images that provide a high dimensional dataset on which phenotypes are learned [30]. These data are used to classify cellular response to various treatment conditions and infer biochemical networks through gene depletion studies [31].

1.2 Biological systems modeling

Understanding the dynamics of biochemical networks is a major goal of systems biology. Networks are built through various biochemical assays of protein-protein interactions including the image-based approaches to inferring biochemical networks previously discussed in Chapter 1.1 [31]. As these networks are frequently quite large and complex, computational modeling approaches are required to simulate and understand their behavior. Historically, most of this analysis has been conducted using homogeneous methods in which chemical species are assumed to be well mixed. Such methods include systems of ordinary differential equations (ODEs) and the Gillespie method [32,33]. These methods can be extended to include compartmental models in which homogenous computational “compartments” that define which molecules can interact with each other. These methods are extremely computationally efficient, solving large systems of equations in seconds to minutes, leading to their widespread popularity. These methods are useful and appropriate for studying some systems in which the copy number of each species is large and compartments are expected to be reasonably well mixed.

Unfortunately, as previously mentioned, the heterogeneous nature of cells is critical to their function and copy numbers of molecules can be very low and varies significantly between cells [34]. As such, in the last twenty years, significant efforts have been made to develop spatial modeling tools for these biochemical systems. These spatial modeling tools include two main classes of spatially resolved modeling techniques. The first uses partial differential equations to simulate locally resolved concentrations. This is the technique employed by the Virtual Cell (VCell) among others [35,36]. This method is mathematically similar to ODE methods however systems of differential equations in this case are solved with respect to both time and space resulting in smooth continuous concentration gradients over the geometries. These methods are most appropriate when spatial heterogeneity is believed to be important but copy numbers are still relatively high such that stochastic effects are believed to be small.

The second class of spatially resolved methods is agent-based methods used by programs such as Monte Carlo Cell (MCell) and Smoldyn [37-41]. These methods simulate each molecule individually and evaluate their diffusion and probability of interactions on a per-particle basis for each time step. This class of simulator is therefore solved in a discrete fashion and requires very small time steps, typically $10^{-9} – 10^{-4}$s, depending on the time scale of the interactions happening in the specific system. They are extremely computationally intensive but have a higher spatial resolution. These methods are therefore most appropriate for systems where there are small numbers of heterogeneously distributed molecules interacting. In this case both spatial and stochastic affects are important to the behavior of the system; however to determine the expected behavior of the system multiple random initializations of the simulation are required further adding to the computational cost to perform these simulations.

At a still higher level of spatial resolution, molecular dynamics describing groups of atoms or individual atoms within a molecule may be performed. These systems describe the interactions between very small subunits in the molecule of interest and therefore are only appropriate for describing the interaction of a very small number of proteins. Though these simulations can be used to study local environmental effects such as trans-membrane domains [42]. At this scale organelle and cellular organization usually cannot be studied due to computational cost.

Despite the development of these spatially resolved simulation tools, a vast majority of cellular modeling continues to be homogenous due in part to the limited number of realistic geometries available for simulation and the inability to efficiently study cellular response using targeted geometries and organizations. When performing spatially resolved simulations, geometries are often either hand segmented or manually fabricated, both of which are tedious processes. Additionally, these simulation tools currently require a large amount of training to properly and efficiently use.

1.3 Modeling neuronal cells[1]

Currently an area of major interest in the field of biology is the understanding of neurons and their interactions particularly in relation to the human brain. These efforts include the NIH large-scale initiative on Brain Research through Advancing Innovative Neurotechnologies (BRAIN initiative) and the Human Brain Project (HBP). This has led to a large surge in efforts to model neurons. As discussed above, generative models and biochemical simulations within model instances can provide key insight into biological processes. Due to the highly complex morphologies of neurons specialized models are needed to describe neuronal shapes.

Previously there have been several efforts towards building generative models of neurite shape. The NeuroMorpho (previously L-Neuron) software package uses a set of recursive rules to describe dendritic geometry and topology by correlating morphological parameters such as branch diameter and length.

[1] This section describes joint work with Armaghan W. Naik, Rebecca Elyanow, Xuexia Jiang and Robert F. Murphy and is being

The system implements three algorithms: Hillman, Tamori, and Burke, to describe the branching length and angles of the dendritic trees [43-45]. The TREES software reconstructs the neuronal branching under the assumption that dendritic trees connect synaptic inputs to the dendritic root using a minimal total length of wiring, and performs a local optimization of total wiring and conduction distances [23]. Other softwares such as Neugen and Netmorph are aimed at the generation of morphologically realistic, large-scale neural networks in 3D [46,47]. Although the LDDMM approach in CellOrganizer could also be used to model the complex morphologies of neurons in theory, it is unlikely that cells could be properly registered leading to inaccurate models.

Neuronal model instances can be imported and run in simulation programs such as Genesis, NEURON, or MCell [40,48-50]. These tools have been successfully applied to modeling complex multi-cell models of neuronal networks and show promise in aiding the understanding of neurons at the single and multicellular levels [51,52]. Although each of these methods is capable of reproducing statistically plausible neurites under various constraints they do not allow for the direct modeling of nuclear shape, soma shape, and protein distributions.

1.4 Thesis contributions

The overall goal of this work is to create methods for the purposes of modeling complex cellular shapes and organizations over time and space. Towards this goal we first extend our previous efforts in generative modeling with the CellOrganizer system to create dynamic generative models to predict changes in cellular morphologies over time. We implement several methods for modeling this shape change and evaluate each method for a set of c2c12 cells. Further, we demonstrate that building such dynamic models of cell transition is possible using static cell images provided a temporal marker of cell state. Second, we design and implement a high-throughput computational pipeline for utilizing the generative model instances from CellOrganizer to simulate spatially resolved biochemical systems for the purposes of studying the impact of cellular organization on cell dynamics. Lastly, we expand the capabilities of CellOrganizer to include the ability to learn models of the complex structures of neuronal cells and synthesize realistic *in silico* instances of neurons for use in subcellular protein modeling. Taken together, this work enables the study of cell shape and organization and its impact on cellular response at high resolutions.

Chapter 2. Dynamic generative models of cell and nuclear shape[2]

2.1 Introduction

Cellular shape and organization play a key role in cellular behaviors, including cellular division, and cellular motility [1-3]. Diseases such as cancer often impact these processes and cellular shape is known to correlate strongly with disease progression including metastasis [5]. By building models of cellular shape and organization we are able to gain an understanding of how these cells behave at different points in the cell cycle [53], a disease progression, and under various conditions such as drug treatments. Previous efforts have shown that classification based approaches are used to distinguish cell types, protein localizations, diseased cells and cell fates [8-10].

Recently, statistical generative models based on distributions of cellular morphology and organization have been used to study the underlying population variances [19-22]. The current work extends our previous efforts in generative modeling using CellOrganizer (cellorganizer.org) to model cellular change over time. For this study we focused on the Large Deformation Diffeomorphic Metric Mapping (LDDMM) [26] with multidimensional scaling (MDS) approach to cell shape modeling as it naturally separates cells along their major modes of cell shape variance making it suitable for building models of shape change [20,25,26].

[2] This chapter describes joint work with Gregory R. Johnson, and Robert F. Murphy and is being actively prepared for journal publication.

This method first computes the distance between cell segmentations via non-rigid image registration using LDDMM. The field of non-rigid image registration has been shown to be a successful method for registering complex shapes resulting from MRI data of brains [54]. Many of the available for this deformation based transformation, including physics-based transformations and Elastic-type models generally produce reasonable transformations, but do not conserve shape topology. Because these models were intended to create reasonable cells at intermediate points along the transformation, preserving topology of the cells was required. We therefore chose to use the Advanced Normalization Tools (ANTs) family of methods that do preserve topology such as the symmetric image normalization method (SyN) [55]. These methods have well-behaved solutions with bounds on distance in deformation space and regularity and consistently perform well relative to other registration methods in the field [56]. The LDDMM method extends the classic SyN by using an "asymmetric" transformation which has been shown to further increase performance of this method [57].

Multidimensional scaling (MDS) was then used to reduce the dimensionality of the shape space. This method was chosen as it is designed to best preserve the global pair-wise distances measured by the LDDMM method in the lower dimensional embedding. MDS accomplishes this by selecting the top eigenvectors and selecting eigenvalues in descending order such that the eigenvector (and corresponding eigenvalue) explaining the greatest variance over the distance matrix is selected for each dimension 1-n. This approach is similar to kernel Principal Component Analysis (PCA) applied to distance matrices and creates a "shape space" that describes the major modes of variation in shape.

In contrast, other dimensionality reduction techniques such as locally linear embedding (LLE) or Laplacian eigenmaps preserve only local properties of the space with the claim that global layout will be preserved through this local technique as well. One advantage of these techniques is that they are more robust in the case where sparsely sampled noisy data creates "short-circuiting" in the manifold where points in different folds of a manifold are closer than neighboring points. Despite this perceived advantage however, methods such as LLE have empirically performed poorly on biomedical data possibly because these techniques tend to collapse large portions of data and perform poorly when local topology of the manifold is complex such as when holes are present making them a poor choice for the task of building generalizable models of cell shape [58].

The shape space resulting from this combination of LDDMM and MDS was then used to synthesize cell and nuclear shapes *in silico* by selecting a point that lies within the convex hull of observed cells by morphing nearby observed cells to create the novel shape at the selected point.

Although learning models of phenotypes has been useful in identifying disease, these efforts have been performed using single images of cells or tissues from an asynchronous population of cells. These methods therefore struggle to capture the dynamic interplay between cell states. In this work we propose using a set of cells of a known temporal sequence (such as a movie) to construct generative models capable of predicting the progression of cells through the set of possible shapes (and therefore cell states) and synthesizing potential cell instances along these trajectories. These capabilities allow us to study the mechanism of cell shape change under various conditions.

Unfortunately, due to limitations in current technologies it is not always possible to take movies of cells. Often cell fixation is used to prepare fluorescently labeled cells as in the case of antibody tagging. Even in the case of transfections, gene tagging or dyes which are used on live cells, there is a limit to the number of images that may be collected of a single cell before phototoxicity significantly impacts cell behavior eventually leading to apoptosis [59].

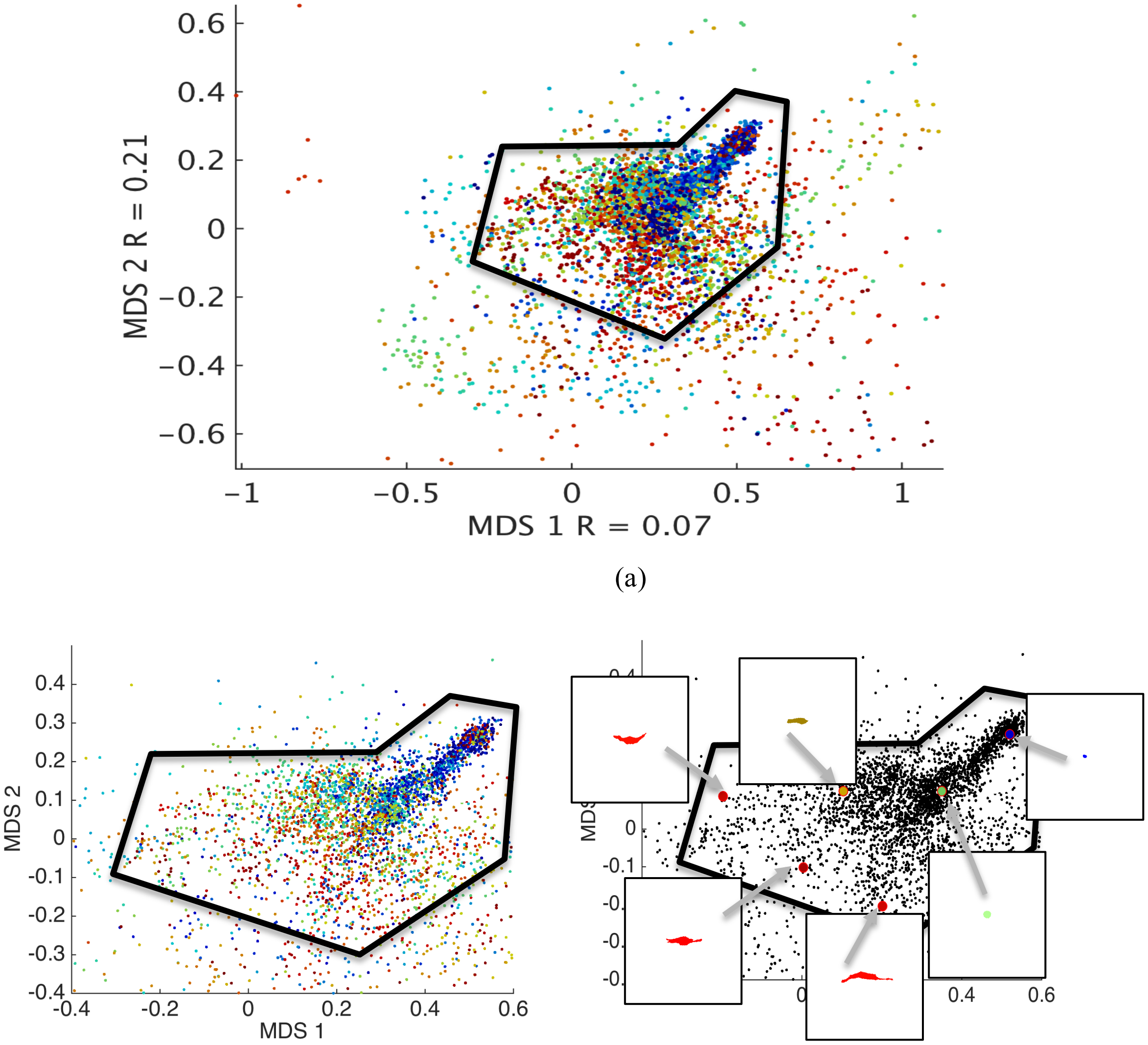
In some cases these limitations are overcome by synchronizing cells to be reasonably homogeneous with regards to their position in cell cycle [60]. Taking sets of images after the addition of external stimuli such as a drug is another way to synchronize cell response for the purposes of learning these transition models. If this perturbagen-based method is used cell transitions are simulated from temporal snap shots of cells in a population. Even in cases where cells cannot be synchronized we may be able to use external markers of temporal order such as DNA content [61,62] or FUCCI [63] markers to create a general registration of temporal order for asynchronous populations of cells and learn meaningful transition models.

2.2 Shape transition models

For this work we analyzed a set of 2D images of c2c12 mouse myoblast cells undergoing cell growth and division acquired on ZEISS Axiovert 135TV phase contrast microscope at 5X magnification over 80 hours at 5 minute intervals [64]. Each cell was segmented and tracked through the field as described in [65]. Once segmented, the movie was separated into cell "traces" consisting of a set of segmentations for a single cell between cell creation and the next detected cellular division, normalized to be a relative temporal position in the cell-trace lifespan (0-1). For this analysis we used the first 33 traces corresponding to 5067 segmented shapes from 572 total frames (approximately 48hrs). We learned a model for cell shape by measuring the distance between cell segmentations using the LDDMM approach. This distance matrix was then used to create a shape space using multidimensional scaling (MDS). We were then able to parameterize likely cell transitions by learning the distributions of transitions within the shape space resulting from the LDDMM approach applied to time-series imaging data (movies).

When fitting models we discarded any traces containing less than 20 frames as they did not contain sufficient information for training and were likely partial traces in which the cell was either lost, left the field or was created near the end of the considered frames. We also discarded traces containing greater than two standard deviations above the mean length for the set of traces because that is greater than the expected cell cycle variation and thus likely contain mis-segmentations. This resulted in discarding the 28th trace containing 429 frames, and the 8th, 9th and 17th traces containing 4,5, and 18 frames respectively. Further, individual segmentations lying outside of three standard deviations from the center of the shape space were discarded, as they were likely segmentation errors. Traces with missing frames resulting from this pruning were still considered as a single continuous trace with temporal registration adjusted according to any deletions.

The cell shape segmentations from these frames were used to construct a shape space using the LDDMM method from the CellOrganizer (v2.1) system for building statistical generative cell models [20,25,26]. By using MDS we created a two-dimensional embedding of the distance matrix to make visualization and model interpretation easy (Fig 2.1). The Pearson correlation coefficient (R) of time with each MDS component can be seen along the respective axis. For this data, cells progress primarily along the second component as indicated by the transition of color from blue to red in Fig 2.1 and the correlation coefficient for that axis.



(a)

(b)

(c)

Figure 2.1: Two-dimensional embedding of the c2c12 shape space where dots represent cells from the 4466 segmented cell shapes (a). The black box indicating the region containing the majority of the cell density is expanded in (b). This region shows a trend of cells to move downward and to the left in the shape space as indicated by the colors representing relative position in cell cycle (0-1) from blue to red. This motion through shape space corresponds to cells growing in size and elongating as they mature as seen in representative shapes selected from throughout this region shown in (c). Pearson's correlation coefficients of the first two dimensions using multidimensional scaling (MDS) with time are displayed on each axis.

Many cellular imaging data are composed of single time points of a population of cells at different points in a process of interest such as cell cycle, disease progression, or drug response. Although we cannot directly learn shape transitions using these data, in some cases these asynchronous data contain a marker that is used for temporal registration. For this work we used the LDDMM model of the set of 207 3D HeLa cells included in CellOrganizer (v2.1) to demonstrate the application of these methods to asynchronous data. These data include a marker of DNA content, which is known to increase over the course of the cell cycle. We used this temporal marker to label an LDDMM shape space trained on the joint cell and nuclear shapes of this asynchronous population with their relative location in cell cycle (Fig 2.2) where color indicates the normalized DNA content with cooler colors representing lower DNA content (earlier time points) and warmer colors representing higher DNA content (later time points). These data have significantly smaller correlation to DNA content than the correlation between c2c12 cell shape and time, shown in Fig 2.1, as indicated by the lower correlation coefficients on each axis. This low correlation may be due to noise in the DNA content measure caused by the sparse sampling of the HeLa data. Alternatively it may represent the true measure of the first two components of HeLa cell shape correlation with the cell cycle. If so these data indicate that HeLa cells vary less predictably over the cell cycle compared to c2c12 cells.

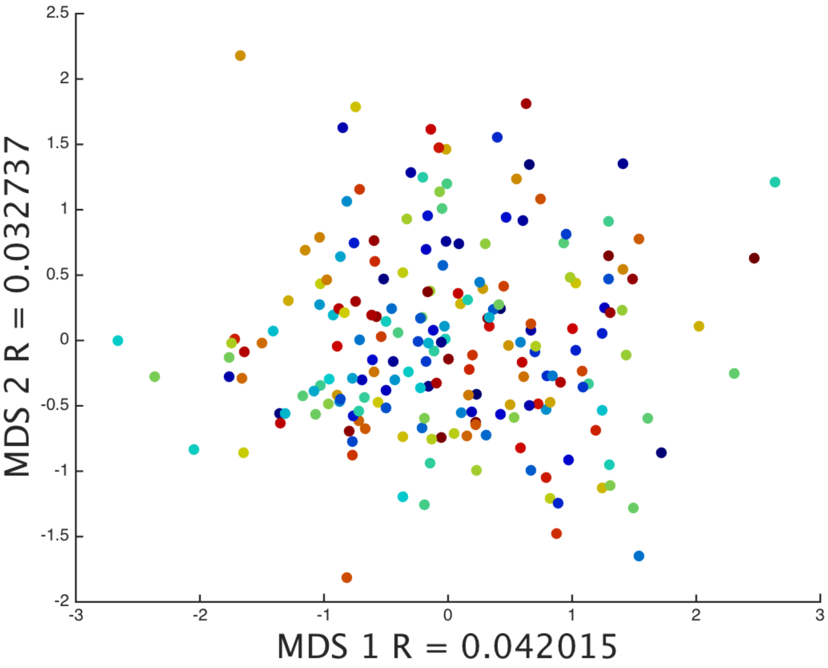


Figure 2.2: HeLa shape space colored by DNA content for 3D HeLa cells trained on the joint model of cell and nuclear shape where points represent cells and warmer colors indicate higher DNA content.

2.3 Building shape transition models

The idea of dynamic models for cell shape was first discussed in Buck et al. [25] where it was proposed that random walks in shape space could be used to generate dynamic models of cell shape. These random walks were generated by sampling a location in shape space nearby the current location with a fixed step size and a direction sampled from the uniform random distribution. As demonstrated in Fig 2.1 the c2c12 cell shape is non-random with respect to time as indicated by the correlation coefficient particularly along the second component of the MDS. This led us to pursue more sophisticated methods of cell shape transition modeling.

2.3.1 Training transition models using synchronous cell movies

The c2c12 cells in this study have a greater probability of lying in certain regions of the shape space based on their position in the cell cycle as measured by relative time since division and tend to move from top to bottom along the second MDS component and right to left along the first as indicated by the transition in color within the shape space shown in Fig 2.1. Stronger correlations between cell shape and relative life cycle position are often present for a single trace; this can be seen in Fig 2.3 where the correlation coefficient between MDS2 and time is quite high. This suggests that single cells may be more constrained than the general population. This may be due to local environment, physical constraints (crowding) or heterogeneous states occupying the same regions of our low dimensional shape space.

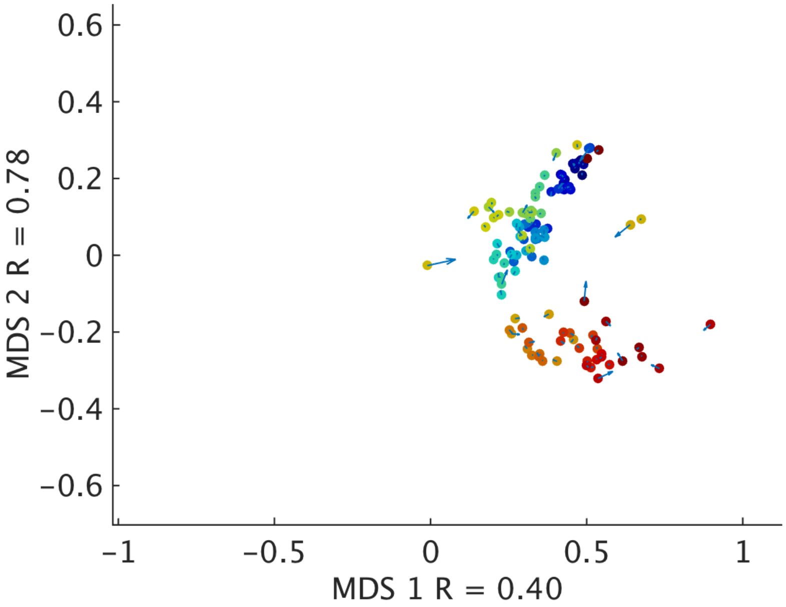


Figure 2.3: Two-dimensional embedding of the c2c12 shape space where dots represent cells from a single trace (movie) of segmented cell shapes. Colors represent relative position in cell cycle (0-1) from blue to red and Pearson's correlation coefficients of the first two dimensions using multidimensional scaling (MDS) with time are displayed on each axis. Vectors indicate direction and magnitude of cell shape transitions between a pair of sequential frames in the trace.

When provided with a synchronous movie of cell shape change, the most naïve form of cell shape prediction would be to use a semi-supervised approach in which cell trajectories are sub-sampled and the missing intermediate cell shape is predicted by a linear regression between the two observed shapes at t and t+2Δt. This model would be accurate if the cell shape changes were locally persistent along one direction of the shape space. Accuracy of this method was estimated for the set of observed c2c12 cell traces by measuring the co-linearity of three consecutive shapes in a trajectory on the shape space embedding. This co-linearity was measured by taking the magnitude of the angle between two consecutive vectors (0-180, eq 2.1).

$$[2.1] \quad A_{i,j} = a\cos\left(\frac{\Delta x_i \cdot \Delta x_j}{\left\|\Delta x_i\right\|_2 * \left\|\Delta x_j\right\|_2}\right) * \frac{180}{\pi}$$

In a perfectly persistent system, each angle would be zero and the average of angles within the trace would be zero. For the c2c12 dataset the average angle between two consecutive cell transitions was found to be 102.68 degrees with a standard deviation of 54 degrees. This is relatively close to 90 degrees that is expected when sampling the angle magnitude (0-180) from the uniform random distribution. We further analyzed each trace searching for enriched regions of linearity of response (Table 2.1). No significant enrichments were found for these data despite some cell traces having substantially more constrained total responses than others. These data indicate that for this dataset there are no strongly linear transitions in our shape space when sampling shape change every 5 minutes and therefore the naïve linear interpolation approach described above would be a poor approach for modeling these data (though it may be appropriate for other data). This is likely indicative of cells extending and retracting pseudopodia randomly, a behavior observed nonmotile cells [66]. Despite this short-time random behavior at longer time scales there is a general shift in cell shape (Fig 2.1).

Table 2.1: Linearity measurements between frame sets for 32 traces in the c2c12 dataset. Angle magnitudes are in degrees where 0 is linear, 180 is anti-linear and 90 is the expected value of uniform random sampling. Gray rows represent cell traces that were eliminated from the analysis for having less than 20 cells (8,9 and 17) or being more than two standard deviations above the mean length (28).

Trace	mean	median	std
1	102.14	107.49	54.659
2	102.47	115.27	54.347
3	108.4	122.11	51.671
4	101.49	113.65	53.413
5	98.306	111.65	57.252
6	108.43	116.61	50.821
7	106.74	119.05	53.642
8	NA	NA	NA
9	NA	NA	NA
10	100.85	106.79	53.666
11	151.37	147.31	22.018
12	114.4	130.35	54.568
13	104.99	118.89	55.3
14	100.78	105.64	56.944
15	103.24	99.337	46.291
16	101.26	104.7	51.139
17	NA	NA	NA
18	98.485	102.81	53.728
19	101.38	102.3	46.723
20	77.684	50.334	62.721
21	97.484	97.968	57.66
22	99.358	102.35	50.815
23	102.41	104.43	50.298
24	104.54	113.86	52.818
25	104.14	114.19	54.748
26	108.36	123.6	51.469
27	94.894	105.53	59.451
28	NA	NA	NA
29	113.06	126.88	52.715
30	104.23	107.52	53.148
31	104.48	113.51	56.683
32	98.866	104.72	54.58
Total	**102.68**	**110.23**	**54.017**

An alternative approach is to utilize information about transitions from nearby frames in the shape space. These frames may not necessarily be temporally near the current frame, and may be from the current trace or another trace. This approach uses a Markov-process belief that cell shape transitions are dependent on the current state of the system and not necessarily the previous states. In other words, how a cell arrived at a given state is not considered. Shape transitions using this approach are predicted by fitting a smoothed regression model over the set of observed transitions. In this work we explore several ways to parameterize regressive Markov process models of cell shape change.

For each regressive model discussed in this work we used linear locally weighted scatter plot smoothing (LOWESS) regression, as it does not require the specification of a function to fit the data. This regression model consists of a number of piecewise linear components. One drawback of this method is that it will become less accurate on the edge of the fit where data is more sparsely sampled. We were willing to incur this penalty as we expect few cells to be located on the edges of the shape space and believed the advantages of the model exceeded the drawbacks. It is important to note however optimizing the type of regression performed for each of these models may significantly improve model performance in the future. The only parameter associated with LOWESS models is the span over which the local smoothing is performed. The span for each of the LOWESS models discussed in this work minimize the mean squared error (MSE) based on 10 fold cross validation optimized using the fminsearch built-in function in MATLAB (2014b). All noise terms in each model were assumed to be normally distributed with mean zero and variance equal to the variance of the response variable for which the regression model is being sampled unless otherwise specified.

2.3.1.1 Magnitude-direction model

One way to decompose movement is in terms of the magnitude of displacement and the direction along which the displacement occurs. This model seemed favorable as one of the ways to accumulate drift of cell shape over time would be a correlation between the size of the shape change and the direction leading to a bias in the total shape change. In the magnitude-direction model we decomposed transition vectors into magnitude and angle (-180-180 degrees from x-axis) for each observed transition (frame pair) in cell shape (Fig 2.4 a-b). Surprisingly, we found the correlation between these two response variables to be insignificant (R=0.0063, Pearsons correlation coefficient, p=0.67). We therefore modeled them independently using separate LOWESS models.

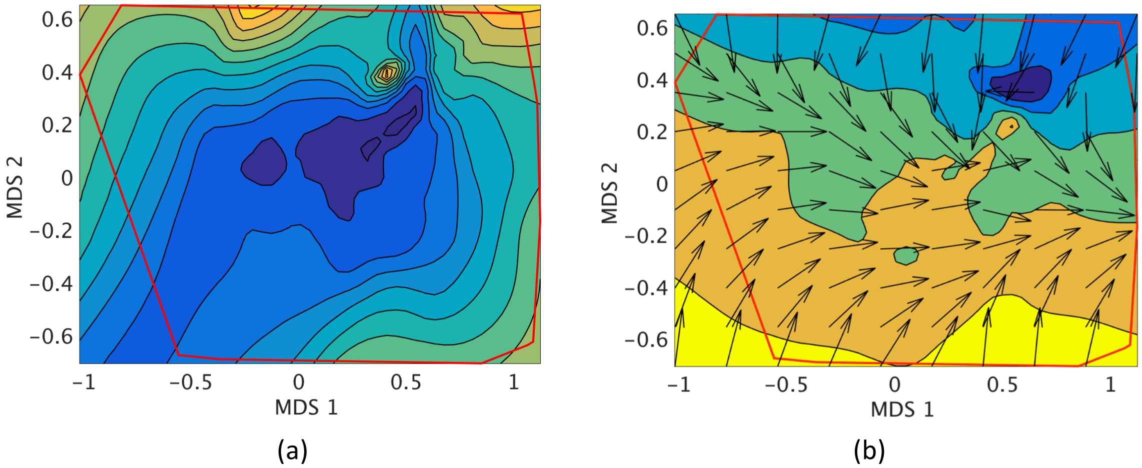


Figure 2.4: Contour plots for LOWESS regression model fits of the c2c12 shape space where warmer colors represent higher values. The magnitude model (a) predicts magnitude of shape change between consecutive frames. The direction model (b) predicts the direction of shape change where arrows indicate the corresponding direction at a point. Red boxes show the convex hull of the shape space where synthetic cells can be sampled. These fits have significance p<1e-20 as compared to a the expected fit for permuted labels.

2.3.1.2 Dimension-displacement model

An alternative method for decomposing motion is to model the displacement of adjacent frames in each dimension of the model. In this case, for two dimensions, we fit two LOWESS models to the displacements in MDS 1 and MDS 2 from Fig 2.1 respectively (Fig 2.5 a-b). Not surprisingly, there was significant correlation between the two-dimensional displacements modeled (R=0.2, p<1e-40). For simplicity, we modeled displacements independently in this initial model, though future iterations of this method will utilize a dependent structure between each dimension.

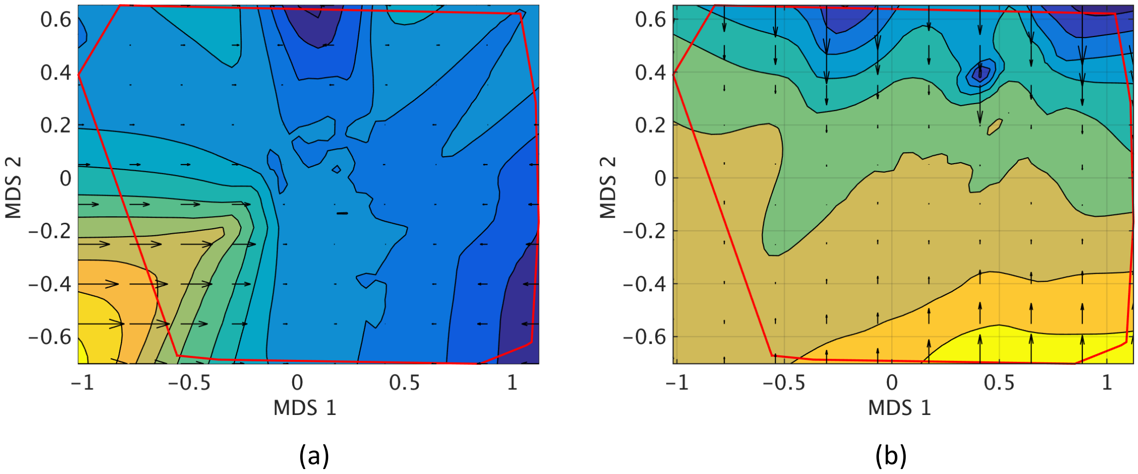


Figure 2.5: Contour plots for LOWESS regression model fits of the c2c12 shape space where warmer colors represent higher values. The models of ∂MDS 1 (a) and ∂MDS 2 (b) predict the displacement in shape space along each MDS component. Arrows indicate the relative direction and magnitude of shape change and red boxes show the convex hull of the shape space where synthetic cells can be sampled. These fits have significance p<1e-20 as compared to a the expected fit for permuted labels.

For this method, we also fit models per-trace to demonstrate the high correlation of shape change per-trace. This correlation indicates that single cells are relatively constrained in shape change as compared to shape change across the population as measured by the Pearson correlation coefficient between the per-trace fit and the observed transitions for that trace (Table 2.2). Note this correlation differs slightly from the per-component correlation seen in Fig 2.1 and Fig 2.3 as this is the correlation of the ∂MDS 1 and ∂MDS 2 with the shape space. Together with the high per-trace dimensional correlation coefficients (Fig 2.3) these data suggest that there exists strong per-trace correlations and that single cells are significantly more constrained in their shape changes than the population. This per-trace information could be used in the future to design more accurate models. This method also has the advantage that it is easily conditioned and expanded to an arbitrary number of dimensions.

Table 2.2: Pearson correlation (R) values for the change in MDS 1 and MDS 2 based on fit LOWESS dimensional displacement models. Models built per-trace obtained significantly higher correlations than the model trained over the entire population of cell transitions. Gray rows indicate data not used in building the total model. Traces 8, 9 and 17 were not fit because they had less than 20 cells.

Trace	∂MDS 1	∂MDS 2
1	0.61	0.57
2	0.43	0.49
3	0.37	0.4
4	0.47	0.59
5	0.68	0.75
6	0.89	0.74
7	0.31	0.33
8	NA	NA
9	NA	NA
10	0.4	0.44
11	0.36	0.55
12	0.83	0.72
13	0.75	0.83
14	0.48	0.36
15	0.44	0.39
16	0.67	0.69
17	NA	NA
18	0.34	0.48
19	0.83	0.71
20	0.66	0.54
21	0.5	0.38
22	0.28	0.41
23	0.46	0.52
24	0.61	0.59
25	0.59	0.53
26	0.44	0.64
27	0.33	0.41
28	0.45	0.44
29	0.39	0.37
30	0.46	0.73
31	0.53	0.59
32	0.27	0.28
Total	**0.09**	**0.14**

2.3.1.3 Gradient ascent model

In this approach a regressive (LOWESS) model is fit to the expected value of the response variable. For the c2c12 cells we modeled the relative time since division as the response variable (Fig 2.6 e). The gradient of this expected value model at a given point was used to determine the direction of cellular motion. This results in cells climbing the temporal gradient in the case of the c2c12 cells.

For the synchronous c2c12 cell movies, the magnitude of each step was modeled by the fit of transition magnitudes from the magnitude-direction model (Fig 2.4 b). A LOWESS regression model on the angle between the observed transition and the expected transition fits was used to estimate noise along the temporal gradient (Fig 2.6 b). This noise was sampled from a normal random distribution with mean of the regression model and deviation based on the standard deviation of the fit inter-vector angles.

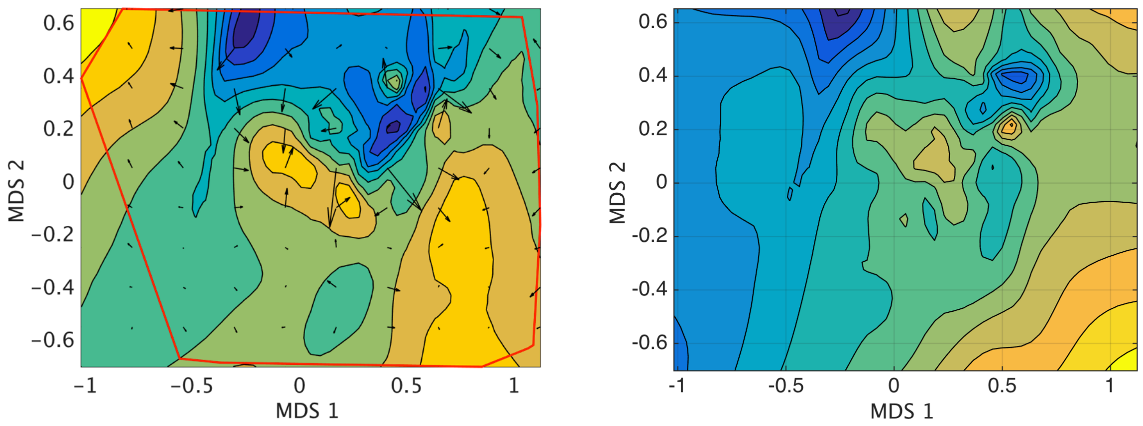


Figure 2.6: Contour plots for LOWESS regression model fits of the c2c12 shape space. The normalized cell cycle position (0-1) fit (a) predicts temporal position of cells where warmer colors indicate later time. Arrows indicate the relative direction and magnitude of the gradient at a location. The red box shows the convex hull of the shape space where synthetic cells can be sampled. Regression model of the fit angle (Fig 2.4b) to observed angle used to estimate noise at a location (b). These fits have significance p<1e-20 as compared to a the expected fit for permuted labels.

2.3.2 Training transitions models for asynchronous data

We now consider the asynchronous 3D imaging data. For this data there is no image pair information to induce vectors as shown in Fig 2.4 and Fig 2.5. The initial fit in the gradient ascent model described in section 2.3.1.3 however ignores information about frame sequence, fitting a model to the general response variable. In the case of the c2c12 data this response variable was the relative position between cell divisions. For the set of asynchronous 3D HeLa cells, DNA content can be used in the same fashion to build a LOWESS model predicting the DNA content of the cell given its shape (Fig 2.7). This fit shows a general increase in DNA content diagonally from right to left as indicated by the increase in warmer colors in the LOWESS fit contour plot in Fig 2.7. To estimate the magnitude of shape change for these asynchronous HeLa images we estimated the magnitude of the shape change to be ten times the range of the shape embedding divided number of frames in the cell cycle for a given sampling rate. This allowed us to sample at an arbitrary frame rate when synthesizing movies of cell shape change.

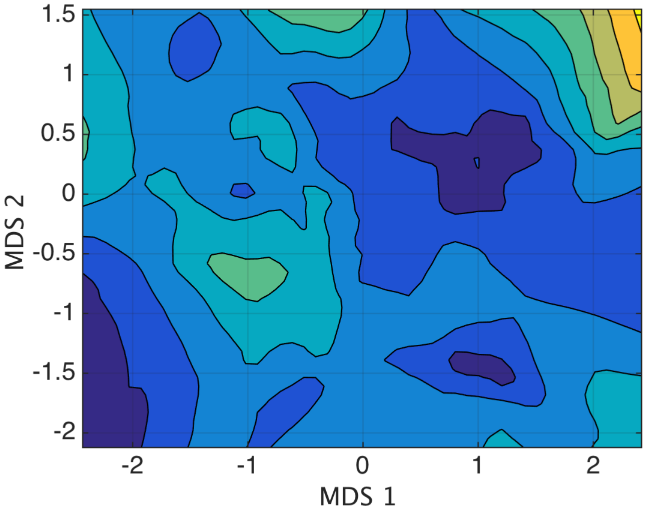


Figure 2.7: LOWESS regression model predicting DNA content trained using 207 3D HeLa cell images. Here warmer colors indicate higher predicted DNA content and therefore later time points.

For these asynchronous data there is no principled way to fit noise to the magnitude of the shape transitions. In this work we assumed the magnitude would be scaled by a normal distribution with mean 0 and standard deviation equal to the magnitude as this was empirically found to produce reasonable synthetic cell traces however future work should be done to optimize this parameter selection.

Logically, if a region of the shape space is densely populated, cells spend more time in that state and therefore should more likely travel in that direction. To achieve this we scaled magnitude by the density, multiplying by the change in density for increasing densities and dividing by the density differential when decreasing density. This scales the step size by the probability of seeing a cell at that location in shape space. The distribution of the noise in the direction of gradient ascent is assumed to be normally distributed about zero and the variance of this noise is conditional on the location in shape space as scaled by a second regressive LOWESS fit of the residual of the original model fit to DNA content such that regions with poor fits are likely to be poorly estimated by the gradient of the DNA content and thus more noisy.

2.4 Synthesizing hypothetical cell shape transitions

When building models it is important to assess the ability of the model to describe the variance within the data. For this work, there are two ways to measure whether a method successfully models cell shape transitions: measuring model residuals and measuring the ability of the model to create realistic shape transitions.

2.4.1 Measuring model fit

The significance of model fits was assessed using analysis of variance (ANOVA) tests using the sum of squares approach. First, the residual of the model was calculated by the sum of squared errors (SS_E) (eq 2.2) where the squared difference between y_i, the observed value of the response variable, and $\hat{y}_i$, the estimate from the regression at point i is summed over all observations. This is an estimate of goodness of fit for the model.

$$[2.2]\, SS_E = \sum_{i=1}^{n} \left(y_i - \hat{y} \right)^2$$

Next the regression sum of squares (SS_R) is computed (eq 2.3) to estimate the difference of the regression estimate for each point $\hat{y}_i$, and the mean of the observed response variables $\bar{y}$ corresponding to a flat model that would arise from a random assignment of labels.

$$[2.3]\, SS_R = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$

The f_0 test statistic was computed by dividing the regression sum of squares by the sum of squared errors (eq 2.4) to test the significance of the model correcting for the number of degrees of freedom (1 and n-2 respectively).

$$[2.4]\; f_0 = \frac{SS_R/_1}{SS_E/_{(n-2)}}$$

We then use the f_0 statistic to test the null hypothesis that the model is uniformly distributed, which is the expected model for a random assignment of points (eq 2.5). In this method accounting for the number of degrees of freedom is essential as it corrects for variations from the null model introduced by over fitting.

[2.5] $p = 1 - P(F \leq f_0)$

All the models discussed produced significant model fits compared to random for each component (Table 2.3). Unfortunately these p values and f_0 statistics are not directly comparable across models to determine the "best" model as the parameters of each model vary.

Table 2.3: f_0 values for models fit by each of the methods discussed in section 2.3. Higher f_0 values correspond to greater significance over random (f_0=0.5). The p values indicate significance of LOWESS regressive fit over the expected fit of randomly permuted labels.

Method	**Response variable**	**f_o value**	**p value**	**Dataset**
Gradient descent	Time	948	p<1e-20	c2c12
Magnitude-direction	Magnitude	1075	p<1e-20	
	Direction	285	p<1e-20	
Dimension displacement	∂MDS 1	453	p<1e-20	
	∂MDS 2	715	p<1e-20	
Gradient ascent	DNA content	26.9	p=5.2e-7	3D HeLa

2.4.2 Evaluating synthetic cell movies

Using the models trained for cell shape transitions described in section 2.3 for both synchronous and asynchronous cells we synthesized *in silico* cell shape transitions. At each time point, the model of shape transition is sampled to select a new location within the shape space. This process is repeated for the number of desired frames in the synthetic cell shape movie. Sampled locations correspond to potential cell and/or nuclear shapes that are synthesized by interpolating cells in the current Delaunay triangle of experimentally observed cells as described by Rhode et al. [24].

Here we examined each proposed model's abilities to generate synthetic cell shape change in the shape space by performing leave-one-out cross validation (LOOCV) where a single cell trace from the set of 33 traces was held out of the data used to train the dynamic models. The starting point for the held out trace was used to seed the synthetic cell trace and frames were synthesized for the number of frames contained in the held out trace. For each held out frame the mean squared error (MSE) defined by the distance between the held out frame's location and the corresponding synthetic frame's location in shape space was measured (Table 2.4). These regressive models are deterministic when sampled purely from the fits; however, by introducing a noise term as discussed in section 2.3 we created a probabilistic model. We therefore generated 100 synthetic cell traces for each held out trace to account for the probabilistic variance in the models. Accuracy was measured as the average mean squared error across all frames for all synthetic cell traces in the leave-one-out cross validation. The accuracies for the four approaches applied to the c2c12 data are seen in Table 2.4 below. In the future, the MDS over the shape space should be performed prior to training the dynamic model for each set of training data. In this case however comparing MSE across traces must be done more carefully as the size of the constraining convex hull will change depending on the MDS performed when holding out a trace. This may also result in parts of the held out trace lying outside the convex hull of the shape space for the observed cells. Here we used a single shape space construction to allow us to directly combine results from the set of held out traces.

Table 2.4: Average accuracy for synthetic cell traces from LOOCV using synchronous models of c2c12 mouse myoblast cell traces where each of the 28 traces was held out in turn. One hundred random seeds were used to create one hundred synthetic traces for each of the 28 folds seeded from the initial frame in the held out experimentally observed cell trace (2800 synthetic traces total). Mean squared error (MSE) is measured as the distance in the 2D MDS embedding of the shape space. A trace is considered to fail if it was not able to generate a subsequent point in a trace after 100 attempts. Partial traces resulting from failures were not considered in the MSE calculation.

Method	Mean squared error	%Success
Random	0.964	97.0
Gradient ascent	0.259	72.1
Magnitude - direction	0.274	97.0
Dimensional displacement	0.128	97.0

In the case of the “magnitude – direction” and “dimensional displacement” models we were able to directly sample from the fit models to obtain the sequence of points. This is in contrast to the “gradient ascent” approach which does not directly sample the cell transition but uses the gradient of the model of the response variable (time or DNA content) to choose a direction in which to change shape. Magnitude and noise are then sampled as discussed in section 2.3.

Shapes can only be synthesized inside the convex hull of the experimentally observed shapes. As a result, occasionally a synthetic walk will become stuck in a region of the shape space and be unable to sample another step. In cases where a point outside the convex hull of observed shapes is sampled that point is rejected and the point is resampled from the model (with noise). If a valid point cannot be found in 100 samples, the method will return an abbreviated walk and notify the user that it was unable to generate the whole walk. For the accuracy assessment in Table 2.4 we considered the inability to generate the full length of a walk to be a failure since we know that at least one walk exists that is of the full length (the walk we are attempting to imitate). If a synthetic trace failed to generate the full length of the requested trace it was eliminated from the accuracy calculation so as not to bias the analysis based on results from early successful frames that are more likely to be close to the corresponding observed frame.

Note that the gradient method has a proportionally larger likelihood of failing than other methods. This is due to the fact that the direction of the model is largely determined by the gradient and the gradient on one edge of the model is strongly divergent causing the model to attempt to exit the convex hull frequently. If the user is willing to accept abbreviated runs, or can afford to restart the sampling to achieve a more realistic run, this method may be acceptable, however the modeler should be aware of this potential draw back when selecting a model. The transition-based models will tend towards the more populated regions of shape space by construction since the cell transition of a cell at the edge of the shape space must be back into the shape space. This leads to a much lower rate of failure for these models.

As seen in Table 2.4, the model with the lowest mean squared error for the synchronous models of c2c12 cells is dimensional displacement method. An example synthetic trace can be seen for this model in Fig 2.8 below. This cell trace exhibits behavior that is qualitatively reasonable covering a reasonable fraction of the shape space and a noisy movement from upper right to bottom left with local transitions appearing reasonably random as with the original data.

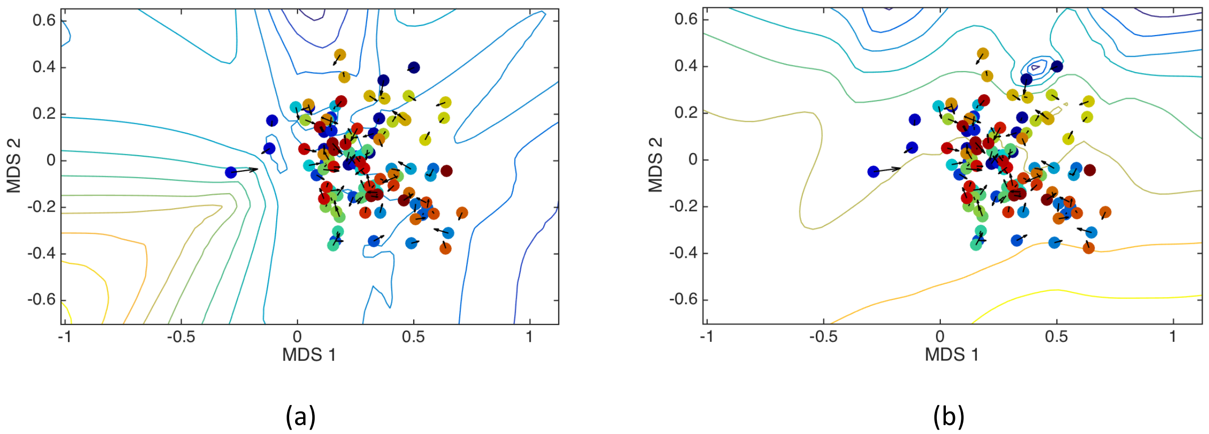


Figure 2.8: Example synthetic cell trace of 100 frames generated from the dimensional-displacement model of the c2c12 shape space super imposed on the fits of ∂MDS 1 and ∂MDS 2 are shown in (a) and (b) respectively. Colors of points indicate relative time from blue to red while vectors represent direction and magnitude of shape change.

For asynchronous shapes, the gradient ascent method was used for the 3D HeLa data. HeLa cells divide approximately every 22 hours, and nuclear DNA content peaks around 15 hours [61]. We therefore set the mean magnitude of the cell shape change to be the range of the shape space divided by the number of frames in a cell cycle assuming a sampling of 5 minutes (264 frames) times 10. This resulted in a step size of 0.161 in the MDS space. We then generated synthetic movies for 180 frames (15hrs) as that is the estimated length of time for the maximum of nuclear DNA synthesis [61].

This model produces plausible shape transitions (Fig 2.9), as the cell traverses a reasonable fraction of the shape space over the course of one "cell cycle", and different regions of time (colors) in the synthetic cell trace proceeds in a noisy but directed fashion towards the region of high DNA content near [-1,-0.5]. Unfortunately, the accuracies of these traces cannot be directly estimated without further experimental data. Currently, only about 1 cell trace worth of frames (207) have been sampled for this shape space, therefore more densely sampling the shape space will lead to improved models for these data and an improved understanding of the proper model parameters.

To quantitatively assess the accuracy of these asynchronous models and further optimize the parameter selection methods the step size and noise parameters need to be selected such that on the average, a large set of synthetic traces recapitulates the distribution of cells throughout the shape space. The model fit on the DNA content and in the naïve case where no noise is present does not give this distribution of cells due to tendency of gradient ascent to converge to local maxima. This issue is accentuated by the maxima and minima induced by outliers on the edge of the shape space.

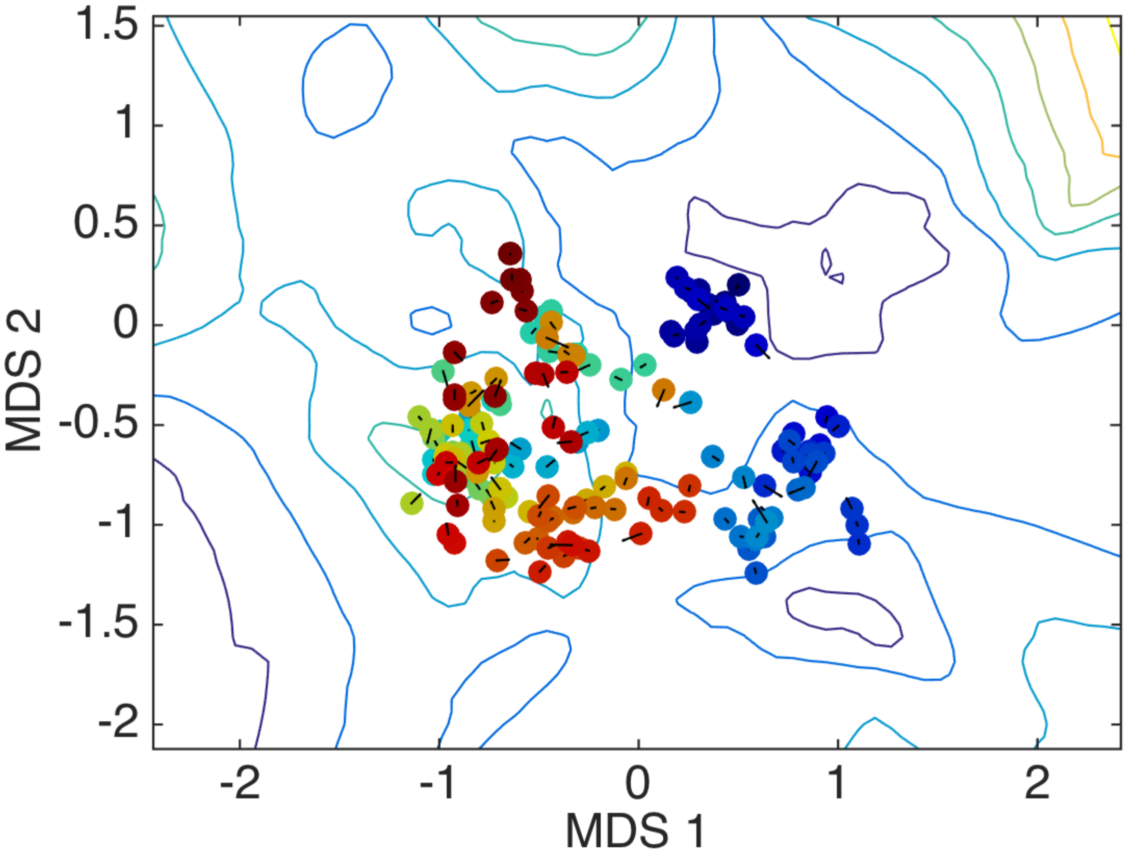


Figure 2.9: Example synthetic cell trace of 180 frames generated from the gradient ascent model of the HeLa shape space super imposed on the fit of DNA content. Colors of points (cells) indicate relative time from blue to red while vectors represent direction and magnitude of shape change.

2.5 Discussion

This work enables the training and synthesis of dynamic generative models. In this chapter we present four methods for learning cell shape transitions and generating probabilistic walks through a shape space using a Markov process. We further built models using various types of markers for temporal registration demonstrating how future users may model a response variable of their choice.

We applied these methods to a set of synchronous cell images from a movie of c2c12 cells as well as 3D asynchronous data of HeLa cells. For both datasets we showed that our models are statistically more likely than random sampling and for our synchronous data we demonstrated that the models presented outperform a random sampling of the data when generating synthetic movies of cells. These models are still very new and can certainly be improved upon however these efforts demonstrate that building generative regressive models of cell and nuclear shape change is a promising approach to creating accurate dynamic generative models.

The best model for the c2c12 data set was the dimensional displacement model. In future works we will use conditional sampling of dimensional displacements and extend the model to higher dimensions to improve this modeling framework. These higher dimensional models will provide increased accuracy, as they will account for more variance in cell shape. As seen in Fig 2.10, for the c2c12 data four components can be used to describe the vast majority of the cell shape variance though this cut-off is dependent on the cell type and conditions for which the shape space is built. We propose that future work utilizes an automated cutoff as a percentage of the variance accounted for in the shape space to determine the number of components. The methods outlined in this work are easily extended to perform similar modeling in an arbitrary number of dimensions.

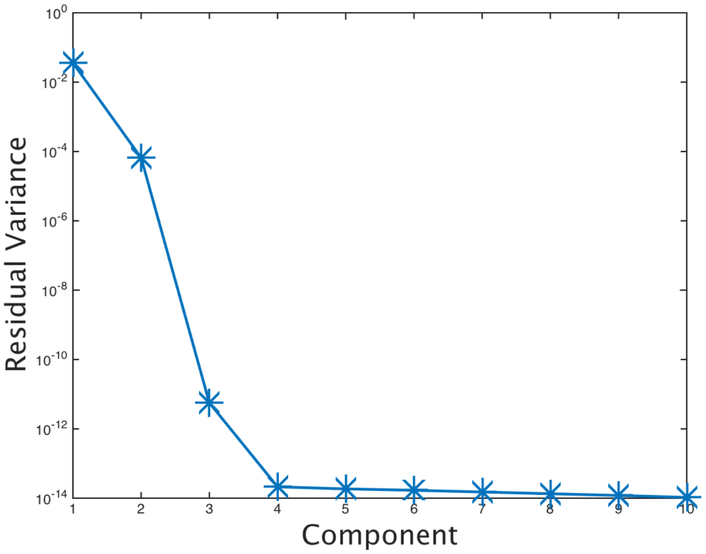


Figure 2.10: Residual variance of the distance reconstruction as a function of the number of dimensions (components) in the reconstruction for the LDDMM-MDS approach included in CellOrganizerv2.1.

Selection of regression method, and improved modeling of noise are other ways in which these methods could be improved. It is important to note that although the dimensional displacement model had the lowest mean squared displacement for these data, it is not guaranteed to be the best model for all cell transitions. As such the further development of each method will be pursued.

Further improvements to these approaches may be possible by taking advantage of per-trace constraints that produce highly accurate models. One area that merits further exploration is whether there are multiple categories of cell traces from the same population indicative of modes of cell behaviors in a single population. In the case of the c2c12 data this may be due to the time spent in different stages of the cell cycle. Developing methods to recognize these modes in cell behavior may lead to novel understanding of heterogeneous populations such as those found in tumors.

This dynamic modeling approach could also be applied to other generative parameters within CellOrganizer such as the number of endosomes in the cell upon addition of EGF. Ultimately the goal is to create conditional models that describe the response of multiple subcellular compartments and proteins together. This conditional modeling framework allows for the exploration of how compartments or proteins interact and whether some compartments, such as nuclear shape more highly correlate with response variables than others such as cell shape.

Applying these dynamic models to study cell transitions between states of cellular response to disease and treatments will provide novel understanding of disease mechanisms. Synthetic movies of cells applied in systems modeling frameworks, will further increase our ability to understand these dynamics.

2.6 Additional required work

Some steps remain to prepare this work for publication; optimize asynchronous parameter selection and perform accuracy assessment, evaluate the accuracy of the asynchronous parameterization on the c2c12 data including sub-sampling the c2c12 data to contain a similar number of cell segmentations as the HeLa data, and rebuild shape spaces for each held out trace in LOOCV when measuring accuracy.

In a subsequent publication we will demonstrate extension of the dimensional displacement model to arbitrary MDS dimensions, develop a conditional formulation for the dimensional displacement model, build models at varying time scales and demonstrate the relationships between these models, and build hierarchical models using both time-varying and single-trace information to constrain the models.

Chapter 3. High-throughput spatially resolved modeling with realistic cellular geometries[3]

3.1 Introduction

The rapidly growing field of systems biology strives to characterize biological systems and their behaviors under various conditions by leveraging computational methods. These approaches can be used to study a variety of scales from populations, to organisms, down to the cellular and sub-cellular level.

At the cellular and sub-cellular levels systems, models are built using results from a variety of biochemical experiments including imaging, sequencing, and immunoprecipitation. In the last 15 years, many technological advances have allowed for the generation of larger and richer datasets than ever before. Currently, the majority of biochemical simulations assume a population is well mixed and therefore has a uniform concentration within a given reaction compartment. We will refer to these methods as homogeneous approaches. Examples of these include systems of ordinary differential equations (ODEs) and the Gillespie method for stochastic simulations. These methods are generally appropriate when investigating average population behaviors from a set of cells or organisms, when the number of molecules is high or when the reagents in the system are believed to be well mixed. The key advantage of these approaches is that they are computationally inexpensive and therefore allow for various forms of analysis including bifurcation analysis, parameter estimation from data, and spatial parameter sweeps [32].

[3] This chapter describes joint work with Jose-Juan Tapia, James R. Faeder and Robert F. Murphy and is being actively prepared for journal publication.

It is well known however that cells are highly spatially heterogenous containing various membrane bound organelles, protein aggregates (such as P-bodies) and chemical gradients such as those involved in chemotaxis. All of these forms of subcellular organization have key impacts on cellular behavior [67]. Often in systems with such high spatial diversity, assuming a well-mixed system will lead to severely inaccurate simulation results and in some cases completely divergent behaviors [68-72]. As such, to accurately understand the behavior of these subcellular systems, disorders in them and how to treat such disorders, it is sometimes necessary to use spatially resolved approaches such as particle-based simulations or partial differential equations (PDEs) [35,36,40,41].

Cellular behaviors that are known to be correlated with spatial organization include endocytosis in response to EGF or other signaling molecules [73,74], cellular motility [3,75], and cell fate [1,2,5]. A recent study by Yin et al. [76] isolated genes responsible for drastic changes in cellular morphology demonstrating that changes in a small number genes can drastically impact cell shape and subsequently behavior. Inversely, work by Gabella et al. [77] suggests that cell contact angle controls cell motility, suggesting that spatial structure is also directly impacting cellular dynamics. These studies have clearly shown that cells respond to various conditions resulting in changes in cell shape, nuclear shape and organization, and that these changes are correlated with changes in cellular behavior.

To better analyze the question of how these spatial organizations impact cellular behavior *directly* spatially resolved *in silico* models represent an alternative to experiment. Such *in silico* simulations have the advantage that they allow the scientist to control for all conditions of the system of study isolating these spatial regulatory systems. Some efforts have been made to use such systems to study the direct impact of spatial organization on cellular behavior, however, the majority of such studies have been limited to a small number of hand drawn or hand segmented geometries out of necessity[78].

One approach to obtaining realistic geometries is to use automated segmentation algorithms to segment realistic geometries for use in simulation [26,79-82]. One such study, Sbalzarini et al. [80], demonstrated that the diffusion of surface proteins on the ER was directly dependent on the complexity of the endoplasmic reticulum (ER) network in the cell by simulating recovery after photobleaching under various amounts of complexity and fitting learning parameters that recapitulated experimentally observed recovery rates. Another example by Slepchenko & Loew [81] used automatically segmented neuronal morphologies to study calcium dependent dynamics and fit model parameters. Both of these studies demonstrated the utility of these automated segmentations in simulation for learning parameters from data and showed that the complex morphologies of cells and organelles play an integral role in determining cellular behavior.

Though automated segmentations can provide high quality segmentations they often require tuning for a dataset limiting their applicability across large amounts of data. Further, this method and manual segmentation are limited to the morphologies observed experimentally. These images are both technically difficult and expensive to collect limiting the number of high-quality realistic geometries available using this method. In addition, the use of these automated segmentations restricts one to the set of experimentally observed morphologies and organizations. In many cases it may be desirable to study specific instances of morphology and organization for the purposes of understanding rare events. These spatial scenarios may be sparsely populated or unpopulated in the experimental space.

Recent efforts in the development of generative models capable of generating realistic *in silico* geometries for cells and their organelles provide a promising solution to this problem as large numbers of realistic *in silico* cellular instances can be generated with little to no manual effort [19-22,24,25,83]. The open source CellOrganizer system (http://cellorganizer.org) incorporates many of these approaches.

These models are learned by fitting the statistical distribution of spatial organizations from realistic segmentations of experimentally observed cells. This approach provides a means of analyzing the distribution of spatial organizations present in a given set of experiments and sampling realistic instances for study specific spatial organization scenarios. In this work we combine these statistical generative models, learned from fluorescent microscope images, with biochemical models to create a high throughput computational pipeline for spatially resolved simulations of biochemistry in realistic geometries capable of learning the dependence of cellular response on spatial parameters of the cell.

Once *in silico* spatial geometries and biochemistry are obtained a choice of spatially resolved simulation method must be made. Partial differential equations (PDEs) may be used to perform spatially resolved simulations by representing chemical species as a continuous gradient within a cell or a compartment. These methods are used by simulation engines such as the Virtual Cell (VCell) and can be performed in a deterministic or stochastic manner [35,36]. Brownian dynamic simulators are a class of stochastic simulation tools that include Monte Carlo Cell (MCell) [40] and Smoldyn [37] . These tools utilize continuous space with discrete time-steps. One limitation of these methods is that they may miss potential interactions when the time-step is too large [84]. This can cause inaccurate and even divergent simulation results. Exact stochastic simulation algorithms that utilize a spatial formulation of the Gillespie algorithm may also be used to simulate spatially resolved biochemical networks. These algorithms have traditionally been too computationally intensive to perform at a large scale as the computational cost scales linearly with the number of reactions in a network. Recent work by Ramaswamy et al. [85] has improved the efficiency of these approaches such that these exact simulations can be performed with complexity that at worst scales linearly with the number of species in a reaction network. This number is generally much lower than the number of reactions, enabling the use of these exact methods for larger systems.

3.2 Creating a high throughput modeling pipeline

Using generative models from CellOrganizer trained on fluorescent microscope images we are able to easily create a nearly unlimited number of realistic synthetic cells *in silico* [19-22,24,25,83]. In this chapter I describe creating a pipeline (Fig 3.1) that combines these geometries with biochemical models to perform high-throughput simulations in a spatially resolved manner at the sub-cellular. An important advantage to using generative models in CellOrganizer is that we can sample from specific areas in the set of all possible cells and subcellular organizations describable by these models. This allows us to investigate specific shape and organization changes that may be associated with a particular diseased state, or perform general parameter sweeps for a specific components of cellular organization such as number of endosomes or cellular morphology as demonstrated in this study.

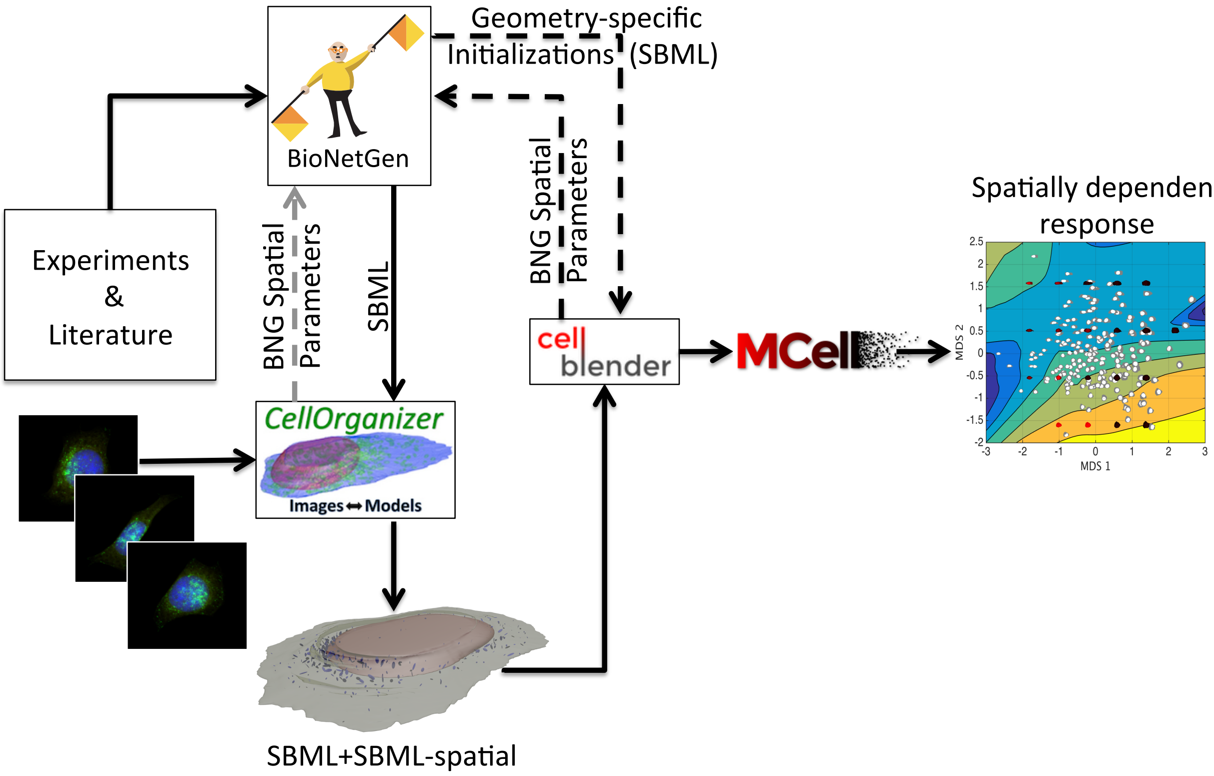


Figure 3.1: The high throughput spatial modeling pipeline uses BioNetGen (BNG) to create biochemical systems models. These biochemical models are then fed to the CellOrganizer system using the Systems Biology Markup Language (SBML). CellOrganizer uses statistical generative models of cellular compartments trained with fluorescent microscope images to generate spatial instances for the necessary compartments in the model using the SBML-spatial format to transfer spatial instances into CellBlender/MCell. These instances can optionally be used to create BNG parameter files for the surface area and volume of each compartment using either CellBlender or CellOrganizer (dashed lines). By combining the SBML-spatial instances with a geometry-specific initializations in CellBlender we can generate and run a large number of instances using MCell and study geometry-dependent responses for a given biochemical system. Here black lines indicate paths utilized in this work, gray lines indicate paths not utilized in this analysis and dashed lines indicate optional paths within the pipeline.

One method for modeling cells available in CellOrganizer uses a large deformation diffeomorphic metric mapping (LDDMM) [24,25,83]. In this approach, cells are first aligned using non-rigid image registration and then morphed using interpolation to measure the relative distance between shapes. These distances can then be projected into a low-dimensional space referred to as a “shape space”. Cells can then be generated *in silico* by interpolating between nearby cells. In this work, we constructed a shape space of 3D images of HeLa cells included in the CellOrganizer (v2.1) distribution using the aforementioned LDDMM approach. Because this approach uses an embedding that separates cells along their axis of greatest variance, it is a natural choice for studying the impact of cellular shape variance.

We combine these generative model instances with biochemical models written in BioNetGen using the Systems Biology Markup Language (SBML) and its spatial extension (SBML-spatial) to generate a large set of spatially resolved simulations that allow us to study the dependence of cellular response on cell shape and organization, something we believe was not possible prior to the creation of this pipeline.

3.2.1 Biochemical modeling

To demonstrate the utility of this pipeline we used the model signal transduction system presented by Harris et al. [86]. A schematic of the system is shown in Fig 3.2 below. This system was chosen because it contains several geometric compartments interacting via a somewhat complex system of 354 reactions. The model was a generalized version of a signal transduction network. For this work we modified the initial concentrations based on experimentally observed concentrations for the EGF-EGFR system (Table 3.1) to obtain a realistic number of molecules found using BioNumbers [87]. Reaction rates were left at the generalized values presented by Harris et al. and the results of this work are meant to demonstrate the utility of the computational pipeline (Fig 3.1) for spatial modeling of biochemical systems.

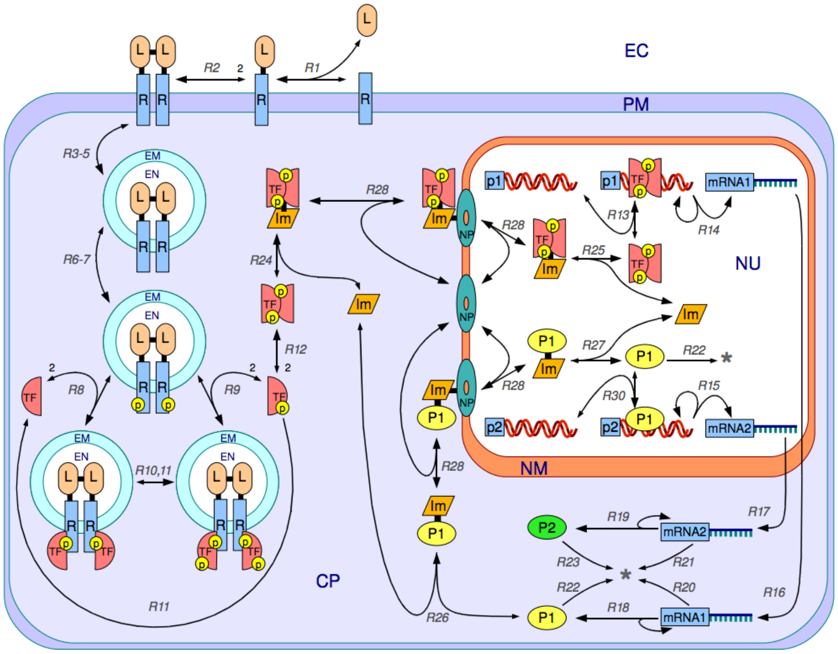


Figure 3.2: Reaction network diagram as first appeared in Harris et al. [86] of the model signal transduction network used to demonstrate high throughput spatial modeling.

In this model, receptors (R) on the plasma membrane (PM) react with ligands (L) in the extracellular space (EC) to form a receptor-ligand dimer (R1-2). This dimer is then internalized to the endosomal membrane (EM) through transport reactions (R3-5). Once internalized, the receptors are phosphorylated through a 0^{th} order reaction in the cytoplasm (CP) (R6-7). The phosphorylated receptor binds with transcription factors (TF) (R8), which are then also phosphorylated (R10-11). Phosphorylated TF is then released into the CP (R9) where it freely diffuses and forms a dimer (R12). This dimer is bound by importin (Im) (R24) and binds to nuclear pores (NP) (R28) at the nuclear membrane (NM). The complex is imported into the nucleus (NU) (R28) where the Im unbinds (R25) and the TF dimer binds DNA and transcribes mRNA1 (R13,R14). This mRNA1 then diffuses to the CP where it translates protein1 (P1) (R16,R20,R22,R18). P1 is then bound by Im and imported into the NU where it transcribes mRNA2 (R26-30,R15). Lastly this mRNA2 diffuses to the CP and translates protein2 (P2) (R17,R19,R21,R23).

Table 3.1: Initial concentrations for seed species used for high-throughput modeling in the signal transduction network first presented by [86].

Species	Value	Units	Citation
Ligand	9.00e-9	M	[88]
Receptors	7.86e-23	mol/(μm^2)	[89]
Transcription Factor	1.55e-8	M	[90]
Nuclear pore	1.23e-23	mol/(μm^2)	[91]
Importin	1.70e-8	M	[92]

3.2.2 Geometric modeling

Using the joint cell and nuclear LDDMM shape space of 3D HeLa cells included in CellOrganizer (v2.1), we generated a library of cell and nuclear geometries which we refer to as cellular "frameworks" sampled from a grid on the 2D multidimensional scaling (MDS) embedding of the shape space (Fig 3.3) [24,25,83].

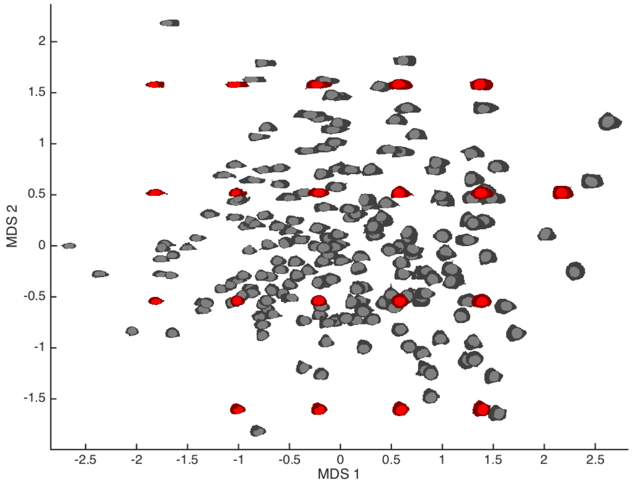


Figure 3.3: A two-dimensional embedding of the shape space constructed using the LDDMM approach in CellOrganizers (v2.1) for 3D fluorescent microscopy images of HeLa cells. Sampled synthetic cells are shown in red.

These cells were then filled with endosomes using the endosomal model provided in CellOrganizer (v2.1) trained using images of fluorescently labeled transferrin receptor (TfR) in HeLa cells [19]. Note that the TfR protein expression was not used in the study of EGF-EGFR dynamics here but purely to determine the spatial organization of endosomes within the observed HeLa cells. The number of endosomes was fixed to the mean of the cumulative distribution function (CDF) of the log-normal distribution of endosomal frequency learned by CellOrganizer (v2.1) to control fluxuations in dynamics introduced by varying the number of endosomes within the cell. This resulted in 481 endosomes per cell.

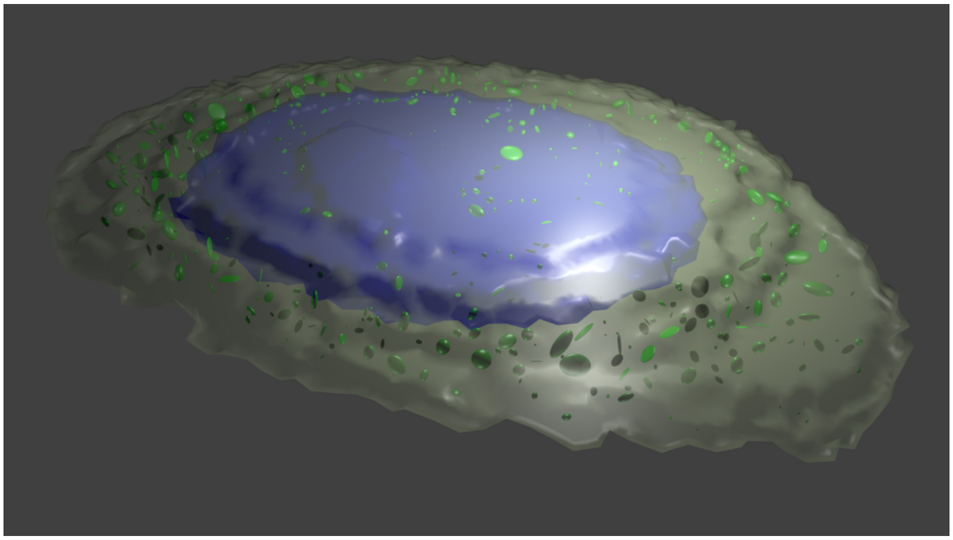


Figure 3.4: Example 3D cellular instance containing 481 endosomes used for simulations. Here the nucleus is shown in blue, endosomes based on GFP tagged TfR are shown in green and the cytoplasm is shown in yellow.

To analyze the impact of the random size, and endosome placement selections made by CellOrganizer, 481 endosomes were simulated for two random seeds for a selected cellular framework (cell and nuclear instance). When using this pipeline to study physiological models and build conclusions about cellular dynamics we strongly recommend that this analysis be done for at least a handful of geometric seeds for each cell framework. This ensures that one area of the shape space is not more highly sensitive to these fluctuations than others. Due to the high computational cost of performing these added simulations we limit this analysis to a single cell framework for this work.

To demonstrate the utility of the pipeline to assess the impact of another spatial parameter, the number of endosomes synthesized was varied within a given cellular framework. Endosome numbers were selected to be the mean of the distribution of number of endosomes learned using CellOrganizer, the 0.1 ("low") and 0.9 ("high") levels of the cumulative distribution function of the distribution corresponding to the minimum and maximum endosomal frequencies for the model allowed by CellOrganizer on synthesis. These three endosomal frequencies correspond to 222, 481, and 808 endosomes. It is important to note that the change in the number of endosomes did not impact the internalization rate for the receptor-ligand complex in this study as we wished to isolate geometric contributions. However, this internalization rate would be more realistically modeled to be proportional to the number of endosomes.

3.2.3 Computational automation

We developed a high-throughput pipeline for using the synthetic cells generated with CellOrganizer to create spatially resolved biochemical simulations (Fig 3.1). This pipeline allows the user to sample in a targeted manner from the distribution of possible cell shapes and organizations to probe certain conditions and spatial changes directly while controlling other variables. Combining these *in silico* instances with biochemical models of cellular behavior we can simulate systems of biochemical reactions and analyze the direct impact shape and spatial variance have on system behavior. To create our biochemical models we chose to use BioNetGenv2.2.6 to take advantage of the compact rule-based modeling system, however the computational pipeline presented here can be used with any pre-existing SBML formatted compartmental biochemical network.

Biochemical models were moved between BioNetGen, CellOrganizer, and CellBlender using the Systems Biology Markup Language (SBML) and making use of libSBMLv5.10 [93]. This standard is widely used in systems biology, however it previously lacked the ability to describe spatial instances (geometries).

3.2.3.1 SBML-spatial development

Transportation of biochemical models between software tools has been standardized using SBML, however there remains no standardized way for transporting spatial instances. The increasing use of spatial modeling has lead to the demand for such a tool. As a result, an extension to SBML to communicate spatial instances was first proposed by Jim Schaff (http://sbml.org/images/1/15/SBMLSpatialProposal_2011_04_20.pdf.). Over the past 3 years there have been efforts to improve this proposal to create a specification capable of characterizing a large array of spatial instances. Despite this proposal there remained only two-dimensional circle-in-circle constructed solid geometry (CSG) instances of SBML-spatial created by Frank Bergmann that were out of date and no longer matched the specification. In this work we implemented the first instances of three-dimensional SBML-spatial. In doing so we developed standards files for CSG and "Parametric" (mesh) geometries that can be used for testing SBML-spatial parsers. This work raised several previously unseen issues with the SBML-spatial specification leading to modifications of the specification. Most notably we contributed the novel class of geometry definition allowing the user to create "Mixed" geometries containing instances from any of the other geometry classes. This contribution allowed for increased compactness in communicating these models in cases where parts of a geometric instance such as a cell membrane are morphologically complex and thus described well with a Parametric geometry while other parts of the geometric instance are better described by other methods such as our ellipsoidal endosomes which are easily described using a CSG. We then generated a standard file for this new class of geometry as well. These contributions are reflected in the updated draft of the SBML-spatial specification (SBML-spatialv0.88,v0.89).

In addition to designing simple standard files, we also created a MATLAB based code within CellOrganizer (v2.1) to generate SBML-spatial instances in accordance with v0.89 of the SBML-spatial specification for a set of geometries. This SBML-spatial writer was the first tool capable of producing such files and was linked into generative model synthesis in CellOrganizer (v2.1) allowing for the automated creation of SBML-spatial instances. This tool is capable of writing SBML-spatial files describing instances of nuclear shape, cell shape, and vesicular organelles using the Mixed geometry class. In addition, this tool may be used to create SBML-spatial instances from any provided segmentations using Parametric (mesh) geometries.

We then designed and implemented a python interpreter to import these SBML-spatial instances into CellBlender, the interface for MCell. This was accomplished by writing a python plug-in packaged within Cellblender and available through the CellBlender code repository (code.google.com/p/cellblender). Once imported to Blender these instances were translated into MCell simulation files using CellBlender.

To further automate the process of spatial instance generation we also developed a method to read an SBML biochemical model and identify the necessary “compartments” corresponding to organelles within the cell. This method then searches the CellOrganizer model database for models of the required compartments. If all the necessary compartment models are found we automatically generate geometries for the required compartments and append the resulting SBML-spatial instances to the existing input SBML to create fully specified spatial models for the given system. This allows a user with a pre-existing SBML model to rapidly create SBML-spatial instances containing the biochemistry of their system.

3.2.3.2 Model initialization

To initialize the number of molecules present in the MCell simulation, a geometry specific biochemical simulation file was generated in BioNetGen using the spatial parameters (compartment surface area and volumes) of a particular geometric instance. These BioNetGen “spatial parameter” files were created by importing geometries into CellBlender and exporting the resulting mesh analysis for volume and surface area in a BioNetGen formatted text file. CellOrganizer can also be used to create these spatial parameter files, however due to slight differences between the MCell calculations and CellOrganizer calculations on meshes, we chose to use the MCell calculation to ensure initial concentrations were accurate. This allowed for the rapid feedback to the biochemical model for compartment volume, and surface area as BioNetGen cannot yet parse SBML-spatial files. This is particularly useful when “priming” the simulation where a spatially resolved simulation may be started at some time after the system initialization based on results from the homogeneous simulations within BioNetGen. If no priming is required and the model contains concentrations instead of molecule counts, this feed back, shown with dashed lines in Fig 3.1, can be avoided entirely. In this case the SBML+SBML-spatial file generated by CellOrganizer can be used to generate the MCell simulation directly.

3.2.4 Simulation Settings

The simulations presented in this work were run using Monte Carlo Cell (MCell v3.2.1), a stochastic particle based modeling system[40,41]. CellBlender, a plugin to the Blender rendering program written in python was used as a preprocessor for MCell to import both the biochemistry from the SBML file and the spatial instance from the SBML-spatial using the importer described in section 3.2.3 above. Currently CellBlender is the only software to our knowledge capable of importing SBML-spatialv0.89 instances using the aforementioned importer, however VCell is currently working on implementation of the updated specification through libSBML. To assist us in setting up MCell simulations for each of our 20 geometric and cell-specific biochemical simulations we created an external python script to call Blender as a python application to set CellBlender parameters specific to our biochemical system and desired simulation settings.

Partitions in MCell divide the reaction volume into subvolumes using a set of planes. This allows MCell to consider interactions within these subvolumes in parallel. Use of appropriate partitions for a model system in MCell is crucial for optimal performance and can increase speeds by more than 100 times. Unfortunately there is not currently a way to automatically select an optimal partition size in MCell. Partition size should not be smaller than the diffusion fastest distance (diffusion coefficient*step size) however often the limiting factor in choosing partition size is the memory of the computer. If a small partition size is chosen it creates a large number of partitions and begin to slow the computation down as it runs out of memory. For these simulations, partitions in MCell were set at 0.1µm in each dimension to avoid having to consider interactions throughout the whole cell. This setting was empirically chosen based on performance of the system for a cell near the mean of the shape distribution.

The free receptor (species “S2”) was defined as “target only” because it was only involved as a target in ligand binding. This reduced the number of interactions between ligand and receptor by a factor of 2 and provided considerable increase in performance. This setting is recommended for all such molecules or species within a system that have a single interaction partner. When choosing the molecule to make target only, choosing the most abundant species will maximize performance.

Visualization data, consisting of the location of each particle for every time step, was turned off for these simulations as the file sizes generated for this quickly become enormous these locations are not necessary for analysis.

The time step was set to 5e-6s and the duration was set to 4e+7 time steps (200s), though only about 1e7 steps (50s) finished in the allotted wall time of 140 hours. This time step was chosen to be as large as possible while avoiding under sampling issues arising from reactions having a greater than 1 probability per time step and some species having lifespans of less than 50 iterations which is not recommended according to the MCell user manual as this can lead to instability in dynamics.

Once these settings were automatically set using our python script, the systems were then exported from CellBlender for simulation in MCell. Though MCell can be run directly within the CellBlender interface and could have been called directly using this script, we used a computing cluster for efficiency and therefore we created separate MCell submissions for our specific cluster. Our script also added “checkpointing” every 6 hours so that they could be restarted as needed based on crashes or cluster usage on the shared resource.

A total of six (6) random seeds were used per geometry for particle initialization to control for variation from the stochasticity of the MCell simulations. When using this pipeline for computational modeling we recommend more random seeds be used, as the stochastic effects can be quite large particularly for molecules with lower copy numbers (Fig3.5 (m-o)). This of course comes at a computational cost that is multiplied by the number of geometries considered.

For these analyses we simulated a total of 20 cellular frameworks containing 481 endosomes each for a total of 400 seconds as well as two instances containing low (222) and high (808) endosome concentrations respectively for a cell framework and an additional geometric seed for a selected cell framework with 481 endosomes. This resulted in a total of 138 simulations (23 geometries*6seeds). We distributed these simulations on our computer cluster using running 3 simulations per node with the limiting factor being the 8-16GB of memory available per node. Each simulation took about 4 days to complete.

3.3 Demonstration of high throughput modeling pipeline

In this section we analyze the results of our computational pipeline described in the previous section on the generalized signal transduction model presented by Harris et al. [86]. Here we show that spatially resolved models vary significantly from results of compartmental ODE simulations performed using BioNetGen and that the results of these simulations can be used to build models of cellular response dependent on spatial organization.

3.3.1 ODE – Monte Carlo simulation comparison

By comparing spatially resolved simulation results to compartmental ODE models we can assess the impact of stochasticity and cellular organization various parts of the biochemical system. In our model, the initial reactions in our signal cascade appear to have very little stochastic variance but vary widely from the predictions of the ODE model (Fig 3.5). These discrepancies are likely due to slight errors in the extremely rapid internalization rate for the receptor-ligand complex though other contributing factors may include geometry of the cell surface, edge effects of the bounding box, or the under sampling of potential reactions due to the large step size used in MCell.

In contrast to reactions involving abundant upstream molecules, downstream products have larger variances due to the cumulative stochastic affects propagating through the reaction network (Fig 3.5). Some system dynamics, particularly those involving transcription factor interaction at the endosome vary widely from the compartmental ODE predictions suggesting that spatial organization plays a major role in this part of the biochemical network. Interestingly, the reaction dynamics are closer to the ODE models for some downstream products, such as number of transcription factors in the nucleus, suggesting there is a reaction-limited rate limiting step likely related to the transport of transcription factors into the nucleus. These data suggest that ODE simulations may provide sufficient accuracy for some species and that the more downstream species in the biochemical network do not necessarily benefit from spatial simulations in cases where reaction-limited rate limiting steps are directly upstream. However, some chemical reactions within the system are largely influenced by space and spatially resolved modeling reveals statistically significantly different dynamics. Ultimately this suggests that a hybrid simulation approach combining agent based models with homogeneous approaches such as ODE or Gillespie simulations [94], or targeted simulations using approaches such as weighted ensemble [95] for some molecules and reactions may be useful for optimizing simulation speed in the future so that computing power can be focused on the species of interest.

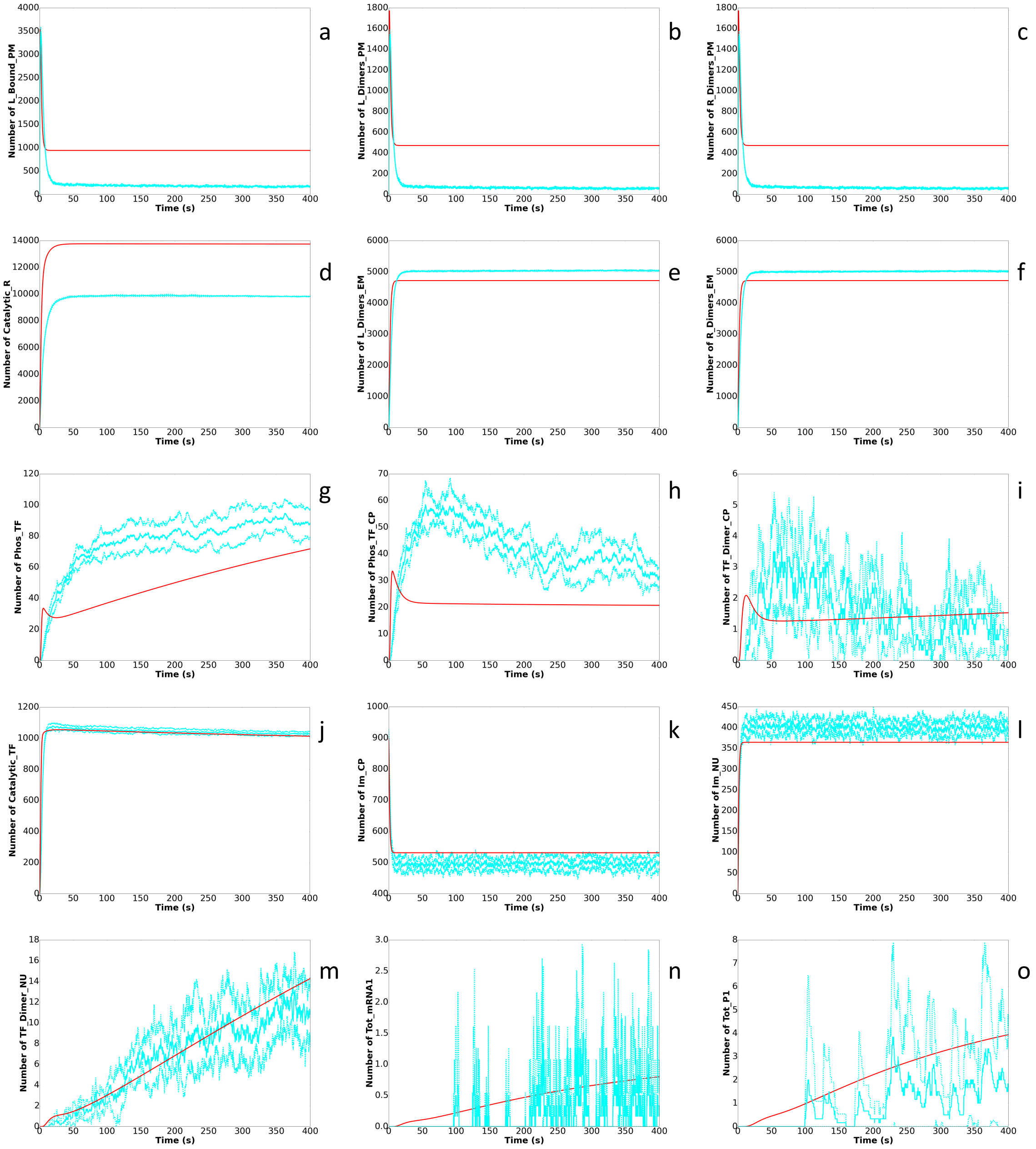
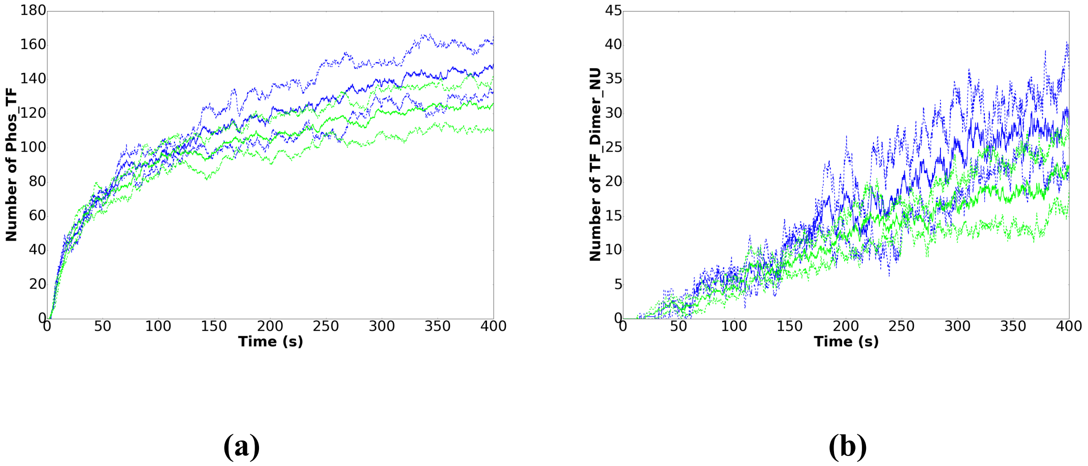


Figure 3.5: Comparison of compartmental ODE simulations performed using BioNetGenv2.2.6 (red) and spatially resolved, stochastic, particle based MCell simulations (cyan) where dotted lines indicate +/-1 standard deviation from the mean of six (6) random initializations. Plots are arranged as the in the order in which the species are produced in the network. Receptor-ligand dynamics (a-c), internalization (d-f), transcription factor phosphorylation (g-j), nuclear transport (k-m), and P1 production (n-o).

Another important feature of the dynamic response of the cell is the different time scales on which different species interact (Fig 3.5). While ligand and receptors have reached equilibrium by 25-50s, other species such as transcription factors in the nucleus, and P1 are still far from equilibrium. In fact, over the 400s of simulations for this cell, we do not see the creation of the final product in the reaction network, P2. This is an important factor to keep in mind when designing high-throughput simulations. If the important dynamics of the system are occurring only after a certain amount of time, "priming" the system by simulating using fast, compartmental ODE or Gillespie based approaches and importing the results after some initial time may be appropriate and save large amounts of computational efforts. This option is indicated by the dashed arrows in Fig 3.1.

3.3.2 Impact of geometric seed

In CellOrganizer (v2.1) the spatial locations and sizes of endosomes are currently drawn from a distribution that is independent of cell shape, endosome size and number of endosomes sampled. To control for the impact of these random scaling and placements, two random seeds for endosome synthesis were simulated for a selected cellular framework. Although there is a slight impact of these random seeds on transcription factor phosphorylation and internalization, these differences lie well within one standard deviation of the measured variance for each geometric seed and are therefore assumed to be noise (Fig 3.6). These represent the largest impact of the change in geometric seed and other species counts remained nearly identical (data not shown).



(a)

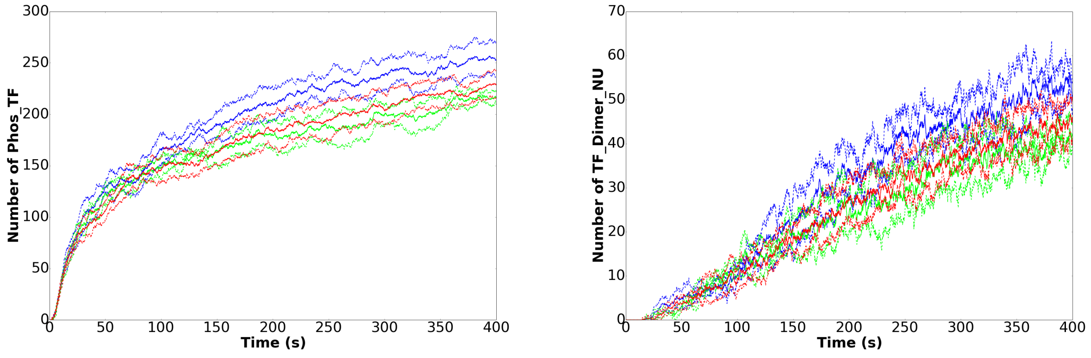
(b)

Figure 3.6: Impact of random geometric seed determining specific location and size of endosomes for a given cell framework for the phosphorylation of transcription factors (a) and the number of transcription factor dimers present in the nucleus (b). Colors indicate two separate geometric initializations of 481 endosomes placed within a fixed cell framework. Differences are within +/- 1 standard deviation indicated by the dashed lines and are therefore interpreted as noise.

3.3.3 Endosome concentration

Cellular signaling through molecular transport via endocytosis is a crucial cellular function that is often a key pathway targeted in disease[96-98]. Here we used CellOrganizer to probe endosome count as an example of another parameter that may be of interest and can be investigated in a targeted manner using this pipeline. We wished to isolate the impact of change in endosome numbers and therefore did not change the internalization rates of the receptor-dimer complex. This means that the "low" numbers of endosomes will have ~4x the concentration of the receptor-ligand complex than the "high" numbers of endosomes.

Interestingly, despite having a lower total number of endosomes, the "low" endosome concentration actually produced a higher signal for the phosphorylation of transcription factors, the event that takes place on the endosomal membrane. This greater phosphorylation is also propagated to the number of phosphorylated transcription factor dimers present in the nucleus (Fig 3.7). We interpreted this unexpected increase in signal to be a reflection of the increased concentration of receptor-ligand complex per endosome. This increased concentration increases the probability of transcription factor phosphorylation upon the interaction of a transcription factor with an endosome. This impact is most pronounced between the "low" concentration (222 endosomes) to the other two concentrations as these impacts are likely offset by the increased number of endosomes for the transcription factors to interact with in the "high" endosomal concentration (808 endosomes). The difference between the "mean" concentration of endosomes and "high" concentration of endosomes is negligible and cannot be distinguished from stochastic noise.



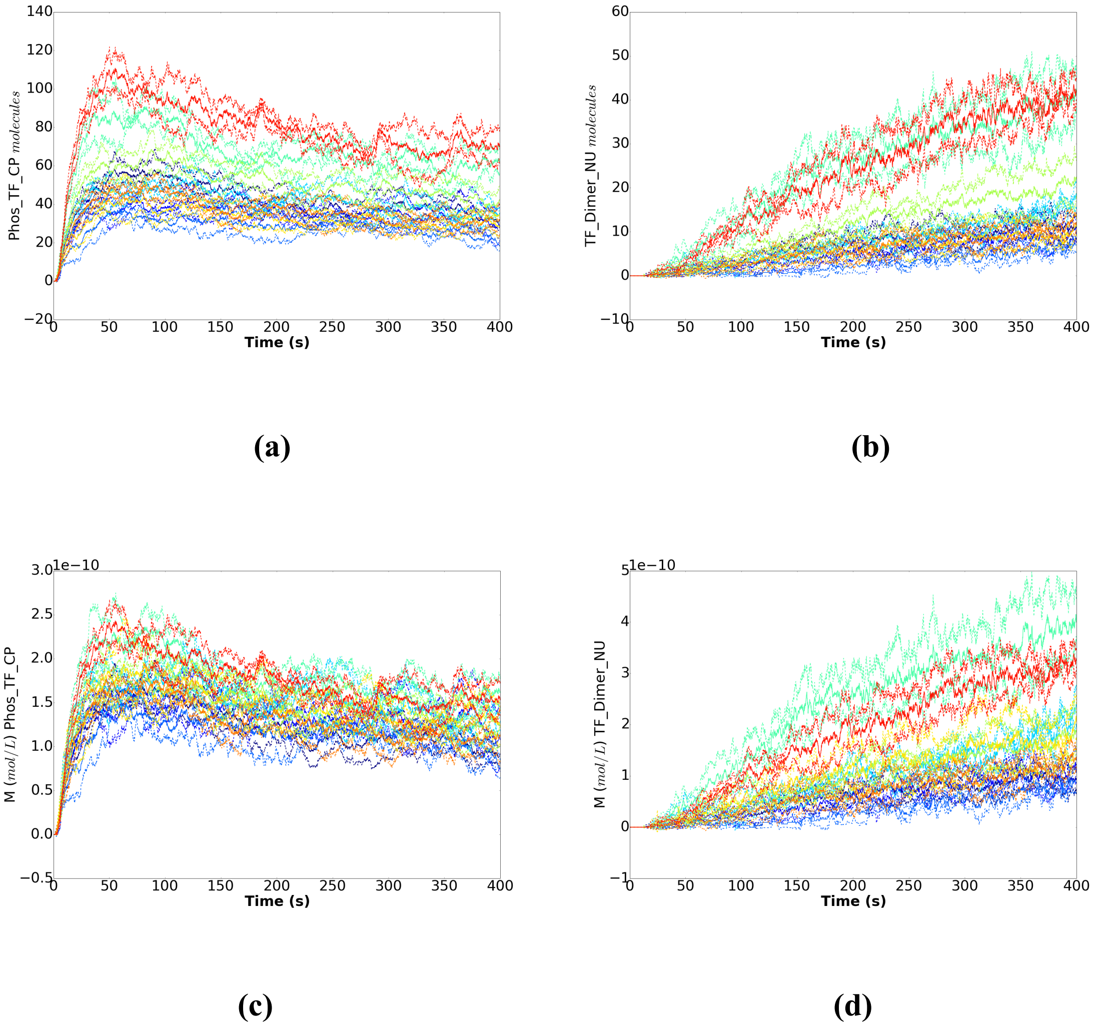
(a)

(b)

Figure 3.7: Impact of varying endosome concentrations on transcription factor signaling while holding internalized receptor-ligand complex concentration constant show increased transcription factor phosphorylation for low concentrations of endosomes (222 endosomes, 0.1CDF) (blue), as compared to the mean (481 endosomes), and high (808 endosomes, 0.9CDF) endosomes concentrations, green and red respectively. Dashed lines correspond to +/- 1 standard deviation around the mean molecule counts for six (6) MCell initializations.

3.3.3 Cell framework dependent response

Using CellOrganizer we were able to generate a library of cell frameworks from throughout a shape space. These sampled synthetic cells (Fig 3.3), were then simulated for 400s using MCell as described in section 3.2 above to determine the impact of spatial variance of cells within a population. When looking at raw expression levels (molecule counts) these data capture the impact of changing surface areas and volumes, and relative sizes of the nuclear and cytoplasmic volume (Fig 3.8 a-b). Although these plots show the raw expression of a given species, in some cases it may be desirable to analyze the concentration of a given species to eliminate direct contributions from differences in the surface area and volumes of cells on cellular dynamics. To do this we examined concentration of molecules as the initial concentration of all species was fixed across morphologies accounting for differences in surface are and volume. This analysis isolates the spatial contributions surface area to volume ratio and diffusion dependent dynamics over the set of synthetic geometries (Fig 3.8 c-d). As seen in Fig 3.8 significant differences in response are present dependent on the cell framework both when considering raw molecule counts and concentrations.



(a)

(b)

(c)

(d)

Figure 3.8: Varying expression levels for phosphorylated transcription factors in the cytoplasm and nucleus (left and right columns respectively) within realistic cellular geometries show the raw expression differences (a-b) and differences in concentrations (c-d) for the set of synthetic cells sampled from the 3D HeLa shape space. Here color represents different cells and dashed lines represent +/- 1 standard deviation over 6 random initializations of MCell.

3.3.4 Modeling shape dependent response

As we demonstrated in the previous section, different cells express different levels of a certain species. Though it is interesting to know this, we would like to determine **why** these differences occur and if they can be predicted. Indeed, certain species and biochemical systems may have a smooth response across a spatial parameter of interest, (cell shape, number of vesicles etc.) allowing a regressive model of response to be learned for these species/systems. Using these models we may discover potential novel targets for drug treatment of pathways that impact spatial organization but may not directly participate in the diseased system biochemically or gain an understanding as to what drives certain changes in cellular behavior. To demonstrate this we show the expression level of a given observable across the shape space and fit a regressive model capable of predicting relative expression of a species given cellular shape (Fig 3.9). For this analysis we chose to analyze the concentration of dimerized transcription factor in the nucleus at 400s. It can be seen from this fit that elongated cells have lower expression of this species than rounded cells towards the bottom of the shape space. The below model demonstrates predicting a particular species at a specific time, however this approach can be easily expanded to predict the full response of the molecule over time by fitting a function to the molecular response and predicting the coefficients of the response function.

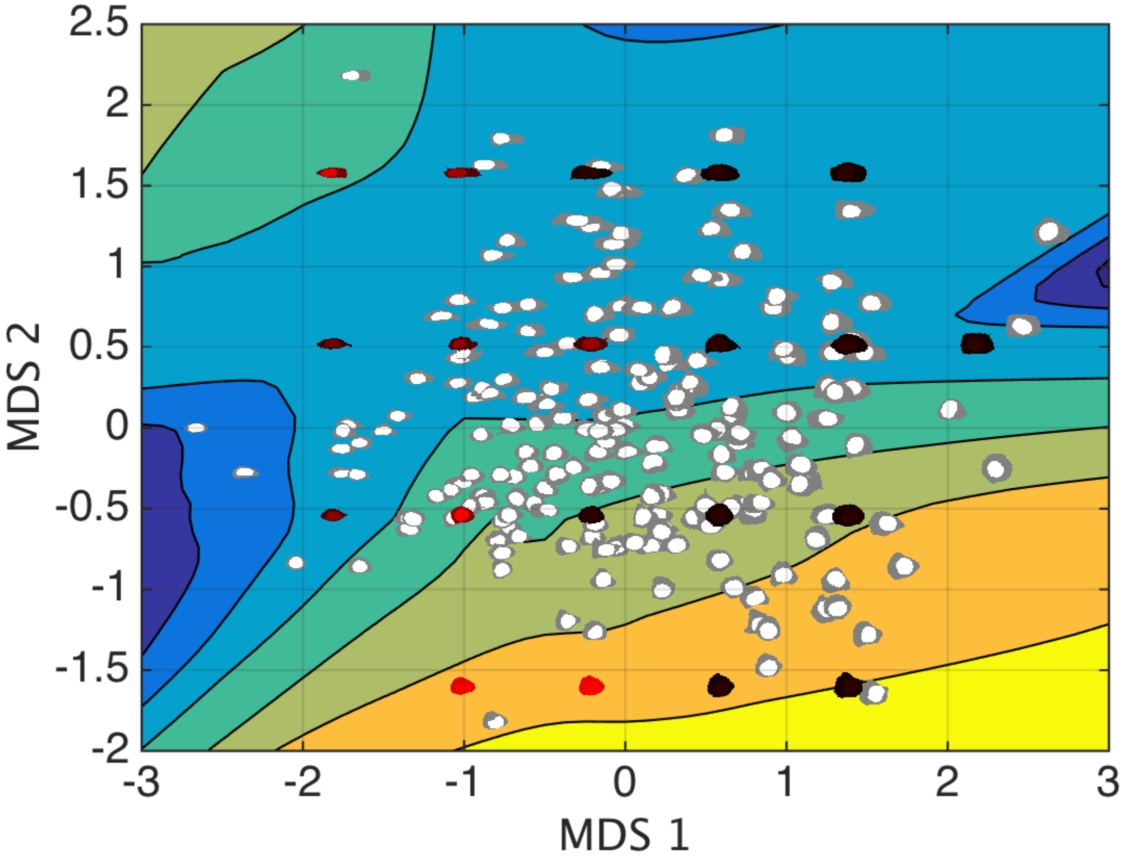


Figure 3.9: Example regressive contour plot for predicting cellular response dependent on cell and nuclear shape using a locally weighted scatter plot smoothing (LOWESS) model for the expression levels of dimerized transcription factors in the nucleus at t=400seconds with warmer colors indicating higher expression levels. Synthetic geometries are shown in red where intensity is proportional to expression level. Black cells were excluded from analysis due to lack of data and were not considered in the model fit.

3.4 Discussion

The work described here enables spatially resolved modeling to be performed in high throughput with minimal manual effort. In this work we have developed a pipeline for these spatially resolved simulations and demonstrated that it can be used to study complex multi-compartmental biochemical systems and their spatial dependencies. Particularly we demonstrated that this pipeline allows us to learn regressive models of cellular response as a function of spatial organization, a result that was previously not possible due to the limited number of realistic geometries available for spatially resolved models.

Although this pipeline allows modelers to study cellular dynamics in several previously inaccessible ways, there remains a large amount of work to be done to improve this type of modeling and simulation. Currently these stochastic spatially resolved simulations are computationally expensive taking several days or even weeks to run and requiring multiple runs to account for the random initializations. As such, methods to increase this simulation efficiency are desperately needed to maintain tractability of spatial modeling. Future investigations may be able to use compartmental information, molecule concentration, shapes of response curves, and on-line comparisons to approaches such as compartmental ODEs (Fig 3.5) to predict the significance of spatial organization for systems and species. These predictions could be used to automatically refine the simulations for cases where a reaction is predicted to be highly spatially dependent while simultaneously abstracting other reactions in the network using much faster ODE or Gillespie style simulations where spatial effects are not predicted to significantly contribute to a species behavior. This hybrid modeling system would allow for much more rapid simulations while maintaining high amounts of spatial realism when and where it is interesting and important.

We also showed that spatially resolved simulations significantly differed from compartmental ODE predictions generated using BioNetGen and that this pipeline can be used to study various spatial parameters including cell and nuclear shape and endosome concentration. Based on the illustrative results presented in this work, we believe that this pipeline can be used to reveal novel dynamics arising from the statistically significantly different dynamics between well mixed (ODE) and spatially resolved (agent based modeling) approached. Further, it may be possible to predict missing members of a biochemical network not possible with homogeneous modeling approaches leading to a better understanding of the underlying biochemistry within the cell. The increased understanding of the impact of spatial organization on the cells coupled with the responsiveness of a system to changes in subcellular structure, and spatial properties (such as diffusion coefficients, internalization rates etc) may also lead to novel treatment pathways through yet to be explored spatially driven mechanisms.

Overall, this computational pipeline takes spatially resolved modeling from a field that requires large amounts of time, manual input, and expertise and makes it highly accessible. Using this system we can transform a compartmental ODE based biochemical model into a fully spatially resolved set of simulations with very little effort. By lowering this barrier for creating spatially resolved simulations and continuing to improve simulation engines, we hope to create a new standard in which realistic spatially resolved simulations are standard and expected when researching systems biology.

[During final editing, after the thesis defense, it was discovered that there were errors in the SBML to MCell converter. These errors involved improper unit conversion of surface reaction rates and explicit molecule counts to concentration. This latter error, caused by improper volume exclusion, caused variable initial concentrations of species dependent on the volume ratios of the cell, nucleus and endosomes. These errors cause errors in the quantitative results presented in this chapter that may also change the conclusions drawn here. However, they do not change the utility or novelty of the methodology presented in the chapter. This includes the development of the computational pipeline for high throughput spatial modeling and the ability to study and learn regressive models of dynamic response conditional on cell shape and organization. Both errors have been corrected and new simulations will be performed prior to the submission of this work for publication.]

3.5 Additional required work

Some steps remain to prepare this work for publication; analysis of results from revised models, and completion of simulations for all sampled cells in the shape space. In a subsequent publication we will demonstrate the pipeline's ability to model and predict real biological network dynamics for changes in cell morphology and organization resulting from gene regulation related to disease.

Chapter 4. Generative models of neuronal shape for use in subcellular modeling[4]

4.1 Introduction

Currently an area of major interest in the field of biology is the understanding of neurons and their interactions particularly in relation to the human brain. These efforts include the NIH large-scale initiative on Brain Research through Advancing Innovative Neurotechnologies (BRAIN initiative) and the Human Brain Project (HBP). This has led to a large surge in efforts to model neurons.

The complex morphologies of neurons and the intricate networks they form are vital to their function, and understanding both is necessary for understanding the various neurodegenerative diseases and disorders [99]. Unfortunately, studying these cells *in vivo* is extremely difficult and often requires the sacrifice of the subject. Additionally, disruption of the native state of the cell during sacrifice is a concern when studying expression within these cells. Common approaches to imaging neurons *in situ* involve staining a sacrificed animal brain for proteins of interest and thinly slicing it so it can be imaged. Processing and reconstructing tissue slices is a difficult problem that is far from trivial limiting the amount of high-quality high-resolution neuronal imaging data [100]. Crucially, these traditional approaches cannot be used to study the dynamic behaviors and interactions of neurons at the cellular and sub-cellular levels *in situ* as they require animal sacrifice. As such, traditional data acquisition methods

[4] This chapter describes joint work with Armaghan W. Naik, Rebecca Elyanow, Xuexia Jiang and Robert F. Murphy and is being actively prepared for journal publication.

will not suffice and there is a need for *in silico* modeling of neurons at multiple scales as a means of studying neuronal behavior.

Towards the goal of understanding neuronal dynamics, modeling tools such as NEURON, Genesis, Netmorph, and MCell have been developed to simulate neurons and their interactions over time in a spatially resolved manner [40,48-50]. These tools use rules defined over a set of neuronal geometries to simulate interactions at the subcellular, cellular and multi-cellular level. They have been successfully applied to modeling complex multi-cell models of neuronal networks and show promise in aiding the understanding of neurons at the single and multicellular levels [51,52]. Unfortunately, as mentioned, image acquisition of these neuronal geometries and networks of neurons is both expensive and not scalable.

To address this issue, generative models present one approach to modeling neuronal morphologies *in silico*. This is because these models can be trained using a limited number of experimentally observed cells and used to generate novel *in silico* instances representative of the observed neurons by bootstrapping the parameters of the learned models. This is especially true for models that combine information on organelles or proteins that were not or cannot be imaged simultaneously in the same cells. To date, there have been several approaches taken to generatively modeling neurons.

The NeuroMorpho (previously L-Neuron) software package uses a set of recursive rules to describe dendritic geometry and topology by correlating morphological parameters such as branch diameter and length [43]. The system implements three algorithms: Hillman, Tamori, and Burke, to describe the branching length and angles of the dendritic trees. The TREES software offers another approach which reconstructs the neuronal branching under the assumption that dendritic trees connect synaptic inputs to the dendritic root using a minimal total length of wiring, and performs a local optimization of total wiring and conduction distances [23]. Neugen and Netmorph are two other software packages aimed at the generation of morphologically realistic, large-scale neural networks in 3D [46,47].

Although each of these methods is capable of reproducing statistically plausible neurites under various constraints they do not allow for the direct modeling of nuclear shape, soma shape, and protein distributions. The generative modeling platform developed in our lab, CellOrganizer is capable of learning these conditional multi-component models for cellular frameworks, subcellular structure and protein distribution from imaging data. Currently, our models of cell framework consist of two classes, parametric and nonparametric. The parametric models are restricted to modeling star-polygon shapes and are therefore not suitable for modeling neurons. The nonparametric large deformation diffeomorphic metric mapping (LDDMM) approach in CellOrganizer can be used to model the complex morphologies of neurons in theory, however it is unlikely that cells could be properly registered, or that this deformation based approach would accurately describe the differences between neurons.

This work aims to build multi-component models of neurons within the CellOrganzier platform with an ultimate goal of the subcellular modeling of neuronal structure, a currently under represented scale of modeling in the ongoing neuronal modeling efforts (Fig 4.1).

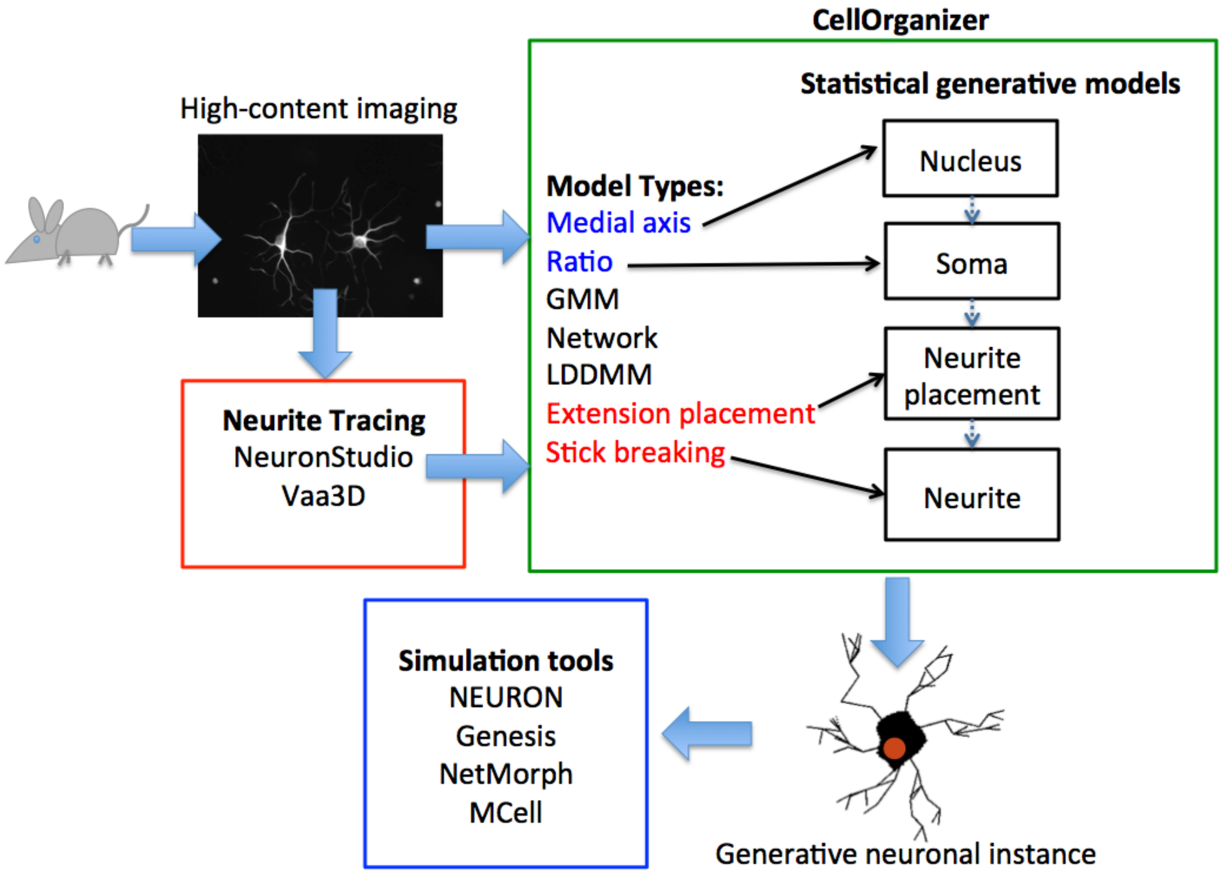


Figure 4.1: Computational pipeline for generative neuronal modeling using CellOrganizer presented in this work. Large arrows indicate direction of workflow. Cells taken from organisms were imaged and Neurites were traced using NeuronStudio (red box). Both the original images and the .swc trace files were passed into CellOrganzier. Various model types from within CellOrganizer were used to model components of neurons. This work utilized the previously developed medial axis and ratio models (blue) as well as the extension placement and stick breaking process models (red) that were developed specifically for this work. Models types listed in black are available within CellOrganizer (green box) but were not used in this work. Dashed arrows between trained compartment models (black boxes) indicate the conditional structure of the model. These models were used to generate multi-compartmental, *in silico*, neuronal instances. The resulting instances could be used in future work to perform simulations using a number of tools (blue box) and include models of subcellular components learned from imaging data.

4.2 Training neuron models

4.2.1 Parameterizing Neurites

Due to the limitations and assumptions made by prior neurite models to most efficiently accomplish their task, we chose to construct a new parameterization for neurite generation. A number of possible neurite parameterization strategies exist as seen in past work. The problem is simply how to properly distribute branch points and lengths. Many of the existing methods for modeling neurites are based on inferred biological rules (L-Neuron, NetMorph, NeuGen) that may not generalize well to all neuron types in an attempt to bring physiological relevance to a highly artificial process. Neurite models were learned from standard .swc files of traced neurons used by many of the databases of neuron structure such as NeuroMorpho. For this work we did not consider the non-trivial task of neurite segementation and tracing.

To parameterize each neurite we used the tree-structured stick breaking process which has a long history of use as a statistical tool [101]. A stick breaking process begins with a fixed unit length that is fragmented according to values drawn from a beta distribution. We cannot claim any biological relevance for generating a neurite using this process as neurons do not grow by breaking a fixed length into fragments, however the process is sufficient for describing the shape of a neurite in both two and three dimensions. An additional benefit of using the tree structured stick breaking process to parameterize neurites is that it allows for multiple types for branching events (bifurcation, trifurcation, etc.) which previous models do not allow but are often seen at the imaging resolution.

We defined segments of neurites as lengths bounded by neurite origin, branch points, and termini (Fig 4.2 a). The stick breaking process assumes a unit length and draws partitioning parameters β from the beta distribution Be(1, α) with α learned from data to produce fragments π (eq 4.1).

$$[4.1] \qquad \beta \sim Beta(1,\alpha) \qquad \pi = \beta_i \prod_{k=1}^{i-1}(1-\beta_k^{'})$$

To define a neurite with the tree-structured stick breaking process we alternated between allocating fractions of the remaining length (φ) to a segment, and stick breaking events (ψ_i) at neurite branch points (Fig 4.2 c).

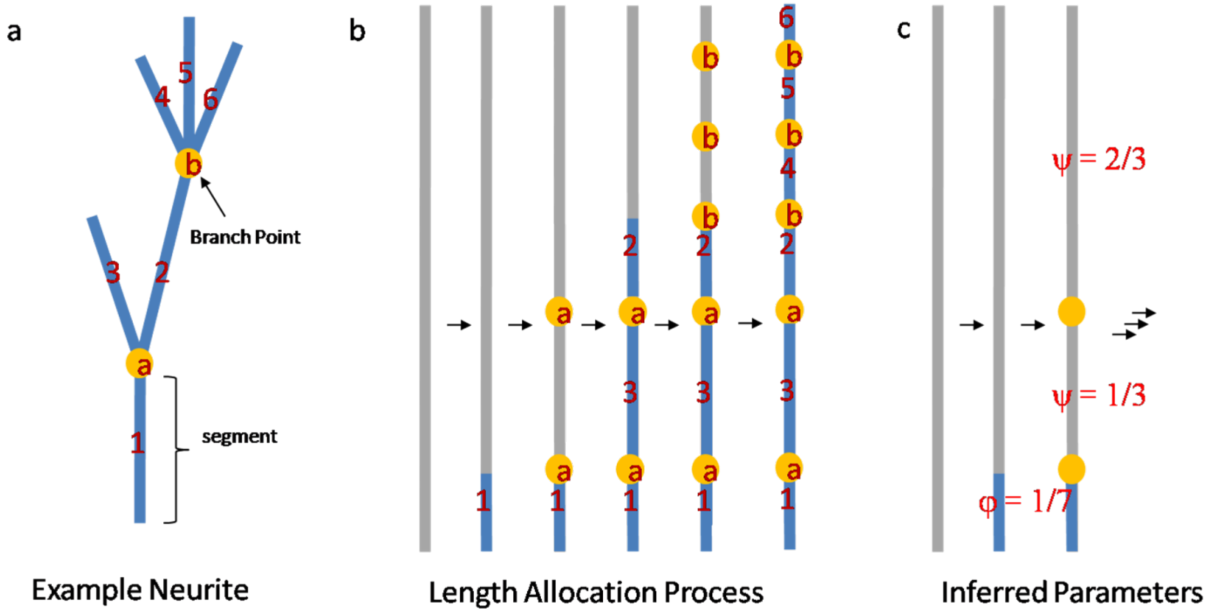


Figure 4.2: Example stick breaking process used for neurite parameterization. An example neurite with numeric segments (blue lines) and alphabetic branch points (yellow dots) (a). The neurite is broken along branch points and arranged along a unit "stick" (b). The neuron is then parameterized by the sequential branch points (c) where φ and ψ represent the percent of unit length used and the fraction of remaining length at which the stick is "broken".

Segment lengths vary with respect to depth ε, measured by the number of progenitor segments connecting to the soma. This captures variation in length between proximal and distal segments. Two separate beta distributions parameters, α for segment length and γ for branching, are learned from data (eq 4.2).

$$[4.2] \quad v_i \sim Beta(1,\alpha|\varepsilon|) \qquad\qquad \psi \sim Beta(1,\gamma)$$

For each neurite segment we observed π_i, the normalized segment length, such that $\Sigma\, \pi_i = L$. We used $\pi_{i|\varepsilon|}/L$ to learn the distribution of $\nu\ |\varepsilon|$, the partitioning coefficient for segment length drawn from the beta distribution, and $\alpha|\varepsilon|$ used for defining the beta distribution. For each segment we also observed $\pi_{>i}$, all children segments of π_i (green box). For each child segment of π_i, $\pi_{i'}$ we observed $\pi_{\geqslant i'}$, all processes starting with $\pi_{i'}$. We defined the partitioning coefficients to be $\psi_i = \frac{\pi_{\geq i'}}{\sum \pi_{\geq i'}}$ as seen in (Fig 4.3). The distribution of these partitioning coefficients for branching was modeled by the beta distribution (eq 4.2), and the fit was used to learn γ, the parameter of the beta distribution.

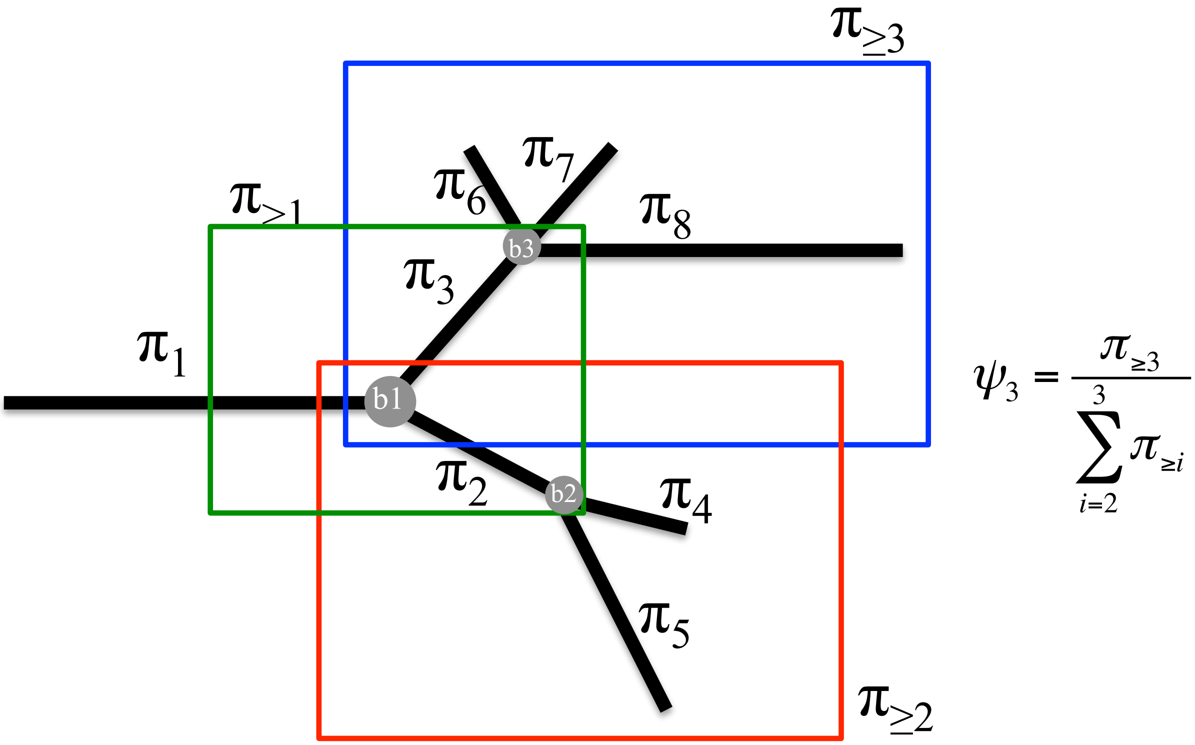


Figure 4.3: Example partition coefficient calculation of ψ_3 at the branch point b1 is performed computing the fraction of neurite mass from π_3 down ($\pi_{\geq 3,}$ blue box) compared to the total mass of the neurite past branch point b1 (blue box+red box). This coefficient is calculated for all branching events in a similar fashion such that each branch event has a number of coefficients equal to its children.

We defined φ to be the fraction of remaining length allocated to each child process φ (eq 4.3). To determine the total length of a child segment, π_i was drawn from the tree structured stick breaking construction (eq 4.4).

$$[4.3] \quad \varphi_i = \psi_i \prod_{k=1}^{i-1} (1 - \psi'_k)$$

$$[4.4] \quad \pi_i = v_i \varphi_i \prod_{i'<i} \varphi_{i'} (1 - v_{i'}) \ , \quad \pi_0 = v_0$$

We represented the branch angle as a depth-independent gaussian trained on observed branch angles. This work used a uniform segment thickness and total neurite length was represented as a gaussian trained on observed neurite lengths. Neurite count was represented as a poisson distribution trained on observed neurite counts.

4.2.2 Training soma and nuclear models

The neuronal soma is relatively small and simple compared to the long dendrites and axons extending from it. The majority of prior work has ignored soma morphology using a single point or a small circle/sphere to represent the cell body. If desired, a similar approach can be used for this work in CellOrganizer by specifying only a stick-breaking model at the time of synthesis in CellOrganizer.

Alternatively, CellOrganizer (v2.1) already contains parametric models capable of describing cell and nuclear shape [19,22]. These approaches can be applied to the modeling of the nucleus and soma segmented from neuronal images. In this work we applied low-pass filtering followed by active contour segmentation to obtain segmentations of the nucleus and soma. Though pieces of neurites may be left from this segmentation, the center of the soma is identified as being the location within the segmentation that is maximally distant to any border. Neurites remaining from this segmentation are trimmed naturally by the parameterization as the medial-axis and ratio models currently available in CellOrganizer can only describe star polygons.

Training these models requires high-resolution images of neurons from sources such as fluorescent microscopy imaging. In this work we used set of 2D fluorescent microscopy images of neuronal cells from the Cell Image Library (CIL) collected by Rusielewicz T. M-VC (CIL: 40358,40359,40360, and

40361) to test the models. . As these data are from cells grown in culture, we assume that projections are approximately two-dimensional compared to the scale of the extensions.

4.2.3 Training extension placement models

Using the models of neurites soma and nuclei described in sections 4.2.1 and 4.2.2 respectively, we can fully describe the volume of neurons, however to completely encode the morphology we need a model describing the conditional structure between these two components. To do this we first identified neurite location on the soma from the 2D CIL data by searching the local area around our identified soma for regions of high intensity likely to be neurites (Fig 4.4).

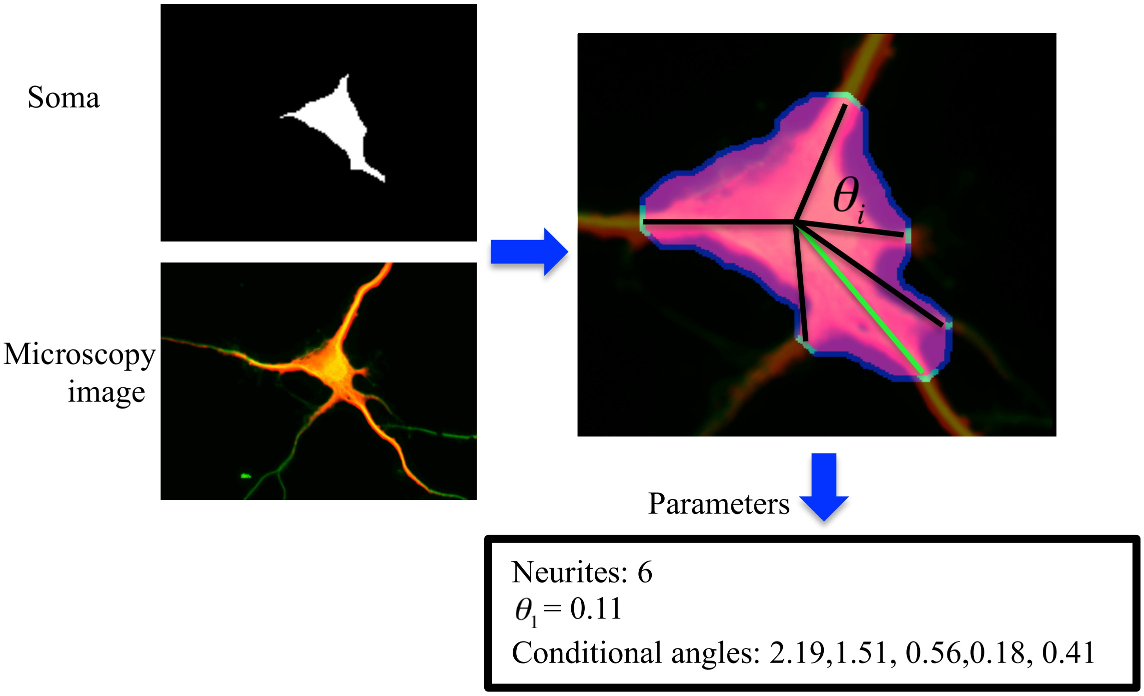


Figure 4.4: Neurite parameterization. Soma are segmented from a neuronal image. Regions of high intensity surrounding the masked soma are considered neurites. A conditional angle, θ_i is calculated between each of the identified neurites (2-n). The angle to the first neurite (green) is measured as an angle from the major axis.

To parameterize the location of neurites, we needed to model the angle between each neurite θ_i for a given image. Ideally we would model each θ_i individually, however this requires models be built for each neurite frequency meaning that separate training data of neurons containing 1 to n neurites are needed. Due to the difficulty in imaging neurons and their variability, there are often a very limited number of geometries for a given cell type and condition making this method impractical. We therefore parameterized neurite placement using a maximal-spacing model where neurites are positioned at every 2π/n radians where n is the number of neurites. This method allowed us to combine information from neurons with varying numbers of neurites. We measured the angle of the first neurite, θ_1, from the major axis of the cell. All subsequent inter-neurite angles θ_i for i=2 to n were modeled as percent deviation ω_i from the expected maximal-spacing model (eq 4.5), where the distribution of these deviations was used to define a normal distribution.

$$[4.5] \quad \omega_i = \frac{\theta_i}{\left(\frac{2\pi}{n}\right)}$$

4.3 Synthesizing neurons

To generate synthetic neurons, nuclei and soma were synthesized according to the procedures outlined in Zhao et al. [22] using the data from the CIL. These images were also traced using NeuronStudio (v0.9.92) to create standard .swc files for learning the neurite models. The placement of the first neurite was sampled from a normal distribution learned from the observed values of θ_1, the angle between the major axis and the location of the first neurite. Subsequent angles (2-n) were sampled from a normal distribution of percent deviation ω_i from the maximal-spacing model relative to the angle from the previous neurite.

Neurites were generated via the Pòlya urn process [102] using the parameters inferred through the stick breaking process (Fig 4.5). This method provides a way of sequentially selecting events from a set of possibilities. In our case we select whether to grow the current neurite branch, proceed to an existing branch deeper in our tree, or create a new branch. This method requires an additional parameter N, which is effectively the sampling resolution. This parameter was set to the length of the sampled neurite (µm). At each branch point and termini, each 0.1 µm fragment of the total neurite length was given a probability of either staying in the current segment, creating a new child branch or continuing to a child of the current segment. These probabilities are evaluated sequentially and the last condition is obviously not possible for leaf nodes where failing to stay at a leaf will always create a new branch. Starting at the neurite origin as determined by the extension placement model, the probability for staying in the current segment and extending it was defined by eq 4.6.

$$[4.6]\ \ P(stay) = \frac{f_{stay} + 1}{f_{stay} + f_{cont} + \alpha + 1}$$

Where f_{stay} is the number of fragments that have stayed at the current segment, f_{cont} is the number of fragments that were assigned to a deeper segment in the tree, and α is the segment length parameter of the beta distribution empirically sampled from the distribution of observed values. If a fragment did not stay at a particular branch, the probability of creating a new branch was defined by eq 4.7 where γ is the branching parameter of the beta distribution defined in eq 4.2.

$$[4.7]\ \ P(branch) = \frac{\gamma}{f_{cont} + \gamma}$$

If a fragment neither stayed nor created a new branch at the terminus of a given segment the fragment continued to one of the children branches where the choice of which branch was based on the probabilities as defined in eq 4.8 where $f_{i,cont}$ +$f_{i,stay}$ is the total mass of fragments currently on the current segment and $f_{i'cont}$ is the number fragments that have continued past the specific child.

$$[4.8]\ \ P(continue_{\pi_i}) = \frac{f_{i,count} + f_{i,stay}}{f_{i',cont}}$$

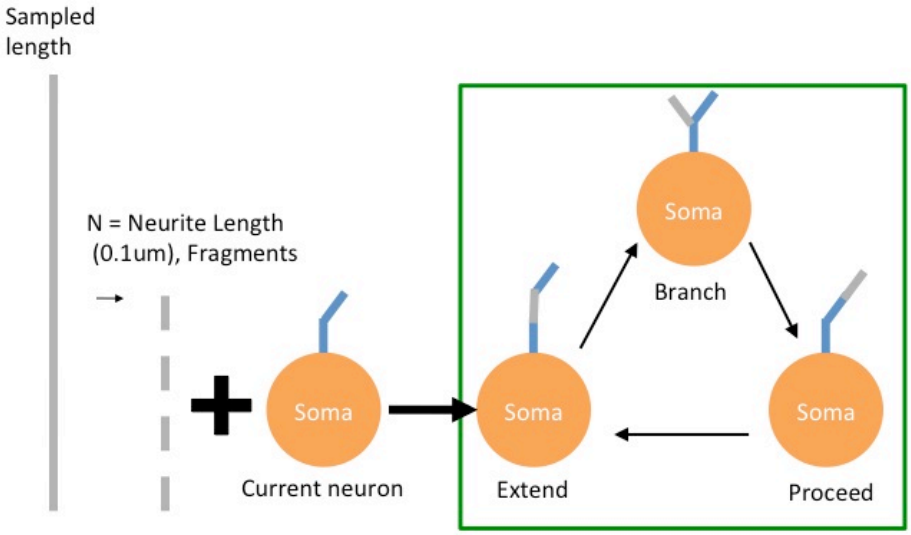


Figure 4.5: Example neurite synthesis process using a Pòlya urn process. Neurite fragments from the sampled neurite length (gray bars) and sequentially added to the current neuron. Each segment evaluates whether to extend, branch or proceed to an existing child (green box).

After a single loop through all fragments, we pruned the neurite such that terminating segments with length less than the mean of all segment lengths – 2 standard deviations of segment lengths are removed. The removed fragments were sent through the polya urn process for the current neurite again with P(branch) = 0 to preserve the original total length. A branching angle and was drawn for every segment. There exists an unknown physiological limit for neurite branching. Since few branching events above trifurcations can be observed, we chose to generate neurites under a branching limit of 3 meaning that bifucations and trifucations are both possible however this method is capable of generating higher order branching events if the user changes this limit within CellOrganizer.

4.3.1 Example synthetic neurons

To demonstrate the utility of these models, we learned models using the publicly available data of oligodendrocytes from the Cell Image Library (CIL). This data consisted of four 2D fluorescent microscopy fields of rat (Rattus norvegicus) oligodendrocytes purified from P2 cortex. Cells were stained with Phallodin-FITC (actin), alpha-tubulin, and DAPI and imaged using a Leica DMI4000B microscope at 63x resulting in a voxel size of 0.1µm. This dataset was chosen as it contained multiple channels allowing us to build multi-component models of subcellular structure. Figure 4.5 below shows an example image from the oligodendrocytes images used, and resulting images of neuron morphology from the models described here.

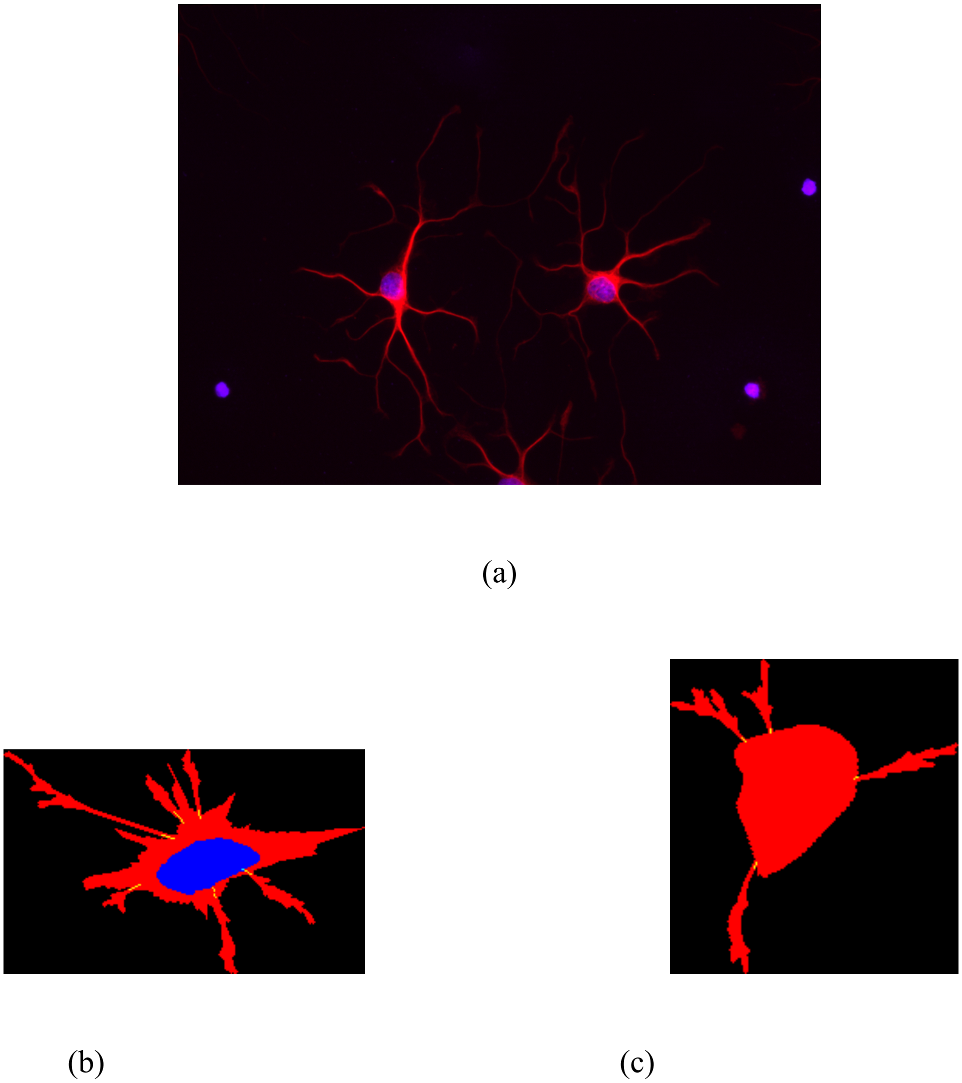


Figure 4.6: Comparison of real oligodendrocytes (a) from the Cell Image Library stained for DAPI (blue), and alpha-tubulin (red) (a) were used to train three generative models of neuron shape and organization. Multiple component model of neuronal organization using the medial axis model for nuclear shape (blue), the ratio model for soma shape combined with the stick breaking process for neurite morphology (red), and the extension placement model (yellow) (b). In the absence of a nuclear marker, the medial axis model was used to model soma shape while maintaining models of neurite location and morphology as before (c).

The CellOrganizer framework previously required nuclear labels to build models of cell shape. Because imaging data of neurons often does not contain nuclear stains we sought to develop more flexibility within CellOrganizer. Though models in CellOrganizer were initially designed to parameterize a specific compartment, there is no reason that they could not be applied to other compartments assuming the conditional and morphological requirements of the model are met. To take advantage of this potential flexibility, we redesigned the model structure in CellOrganizer to be modular such that compartments can be modeled by any of the available methods as specified by the user. This allowed us to apply previously developed modeling methods to a wider range of data. Specifically, we applied the medial axis model, previously used to model nuclear morphology to describe soma shape in the absence of a nuclear marker (Fig 4.6 c). Results of this method produce qualitatively reasonable soma as seen in Fig 4.6. Each method (Fig 4.6 b-c) has distinguishable traits of the model. The ratio model (b) is jagged as it samples the cell-nuclear ratio for each radial location while the spline fit (c) produces a smoothed fit of shape.

4.4 Evaluation of neuronal modeling

Unfortunately, there is a very limited number of publicly available high quality images of neurons. Often, as in the case of the oligodendrocytes from the Cell Image Library (CIL), collections of images contain only a few fields for a given experimental condition and cell type and combining experiments across labs, conditions, and cell types is not appropriate. In the case of the oligodendrocytes, only four fields containing 6 soma segmentations were available. Due to this limited amount of data we were unable to assess the accuracy of our multi-compartmental models as we have in our previous generative modeling efforts [19,22].

There are however large databases of neuronal tracings available on databases such as NeuroMorpho and Cercal DB, Cell Centered Database, and Fly Circuit containing neuronal traces (.swc files) [45,103]. In this work we used the NeuroMorpho database to obtain a large of .swc image reconstructions of 8 different neuron classes. We used the stick breaking process model of neurite structure to parameterize 8 neurite types with 8 features; 3 for dimensional angle, the mean and standard deviations of ψ_i , and α and the total length of the neurite. We then used a multiclass Support Vector Machine (SVM) classifier to predict neurite type based on the extracted generative features with 5 fold cross validation (Table 4.1). This demonstrated that our neurite parameterization is capable of distinguishing several morphologically distinct neurite types with high accuracy (mean = 90%). Post hoc analysis revealed that the most salient features in this analysis was α with total length also being very defining for some cell types such as Purkinje which tend to be much longer than other neurite classes.

Table 4.1: Average classification confusion matrix of neurites using generative model parameters extracted from .swc files of neuronal reconstruction from the NeuroMorpho database for five-fold cross validation using a multi-class Support Vector Machine (SVM). Diagonal elements are the average classification accuracies while off-diagonal values show confusion between classes. This data demonstrates that the model parameterizations are capable of describing neurites of a particular class. Numbers below each neurite type indicate the total number of reconstruction files used for this analysis. The low classification accuracy of Martone Purkinje neurites is likely an artifact of class imbalance. Total neurite classification accuracy was 90%.

	Allman Bipolar	Ascoli Moto-Neuron	Korte Pyramidal	Lee Pyramidal	Martone Medium Spiny Cell	Monyer Small Axonless	Martone Purkinje	Yuste Interneuron
# of neurons	[29]	[17]	[36]	[90]	[6]	[6]	[4]	[29]
Allman Bipolar	0.9	0	0	0.01	0	0	0	0
Ascoli Moto-neuron	0	0.95	0.05	0	0	0	0.17	0.03
Korte Pyramidal	0	0.05	0.87	0	0	0	0.33	0
Lee Pyramidal	0.1	0	0	0.99	0.17	0	0	0.03
Martone Medium Spiny Cell	0	0	0.03	0	0.5	0	0	0.07
Monyer Small Axonless	0	0	0	0	0	1	0	0
Martone Purkinje	0	0	0.05	0	0	0	0.5	0
Yuste Interneuron	0	0	0	0	0.33	0	0	0.87

4.5 Discussion

This work enables the generative modeling of neurons and their subcellular components using the open source CellOrganizer system. Currently, this pipeline requires two inputs; imaging data and a neurite tracing file generated using an external software tool.

In this work, we adapted existing methods in CellOrganizer for nuclear and cell shape modeling to model the nucleus and soma of neurons. We also developed two novel parameterizations to describe neurites and their locations around the soma. Although there have been several previous efforts to build generative models of neurons [23,43,46,47,50], to our knowledge this is the first work that incorporates conditional subcellular structure into the modeling framework.

We applied this multi-component approach to a set of 2D fluorescent microscopy images of oligodendrocytes from the Cell Image Library to demonstrate the ability of these conditional component models to generate realistic neuronal shapes. We further demonstrated that our neurite parameterization was capable of distinguishing a set of 8 neurite classes downloaded from the NeuroMorpho database with an overall accuracy of 90% using a multiclass SVM classifier

Despite these promising results, this work is still in the very early stages and there are many ways in which these models can be improved. Currently, our parameterization of neurites ignores curvature and thickness. Including these parameters in the model would create much more realistic looking structures necessary *in silico* simulations and will be supported in future versions of CellOrganizer. The generated neurites also appear to be too compact suggesting that the learned branching angle is too conservative. This is likely due to the "y" shape observed in many branching events where segments branch at a large angle but curve back towards the branch point before the termination of the segment leading to a smaller estimated angle when calculated over the length of the neurite. In the future, using a parameterization that accounts for the curvature throughout the neurite conditional on a points relative location to a branch point should improve this issue. Additionally, expanding the stick breaking process and extension localization models to three dimensions is trivial and will be included in a subsequent version of CellOrganizer.

In some cases in the NeuroMorpho tracing data, we observed depth-dependent branching probabilities suggesting that a depth-dependent γ may improve the model over the global γ used in this work. Modeling neurite structure using a fit for the length parameter alpha rather than sampling it empirically would enhance the compactness and scalability of the model.

In this work we also made two arbitrary choices that should be motivated by physiological parameters in the future. First we assigned a conservative branching limit of 3 to neurites. Ideally this could be set at the physiologically relevant limit but such a limit is unknown. Empirically we observed as many as 6 branching events at a single junction, however, this limit should be informed by the distributions of observed branching of a given set of neurites. Second, the lower limit of segment length should be learned from data rather than arbitrarily to the sampling resolution (0.1µm) as there is likely a physiological limit for the minimum size of a stable neuronal projection.

As discussed, there have been many previous efforts to model neurons and neurites specifically. Future versions of CellOrganizer will support instances from these various modeling frameworks in addition to reconstructions of neurons from .swc files in order to allow modeling of subcellular components and protein distributions within these instances learned in CellOrganizer. This work was done in 2 dimensions, however both novel models presented here can be applied in either 2 or 3 dimensions and future versions of CellOrganizer will include support for neurite modeling in 3D as well.

With this work we aimed to build a flexible tool for modeling many conditional cellular components describing neuronal morphology and their subcellular structure. The models presented were applied to neuronal modeling, however these models may be useful for modeling other cells with complex morphologies that lie outside the space of star polygons currently supported via the parametric models in CellOrganizer. Where data is available, dendrites should be modeled separately from axons, as they possess distinct morphological features. In this work we ignore this caveat, as oligodendrocytes do not contain axons. Ultimately we hope these models of neurons and their cellular components can be useful in simulation tools such as NEURON, Genesis and MCell to gain an understanding of neuronal organization and behavior.

4.6 Additional required work

Some steps remain to prepare this work for publication; model neurite thickness and curvature, support whole cell models learned from swc tracing files, quantitatively evaluate model fits of soma and nucleus, and model two large sets of neurons (at least 20 cells each) to demonstrate the differences between these cells and evaluate fits. In a subsequent publication we will extend these works to three-dimensional models, and develop models for the organelle localizations in neurites to extend the multi-component model presented here.

Chapter 5. Conclusions and future work

The goal of cellular systems biology is to understand the behavior and interaction of cells and their components. These interactions are dynamic and complex both temporally and morphologically. This thesis focused on developing models and methods for describing these morphological and biochemical complexities and dynamics using conditional generative models learned from imaging data.

5.1 Summary of results

In Chapter 2, several models describing cellular dynamics were developed. These models were subsequently applied to both synchronous frames of a movie and asynchronous snapshots containing a marker of temporal registration. This was done using the large deformation diffeomorphic metric mapping (LDDMM) method and was implemented within the CellOrganizer framework. We demonstrated that these not only captured significant variance over the dataset on which they were trained, but that they performed significantly better than a random walk when synthesizing novel movies describing cellular morphology changes with the most successful approach being the model of dimensional displacement conditionally dependent on location in the shape space. These models can be used to study of cellular dynamics in ways previous static modeling approaches could not capture, namely how cells transition between states including phases in cell-cycle, disease progression or treatment.

In Chapter 3, a high-throughput pipeline for spatially resolved modeling using physiologically plausible geometries and organizations generated using CellOrganizer was built. We demonstrated the utility of this pipeline by performing simulations on a number of cells from throughout an LDDMM shape-space and showed that the impact of cellular morphology on cellular response could be predicted using regression models over the set of outputs from MCell simulations. We also explored the impact of number of endosomes present in a cell to demonstrate additional spatial parameters that can be studied with this system. This system allows users to study the impact of cellular shape and organization on the biochemical interactions taking place within the cell to gain an understanding in what role this organization plays in cellular behaviors and disease. *[As stated at the conclusion of the chapter, an error was found in final editing that will change the quantitative results; however all of the methodological contributions discussed here remain valid.]*

In Chapter 4, the capabilities of CellOrganizer were expanded to include modeling of neuronal morphologies and their subcellular organization. We developed two models to describe neurite branching and neurite placement and used existing models in CellOrganizer to model nuclear and soma shape. This work enabled the flexible conditional modeling of neurons and their subcellular structures from imaging data. Despite previous approaches to generative modeling of neurite structure and neuronal morphology, this work represents the first use of multi-component conditional models that incorporate subcellular structure.

5.2 Thesis Contributions

1. We developed methods for the generative modeling of cellular dynamics at arbitrary temporal resolution learned from both synchronous movies and asynchronous cell populations containing a marker of temporal registration.

2. We demonstrated the ability of these models to generate *in silico* cell movies with higher accuracy than a random walk.

3. We implemented a high-throughput semi-automated pipeline for the spatially resolved modeling of cells and their components.

4. We demonstrated that this high-throughput pipeline could be used to study the impact of morphological and organizational variability within cells.

5. We made significant contributions to the development of the spatial extension to the Systems Biology Markup Language (SBML-spatial).

6. We implemented the first instances of SBML-spatial.

7. We implemented write support for SBML-spatial (v0.89) in CellOrganizer.

8 We implemented read support for SBML-spatial (v0.89) in CellBlender, an MCell interface.

9. We demonstrated the use of regressive modeling to predict dynamic response based on cellular morphology using results of spatially resolved simulations.

10. We developed a generative neurite parameterization using a stick breaking process.

11. We demonstrated this neurite parameterization's ability to distinguish between neurite types.

12. We developed a generative parameterization for the conditional placement of neurites around the soma.

13. We demonstrated the ability of CellOrganizer to learn and synthesize multi-component models from neuronal images and tracing data including nuclear morphology.

5.3 Future directions

This work is pursuing the ultimate goal of being able to model and understand any cell and its response to any condition. Particularly, we worked to enable the study of if and when cellular organization is a primary determinant of this response.

Towards that goal, dynamic models should be improved by expanding the modeling methods discussed to higher dimensionalities of the parameter space. This will allow for more accurate models that capture a larger amount of the variance across cells. The dimensional displacement model, the most accurate model tested, should be conditioned such that the displacement in the n^{th} dimension depends on displacement in all previous dimensions. In future work these dynamic models will be used to study the cellular response to specific stimuli such as a drug treatment.

Future directions include the building of dynamic models from different time scales and comparing cellular trends and accuracy of models across time scales by super-sampling or sub-sampling the frame rate. Using these different scaled models, hierarchical models that allow us to describe gross cellular changes such as division and rapid dynamics such as pseudopodia extension and retraction can then be built. We can also imagine that the modeling of single cells, which appeared more constrained in our models than the general population, with an approach that takes advantage of these single-cell constraints could be highly accurate. Along this vein we could develop models to identify and describe modes of cellular behavior and transitions within a heterogeneous population such as a tumor.

The development of conditionally dependent dynamic organelle models is another key future direction of this work. By modeling the dynamic relationship between cellular components and protein distributions we could gain a high-resolution understanding of how cells and the proteins within them function and interact.

These dynamic generative models have the potential to be useful in biochemical systems modeling where simulation tools such as VCell and MCell are working to support dynamic shape models. Simulating biochemical interaction networks within these realistic dynamic geometries will allow for the simulation of systems that were previously impossible to study using these techniques including endocytosis, cell migration, cell growth and cellular division. These dynamic models may also be useful in the future for inferring reaction networks from images as Evans et al. [31] explored.

This direction leads directly into the work done in Chapter 3. Further automation of the high-throughput spatial modeling pipeline should be explored. As mentioned, inferring reaction networks directly from imaging data may some day allow CellOrganizer to learn and generate whole spatially realistic biochemical simulation initializations internally. More immediately, we can begin to study the impact of spatial organization and morphology on known biochemical networks using the pipeline presented in Chapter 3. To do this we should look to large databases of biochemical systems models such as BioModels and explore the impact of spatial organization on biochemical response. This can be done by automatically searching for systems containing compartmental models where there exists a trained model within CellOrganizer for each necessary compartment such that complete spatially resolved SBML+SBML-spatial instances can be generated. We can then automatically generate parameter sweeps of morphology and organization and build models of the dependence of cellular response on organization.

Due to the vast number of spatial parameters that may be of interest in various systems, we envision a continuous simulation system that is capable of scraping biochemical models from databases such as BioModels. The pipeline could then translate these models into compartmental systems models, generate *in silico* geometric instances for compartments from CellOrganizer in an intelligent fashion (e.g. a parameter sweep) and simulate combined biochemical and spatial instances. Analysis of spatial organization's impact on the system could be performed automatically and statistically significant results could be sent to a database of responses. These responses could also be used to fit regressive models predicting the spatially dependent response of the biochemical species that was significantly dependent on spatial organization over the spatial parameter on which it depends.

Currently the pipeline uses MCell to perform simulations, however as systems begin to support the newly developed SBML-spatial package we should be able to apply this pipeline to additional modeling tools such as VCell, NEURON and Genesis.

High-throughput simulation within NEURON or Genesis requires a large number of neuronal spatial instances. This thesis represents the beginning of the work to develop multi-component models of neurons and their subcellular structure for use in this automated simulation pipeline. Before these models can be used in such a capacity they need to be improved by adding models for neurite curvature and thickness. These models also require more rigorous validation to determine their physiological realism before they can be used to study high-throughput spatially dependent biochemistry.

The high-throughput pipeline described in this work can function both as a simulation tool to learn about biochemistry and as a validation tool in the cases where the systems dynamics are well known.

5.4 Availability

All the code for work in this thesis is open source and will be publicly available via an upcoming release of CellOrganizer (v2.2) at cellorganizer.org.

Works Cited

1. Partin AW, Schoeniger JS, Mohler JL, Coffey DS (1989) Fourier analysis of cell motility: correlation of motility with metastatic potential. Proc Natl Acad Sci U S A 86: 1254-1258.
2. Weeraratna AT, Jiang Y, Hostetter G, Rosenblatt K, Duray P, et al. (2002) Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma. Cancer Cell 1: 279-288.
3. Mitra SK, Hanson DA, Schlaepfer DD (2005) Focal adhesion kinase: in command and control of cell motility. Nat Rev Mol Cell Biol 6: 56-68.
4. Partin MH (1987) A preliminary conceptual framework for the design, development, and use of client-oriented information systems in health. J Med Syst 11: 205-217.
5. Belletti B, Nicoloso MS, Schiappacassi M, Berton S, Lovat F, et al. (2008) Stathmin activity influences sarcoma cell shape, motility, and metastatic potential. Mol Biol Cell 19: 2003-2013.
6. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, et al. (2006) Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. Nat Methods 3: 525-531.
7. Schweizer J, Loose M, Bonny M, Kruse K, Monch I, et al. (2012) Geometry sensing by self-organized protein patterns. Proc Natl Acad Sci U S A 109: 15283-15288.
8. Boland MV, Markey MK, Murphy RF (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 33: 366-375.
9. Ahmad SM, Busser BW, Huang D, Cozart EJ, Michaud S, et al. (2014) Machine learning classification of cell-specific cardiac enhancers uncovers developmental subnetworks regulating progenitor cell division and cell fate specification. Development 141: 878-888.
10. Ozolek JA, Tosun AB, Wang W, Chen C, Kolouri S, et al. (2014) Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. Med Image Anal 18: 772-780.
11. Murphy RF (2004) Automated interpretation of protein subcellular location patterns: implications for early cancer detection and assessment. Ann N Y Acad Sci 1020: 124-131.
12. Murphy RF (2005) Cytomics and location proteomics: automated interpretation of subcellular patterns in fluorescence microscope images. Cytometry A 67: 1-3.
13. Murphy RF (2005) Location proteomics: a systems approach to subcellular location. Biochem Soc Trans 33: 535-538.
14. Murphy RF (2008) Automated Proteome-Wide Determination of Subcellular Location Using High Throughput Microscopy. Proc IEEE Int Symp Biomed Imaging 2008: 308-311.
15. Murphy RF (2010) Communicating subcellular distributions. Cytometry A 77: 686-692.
16. Li J, Newberg JY, Uhlen M, Lundberg E, Murphy RF (2012) Automated analysis and reannotation of subcellular locations in confocal images from the Human Protein Atlas. PLoS One 7: e50514.
17. Newberg J, Hua J, Murphy RF (2009) Location proteomics: systematic determination of protein subcellular location. Methods Mol Biol 500: 313-332.
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
19. Peng T, Murphy RF (2011) Image-derived, three-dimensional generative models of cellular organization. Cytometry A 79: 383-391.
20. Peng T, Wang W, Rohde GK, Murphy RF (2009) Instance-Based Generative Biological Shape Modeling. Proc IEEE Int Symp Biomed Imaging 5193141: 690-693.
21. Shariff A, Murphy RF, Rohde GK (2010) A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images. Cytometry A 77: 457-466.
22. Zhao T, Murphy RF (2007) Automated learning of generative models for subcellular location: building blocks for systems biology. Cytometry A 71: 978-990.

23. Cuntz H, Forstner F, Borst A, Hausser M (2010) One rule to grow them all: a general theory of neuronal branching and its practical application. PLoS Comput Biol 6: e1000877.
24. Rohde GK, Ribeiro AJ, Dahl KN, Murphy RF (2008) Deformation-based nuclear morphometry: capturing nuclear shape variation in HeLa cells. Cytometry A 73: 341-350.
25. Buck TE, Li J, Rohde GK, Murphy RF (2012) Toward the virtual cell: automated approaches to building models of subcellular organization "learned" from microscopy images. Bioessays 34: 791-799.
26. Khan A, Aylward E, Barta P, Miller M, Beg MF (2005) Semi-automated basal ganglia segmentation using large deformation diffeomorphic metric mapping. Med Image Comput Comput Assist Interv 8: 238-245.
27. Dixit R, Cyr R (2003) Cell damage and reactive oxygen species production induced by fluorescence microscopy: effect on mitosis and guidelines for non-invasive fluorescence microscopy. Plant J 36: 280-290.
28. Hu Y. O-HE, Hua J., Nowiki T. S., Stoltz R., McKayle C., and Murphy R. F. (2010) Automated analysis of protein subcellular location in time series images. Bioinformatics 26: 1630-1636.
29. Lacayo CI, Pincus Z, VanDuijn MM, Wilson CA, Fletcher DA, et al. (2007) Emergence of large-scale cell morphology and movement from local actin filament growth dynamics. PLoS Biol 5: e233.
30. Bakal C, Aach J, Church G, Perrimon N (2007) Quantitative morphological signatures define local signaling networks regulating cell morphology. Science 316: 1753-1756.
31. Evans L, Sailem H, Vargas PP, Bakal C (2013) Inferring signalling networks from images. J Microsc 252: 1-7.
32. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2004) BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. Bioinformatics 20: 3289-3291.
33. T. GD (1977) Exact stochastic simulation of coupled chemical reactions J Phys Chem 81: 2340-2361.
34. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440: 358-362.
35. Loew LM, Schaff JC (2001) The Virtual Cell: a software environment for computational cell biology. Trends Biotechnol 19: 401-406.
36. Moraru, II, Schaff JC, Slepchenko BM, Blinov ML, Morgan F, et al. (2008) Virtual Cell modelling and simulation software environment. IET Syst Biol 2: 352-362.
37. Andrews SS (2012) Spatial and stochastic cellular modeling with the Smoldyn simulator. Methods Mol Biol 804: 519-542.
38. Andrews SS, Addy NJ, Brent R, Arkin AP (2010) Detailed simulations of cell biology with Smoldyn 2.1. PLoS Comput Biol 6: e1000705.
39. Dematte L (2012) Smoldyn on graphics processing units: massively parallel Brownian dynamics simulations. IEEE/ACM Trans Comput Biol Bioinform 9: 655-667.
40. Kerr RA, Bartol TM, Kaminsky B, Dittrich M, Chang JC, et al. (2008) Fast Monte Carlo Simulation Methods for Biological Reaction-Diffusion Systems in Solution and on Surfaces. SIAM J Sci Comput 30: 3126.
41. Stiles JR, Van Helden D, Bartol TM, Jr., Salpeter EE, Salpeter MM (1996) Miniature endplate current rise times less than 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle. Proc Natl Acad Sci U S A 93: 5747-5752.
42. Callenberg KM, Latorraca NR, Grabe M (2012) Membrane bending is critical for the stability of voltage sensor segments in the membrane. J Gen Physiol 140: 55-68.
43. Ascoli GA, Donohue DE, Halavi M (2007) NeuroMorpho.Org: a central resource for neuronal morphologies. J Neurosci 27: 9247-9251.

44. Halavi M, Polavaram S, Donohue DE, Hamilton G, Hoyt J, et al. (2008) NeuroMorpho.Org implementation of digital neuroscience: dense coverage and integration with the NIF. Neuroinformatics 6: 241-252.
45. Nanda S, Allaham MM, Bergamino M, Polavaram S, Armananzas R, et al. (2015) Doubling up on the Fly: NeuroMorpho.Org Meets Big Data. Neuroinformatics 13: 127-129.
46. Eberhard JP, Wanner, A., & Wittum, G. (2006) NeuGen: A tool for the generation of realistic morphology of cortical neurons and neural networks in 3D. Neurocomputing 70(1-3): 327-342.
47. Koene R. A. TB, van Hees P. (2009) NETMORPH: A Framework for the Stochastic Generation of Large Scale Neuronal Networks With Realistic Neuron Morphologies. Neuroinformatics 7(3): 195-210.
48. Carnevale NT, Hines ML (1997) The NEURON simulation environment. 9: 1179-1209.
49. Hines ML, Morse TM, Carnevale NT (2007) Model structure analysis in NEURON : toward interoperability among neural simulators. Methods Mol Biol 401: 91-102.
50. Adorjan P, Barna G, Erdi P, Grobler T, Kepecs A, et al. (1996) Multicompartmental modeling of hippocampal pyramidal cells and interneurons with the GENESIS software tool. Neurobiology (Bp) 4: 247-249.
51. Schneider CJ, Bezaire M, Soltesz I (2012) Toward a full-scale computational model of the rat dentate gyrus. Front Neural Circuits 6: 83.
52. Santhakumar V, Aradi I, Soltesz I (2005) Role of mossy fiber sprouting and mossy cell loss in hyperexcitability: a network model of the dentate gyrus incorporating cell types and axonal topography. J Neurophysiol 93: 437-453.
53. Zhong Q, Busetto AG, Fededa JP, Buhmann JM, Gerlich DW (2012) Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. Nat Methods 9: 711-713.
54. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, et al. (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54: 2033-2044.
55. Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12: 26-41.
56. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, et al. (2009) Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46: 786-802.
57. Avants B. B. TN, and Song G. (2009) Advanced Normalization Tools (ANTS). Insight J: 1-35.
58. L J P van der Maaten EOP, H J van der Herik (2009) Dimensionality reduction: A comparitive review. Tilburg University Technical Report TICC-TR: 2009-2005.
59. Buck TE, Rao A. Coelho L. P., Murphy R. F. (2008) Cell cycle dependence of protein subcellular location inferred from static, asynchronous images. 1016-1019.
60. Harper JV (2005) Synchronization of cell populations in G1/S and G2/M phases of the cell cycle. Methods Mol Biol 296: 157-166.
61. Posakony JW, England JM, Attardi G (1977) Mitochondrial growth and division during the cell cycle in HeLa cells. J Cell Biol 74: 468-491.
62. Murata S, Herman P, Lakowicz JR (2001) Texture analysis of fluorescence lifetime images of AT- and GC-rich regions in nuclei. J Histochem Cytochem 49: 1443-1451.
63. Sakaue-Sawano A, Kurokawa H, Morimura T, Hanyu A, Hama H, et al. (2008) Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. Cell 132: 487-498.
64. Yin Z, Kanade T, Chen M (2012) Understanding the phase contrast optics to restore artifact-free microscopy images for segmentation. Med Image Anal 16: 1047-1062.
65. Bise R, Kanade T, Yin Z, Huh SI (2011) Automatic cell tracking applied to analysis of cell migration in wound healing assay. Conf Proc IEEE Eng Med Biol Soc 2011: 6174-6179.

66. Brahmbhatt AA, Klemke RL (2003) ERK and RhoA differentially regulate pseudopodia growth and retraction during chemotaxis. J Biol Chem 278: 13016-13025.
67. Bayer N, Schober D, Prchla E, Murphy RF, Blaas D, et al. (1998) Effect of bafilomycin A1 and nocodazole on endocytic transport in HeLa cells: implications for viral uncoating and infection. J Virol 72: 9645-9655.
68. Fink CC, Slepchenko B, Moraru, II, Watras J, Schaff JC, et al. (2000) An image-based model of calcium waves in differentiated neuroblastoma cells. Biophys J 79: 163-183.
69. Kholodenko BN, Hancock JF, Kolch W (2010) Signalling ballet in space and time. Nat Rev Mol Cell Biol 11: 414-426.
70. Hernjak N, Slepchenko BM, Fernald K, Fink CC, Fortin D, et al. (2005) Modeling and analysis of calcium signaling events leading to long-term depression in cerebellar Purkinje cells. Biophys J 89: 3790-3806.
71. Cowan AE, Moraru, II, Schaff JC, Slepchenko BM, Loew LM (2012) Spatial modeling of cell signaling networks. Methods Cell Biol 110: 195-221.
72. White DE, Kinney MA, McDevitt TC, Kemp ML (2013) Spatial pattern dynamics of 3D stem cell loss of pluripotency via rules-based computational modeling. PLoS Comput Biol 9: e1002952.
73. Sorkin A, von Zastrow M (2009) Endocytosis and signalling: intertwining molecular networks. Nat Rev Mol Cell Biol 10: 609-622.
74. Sorkina T, Richards TL, Rao A, Zahniser NR, Sorkin A (2009) Negative regulation of dopamine transporter endocytosis by membrane-proximal N-terminal residues. J Neurosci 29: 1361-1374.
75. Booth-Gauthier EA, Du V, Ghibaudo M, Rape AD, Dahl KN, et al. (2013) Hutchinson-Gilford progeria syndrome alters nuclear shape and reduces cell motility in three dimensional model substrates. Integr Biol (Camb) 5: 569-577.
76. Yin Z, Sadok A, Sailem H, McCarthy A, Xia X, et al. (2013) A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol 15: 860-871.
77. Gabella C, Bertseva E, Bottier C, Piacentini N, Bornert A, et al. (2014) Contact angle at the leading edge controls cell protrusion rate. Curr Biol 24: 1126-1132.
78. Neves SR, Tsokas P, Sarkar A, Grace EA, Rangamani P, et al. (2008) Cell shape and negative links in regulatory motifs together control spatial information flow in signaling networks. Cell 133: 666-680.
79. Sbalzarini IF (2013) Modeling and simulation of biological systems from image data. Bioessays 35: 482-490.
80. Sbalzarini IF, Hayer A, Helenius A, Koumoutsakos P (2006) Simulations of (an)isotropic diffusion on curved biological surfaces. Biophys J 90: 878-885.
81. Slepchenko BM, Loew LM (2010) Use of virtual cell in studies of cellular dynamics. Int Rev Cell Mol Biol 283: 1-56.
82. Coelho LP, Shariff A, Murphy RF (2009) Nuclear Segmentation in Microscope Cell Images: A Hand-Segmented Dataset and Comparison of Algorithms. Proc IEEE Int Symp Biomed Imaging 5193098: 518-521.
83. Murphy RF (2012) CellOrganizer: Image-derived models of subcellular organization and protein distribution. Methods Cell Biol 110: 179-193.
84. Takahashi K, Arjunan SN, Tomita M (2005) Space in systems biology of signaling pathways--towards intracellular molecular crowding in silico. FEBS Lett 579: 1783-1788.
85. Ramaswamy R, Gonzalez-Segredo N, Sbalzarini IF (2009) A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks. J Chem Phys 130: 244104.

86. Harris LA, Hogg, Justin S., Faeder, James R. (2009) Signal transduction with receptor internalization and transcriptional regulation. In Proceedings of the 2009 Winter Simulation Conference: 908-919.
87. Milo R, Jorgensen P, Moran U, Weber G, Springer M (2010) BioNumbers--the database of key numbers in molecular and cell biology. Nucleic Acids Res 38: D750-753.
88. Kholodenko BN, Demin OV, Moehren G, Hoek JB (1999) Quantification of short term signaling by the epidermal growth factor receptor. J Biol Chem 274: 30169-30181.
89. Berkers JA, van Bergen en Henegouwen PM, Boonstra J (1991) Three classes of epidermal growth factor receptors on HeLa cells. J Biol Chem 266: 922-927.
90. Li JJ, Bickel PJ, Biggin MD (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2: e270.
91. Oberleithner H, Brinckmann E, Schwab A, Krohne G (1994) Imaging nuclear pores of aldosterone-sensitive kidney cells by atomic force microscopy. Proc Natl Acad Sci U S A 91: 9784-9788.
92. Hubner S, Xiao CY, Jans DA (1997) The protein kinase CK2 site (Ser111/112) enhances recognition of the simian virus 40 large T-antigen nuclear localization sequence by importin. J Biol Chem 272: 17191-17195.
93. Bornstein BJ, Keating SM, Jouraku A, Hucka M (2008) LibSBML: an API library for SBML. Bioinformatics 24: 880-881.
94. Solovyev A, Mi Q, Tzen YT, Brienza D, Vodovotz Y (2013) Hybrid equation/agent-based model of ischemia-induced hyperemia and pressure ulcer formation predicts greater propensity to ulcerate in subjects with spinal cord injury. PLoS Comput Biol 9: e1003070.
95. Zwier M.C. AJL, Kaus J.W., Pratt A.J., Wong K.F., Rego N.B., Suárez E., Lettieri S., Wang D.W., Grabe M., Zuckerman D.M., and Chong L.T. (2015) WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. J Chem Theory Comput 11: 800-809.
96. Maxfield FR (2014) Role of endosomes and lysosomes in human disease. Cold Spring Harb Perspect Biol 6: a016931.
97. Murphy RF, Powers S, Cantor CR, Pollack R (1984) Reduced insulin endocytosis in serum-transformed fibroblasts demonstrated by flow cytometry. Cytometry 5: 275-280.
98. Roederer M, Murphy RF (1986) Cell-by-cell autofluorescence correction for low signal-to-noise systems: application to epidermal growth factor endocytosis by 3T3 fibroblasts. Cytometry 7: 558-565.
99. Cajal S. Ramon y AL (1894) Les nouvelles idées sur la structure du système nerveux chez l'homme et chez les vertébrés. Reinwald, Paris.
100. Helmstaedter M, Briggman KL, Denk W (2011) High-accuracy neurite reconstruction for high-throughput neuroanatomy. Nat Neurosci 14: 1081-1088.
101. Adams R. P GZ, Jordan M. I (2010) Tree-Structured Stick Breaking for Hierarchical Data. Advances in Neural Information Processing Systems, arXiv:10061062 23: 19-27.
102. R. G (1993) Martingale Functional Central Limit Theorems for a Generalized Pòlya Urn. The Annals of Probability 21: 1624-1639.
103. Parekh R, Ascoli G. R. (2014) Neuronal Morphology goes Digital: A Research Hub for Cellular and System Neuroscience. Neuron 77: 1017-1038.