# A Modeling Study to Characterize Microtubule Mechanisms of Dynamic Instability: Connecting Micro-Level Tip Structures to Macro-Level Phases

**Shant M. Mahserejian**

**Publication Date**

13-04-2017

# A MODELING STUDY TO CHARACTERIZE MICROTUBULE MECHANISMS OF DYNAMIC INSTABILITY: CONNECTING MICRO-LEVEL TIP STRUCTURES TO MACRO-LEVEL PHASES

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Shant M. Mahserejian

---

Dr. Mark Alber, Director

Graduate Program in Applied and Computational Mathematics and Statistics
Notre Dame, Indiana

April 2017

# A MODELING STUDY TO CHARACTERIZE MICROTUBULE MECHANISMS OF DYNAMIC INSTABILITY: CONNECTING MICRO-LEVEL TIP STRUCTURES TO MACRO-LEVEL PHASES

Abstract

by

Shant M. Mahserejian

Microtubules (MTs) are cytoplasmic biopolymers that are common in eukaryotic cells. The MT is assembled by $\alpha\beta$ tubulin dimer subunits that can be in either a GTP- or GDP-bound nucleotide state. These dimer subunit connect with longitudinal bonds to form linear strands called protofilements (PFs). Lateral bonds connect 13 PFs together to form the tube-like structure of a MT. GTP-bound subunits collect near the MT tip region to form a GTP-cap, which helps maintain the bonds that hold the MT structure intact. Losing the GTP-cap exposes GDP-bound subunits which are more likely to break their bonds, and promote subunits to detach from the MT structure. The MT length changes in time by undergoing spontaneous switches between periods of sustained growth and rapid shortening, which characterize the behavior called dynamic instability (DI).

The molecular reactions that drive MT dynamics primarily affect the tip portion of the structure. Therefore, a study of the connection between MT tip structures and macro-level phases is needed to gain a better understanding of the mechanisms that drive phase changes in DI. Laboratory conditions limit the level of detail that can be experimentally collected from MT structures. Computational models are a vital tool that provide this level of information, and they have helped understand

how molecular level reactions alter the micro-level MT structure, which drives the MT length changes observed at the macro-level. The detailed 13-PF MT model was capable of running long-time simulations that display DI behavior with a low computational cost, but it made use of an approximation that skips over MT structural states. This study first develops the extended 13-PF MT model in order to simulate a biochemically exact trajectory of all the MT structural states resulting from possible reactions events. Then, the minimal MT structure that includes the lateral bond is considered to present the simplified 2-PF MT model, a novel consideration which helps make calculations of the MT tip structure features more feasible while successfully simulating DI behavior.

The high frequency and low amplitude fluctuations present in simulated MT length history data make it difficult to pinpoint where DI phases begin and end, and where phase transitions occur. To this end, an unsupervised machine learning method based on $K$-means clustering is presented to identify, classify, and analyze macro-level phases present in MT length history data. Application of this method revealed an intermediate phase called "stutters", during which the rate of MT length change is smaller in magnitude compared to classically recognized growth and shortening phases. Additionally, stutter phases commonly appeared as a transitional phase during catastrophe events, between growth and shortening phases. This indicated that before a catastrophe event takes place, a MT is likely to first undergo structural changes that do not alter the MT length, which result in structural configurations prone to entering a period of rapid depolymerization. The proposed DI phase classification method now can identify these periods, which in past experimental studies have been observed, but not separately considered as a unique class of behavior [21]. Furthermore, the stutter events specifically provide a target region to study the mechanisms involved with catastrophe events.

Finally, a supervised machine learning approach called Random Forest was used

to test the ability for micro-level tip structure features to predict their corresponding macro-level DI phases, and to forecast upcoming phase transitions. The results indicated that the GTP-cap size and it's relative position to the cracked tip region are important factors in predicting which DI phase a MT is in. In addition to the GTP-cap size, information on the PF-tip lengths and the dispersion of GTP-bound subunits in the tip region were found to be important in forecasting upcoming phase transitions. Thus, specific MT tip structures and the reaction events that create them are identified as the mechanisms that drive respective transitions between DI phases.

This dissertation is dedicated to my family, past and present, who have sacrificed so much for my opportunity to pursue higher academic degrees, and whose sense of pride has been my joyous reason for doing so.

CONTENTS

FIGURES

TABLES

# ACKNOWLEDGMENTS

I would like to first thank my advisor, Dr. Mark Alber, for his guidance during my endeavors of exploring interdisciplinary science through the lens of applied and computational mathematics.

I would like to thank Dr. Holly Goodson for her countless hours of support and instructions in helping shape part of the ongoing microtubule project into my dissertation work.

I would like to thank Dr. Jun Li for lending his statistical expertise, which helped guide the data exploration of the microtubule studies into uncharted territories.

I would like to thank Dr. Alexandra Jilkine for participating in my defense committee, her helpful comments and suggestions, and for her supplemental instruction that helped me to better learn about mathematical applications in biology.

I would like to thank Dr. Joshua Shrout for his guidance in the experimental studies of bacteria and the data collection involved with microscopy and image analysis during my first years at Notre Dame.

I would like to thank the many professors at Notre Dame that guided my doctoral education, and helped expand my knowledge base in quantitative sciences.

I would like to thank the ACMS Department and staff for their tireless work, which has not only provided access to the many projects offered at Notre Dame, but also made my participation in them as effortless a process as possible.

I would like to thank my past and present colleagues at Notre Dame for their collaborations and friendship, including but not limited to Dr. Ava Mauro, Dr. Erin Jonasson, Dr. Cameron Harvey, Dr. Oleg Kim, Dr. Ali Nemetbakhsh, Dr. Shixin

Xu, Dr. Ling Xu, Dr. Daniel Brake, Dr. Amy Buchmann, Dr. Wenzhao Sun, Dr. Chunlei Li, Dr. Timur Kupaev, Dr. Aboutaleb Amiri, Dr. Michael Machen, Francesco Pancaldi, Martin Barron, Jianxu Chen, Chinedu Madukoma, Bide Xiong, Christopher Ebsch, Daniel Howard, Oyekola Oyekole, Karly Harrod, and Bryant Vande Kolk.

I would like to thank the Armenian community at Notre Dame, including Dr. Ani Aprahamian, Dr. Khachatur Manukyan, Jesse Arlen, Alan Grigorian, Armen Gyurjinyan, Matteos Matigian, Grigory Rchtouni, and Mike Gedjeyan for allowing me to keep in touch with my roots in a place far from home.

I would like to thank the members of Element Band and Modiviccan, who have allowed my musical spirit to live on during my days as a graduate student.

I would like to thank my former advisor Dr. Rabia Djellouli for his continued support and encouragement for my pursuit of a doctoral degree, and my involvement in interdisciplinary research.

I would like to thank my best friend, Dr. Sevan Krikor Gulesserian, for being an example as well as a fellow journeyman as we earned our doctoral degrees.

I would like to thank the love of my life, my fiance, Garine Gilabochian, whose love and undying support played a significant role on a daily basis to maintain my path toward graduation.

I would like to thank my family who have encouraged and cheered on my progress towards graduation, and especially my parents, Hovig and Ani Mahserejian, who have instilled in me the Armenian spirit that has fueled my inner desire and strength needed to tackle the more difficult, but worthwhile challenges in life.

Finally, I would like to thank the Dolores Zohrab Liebmann Fellowship for funding my doctoral studies, and for including my name as part of the Zohrab family story, which continues to inspire my journey as an Armenian citizen of the world determined to do the kind of work that makes a positive impact benefiting all.

# CHAPTER 1

# INTRODUCTION

Microtubules (MTs) are cytoplasmic biopolymers found in eukaryotic cells. As cytoskeletal components, MTs play a crucial role in cell shape structural support, and in cellular processes such as cell division and organelle transportation. MTs collectively perform these functions as part of a dynamic network within the cytoskeleton, but each MT independently undergoes its own dynamic changes. More specifically, they portray a unique behavior called dynamic instability (DI), classically characterized by repeated changes between growth and shortening phases called catastrophes and rescues (see Figure 1.1) [17, 45, 62]. Interrupting this DI behavior in MTs can result in diseases, such as Alzheimer's, Parkinson's, and some forms of cancer [23, 33, 35]. DI behavior is a result of changes to the MT structure, a tube-like polymer whose walls are constructed with protofilament (PF) strands bound together in parallel (see Figures 2.1 and 2.2). Currently, the mechanisms of transitions in DI behavior are poorly understood, and a clear connection between micro-level structures and macro-level DI behavior has not yet been made.

In order to study MT behavior at a deeper level, computational models have accompanied biological experiments to help learn more about the mechanisms leading to the macro-level dynamics characterizing DI behavior. However, the complexity of the MT structure (illustrated in Figure 2.1) needed to be simplified in order to begin modeling the fundamental features of MT dynamics. Earlier, coarse grained models were introduced by reducing the structure to a single-PF, and provided a useful tool for studying bio-polymers in general. In the case of MT structures, which typically

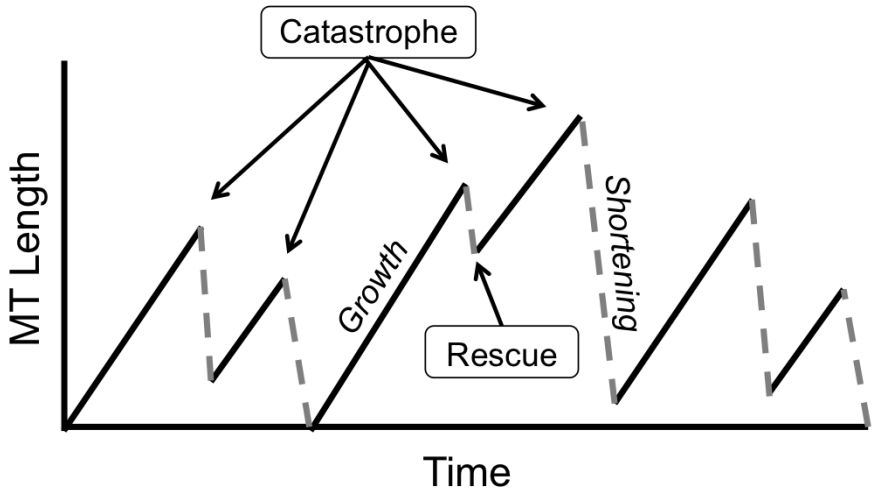



Figure 1.1. An illustration of a microtubule (MT) length history plot demonstrating dynamic instability (DI). The MT length undergoes sustained periods of growth (solid) and more rapid shortening (dashed). The transitions from growth to shortening are called catastrophes, since the MT will likely depolymerize down to having a length near zero. The rare transitions from shortening to growth is called a rescue, since it saves the bioploymer from losing most of its structural mass.

consist of 13 PFs, the course grained models like the one presented in [32, 54] approximate the polymer to have just a single-PF. A subunit in the single-PF model represents the dynamical behavior as an approximation of a row of subunits across all 13-PFs found in a MT. Each subunit can be in one of two nucleotide bounded states: the energy carrying guanosine-triphosphate (GTP) bound state that promotes stable bonds, and the low energy guanosine-diphosphate (GDP) bound state that promotes breaking bonds. The simplified model has been successful in demonstrating DI behavior, while simultaneously making simulations computationally affordable. This was mainly a result of the interaction between polymerization/depolymerization rates from the single-PF structure, and the hydrolysis rates changing the subunit states. Additionally, the resulting correlation found between shortening phases and losing the GTP-cap agreed with experimental observations, which revealed little to no GTP-

bound subunits present during shortening phases [10, 20, 40, 44, 62, 79–81]. This computational model has been beneficial in studying the bulk behavior of MTs as well as their competition for resources among multiple MTs, since it is easy to simulate the dynamics of a large population [54]. Furthermore, this simplification offers an attractive model for studying other biopolymers as well. For example, a slight modification to this model can represent another cytoskeletal component called actin, and can successfully simulate its characteristic dynamic behavior called treadmilling [43]. Recent studies with single-PF models consider the critical concentrations and the parameter ranges typical of the various behaviors that are possible when considering both DI and treadmilling [43].

Upon gaining a better understanding of the fundamental role of the kinetic rates needed to display DI behavior, new models were developed to represent more of the complex features seen in actual MT structures. By including the entire 13-PF structure into the MT model, the lateral bonds that kept the tubular structure's walls intact inevitably had to be considered. The first attempts in using the 13-PF models limited the structure to certain profiles, and only considered certain tip configurations observed during growth and shortening phases separately (see Figure 2.4). These models were used to study the short-term behavior promoted by starting simulations from blunt or tapered MT tips [14, 27, 85]. This limited structure model was a good start, however they were lacking the dynamical considerations to simulate phase transitions due to the computational limitations that only allowed for short-duration simulations.

More recently, a mechanical model presented by Odde et al. added more details to the MT model by allowing for energy minimization to adjust the structural configuration after every dynamical event that added/removed subunits to/from the biopolymer [77]. Though this considered a greater level of detail compared to older models, the computational cost of performing the energy minimization alone makes

it an unfeasible option for studying macro-level phase transitions in MT dynamics. Also, this model made a simplifying assumption, where new subunits form two bonds simultaneously: longitudinal bonds with the subunits below, and lateral bonds with neighboring subunits on one side. This assumption may have been reasonable to implement the simulations and demonstrate the importance of losing a GTP-cap to promote a shortening phase, but it is biochemically unlikely for multiple bonds to form simultaneously, especially when considering the timescale during which energy minimization occurs (see Section 3.5.1 for details). Furthermore, they too limited their scope to the variety in structural features that are observed with MTs in the laboratory, and their studies did not consider long term MT dynamics [63].

One of the key structural features observed in experiments are the strands of individual PFs that protruded from the tip of the MT structure, and create a frayed profile [53]. These individual PF extensions, which resemble "ram horns" due to the internal curvature created by the intrinsic bend found in GDP-bound subunits, can only be created if lateral bonds form or break separately from other bond formations. Recently, a statistical mechanics approach has studied these frayed tip profiles using a limited structure format, by only considering shortening phases [41]. However, for a model to truly consider the full range of MT dynamics that form frayed tips, lateral bonds need to be modeled so that they can break sequentially even during growth phases. The only model, to our knowledge, that makes this consideration is the detailed level 13-PF MT model [56]. This model has successfully simulated long-term DI behavior similar to biologically relevant MT systems, including phase transitions with low computational cost, as a result of micro-level reaction events being considered. In addition, the frayed tip profile is represented through cracks, or laterally unbonded sections between PFs near the tip of the MT structure. The resulting simulated data from this model indicated that the state of the subunits at the bottom of these cracks is important in determining whether the MT would

continue in a growth or shortening phase, or possibly undergo a transition in the short-term.

By tracing the history of computational models that study MT dynamics, the need for exposing the structural complexities associated with dynamic behavior is clear. In particular, the next step would include the development of a model that displays the appropriate level of dynamics, such that enough structural details are visible, while simultaneously generating enough data to reveal DI behavior. In this dissertation, a computational modeling approach is used to meet these requirements, and the resulting *in silico* data is used for analysis.

First, the detailed level 13-PF model is extended by removing an approximation to the treatment of hydrolysis, which was originally introduced to save on computational cost. Hydrolysis events were updated so that they are treated in the same manner as polymerization, depolymerization, lateral bond formation, and lateral bond breakage events. This correctly executes the Gillespie algorithm, where the simulated sequence of reaction events outputs a bio-chemically exact trajectory of MT structural states. Furthermore, the code was optimized to calculate hydrolysis rates more efficiently, bringing performance and computational cost of the exact method to the same level as the approximated version. This extended 13-PF MT model now makes it possible to create long-time simulations, and thus to study a wide variety of structures occurring over many phase transitions, while also providing data on MT structure details that are not available in the laboratory setting.

Second, in order to study the tip region where most of the relevant dynamical changes to the structure occur, a simplification is desired to reduce the sheer number of distinct tip configurations that exist with the 13-PF model, but without losing the complexity and relevance offered by its treatment of lateral bonds. To this end, a 2-PF model is introduced, where the lateral bond and the consequential cracks in the tip region are still present. This model provides a more tractable scenario

for studying possible tip structures, and to identify which patterns are relevant to changes in macro-level dynamics of MTs.

Next, analysis of the macro-level results from the extended 13-PF model and the 2-PF model required a data-driven statistical approach. Prior methods of approximating DI phases made use of a pencil and straight-edge approach to identify periods of consistent growth or shortening, and choosing transition points relied on a human eye. Some automated approaches segmented different periods by seeking a minimum height change based on *a priori* assumptions of what growth and shortening rates would be. When attempting to use these methods in conjunction with the detailed level 13-PF model, the approximation failed to be accurate in several moments of intermediate dynamics, which were observed to have slopes in the length history plots smaller in magnitude than those expected from traditional growth or shortening phases. Also, the resulting approximations identified transition points very poorly, especially since part of the goal of this study is to observe the structural features associated with exact moments of macro-level changes. For this reason, an improved automated method creating a continuous piece-wise linear approximation is needed to adaptively handle the stochastically occurring transitions observed during DI behavior, as introduced in this dissertation. Once periods of consistent behavior were accurately segmented from length history plots, an unsupervised classification method labels each segment into appropriate phases, while being blind to the bi-phase assumptions of past methods. This automated approximation and classification model is applicable to any DI data extracted from either experiment or from simulation. Using this approach on simulated data revealed the existence of a third DI phase, referred here as "stutters", which represent periods of consistent behavior that have a smaller magnitude rate of change to MT length compared to classical growth or shortening periods. Furthermore, stutters frequently appear as a transitional phase between growth and shortening during catastrophe events. This may represent the

"slow-down" period before shortening periods, which have observed but not quantified experimentally [21]. The presented DI phase classification method now allows for the separate treatment of these periods that have been overlooked in past studies.

Finally, in order to make a connection between the detailed level tip structure features measured from the 2-PF MT model (see an example of a 2-PF MT tip structure in Figure 3.3) to the macro-level dynamic phases identified by the aforementioned classification model, a machine learning approach is introduced to develop a prediction and forecasting model. This method of data analysis carefully considers structural features of the 2-PF tip region that are bio-chemically relevant, have been shown to be important in past literature, and are easy to extend to a 13-PF case once shown to be worthwhile in the simplified model. The predictive model makes a connection between the structural features and DI phases, while the forecasting model determines if a sequence of structural features would continue the same phase, or trigger a transition.

The use of computational models in studying the biological system of MTs is crucial in gaining an understanding of the structural mechanisms connected with DI phases, and the causes for DI phase transitions. First, the scope of structural detail provided by simulations is not attainable experimentally. Technological limitations, both in temporal and spatial resolution, do not allow for the real-time tracking of structural configurations, nor the exact organization of subunit states that make up the MT. The computational model following the Gillespie algorithm as presented in this dissertation does provide this level of detail. Using a long-time simulation and observing the structural features during a large number of phase transitions provides an appropriate setting to use a combination of mathematical and statistical modeling techniques to identify the underlying patterns leading to DI phase transitions. Previous work has hypothesized on some structural characteristics that would promote growth or shortening phases, such as the loss of a GTP-cap. However, little is under-

stood about the exact order in which the structural changes occur, and whether those structural changes are responsible for macro-level dynamic changes, or are a result of them. Computational models also allow for flexibility in considering theoretical configurations, adjusting rate parameters, and controlling conditions to test hypotheses that lead to a better understanding of how phase transitions occur. Hence, diving deeper into simulated data during moments near phase transitions, and particularly during intermediate phases that are not quite growth or shortening, is the proposed idea for seeking the mechanisms that dictate DI behavior observed in MT dynamics.

In Chapter 2, the biological details of MTs and the problem that they pose is laid out, which are the important considerations for modeling the MT structure and its dynamics. Chapter 3 outlines the details of the computational model developed and used in this study, how it is implemented, and how this leads to the simplified 2-PF MT version. Chapter 4 describes the first data analysis from a macro-level perspective, which can be used to approximation and classify DI phase segments found in the length history of any data representing DI behavior. Chapter 5 describes the second data analysis approach, which bridges different time scales by predicting macro-level DI phase behavior from the micro-level structural features of the 2-PF MT tip region. Finally, Chapter 6 consolidates the conclusions about how the structural features of a MT tip are related to the length profile changes in time.

# CHAPTER 2

# BIOLOGICAL AND MODELING BACKGROUND

In this chapter, the necessary biological considerations are presented to help the development of the microtubule (MT) models used in this study. An important aspect of the mathematical representation is to utilize results from experimental scenarios, and reproducing them in computational simulations. Computational modeling approaches to verify results are beneficial in providing a deeper understanding of the mechanics involved in the changes that take place in a biological system. Existing models have provided some insight to which MT structures relate to different general dynamics, though they have not yet satisfied a complete understanding of which mechanisms lead to significant shifts during dynamic instability behavior in MTs. In order to ensure the correct level of detail is included in this study, the biological perspective of what is truly understood about MTs is first reviewed in order to develop the computational model, and the subsequent simulated data will then be used for identifying the MT structure features associated to dynamic phase transitions.

## 2.1 Description of Microtubules (MTs)

The MT structure is the central focus of this study. This section reviews the components that make up the MT structure at the dimer subunit level, as well as the dynamic events that lead to altering the structure at this level. These micro-level events accumulate over longer periods of time, and result in changes to MT length that is commonly studied at the macro-level. In this farther vantage point, the length history plot displays the characteristic behavior of MTs called dynamic instability, the

sporadic and sudden changes between periods of growth and shortening. However, it is a sequence of reaction events at the molecular level that alter the structural integrity of the MT structure, and make it more or less prone to dynamic changes as discussed later in this section.

### 2.1.1 The MT Structure

The structure of the MT is a hollow tube, and the tube wall typically consists of 13 linear polymers side by side (see Figure 2.1) [17, 45, 62]. An individual linear polymer, called a protofilament (PF), is built from $\alpha\beta$ tubulin dimers [17, 45, 62]. These $\alpha\beta$ dimers are referred to as subunits, since they are the basic building blocks of a MT, and they are integrated into the polymer structure through longitudinal bonds. The PFs are connected to each other through lateral bonds between neighboring subunits, and forming a lattice: the longitudinal direction through each PF, and the lateral direction through neighboring PFs. The lateral bond between the first and last PFs is called the "seam", where there is a 1.5 dimer shift in the arrangement of subunit neighbors. Thus, the subunits have a helical arrangement as they form the outer walls of the MT's tubelike structure.

The individual $\alpha\beta$ dimer subunits can be in one of two states. While in the cytoplasm, or when the $\beta$ end of the dimer is exposed and unbonded, a guanosine-triphosphate (GTP) molecule (an energy carrier) can bind to the subunit to form the GTP-bound state of a subunit. A short time after a dimer subunit has been internalized into the MT structure, and another subunit has longitudinally bonded above it on the $\beta$ end of the dimer, a guanosine-diphosphate GDP molecule can replace the GTP molecule through a process called hydrolysis, and thus forms a subunit in the GDP-bound state (see Figure 2.2). The structure of the GDP-bound subunit is bent compared to the straighter GTP-bound subunit [76, 77].

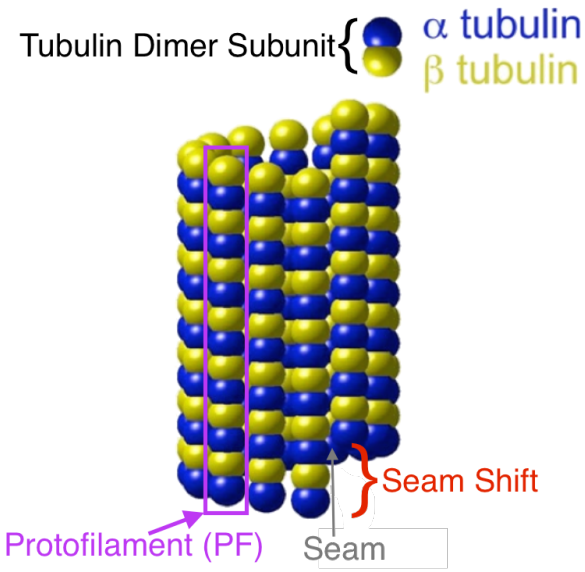



Figure 2.1. A rendering of the fundamentals of a MT structure. Each blue and yellow sphere represent $\alpha$ and $\beta$ monomers respectively. The combined pairing forms the $\alpha\beta$ tubulin dimer subunits, which are the basic building blocks of the MT. Vertical strands of subunits are held together through longitudinal bonds as part of a PF. The PFs are joined together with lateral bonds to form the walls of a tube-like structure. The first and $13^{th}$ PFs come together at a special sequence of lateral bonds called the seam, where their orientation of neighboring subunits is shifted by three monomers, which creates the helical structure in the MT polymer (adapted from [56]).

### 2.1.2 Dynamic Instability (DI) of MTs

One important feature of how the MT structure changes over time is called dynamic instability (DI), the abrupt transition from growth to shortening of the MT length, and vice versa [17, 45, 62]. This behavior is part of a healthy functioning cell, and disrupting it can lead to diseases such as Alzheimer's, Parkinson's, and cancer [23, 33, 35]. During growth, the MT undergoes sustained and consistent lengthening of the its structure, though rates of growth may vary slightly during these periods. In contrast, during shortening, the MT experiences a much quicker rate of change to the overall length, such that the majority of the biopolymer mass present at the start of the shortening period is deconstructed and dispersed into the cytoplasm, and by the

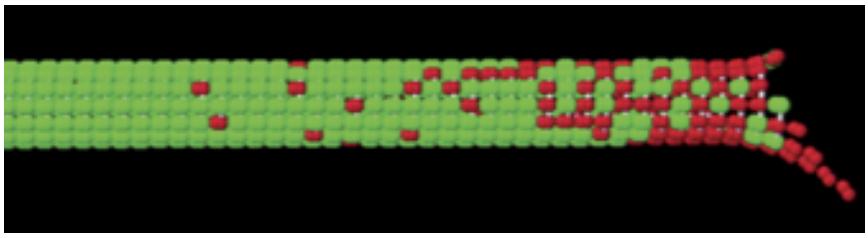


Figure 2.2. MT structures rendered from simulations. Each block represents one $\alpha\beta$ tubulin dimer (referred to as a subunit). The red subunits are GTP-bound, and the green subunits are GDP-bound. GTP-bound subunits that have been more recently incorporated in the MT structure mostly populate the top of the polymer structure, and the older ones in lower portions have a higher likelihood of having undergone hydrolysis, and thus tend to be GDP-bound. The bent conformation of individual GDP-bound subunits are the cause for the curved profile of laterally unbonded sections of PF tips (adapted from [47]).

end a very short MT remains. For this reason, the event when a MT transitions from a growth to a shortening period is described as a "catastrophe", and the rare events when the biopolymer structure is saved from near complete destruction is referred to as "rescue" (see Figure 1.1) [17, 45, 62].

The dynamic instability processes observed on the macro-level of a single MT are the result of the collective micro-dynamics occurring at the subunit level. The bio-chemical reactions of forming and breaking bonds are the molecular level changes fundamentally giving rise to a MT's growth or shortening. Individual GTP-bond subunits are incorporated into the MT structure when a longitudinal bond is created between the new subunit and the top-most subunit of one of the PFs in the MT. Thus, the event of a MT growing by a single subunit is called "polymerization". Once incorporated into the MT, a subunit is allowed to form a lateral bond with an adjacent subunit of a neighboring PF, if it exists there. Lateral bonds help stabilize the MT structure, since they strengthen the ties between an individual subunit and the surrounding MT lattice. Otherwise, any subunit not laterally bonded to its

neighbors is subject to break its longitudinal bond with the subunit below it. Thus, any sequence of subunits can without lateral bonds may detach from the MT structure in an event called "depolymerization". Since a MT can polymerize one subunit at a time, yet lose multiple subunits simultaneously through depolymerization, it's easy to see why growth periods at the macro-level occur with more steady and slower rates when compared to shortening periods [31].

The five possible events that are recognized to alter the MT structure are polymerization, depolymerization, lateral bonds forming, lateral bonds breaking, and hydrolysis, as illustrated in Figure 2.3. From a macro-level perspective, the growth and shortening periods observed through a MT length history profile are most significantly a consequence of polymerization and depolymerization events. However, it should be noted that lateral bonds forming and breaking, and subunits changing states via hydrolysis are all events that surely change the MT structural configuration despite their lack of contribution to a changing MT length. In the rest of this section, the extended effects of these five events on the structural integrity of a MT polymer are discussed, leading to an improved understanding on MT dynamic behaviors at large.

### 2.1.3 Impact of MT Structure on MT Dynamics

The growth and shortening behavior of a MT is a direct consequence of subunits being added and subtracted to the polymer structure. However the likelihood of polymerization and depolymerization depends on the structural integrity of the MT. The state of individual subunits can play a role in the collective stability of the MT as a whole, and hydrolysis is the bio-chemical reaction that controls this feature in MT dynamics. After a short time, the GTP molecules on the newly incorporated subunits are hydrolyzed to GDP molecules. This change in state promotes the affected portion of a PF to bend back and curl away from the central axis of the MT, as seen in MT

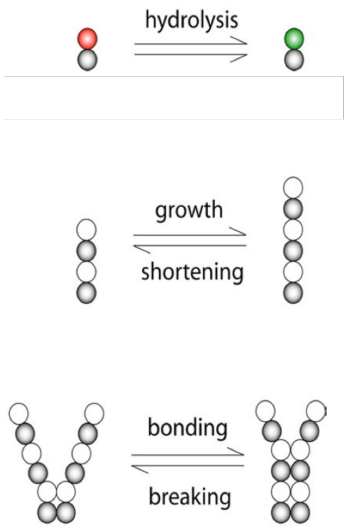
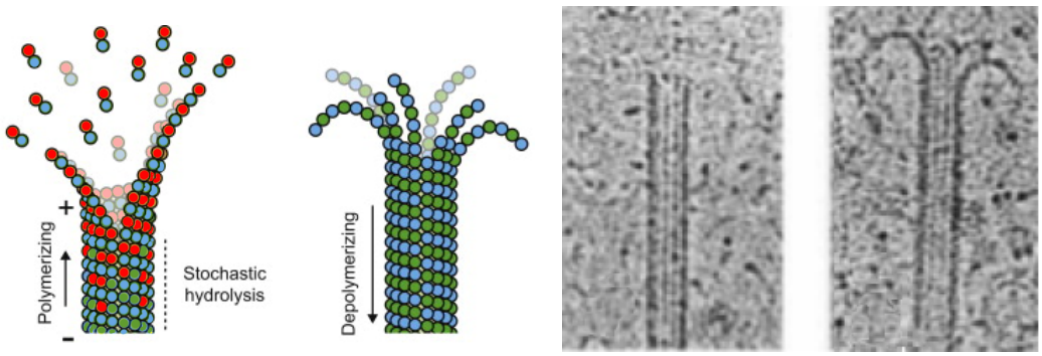



Figure 2.3. There are five molecular-level reaction events considered during MT dynamics that can alter the structure of the biopolymer. Growth (or polymerization) occurs when a single GTP-bound subunits is added to a PF tip, and shortening (or depolymerization) occurs when a sequence of laterally unbonded subunits detach from the MT. Lateral bonds can form or break between neighboring PFs. Hydrolysis is the irreversible event when a GTP-bound subunit (red) transitions into a GDP-bound state (green) (Adapted from [56]).

tips with PFs that fray and curl outward (i.e. they look like rams horns). In contrast, when a PF is comprised of predominantly GTP-bound subunits, the local structure is straighter and more easily allows for the formation of lateral bonds. So, hydrolysis destabilizes the MT structure, because GDP-bound populated ram-horn like PFs are bend father away from the rest of the MT structure, which makes it more difficult for a lateral bond to form between two subunits when at least one of them is in a GDP-bound state. Figure 2.4 illustrates the difference between these structures during growth and shortening. Furthermore, the bent conformation of GDP-bound subunit can an add strain to its existing bonds after undergoing hydrolysis, and this increases the risk of breaking those bonds [76, 77].

Additionally, the timing and irreversibility of hydrolysis leads to an interesting structural orientation of the MT structure. Newly incorporated subunits tend to be in a GTP-bound state, and are added to the top-most positions of a PF. In time, hydrolysis events alter these GTP-bound subunits into a GDP-bound state, which can act as an indicator for age of subunits within the MT structure. Effectively, younger GTP-bound subunits populate the top of the MT, leaving the older GDP-bound subunits to predominantly populate most of the lower MT structure. This characteristic lends to the tendency for a growing MT to have a region near the tip rich with GTP-bound subunits, called the "GTP-cap" [10, 20, 40, 44, 62, 79–81] (see Figure 2.4). MTs with a large enough GTP-cap are somewhat protected from the onset of a catastrophe event, mostly because the GTP-bound subunits are straighter and are more resilient to lateral bonds breaking. MTs without a GTP-cap have a great deal of GDP-bound subunits exposed in the tip-region, which tend to promote breaking of lateral bonds, which then makes depolymerization more likely [10, 20, 40, 44, 62, 79–81]. Hence, MTs with GTP-caps are more stable, prevent the loss of subunits, and instead tend to grow [10, 20, 40, 44, 62, 79–81].

Though hydrolysis helps the MT subunits change into their shaky GDP-bound state, this is countered and controlled by the addition of new GTP-bound subunits from the cytoplasm. Therefore the dynamics of the MT at large lies in the interaction between the available concentration of fresh GTP-bound tubulin dimers and hydrolysis events catching up to the newly polymerized subunits. At tubulin concentrations that are too low, below the critical concentration for elongation, polymerization events are not enough to outlast the hydrolysis events, a GTP-cap is unlikely to form, and therefore the MT length remains near zero. At tubulin concentration that are much higher, above the critical concentration of persistent growth, polymerization events can easily overcome the changes from hydrolysis to maintain an everlasting GTP-cap, so much so that catastrophes are very rare, and MT growth is imminent.




(a) (b)

Figure 2.4. (a) A rendered illustration adapted from [1], and (b) *in vitro* images adapted from[52] of MT structures during growth and shortening. In both subfigures, the MT on the left is growing with a visibly straighter profile due to the presence of a GTP-cap, especially when compared to the shortening MT on the right, which has "ram's horns" structure visible possibly due to sections of laterally unbonded protofilement tips largely populated with GDP-bound subunits.

However, in the range of available tubulin concentration levels between these two critical concentrations, polymerization and hydrolysis compete in a manner where substantially long MT structures may form, however catastrophe events inevitably unravel the components of the biopolymer back into the cytoplasm. The classical DI behavior is observed in this range, and renders MTs to be steady-state polymers. This is distinguishable from equilibrium polymers, which possess an overlap of different interpretations for a single critical concentration value [43].

It is helpful to know about the structural features associated with MT dynamics, such as the presence of the GTP-cap, but there is a deeper connection that is still lacking. More specifically, the mechanisms that underlie the dynamic phase transitions are poorly understood. Much of what is known has been observed experimentally, and several computational models representing MT systems have duplicated these results during growth and shortening phases separately. However, the short-comings of the experimental scenario prevent more detailed observations of the MT structure

at precise moments near significant dynamic changes. Both time and spatial resolutions are exhausted in laboratory conditions, and thus are not reliable to study the interactions between individual subunits and the resulting structural configurations. For this reason, a computational model considering the reactions at this level of detail is desired to surmount the hurdles posed in the laboratory.

## 2.2 List of Assumptions

To develop a computational model representing the MT system at the level where individual subunit interactions are recognized, biological considerations of MTs known to date are used to develop the simplifying assumptions needed to establish the fundamental basis of the desired model. The following list identifies the assumptions in relation to structural or dynamic components, and the justification for their appropriateness to the MT systems relevant to this study:

1. Only a single GTP-bound subunits can attach to a MT: the concentration of GDP nucleotide in the cell or *in vitro* is generally assumed to be low relative to the concentration of GTP nucleotide, and that the rate of exchange by tubulin of GDP for GTP is fast. Therefore, the model assumes that only GTP tubulin can add to the end of a MT [9]. Thus, a polymerization event involves a single GTP-bound subunit to form a longitudinal bond with the top-most subunit of an individual PF tip. With this consideration, the rates of polymerization would depend on the availability of GTP-bound tubulin concentration in the cytoplasm. Finally, it is further assumed that individual GTP-bound subunits are uniformly distributed throughout the cytoplasm, and any fluctuations of the available tubulin concentration levels in the environment are neglected.

2. Fixed levels of free-tubulin concentration: the cytoplasmic environment considered here to provide the free-tubulin pool of subunits is assumed to be held at fixed levels. Since only a single MT is being observed in the simulations, there competition between other MTs, spatial inconsitencies, or other sources of depleting tubulin concentration levels are not considered.

3. Sequential lateral bond formation/breakage: As subunits are incorporated into the MT structure, the lateral bonds with a neighboring subunit are assumed to form only if the neighboring subunits immediately below them have already

formed a lateral bond. Similarly, lateral bonds are assumed to break only if there is a lateral bond missing above it. This is analogous to "pulling a zipper" up and down, where the teeth on both sides represent subunits, and the pull tab is the top-most lateral bond. Those subunits above the top-most lateral bond cannot yet be bonded until the position of the top-most lateral bond moves up to them, and those subunits below the top-most lateral bond cannot break their bond until the position of the top-most lateral bond moves down to them. This configuration creates a "crack" between neighboring PFs in the region above the top-most lateral bond. This assumption is further justified by the existence of the ram-horns in experimental observations, which logically cannot be contemplated without the sequential dynamics of lateral bonds. The likelihood of lateral bonds breaking especially depends on the nucleotide bounds state of the subunits it is connecting, but as well as the existence of lateral bonds existing on the outsides of those same subunits. The possibility of a lateral bond breaking is less likely if both subunits surrounding it are still secured to the MT structure by lateral bonds. This in turn promotes crack formations more as a result of individual PFs extending on their own, rather than lateral bonds being quickly zipped down.

4. Sequential subunit gain/loss: Polymerization and depolymerization events are assumed to occur sequentially, such that a MT can grow or shorten only by gaining or losing subunits at the ends of polymer structure respectively. This assumption is valid since a longitudinal bond needs to form so that the PF within a MT can lengthen. Similarly, a longitudinal bond below a subunit without lateral bonds must be broken so that it (and any string of subunits above it) can be freed from the MT structure. This sequential polymerization and depolymerization also enables the observations of consistent and slower growth rates, as well as the faster and sporadic shortening rates.

5. Random hydrolysis: Hydrolysis events are assumed to occur randomly to those GTP-bound subunits within the MT structure that have formed a longitudinal bond with another subunit above it. This is valid, since the only bio-chemical restriction preventing a GTP-bound subunit from hydrolyzing is that the dimer not be exposed from the $\beta$ end. The non-reversibility of hydrolysis and the choice of an appropriate rate have already been shown to create realistically sized GTP-caps in computational models using the random dynamics, without further assuming a vectorized or sequential restriction [55, 56].

6. The majority of MT dynamics occur in the tip region: This is assumption is quite obvious, and results from the assumptions for the sequential dynamics made above. The definition of the tip-region is clarified in Chapter 5, but this assumption allows us to put attention on the portion of the MT structure where immediate structural changes occur. Subunits are gained and lost from the PF tips, lateral bonds are formed and broken at the interface defined by

the top-most lateral bonds relatively close to the PF tips, and hydrolysis affects GTP-bound subunits predominantly located in the GTP-cap near the MT tip region.

7. The MT Seed: In nature, cytoskeletal polymers tend to have dynamics on both end of the filament structure, at the plus (+) ends where more growth takes place, and minus (-) end where the dynamics are slower as well as subunit loss is more likely. To simplify the region of interest to the MT tip, where most of the interactions take place, the model being developed here assumes an indestructible seed, such that only the dynamics from the plus end to occur, and ignores any dynamics from the minus end. This is analogous to the GMPCPP *in vitro* experimental conditions where an seed is fixed to a glass substrate, and MT dynamics are observed and measured from the highly dynamic plus end [40]. In this instance, a catastrophe event often will precede a shortening period, after which the MT structure will shorten near to or up to the seed, but the bonds and subunits comprising the seed are unaffected. Furthermore, the seed is considered the initial condition and starting point for the computational model, unless otherwise specified.

8. Subunit states alter bond strength: As discussed in the Section 2.1.3, GDP-bound subunits tend to bend more than their GTP-bound counterparts as observed in the ram-horn like structures. The added strain by the bent PF in the presence of a GDP-bound subunit is assumed to aid the breaking of a lateral bond. Also, the bent configuration further exposes a non-laterally bonded subunit connected above out of the MT scaffolding, and hence is assumed to increases the likelihood for the longitudinal bond above a GDP-bound subunit to break. These assumptions facilitates the observations of the stability and preference towards growth provided by MT structures with a significantly sized GTP-cap, versus MT tip structures greatly populated with GDP-bound subunits.

9. Reaction rates depend on the state of the MT structure: The five events being considered to alter the MT structure are understood to take place with certain limitations. Polymerization depends on the availability of GTP-bound subunits to lengthen a MT. Depolymerization requires subunits to be free of lateral bonds in order to facilitate dissociation, and breaking a longitudinal bond depends on the state of the subunit below the bond. Lateral bond formation requires two subunits in neighboring PFs to be present in the space immediately above the top-most lateral bond. Lateral bond breaking depends on the state of the subunits that they are connecting. Hydrolysis depends on the presence of GTP-bound subunits. With the exception of polymerization, the likelihood of these events occurring depends on the MT structure's state at a given moment in time. Furthermore, past MT structures may create limiting conditions for what possible MT structures can form, however only the current MT structure

is assumed to have any effect on what reaction event occurs at any moment.

## 2.3 Computational Models of MTs

In addition to the knowledge base built from experimental methods, a combination of computational and theoretical modeling of MTs has also been an indispensable alternative approach to study MT dynamics at a detailed molecular level. From the pioneering work of Chen and Hill [11, 12, 38], MT modeling evolved from simplified single-PF models [2, 8, 11, 24, 25, 38, 39, 54, 59, 66–68, 73] to more recent multiple-PFs models [4, 12, 34, 47, 55, 56, 58, 76, 77]. In this section, the history of some of these different models is reviewed to gain an understanding for which of their components are important for carrying into newer models, including the computational model being used in this study, which seeks to connect micro-level MT tip structures to macro-level dynamical changes.

### 2.3.1 1-Protofilament Microtubule (1-PF MT) Models

For the most part, single PF models describe the polymerization and depolymerization process of MTs as the addition and detachment of one or multiple subunits, respectively. The corresponding rates for adding and subtracting subunits were deduced approximately from experiments. For the hydrolysis process, two mechanisms have been employed. In the models described in [39, 59, 66, 67], the GTP hydrolysis occurs at only the interface between a GTP-bound subunit and a GDP-bound subunit vertically along a PF, which is called vectorial hydrolysis. Once the hydrolysis boundary (the interface between a GTP-bound subunit and a GDP-bound subunit) catches up to the growth front of a MT, the MT structure will be composed of all GDP-bound subunits, and the MT would be prone to undergo catastrophe. In the other hydrolysis mechanism, called random hydrolysis and used in[2, 8, 24, 25, 54, 68], GTP hydrolysis can occur randomly everywhere except the terminal subunit at the

tip of each PF. When the polymerization rate is larger than the hydrolysis rate, vectorial hydrolysis generates large GTP caps, while random hydrolysis prevents the occurrence of large GTP caps, the latter of which is consistent with the experimental results [10, 20, 64, 78, 79, 81]. In addition, random hydrolysis allows for the existence of the experimentally observed remnants of GTP-bound subunits in the MT lattice [18]. It should be noted that in comparison to vectorial hydrolysis, random hydrolysis allows for GTP-caps that do not have a clear interface between the GTP- and GDP-bound subunits in the MT structure, and this adds to the difficulty of clearly defining a GTP-cap and measuring its size.

### 2.3.2 Detailed Level 13-PF MT Models

Simulations of MT models have the beneficial aspect of providing details that are not available in the laboratory. These computational representations of MTs are brought into the discussion in order to study which structural features play important roles in displaying dynamic instability behavior. While [5, 8, 27] developed stochastic multiple-PF MT models, the individual PFs behaved independently, and no inter-PF lateral interactions were included in those models. To address a deeper level of micro-level details, the mechanochemical model presented in [77, 85] consider energy minimization to determine structural configurations between reaction events. However in [77], lateral bonds are automatically formed when new subunits are incorporated into the MT structure, and the energy minimization step causes the simulation run time to be too long without this simplification that does not allow for the consideration of cracks (laterally unbonded sections between PFs). The approach in [85] allows for curling PFs, but does so without considering lateral bond interactions. Both [77, 85] include more detail than necessary, and in doing so make assumptions that compromise the number of structural configurations covered by the model. Furthermore, these models have simulation run times that are too long to gen-

erate length history plots for studying the connection between micro-level structures along with the macro-level behavior.

Other detailed level computational models treated the MT as a polymer consisting of 13 PFs and independent lateral bonds between them, and utilized a stochastic dynamics to evolve the MT structure, which are much quicker to simulate [47, 56]. These models showed that shortening MTs exhibit deep cracks[56], and if the subunits at the bottom of the cracks are GDP-bound, the MTs are more likely to undergo catastrophe [47]. Furthermore, allowing for cracks between PFs at the MT tip is realistic, since other models without this consideration are not able to explain the existence of the ram-horn structures seen experimentally.

However, the dynamics of the lateral bonds in any 13-PF MT model create complex structures in the, including an overwhelming number of distinct MT tip structural configurations to consider. The significance of the MT tip region was discussed earlier in this chapter in Sections 2.1 and 2.2. The MT tip structure includes a majority of the subunits that dictate the probabilities for which reaction events can take place, therefore once the different MT tip structure configurations are known, the dynamics rates of the MT can be computed. Determining the probabilities of a MT being in different tip configurations during each dynamic phase is an important part of understanding the structural mechanisms that lead to phase changes. Despite the low computational costs provided by the 13-PF models presented in [47, 56], approximating hydrolysis events effectively skips over some structural configurations, and this returns a somewhat incomplete representation of the tip configurations that are of interest.

### 2.3.3 Requirements for a New 13-PF MT Model

Building from basic detailed level 13-PF models, this study seeks to construct a computational model capable of simulating all of the possible structural configu-

rations realized during dynamic instability behavior. Furthermore, a rich data set requires running long time simulations in order to acquire a deeper understanding of which structural configurations are observed during different phases, and which ones are commonly involved in significantly transitioning the dynamics. Such a data set would display hours of DI, with many growth and shortening periods, and rich with catastrophes, rescues, and any other phase transition event that can be contemplated. Though the [77, 85] models provides a level of detail including energy transactions of each reaction event, the computational expense of simulating hours of MT behavior is far too high, especially since the scope of this study is concerned with the MT structures at the end of each event. For this reason, the detailed level 13-PF MT model of [47, 56] is the preferred starting point. To this, improvements are made for acquiring a more detailed output with a reasonable computation time. However, the previous model approximates hydrolysis events separate from the other four reaction events. For the model developed here, an exact method is constructed by omitting any approximations to reaction events, and calculating the occurrence of every possible reaction event at each step of the algorithm in order to simulate an bio-chemically realistic trajectory of MT structural states.

It is important to recognize that the reaction events that create and alter a MT structure follow a Markov process, such that the occurrence any single event depends only on the present state of the MT structure. Furthermore, the lack of dependence on past states qualifies the entire MT structure to carry the memoryless property. So, the computation model developed in this study simulates the MT evolution through a stochastic sequence of events that alter the MT structure one at a time. Kinetic rates for each reaction event dictate the order for this sequence. More specifically, the Gillespie algorithm is utilized to choose the sequence order, especially since it is a common approach used to simulate molecular reactions in biochemistry and epidemiology settings [3, 19, 22, 26, 65]. Also, using the Gillespie Algorithm is considered

an exact method for simulating the trajectory of MT states when approximations are not made [29, 30]. Based on the kinetic rates, times are randomly sampled from exponential distributions for each possible event, and the event with the smallest time is selected to occur at a given step in the simulation. This provides continuous time values between reaction events, associated to every single structural change undergone by the MT being simulated. Chapter 3 describes more of the details for how the knowledge of MT structure and dynamics at the biological level listed in this chapter are used to develop the computational model capable of simulating MT behavior needed for this study.

# CHAPTER 3

# STOCHASTIC COMPUTATIONAL MODELS

A key trait of MT behavior involves the dynamic instability (DI) patterns characterized by the rapid changes observed between periods of growth and shortening, and vice-versa. This display of DI emerges from events on the molecular level, which effectively add/subtract subunits to/from PFs, form/break lateral bonds between PFs, and change the nucleotide-bounded state of dimer subunits. A MT with a fixed seed goes through its dynamic processes by mostly altering the end-region of the biopolymer farthest from the seed, i.e. the tip region. PF and lateral bond construction/deconstruction is assumed to be sequential in nature, in that they only affect the top-most portion of each MT component, and the rest of the MT structure goes mostly unaltered. The older dimer subunits within the MT are located lower in the structure, and they are more prone to have already undergone the change in their subunits bounded state. So, the corresponding hydrolysis events will more likely target newer sections of the MT structure, which form near the top. If we take into consideration the immediate effect of these events, then it is easy to see that most of the changes to the MT structure occur at the tip region, defined by the portion of the MT end including, but not limited to, the segments of PFs lacking a lateral bond. Thus, the desire to peer more closely at the molecular-level dynamics that lead to the macro-level changes as observed in DI behavior begs us to focus on the MT tip region, the portion of the MT structure where all the action occurs. In this chapter, the details for developing a computational model are outlined for studying the MT structure in the tip region at a detailed time scale. In addition, a simplified variant

of the model is presented in order to accommodate studying the structural features that occur in the tip region. By using the the simulated data from the computational model presented in this chapter, the statistical analyses conducted in later chapters on the structural features not available in the laboratory are made possible.

## 3.1 Extending the Basic 13-PF MT Model

This dissertation is interested in further exploring the MT structures and how they relate to changes in MT dynamics and, on a larger scale, transitions seen in DI behavior. Attention is turned on the MT tip region. The reaction events affecting the MT structure shape (dimer subunit gain/loss, lateral bond forming/breaking) result in changes to the MT tip region exclusively. Additionally, when considering the existence of a GTP-cap near the tip of the MT, it's easy to see that most hydrolysis events will also be targeting those GTP-bounds subunits included in the tip region as opposed to those few remaining GTP-bound subunits located in the lower portions of the MT structure. In other words, the tip region is where the most MT structural changes take place, and hence where focus is placed when extracting meaningful information about MT dynamics.

So, a detailed understanding of how changes in the MT tip affect macro-level behavior is desired. However, it is quickly understood that considering the tip region alone fails the memoryless property required for a Markov process. For example, when the tip region interfacing boundary is defined at the position of the top-most lateral bonds, calculating the probability or kinetic rate for a lateral bond breaking event requires knowledge of the two subunits below the tip region, and this information is not available when only tracking the subunits located in the tip. If the tip region cutoff is defined at the position below the top-most lateral bond instead, the kinetic rates are now possible to calculate, however the destination state after a lateral bond breaking event occurs requires knowledge of the states for those subunits two

positions below the current top-most lateral bond. So on, and so forth, this leads to finally requiring the knowledge of the states of the subunits as far down as the MT seed, which at this point is the entire MT structure. Therefore, it is not possible to model the tip region alone as a Markov process. Instead, the entire MT structure is modeled, and the information regarding the tip region is extracted throughout the simulation. For this reason, the existing detailed level computational models are extended to study the structural dynamics that take place in the MT tip region.

The single-PF models in [2, 8, 10, 18, 20, 24, 25, 39, 54, 59, 64, 66–68, 78, 79, 81], and the conclusions reached using them, have inevitably inspired the more recent detailed level models developed using a 13-PF configuration in [5, 8, 27, 47, 55, 56, 77, 85]. The benefits offered by these detailed computational models include the ability to take a closer look at the different sub-structures created and morphed by the individual events which represent biochemical reactions associated with MT dynamics. The bulk behavior is captured by these models in the form of the MT length history, and the corresponding simulations display various forms of DI dependent on model parameters. This is an analogous to the limited scope of the bio-polymer behavior observed *in vitro*. Knowing that the MT tip is the region of interest for studying the mechanisms that dictate key changes in DI behavior poses some limiting factors. Electron microscopy during strictly growth or shortening periods has provided important high spatial resolution information for the existence of ram-horn structures, but accessing this level of detail is not possible during macro-level dynamic changes with this approach[16, 52, 75]. The dynamic imaging techniques used to observe MTs (typically fluorescent tagged tubulin) do not provide detailed structural information at the subunit level. The exact structural configurations of which type of subunits comprise individual PFs are not visible, nor are the lateral bonds between them. This is where the benefits of computational models shine through, by offering an alternative domain to study a biological scenario through a perspective not available

in the laboratory. Therefore, in order to study the MT tip structures and how they are associated to transitional behavior changes in length history, the existing detailed level 13-PF MT computational model is altered and extended to suit the necessary needs for studying this relevant biopolymer sub-structure.

### 3.1.1 The Basic 13-PF MT Model

The detailed level computational model of a 13-PF MT used in [34, 47, 55, 56] was chosen as the starting point in this study for several of its novel properties. First, this model is capable of simulating a full assortment of macro-level dynamic behaviors that emerge from micro-level reactions with either a fixed or varying concentration level for available tubulin dimer pool. This model's simulation using tubulin concentration levels chosen within an appropriate range of critical concentrations can output MT length histories for long time durations with characteristics similar to the length evolution seen with DI in laboratory observations. Second, it is an MT model that represents the lateral bonds independently, without considering an excess level of detailed dynamics. The detailed level model of [77] assumes lateral bonds to form simultaneously with longitudinal bonds once a new subunit is incorporated into the MT structure. The independent behavior of lateral bonds forming and breaking is preferred for the needs of this study since it does not violate the assumption of only allowing one bio-chemical reaction to occur at a time. In regards to the tip structure, the chosen model also allows for the formation of cracks between PFs, and represents them in the form of missing lateral bonds. Cracks described in this manner offer a reasonable method for how the "ram horn" configurations are observed. In comparison, the molecular-mechanical model of [85] relies on the deformations of individual dimers to create PF curls without considering lateral bond interactions.

Additionally, recent experimental observations have shown that there are indeed cracks between PFs, and the absence of cracks adds further support to lateral bond

interactions between neighboring PFs [60]. It should also be added that individual subunit contributions to shaping the MT structure are taken into consideration in the form of kinetic rate constants that depend on the nucleotide bound state of dimer subunits. GTP-bound subunits tend to be straight, and hence are more prone to maintain existing bonds compared to their bent GDP-bound counterparts. Finally, the computational power available now offers more memory per processor in comparison to what was available a decade ago, the time at which the detailed level models were first conceived. These improvements to processing power have not overcome some high computational costs, such as the caveat of the energy minimization calculations in the [77] model, or the calculations involved with the spatial considerations in the [85] model. Instead, new improvements to computer hardware memory certainly benefit the computational speed of implementing a kinetic Monte Carlo simulation, which needs to track the state of the individual subunits in the entire MT structure. This makes the existing detailed 13PF MT model a good candidate to extend beyond it's current approximate implementation for the purposes of studying the structural configurations associated with DI phases and phase transitions.

The basic 13-PF MT model of [34, 47, 55, 56] offers an opportunity to simulate a large amount of data that tracks a MTs structure through its dynamic processes with low computational costs. However, in order to accommodate the need for observing all of the tip structures relevant to key moments of significant dynamic changes, this model must be extended in order to reveal all of the structural states that are possible in a sequence of reaction events. In its current form, the detailed level 13-PF MT model approximates hydrolysis events, by allowing multiple subunits to hydrolyze simultaneously after one of the other four reaction events (lateral bonding/breaking or subunit gain/loss) have occurred. Despite this approximation being a reasonable one, it violates the assumption that only one of the bio-chemical reactions occur at a time. Furthermore, it also creates a scenario where several structural configurations

are skipped over when observing the simulation's output, which in turn counters the intentions of studying the details of the structures that should be observed during MT dynamics. For this reason, the computational model is extended to more carefully implement the Gillespie Algorithm, where hydrolysis reaction events are treated similarly along with the other four dynamic events that alter the MT structure. This extension to the original model is capable of delivering an exact method. For long-time simulation runs, the Gillespie Algorithm can deliver data closely resembling a large number of transitions through a realistic sequence of bio-chemical reactions and structural states by making fewer model considerations, and therefore fewer assumptions. With a large simulation data set, enough information can be gathered to determine which structural states are connected to different phases observed in DI behavior, and even more, which states are more prevalent during transitions between dynamic phases.

### 3.1.2 Extending the 13-PF MT Model to Study the Tip Region

In this study, the tip region structural configuration is of particular interest, since those structures are the most dynamic portion of a MT. A first attempt to define which portion of the MT structure is included in the tip region should reveal some of the complications involved with defining a generic tip region for a 13-PF MT. When considering a MT formed by 13 PFs and 13 lateral bonds between them, the tip region definition takes on many variations depending on where one decides to separate the tip from the rest of the MT structure. Different lateral bond heights allow for sections of PFs to be partially bound, deeming the strict "laterally unbonded" region inapplicable. Allowing for all partially bound PF regions as part of the tip region leaves a large number of PFs and lateral bonds to account for in a single tip structure configuration. For example, if the cutoff is chosen as the shortest lateral bonds height position, then the tip may include a substantial section of PFs that are

laterally bound to their neighbors, a relatively stagnant section of structure when compared to the laterally unbonded parts that are more inclined to change from the MT dynamics. This classical representation of a 13-PF MT serves too complex a scenario for tackling the tip structure study as is.

As a first attempt to study the MT structures at this level, a model using a simpler configuration structure is desired to make studying the tip region more tractable. To this end, a 2-PF MT is presented as the simplest MT structure that takes the lateral bond into consideration: 2-PFs with a single lateral bond between the adjacent subunits. This representation can be perceived as a 2-PF polymer, or as an arbitrary pair of neighboring PFs embedded within a larger 13-PF MT. Although this offers a less complicated scenario than the 13-PF case, it is worth recognizing that the proposed 2-PF case is relevant to actual two-stranded filaments that exhibit dynamic instability behavior, such as a protein called ParM [28]. Since the scope of the study is making reference to conclusions and model parameters associated with prior MT studies, the 2-PF structure and its behavior will be referred to as a special case MT. The dynamics in the model simplification follow directly from the proposed Gillespie Algorithm extension of the detailed 13-PF MT model. However, due to the relevance of the changes enacted, this chapter first discusses the new implementation of the extended 13-PF MT model, and then moves the discussion towards the simplified 2-PF MT version. By laying the groundwork for this computational implementation, these will be the two models referenced throughout the remainder of this study.

## 3.2 Representation of the 13-PF MT Structure

The structure of the MT used in the computational model here follows from the list of assumptions understood from the biological perspective detailed in Section 2.2. Many detailed level 13-PF models agree on only including the helical orientation to create the tube-like structure [5, 8, 27, 47, 55, 56, 77, 85]. However, in the model being

introduced here, the lateral bonds introduced in [34, 47, 55, 56] are also included. This section describes in detail the components that make up the totality of the biopolymer structure for the computational model being presented.

### 3.2.1 Tubulin Dimer Subunits

The MT model developed here considers the smallest structural components, or the "basic building blocks", to be the $\alpha - \beta$ tubulin dimer compounds, which are often referred to as "dimer subunits", or just "subunits". The dimer subunits can be found in one of two forms: GTP-bound, or the hydrolyzed counterpart, GDP-bound (see the red and green blocks in Figure 3.1). Only GTP-bound subunits are able to form a longitudinal bond with a subunit already part of a MT at the very top. Once embedded into the MT structure, a dimer subunit may hydrolyze into a GDP-bound state. The GTP-bound subunits are considered to fill the free tubulin pool with concentrations on the order of micro-molar ($\mu M$) levels in the form of "Free-GTP-Tubulin", such that they are available to polymerize onto the MT structure [31]. When the subunits detach from the MT, GDP-bound subunits are assumed to quickly change back to a GTP-bound state, and therefore replenish the Free-tubulin concentration levels used for simulations [31]. The scope of this study involves the dynamics of a single MT at a time, it does not constitute the need for competition between multiple MTs, and therefore assuming a constant free-tubulin concentration at micro-molar levels is reasonable.

### 3.2.2 Protofilaments (PFs)

A consecutive sequence of subunits held together by longitudinal bonds constitutes a PF structure. These PFs are constructed by allowing only a single GTP-bound subunit to attach the the top most position of the PF during any one polymerization event [56]. However, the deconstruction of a PF is more involved, where any consecu-

tive section of the PF from the top down that is not supported by a lateral bond can detach. This casts light on the expected rate of polymerization to be much slower compared to the rapid rate of depolymerization resulting from the shear difference of the number of subunits exchanged through each event. The sections of PF that are laterally unbonded are significant in their own right, since those subunits are the ones susceptible to depolymerization (see Figure 3.1). For this reason, the term **PF tips** refer to the sequence of subunits that are not laterally bonded to both neighboring PFs. In the 13-PF case, there are situations where sections of PFs have a lateral bond on one side and not on there other, however these are sometimes referred to as sheets of protruding PFs, and are not included in the PF tip definitions *per se*. However, the significance of PF tip is more apparent in the 2-PF model described later.

The inherent chemical structure for the GTP-bound subunits encourages the formation of sections PFs that are straighter than the sections consisting of GDP-bound subunits. This causes the so called "ram-horns", or curved polymer structures as observed in the PF tips. Additionally, in laterally bonded sections of a MT, subunits that hydrolyze into GDP-bound states add strain to the structure, which increases the internal energy [71]. The model being developed here is not concerned with the consequential geometry or the spatial coordinates of individual subunits; only the order of the subunits as they appear in each PF is considered. Instead, the difference in the dimer subunit bounded state is reflected in the dynamics rates associated with different scenarios. These rates defined in Section 3.3.2 reflect that GTP-bound subunits promote a more stable structure and growth of PFs and lateral bonds, whereas GDP-bound subunits tend to destabilize the MT structure by discouraging lateral bond formation and encouraging shortening of PFs.

### 3.2.3 Lateral Bonds

Individual lateral bonds can form, from the base toward the top of the MT, between pairs of neighboring dimer subunits in the same position within their respective PFs. A single new bond can form only in the available space directly above a previously existing bond, and only the single top most lateral bond can break. We refer to the consecutive sequence these bonds all together as the lateral bond (see the white squares in Figure 3.1). The top position of a sequence of lateral bonds can be thought of as a "zipper-head", which as it moves up binds the teeth on the two sides of a zipper. Sections of teeth that are not zipped (above the zipper-head) are free to move around, whereas the zipped sections (below the zipper-head) combine the two sides into a single sheet that is restricted and stabilized. Similarly, the lateral bond stabilizes the PFs between which they form, since only those subunits lacking a lateral bond are allowed to depolymerize, but those subunits with a lateral bond will remain a part of the polymer structure. Also, how high the zipper-head can go depends how much available zipper-teeth there are to bind together. In the case of MTs, a lateral bond's height, determined by the position of the top-most subunits that are laterally bound, can only rise as high as the shorter PF that it brings together. This results from the condition that neighboring pair of subunits need to exists so that a lateral bond can form between them. The laterally unbonded section between two neighboring PFs is called the **crack** (see Figure 3.1), and the **crack depth** is measured by the the shorter of the PFs that extends around the crack section, where future lateral bonds can be formed.

Independently, the lateral bond height can be thought of as a biased random walk on a discrete lattice with a moveable upper boundary defined by the heights of its surrounding PFs. The lattice allows for the lateral bond height to move up only if there are neighboring pairs of subunits immediately above the top-most lateral bond (i.e. if there exists a non-zero crack depth). The bias is due to the higher success rate

of a lateral bonds forming with GTP-bound subunit dimers compared to GDP-bound ones. This puts emphasis on the state of the subunits at the interface between the lateral bond height and the crack, which we refer to as the "gate" [47]. The pair of subunits bound together by the top-most lateral bond are labeled as the **Gate subunits (or G-subunits)**, and the pair of subunits at the bottom of the crack are labeled as **Above Gate subunits (AG-subunits)**. Ultimately, the likely direction of the lateral bond height's movement depends on the bounded state of the G- and AG-subunits. Notice that in configurations that lack a crack between PFs, the lateral bond would be the same height as at least one of the adjacent PFs. In this case, it is not possible to identify AG-subunits, effectively yielding a zero probability for the lateral bond to move upward. The lower boundary of a lateral bond would be located at the very bottom of the MT, and is defined by the MT seed.

Additional support provided by lateral bonds comes when both neighboring lateral bonds are present. The likelihood of a lateral bond breaking diminishes greatly when both of the subunits connected by the lateral bond in questions still have existing lateral bonds to the left and right. This restriction limits how cracks are formed, and prevents their existence to be a result of lateral bonds quickly zipping down. Instead, the rapid breaking of lateral bonds seen during shortening phases are instead more reliant on the nucleotide bound states of the subunits surrounding it. This restriction was introduced in the model used in [34, 47, 55, 56] in accordance with the approximate mechanical constraints measured in simulations.

### 3.2.4 Seam

The lateral bonds connect neighboring subunits that are at the same position with their PFs, except between the first and last PFs. The sequence of lateral bonds between PF #1 and PF #13 is called the **seam**, which creates a shift in subunit neighbors, such that the subunits in PF #1 are bonded and positioned next to those

subunits in PF #13 1.5 above their position. It should be noted that the 1.5 dimer shift is considered in all of the calculations involving the position of the top-most lateral bond in the seam, however the visualization output treats dimers as individual units and thus displays the seam to have a shift with 1.0 dimer height difference. Furthermore, to accommodate for that fact the subunits surrounding the seam are associated with two neighboring subunits each, the kinetic rates for lateral bonds forming and breaking are doubled.

### 3.2.5 MT Seed

In order for the MT structure to take its form, it relies on an initial condition, or a MT seed, represented by the collection of short sections at the bottom of each PF configured with permanently GTP-bound subunits, and lateral bonds between them that cannot break. Consequently, the subunits in the MT seed cannot detach from the polymer structure, and the choice of a GTP-bounded state is because it encourages stability, and hence promotes polymerization at the beginning of the simulation [40]. Since the seed constitutes a permanent part of the structure, the seed height is equivalently the minimum height taken on by the entire MT structure, as well as the individual PFs and lateral bonds. In the 13-PF MT model, the portion of the MT seed associated with PF #1 is shortened by one subunit length compared to the other seed height for the rest of the PFs, as illustrated in Figure 3.1. This is to accommodate the shift in correspondence between neighboring subunits as created by the seam, which in turn creates the helical pattern as expected in the MT structure. Any biasing affects of the seed structure's orientation on the MT configurations are considered minimal, because the seed's purpose is mostly to promote the onset of MT growth. Most of the dynamics and configurations of interest take place when the MT has grown quite long, and the changes to the MT structure are occurring far away from the seed.

### 3.2.6 Complete MT Structure

The complete 13-PF MT structure is a combination of all the components listed in this section. One can imagine that the 13 PFs can be arranged side by side, creating a 2-D lattice when they are held together through a sequence of lateral bonds between them (see Figure 3.1). When they are wrapped around a longitudinal axis parallel to the PFs, the PFs form the walls of a tube-like polymer. The first and $13^{th}$ PFs come together at the seam, where there is a 1.5 dimer shift in arranging the neighboring subunits within these PFs, and this completes the helical orientation of the subunits throughout the MT. At the bottom of the MT structure, the seed creates an indestructible portion of GTP-subunits, that acts both as an initial condition where growth can begin, as well as the minimum structure to which a shortening MT can be reduced. The tip region of the 13-PF MT can be quite complex, where PFs can extrude to different lengths, and cracks between PF tips can add to the complexity of the possible tip structures that can be realized. Furthermore, the tip region and several rows of subunits below it is where a stabilizing GTP-cap can form. Finally, the total length of the MT is measured by averaging over the total length of all the PFs.

As it is described further in Section 3.3, most of the structural changes take place in the tip region farthest away from the MT seed. Even so, it is important to track the entirety of the MT structure and the individual subunits, their nucleotide bound states, and the presence of lateral bonds throughout the evolution of the structural changes that are possible in the presented computational model. Even when the MT grows to lengths far beyond the seed, it is very possible (and expected in some circumstances) for the MT dynamics to result in a structure that is more susceptible to rapid shortening phases, where lateral bonds can break quickly and multiple subunits can depolymerize at a time. In these cases, the complete configuration of the MT structure needs to be tracked to ensure that the rates of the micro-level reaction

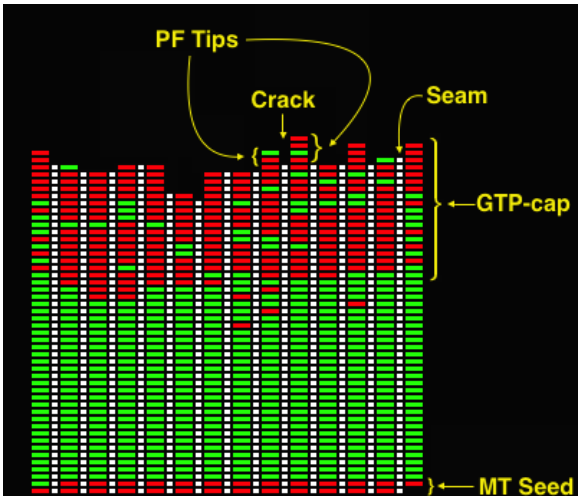



Figure 3.1. A 2D visualization from the computational model simulation. The red and green blocks represent GTP- and GDP-bound subunits respectively. The 13 vertical sequences of subunits are the PFs. The white squares between neighboring PFs are the lateral bonds. The first PF is duplicated on the right of the $13^{th}$ PF to illustrate the shift that occurs at the seam. Laterally unbonded subunits that protruded above the surrounding structure are the PF tips. A crack is created by missing lateral bonds between PFs. The seed is the indestructible portion at the bottom of the polymer structure, and has a shorter height for the first PF to accommodate the shift at the seam. When the tip region is highly populated with GTP-bound subunits, a stabilizing GTP-cap can form without clear boundaries.

events that physically evolve the biopolymer are being determined correctly.

## 3.3 The Extended 13-PF MT Model

In this section, the dynamics included in the computational model are described, such that the individual molecular level reactions that change the MT structure accumulate to resemble the DI behavior as observed experimentally from a macro-level perspective. In order to represent the rules set forth by these reaction events, similar considerations are made here as those in the computational models in [34, 47, 55, 56] to describe the possible changes in PF length and lateral bond heights. The

most significant change to that model is the treatment of hydrolysis events without an approximation. The resulting model will represent the dynamics of a MT structure as a Markovian stochastic process, since the possible reactions only depend on the current state of the MT structure, and the concentration of free-tubulin available for polymerizing the PFs. Once the stochastic dynamics are defined, they can be used to build a computational algorithm to simulate the resulting behavior of this model.

### 3.3.1 Micro-level Reaction Events Determining MT Dynamics

Following through with the reactions that are expected to affect and change a MT structural configuration, the dynamics in the computational model are restricted to the following five events and their respective rules:

1. Polymerization: a single GTP-bound subunit may attach to a MT, by forming a longitudinal bond formation between the top subunit in a PF and an available GTP-bound free-tubulin subunit.

2. Depolymerization: single/multiple subunit(s) may detach from a MT, by breaking a single longitudinal bond between any non-laterally bonded subunit in a PF tip and the subunit below it, thus causing the loss of any number of subunits above the breaking point.

3. Lateral bond formation: a single lateral bond may form directly above the top-most lateral bond only when both PFs have existing subunits there.

4. Lateral bond breakage: the single top-most lateral bond may break, except if this bond is part of the MT seed.

5. Hydrolysis: a single GTP-bound subunit may hydrolyze into a GDP-bound subunit, except for the top-most subunit of any PF, and in the MT seed.

These rules are all based on established MT structure and biochemistry [31].

### 3.3.2 Kinetic Rates for Molecular Reaction Events

The kinetic rate constants, and their respective rates of occurrence for the five events being considered in the MT dynamics are as follows:

1. $k_{poly}^{GTP}$ is the kinetic rate constant of polymerization for a GTP-bound subunit onto a PF (longitudinal bond formation). Since this involves pulling in GTP-bound subunits from the environment, the rate for this reaction should depend on the tubulin levels available in the substrate. The kinetic rate for polymerization is $\kappa \cdot \frac{c}{c + c_{1/2}}$, where $c$ is the concentration of free GTP-bound tubulin available in the surrounding cytoplasm environment, $\kappa$ is the maximum growth rate possible, and $c_{1/2} = \frac{\kappa}{k_{poly}^{GTP}}$ is the concentration that allows for half the maximum growth rate. This follows from the Michaelis-Menten kinetics, used commonly to describe biochemical reactions that involve concentrations in a substrate [61]. Despite using this form in the computations of the model presented here, it is interesting to note for $c_{1/2} = 200\mu M$, the relation $c_{1/2} \gg c$ is valid for the tubulin concentration ranges used in this and similar studies. This relation can reduce Michaelis-Menten form to a linear approximation, which in turn gives us a kinetic rate for polymerization $\approx k_{poly}^{GTP} \cdot c$. This rate is the only dependence on the cytoplasmic tubulin concentration in the model, and $c$ is further considered to be the input parameter that generates the different DI behaviors possible from this model. Only GTP-bound subunits are assumed to form longitudinal bonds with the top-most subunits in a PF, so there is no kinetic rate considered for the polymerization of GDP-bound subunits.

2. $k_{depoly}^{GTP}$ and $k_{depoly}^{GDP}$ are the kinetic rate constants of depolymerization (longitudinal bond breakage). Since the rate of depolymerization only depends on a single longitudinal bond breaking, $k_{depoly}^{GTP}$ and $k_{depoly}^{GDP}$ also directly provide the rates of these processes. In this computational model, the rate of a longitudinal bond breaking below a given subunit is consistent with predictions based on MT structure [69, 82]. If for a given subunit, the adjacent subunit below is GTP-bound, then the PF in this location tends to be straight, and the subunit in question detaches by breaking the longitudinal bond below it with rate $k_{depoly}^{GTP}$. If the subunit below is GDP-bound, then the PF in this location tends to be bent, and the subunit in question detaches with rate $k_{depoly}^{GDP}$ [76, 77]. Because of the different structural tendencies, the relation $k_{depoly}^{GTP} < k_{depoly}^{GDP}$ is maintained, which agrees with experimental observations. The kinetic rate for a single longitudinal bond to break is equivalent to the kinetic rate constant for subunits in PF tips that do not have any lateral bonds, and thus are free to dissociate from the MT structure by breaking a longitudinal bond from below.

3. $\underline{k_{bond}^{TT}(k_{bond}^{TD}, k_{bond}^{DT}, \text{ and } k_{bond}^{DD})}$ are kinetic rate constants of lateral bond formation between the two AG-subunits when they are both GTP-bound (one is GTP-bound and one is GDP-bound subunit, one is GDP-bound and one is GTP-bound subunit, and both are GDP-bound subunits respectively). The kinetic rate is equivalent to the kinetic rate constant for a lateral bond to form between a pair of laterally neighboring AG-subunits exists in the space directly above the top-most position of a sequence of lateral bonds, if they exist. The respective rates for lateral bonds forming at the seam are doubled since they would represent two bonds forming for any one subunit on either side of the seam due to the 1.5 dimer seam shift.

4. $\underline{k_{break}^{TT}(k_{break}^{TD}, k_{break}^{DT}, \text{ and } k_{break}^{DD})}$ are kinetic rate constants of lateral bond breakage between the two G-subunits when they are both GTP-bound (one is GTP-bound and one is GDP-bound subunit, one is GDP-bound and one is GTP-bound subunit, and both are GDP-bound subunits respectively). Breaking of lateral bonds is assumed to be influenced by the nucleotide state of the G-subunits because it has been experimentally shown that GDP-bound subunits have a strong preference for the bent conformation, while GTP-bound subunits are more compatible with the straight conformation parallel to the rest of the MT lattice [31]. The kinetic rate is equivalent to the kinetic rate constant for a lateral bond to break between a pair of laterally neighboring G-subunits, if the lateral bond is not part of the MT seed. However, a geometric penalty is considered, and the rate of breaking any lateral bond is decreased by a factor of $\pi_{break} = 1000$ when there already exist lateral bonds on the farther sides of the subunits connected by the bond in question. This assumption was used to approximate the mechanical constraints of the MT lattice as demonstrated in [34, 47, 55, 56]. The respective rates for lateral bonds breaking at the seam are doubled since they would represent two bonds breaking for any one subunit on either side of the seam due to the 1.5 dimer seam shift.

5. $\underline{k_h}$ is the kinetic rate constant of hydrolysis for a GTP-bound subunit to become GDP-bound subunit. The kinetic rate is equivalent to the kinetic rate constant for an individual GTP-bound subunit to hydrolyze, provided it does not occupy the top-most position of any PF, or it is not part of the MT seed.

These rates can be thought of using arbitrary units of time ($aut$), and $1aut$ is set equal to 1 second ($sec$) for the purpose of comparing to experiments. Thus, all rates used here will have units of $aut^{-1} = sec^{-1}$. Since this is a stochastic model, the rates represent "probabilities per unit time". The actual values used are listed later in this chapter, and a parameter tuning process was performed in [34, 47, 55, 56] for the hydrolysis approximation model, from which the model being developed

here is extended. Ridding the approximation treatment in the new generation of the computational model, however, has little effect on the sensitivity of these rate constants on the simulated behavior generated by the model. This makes sense, because ultimately the hydrolysis approximation was valid, but, for this detailed level study, the exact trajectory of structural configurations is desired. This parameter set is more greatly affected when the 2-PF simplification is introduced in Section 3.8.3, when the necessary alterations are introduced.

### 3.3.3 Rates of Subunit Addition and Loss

At this point, the possible events responsible for changing the MT structure in the computational model have been established, and their corresponding kinetic rates have been identified. However, for polymerization and depolymerization events, it is important to distinguish the kinetic rates of longitudinal bond formation or breakage (with units in 1/time) from the rates of subunit addition or loss (with units of subunits/time). This can be analogous to measuring the velocity of growth or shortening of the PF length, since the subunit heights are effectively considered a unit of measuring the total length of the MT. However, it should be noted that the number of PFs affects how the MT length is measured. Since this study leads to a simplification that reduces the number of PFs considered in the polymer structure, the concept of velocity is therefore avoided in further discussion. In any case, the number of subunits associating and dissociating from the polymer during the described dynamics is preserved in any variation of MT structural configuration.

Polymerization events allow only one GTP-bound subunit to attach onto a PF, therefore the rate of subunit addition, $R_{subunit\ addition}$, follows from the Michaelis-Menten formula used to compute the kinetic rate for polymerizing a single dimer

subunit defined earlier in this chapter:

$$\begin{aligned} R_{subunit\ addition} &= (1\ subunit) \cdot \left( k_{poly}^{GTP} c_{1/2} \left( \frac{c}{c + c_{1/2}} \right)\ sec^{-1} \right) \\ &= \frac{k_{poly}^{GTP}\ c\ c_{1/2}}{c + c_{1/2}}\ \ subunits/sec \\ &\approx k_{poly}^{GTP} c\ \ subunits/sec \quad \text{when} \quad c \ll c_{1/2} \end{aligned} \tag{3.1}$$

Depolymerization events, however, allow for any sequence of laterally unbonded subunits in a PF tip to detach, and effectively be dissociated from the MT structure. Thus, depolymerization is a more involved process, especially when considering the different rates for breaking longitudinal bonds connected to a GTP- and GDP-bounds subunits from below. So, in order to formulate the rate of subunit loss, the states of most of the PF tip configuration plus the top-most subunit with at least one lateral bond must be known. Let $n_{tip}$ be the length of a PF tip. The subunits in the tip are indexed from the top down, such that the $1^{st}$ position corresponds to top-most subunit in the PF tip, and the $n_{tip}^{th}$ position bottom-most subunit in the PF tip, such that subunits 1 to $n_{tip}$ do not have any laterally bonds. Using this indexing, the $(n_{tip}+1)^{th}$ position corresponds first subunit below the PF tip, which is laterally bonded to at least one of it's neighboring PFs. Let $X_i$ represent the nucleotide bounded state of a subunit, GTP- or GDP-bound, in position $i$, where $1 \leq i \leq (n_{tip} + 1)$. Now, consider the following notation for describing a configuration for a sequence of dimer subunits:

$$\begin{aligned} \mathbf{X}^{\mathbf{PF}}_{\mathbf{1},\ldots,\mathbf{n_{tip}}} &= [X_1, X_2, ..., X_{n_{tip}}] \text{ for the subunits in a PF tip,} \\ \mathbf{X}^{\mathbf{PF}}_{\mathbf{2},\ldots,\mathbf{n_{tip}+1}} &= [X_2, X_3, ..., X_{n_{tip}+1}] \text{ for the subunits below those in } \mathbf{X}^{\mathbf{PF}}_{\mathbf{1},\ldots,\mathbf{n_{tip}}}. \end{aligned}$$

So, the kinetic rate for a longitudinal bond to break below any one of the subunits in a PF tip depends on the nucleotide bound states of the subunit below, which are defined in $\mathbf{X}^{\mathbf{PF}}_{\mathbf{2},\ldots,\mathbf{n_{tip}+1}}$. More specifically, if the longitudinal bond between the $i^{th}$ and

$(i+1)^{th}$ subunits were to break, the kinetic rate for that event would be

$$k_{break}^{X_{i+1}} = \begin{cases} k_{break}^{GTP} & \text{if } X_{i+1} \text{ is GTP-bound} \\ k_{break}^{GDP} & \text{if } X_{i+1} \text{ is GDP-bound} \end{cases}$$

For a PF tip of length $n_{tip}$, there are $n_{tip}$-many possible depolymerization events that can take place:

- Case $i = 1$The longitudinal bond between the $1^{st}$ and $2^{nd}$ subunits, $X_1$ and $X_2$, can break with rate $k_{break}^{X_2}$, causing the top subunit, $X_1$ to dissociate from the MT.

- For $i = 2$ The longitudinal bond between the $2^{nd}$ and $3^{rd}$ subunits, $X_2$ and $X_3$, can break with rate $k_{break}^{X_3}$, causing the top two subunits, $\mathbf{X_{1,2}^{PF}} = [X_1, X_2]$, to dissociate from the MT.

- Case $i = 3$ The longitudinal bond between the $3^{rd}$ and $4^{th}$ subunits, $X_3$ and $X_4$, can break with rate $k_{break}^{X_4}$, causing the top three subunits, $\mathbf{X_{1,2,3}^{PF}} = [X_1, X_2, X_3]$, to dissociate from the MT.

- Cases $4 \leq i \leq n_{tip}$ The longitudinal bond between the $i^{th}$ and $(i+1)^{th}$ subunits, $X_i$ and $X_i$, can break with rate $k_{break}^{X_i}$, causing the top $i$-many subunits, $\mathbf{X_{1,...,i}^{PF}} = [X_1, ..., X_i]$, to dissociate from the MT.

- Case $i = n_{tip}$ The longitudinal bond between the $n_{tip}^{th}$ and $(n_{tip}+1)^{th}$ subunits, $X_{n_{tip}}$ and $X_{n_{tip}+1}$, can break with rate $k_{break}^{X_{n_{tip}+1}}$ causing all of the $n_{tip}$-many subunits in the PF tip, $\mathbf{X_{1,...,n_{tip}}^{PF}} = [X_1, ..., X_{n_{tip}}]$, to dissociate from the MT.

Now, it should be clear that for a particular depolymerization event, the rate of subunit loss is the product of the kinetic rate of depolymerization involved with breaking the longitudinal bond and the number of subunits to be dissociated from the MT structure as a result:

$$(i\ subunits) \cdot \left(k_{depoly}^{X_i}\ sec^{-1}\right) = i\ k_{depoly}^{X_i} subunits/sec$$

where $X_i$ is the nucleotide bound state of the subunit below the longitudinal bond that

is breaking. Now, for the entire PF tip configuration $\mathbf{X^{PF}_{1,\ldots,n_{tip}}}$ to detach, knowledge of the configuration of subunits below, $\mathbf{X^{PF}_{2,\ldots,n_{tip}+1}}$, is needed. Let

$$K = \sum_{i=2}^{n_{tip}+1} k_{depoly}^{X_i} = k_{depoly}^{X_2} + k_{depoly}^{X_3} + ... + k_{depoly}^{X_{n_{tip}+1}}, \tag{3.2}$$

such that $K$ defines the total rate for a detachment event to occur, computed as the sum of the possible kinetic rates of depolymerization for a given configuration of subunits $\mathbf{X^{PF}_{2,\ldots,n_{tip}+1}}$. Conditional on a depolymerization event occurring for a PF tip configuration $\mathbf{X^{PF}_{1,\ldots,n_{tip}}}$, the conditional probabilities, $p_i$, for the individual cases $i = 1$ to $i = n_{tip}$ defined above, can be computed as follows:

- Case $i = 1$ The conditional probability for depolymerization of the top-most subunit, $[X_1]$: $p_1 = k_{depoly}^{X_2}/K$
- Case $i = 2$ The conditional probability for depolymerization of the top two subunits, $\mathbf{X^{PF}_{1,2}} = [X_1, X_2]$: $p_2 = k_{depoly}^{X_3}/K$
- Cases $3 \leq i \leq n_{tip}$ The conditional probability for depolymerization of the top $i$ subunits, $\mathbf{X^{PF}_{1,2,3}} = [X_1, X_2, \ldots, X_i]$: $p_i = k_{depoly}^{X_{i+1}}/K$
- Case $i = n_{tip}$ The conditional probability for depolymerization of the entire PF tip, $\mathbf{X^{PF}_{1,\ldots n_{tip}}} = [X_1, X_2, \ldots, X_{n_{tip}}]$: $p_{n_{tip}} = k_{depoly}^{X_{n_{tip}+1}}/K$

Since $i$ subunits are dissociated with probability $p_i$, and given that a depolymerization event is to occur, the conditional expected number of subunits to be lost can be calculated as follows:

$$\sum_{i=1}^{n_{tip}} (i \ subunits)(p_i) = p_1 + 2p_2 + ... + n_{tip}p_{n_{tip}} \ subunits \tag{3.3}$$

Thus, the rate of subunit loss [1] for a particular PF tip configuration can be calcu-

[1]The formulation for the rate of subunit loss is part of an interdisciplinary collaboration pending publication, including contributions from Ava Mauro, Erin Jonasson, and Holly Goodson

lated as the product of the corresponding rate of a depolymerization event (Equation 3.2), and the expected number of subunits to depolymerize (Equation 3.3):

$$\begin{aligned} R_{subunit\ loss} &= \left(p_1 + 2p_2 + ... + n_{tip}p_{n_{tip}}\ subunits\right) \cdot \left(K\ sec^{-1}\right) \\ &= K\left(\frac{k_{depoly}^{X_2}}{K} + 2\frac{k_{depoly}^{X_3}}{K} + ... + n_{tip}\frac{k_{depoly}^{X_{n_{tip}+1}}}{K}\right)\ subunits/sec \\ &= k_{depoly}^{X_2} + 2k_{depoly}^{X_3} + ... + n_{tip}k_{depoly}^{X_{n_{tip}+1}}\ subunits/sec \\ &= \sum_{i=1}^{n_{tip}} ik_{depoly}^{X_{i+1}}\ subunits/sec \end{aligned} \tag{3.4}$$

Clearly, the rate of subunit loss depends strongly on the nucleotide bound states of the subunits in a particular configuration of a PF tip and the subunit below it. This adds further support for requiring a closer study of the configurations involved in the tip region to gain a deeper understanding of what drives MT dynamics. Furthermore, it is interesting to note that if the breaking rate were to be uniform for all subunit types, and $k_{depoly}^{GTP} = k_{depoly}^{GDP}$, and the rate $K_{shorten} = k_{depoly}^{X_i}$ were the same for all $i$, then the rate of subunit loss $R_{subunit\ loss} = \sum_{s=1}^{n_{tip}} sK_{shorten}$ which is in agreement with the shortening term in Equation (2) used for the mean-field study of [55]. However, the model developed here is concerned with a more detailed look at the MT structure, and the different subunit types are always considered to affect the depolymerization rates, which is effectively one way of featuring the differences between the straight orientation of GTP-bound subunits and the bent GDP-bound subunits.

At this point, the rates of subunit addition and loss, $R_{subunit\ addition}$ in Equation 3.1 and $R_{subunit\ loss}$ in Equation 3.4 respectively, have been formulated. These are the rates that alter the length of PFs, and effectively the measured length of the entire MT structure. The MT length is what can be measured in experimental observations, and is the common domain where simulated data can be compared to *invivo* data. More details of extracting macro-level information are discussed in Chapter 4. These two

rates presented in this section are important to describe how the total content of the MT is prone to change for one specific structural configuration. The rate of subunit loss in particular will be used as a feature describing the tip region for the predictive models developed in Chapter 5. However, polymerization and depolymerization are not the only ways of altering the MT structure. Hydrolysis events also change the state of individual subunits, which in turn changes the rate at which other reaction events can occur. Moreover, lateral bonds forming and breaking changes the position of the top-most lateral bond between PFs, which in turn changes the number of subunits that are susceptible to depolymerization. The next section describes the structural states and the transitions between them taking into consideration all of the reaction events considered in the MT dynamics of this computational model.

## 3.4 MT States and the Master Equation

As part of a Markov process, the MT is assumed to make stochastic transitions representing molecular reaction events that step the biopolymer from one structural state to another, without any dependence on older states. A specific combination and the orientation of the structural components in a 13-PF MT described in the previous section are involved in constructing any one MT structural configuration. The possible reaction events that take place depend on the structural components near the MT tip. However, the tip region alone is not enough to define the states of a Markov process. In this section, the MT structural components and the reaction events introduced in Sections 3.2 and 3.3.1 will be used to describe the state space for the model being developed in this study, as well as the transition probabilities between those states in the form of a master equation.

### 3.4.1 The State Space $S$

Random hydrolysis events leave GTP-bound subunits scattered throughout the top portion of the MT structure, and their positions need to be tracked as part of the MT structure in the Markov process. Additionally, lateral bond breaking events extend the cracked portion, and introduce new subunits into the PF tip configurations. If only the tip region was considered the states of the Markov chain, additional information of the older subunits in the MT structure would be necessary to identify the target state after a lateral bond breaking event. Requiring this additional information violates the Markov memoryless property [57]. Thus, the tip region alone is not adequate to represent the individual states at each step of the stochastic process. The same issue of requiring additional information of older subunits would persist even if larger sections beyond the tip region were included in the state's definition, since there is always a non-zero probability that hydrolysis has not changed the nucleotide-bound state of one particular subunit deep in the MT structure, which in turn would alter the transition rates for events that involve that subunit. For this reason, the entire MT structural configurations is chosen to define the states at each step of the Markov process being modeled. This does not change the fact that the tip region is still the focal point dictating how most transitions are to occur. Instead, relevant information about the tip is extracted from the entire MT's configuration as it is being simulated.

Furthermore, from a practical implementation perspective, the computational cost of tracking the entire MT structure at each step is not an issue for the processing memory currently available, and even long time simulations are completed within a reasonable amount of wall-clock time. Especially for the tubulin concentration levels being used as the input parameter in this study, DI behavior is modeled in the simulations, which means that eventually even the longest MT configuration will experience a catastrophe event, triggering a rapid shortening of the polymer

structure, and thus freeing the processor memory of the thousands of subunits being tracked. Of course, lower tubulin concentration levels would make catastrophe events more frequent, which would prevent very long MT structures from being created, and thus speeding up computational time. However, it should be noted that for tubulin concentrations above the critical concentration for persistent growth would make catastrophes infrequent, and this would allow for MT structures to grow very long, ultimately taking up a great amount of processor memory and slowing down computational time.

Let $S = \{\text{all the MT structural configurations}\}$ be the state space for the Markov chain. The minimum structure in $S$ would be the MT seed, and the rest of $S$ would contain a combination of any sequences of GTP- or GDP-bound subunits that construct different PFs on top of the MT seed, and any sequence of lateral bonds between those PFs, given that the top-most position of any sequence of lateral bonds has a position at the same height or below the height of the surround PFs. Based on *a priori* knowledge, there is a higher likelihood for some of the MT structures in $S$ to be observed than others. However, for completeness, the stochastic nature of the Markov chain model being developed requires the infinite number of combinations of possible MT structural configurations to be considered in $S$. To describe the probability of a MT structure attaining any one configuration, the master equation for this Markov process is developed in the remainder of this section.

### 3.4.2 Master Equation

The possible reaction events that would transition a MT configuration in $S$ to another, along with their kinetic rates, were listed in Section 3.3.2. For a general case, the master equation describing the probability of being in any one state is desired, and by nature of a Markovian process, this depends on the likelihood of being able to transition into that state. Let a random variable $X$ represent any MT

configuration in the set $S$. Let $p(x) = p(X = x)$ be the occurrence probability of an arbitrary MT configuration $x \in S$. Then, the following master equation can be formulated:

$$\begin{aligned}
\frac{d}{dt}p(x) = & \sum_{y \in Y_{poly}(x)} p(y) k_{poly}^{GTP} \left( \frac{c c_{1/2}}{c + c_{1/2}} \right) \quad + \\
& + \sum_{y \in Y_{depoly}^{GTP}(x)} p(y) k_{depoly}^{GTP} \quad + \sum_{y \in Y_{depoly}^{GDP}(x)} p(y) k_{depoly}^{GDP} \quad + \\
& + \sum_{y \in Y_{break}(x)} p(y) k_{break}(y) \quad + \sum_{y \in Y_{bond}(x)} p(y) k_{bond}(y) \quad + \\
& + \sum_{y \in Y_{S-break}(x)} p(y) k_{S-break}(y) \quad + \sum_{y \in Y_{S-bond}(x)} p(y) k_{S-bond}(y) \quad + \\
& + \sum_{y \in Y_h(x)} p(y) k_h \quad - \quad p(x) \sum k_* \qquad (3.5)
\end{aligned}$$

where the alternative notations for the kinetic rates represent the following:

$$k_{break}(y) = \begin{cases} k_{break}^{TT} & \text{if } y \text{ has a non-seam lateral bond w/ two GTP-bound G-subunits} \\ k_{break}^{TD} & \text{if } y \text{ has a non-seam lateral bond w/ GTP- and GDP-bound G-subunits} \\ k_{break}^{DT} & \text{if } y \text{ has a non-seam lateral bond w/ GDP- and GTP-bound G-subunits} \\ k_{break}^{DD} & \text{if } y \text{ has a non-seam lateral bond w/ two GDP-bound G-subunits} \end{cases}$$

$$k_{bond}(y) = \begin{cases} k_{bond}^{TT} & \text{if } y \text{ has a non-seam crack w/ two GTP-bound AG-subunits} \\ k_{bond}^{TD} & \text{if } y \text{ has a non-seam crack w/ GTP- and GDP-bound AG-subunits} \\ k_{bond}^{DT} & \text{if } y \text{ has a non-seam crack w/ GDP- and GTP-bound AG-subunits} \\ k_{bond}^{DD} & \text{if } y \text{ has a non-seam crack w/ two GDP-bound AG-subunits} \end{cases}$$

$$k_{S-break}(y) = \begin{cases} k_{S-break}^{TT} & \text{if } y \text{ has a seam lateral bond w/ two GTP-bound G-subunits} \\ k_{S-break}^{TD} & \text{if } y \text{ has a seam lateral bond w/ GTP- and GDP-bound G-subunits} \\ k_{S-break}^{DT} & \text{if } y \text{ has a seam lateral bond w/ GDP- and GTP-bound G-subunits} \\ k_{S-break}^{DD} & \text{if } y \text{ has a seam lateral bond w/ two GDP-bound G-subunits} \end{cases}$$

$$k_{S-bond}(y) = \begin{cases} k_{S-bond}^{TT} & \text{if } y \text{ has a seam crack w/ two GTP-bound AG-subunits} \\ k_{S-bond}^{TD} & \text{if } y \text{ has a seam crack w/ GTP- and GDP-bound AG-subunits} \\ k_{S-bond}^{DT} & \text{if } y \text{ has a seam crack w/ GDP- and GTP-bound AG-subunits} \\ k_{S-bond}^{DD} & \text{if } y \text{ has a seam crack w/ two GDP-bound AG-subunits} \end{cases}$$

$$\sum k_* = \sum \{\text{kinetic rates for all the events that } x \text{ can undergo}\}$$

and where the configurations $y$ in each summation term belong to the following subsets of $S$:

$$\begin{aligned}
Y_{poly}(x) = &\quad \{y \in S |\ y \text{ can polymerize a GTP-bound subunit to form } x\} \\
Y_{depoly}^{GTP}(x) = &\quad \{y \in S |\ y \text{ can break a longitudinal bond above a GTP-bound subunit to form } x\} \\
Y_{depoly}^{GDP}(x) = &\quad \{y \in S |\ y \text{ can break a longitudinal bond above a GDP-bound subunit to form } x\} \\
Y_{break}(x) = &\quad \{y \in S |\ y \text{ can break a non-seam lateral bond to form } x\} \\
Y_{bond}(x) = &\quad \{y \in S |\ y \text{ can form a a non-seam lateral bond to form } x\} \\
Y_{S-break}(x) = &\quad \{y \in S |\ y \text{ can break a seam lateral bond to form } x\} \\
Y_{S-bond}(x) = &\quad \{y \in S |\ y \text{ can form a lateral bond in a seam to form } x\} \\
Y_{h}(x) = &\quad \{y \in S |\ y \text{ can hydrolyze a GTP-bound subunit to form } x\}
\end{aligned}$$

Note that each term $\sum_y p(y)k_i$ represent all the ways that the MT structure can transition into configuration $x$ via the reaction event associated with $k_i$. Furthermore,

the term being subtracted on the RHS of the master equation, $p(x)\sum k_*$, represents all the ways a MT can transition out of state $x$. In typical circumstances, the master equation seen here may be simpler, since there may not exist some MT structural configurations in $S$ that would transition into configuration $x$ via certain reaction events, which would render the corresponding $Y_i(x)$ to be an empty set.

## 3.5 Simulating the Extended 13-PF MT Model

This chapter has developed the structure and reactions rates for the MT system, which represent the states and transitions in a Markov process. In order to model the evolution of a MT structure through time, the represented kinetic chemical reactions are chosen to occur in a stochastic order, where the occurrence probabilities depends on the structural state at a given point in time. For this, a kinetic Monte Carlo scheme resembling the Gillespie Algorithm is utilized to choose the order of reaction events to occur, and the time durations associated with each reaction. The remainder of this section describes the type of Gillespie Algorithm used to implement the detailed computational model simulations, and the benefits of using this method to represent the structural changes that a MT polymer can undergo.

### 3.5.1 The Gillespie Stochastic Simulation Algorithm

In this extension of the detailed level computational model for MT dynamics, the hydrolysis approximation is removed by applying equal treatment to all of the reaction events that can alter the MT structural configuration. By doing so, the goal is to develop an exact method which can simulate the time evolution of transitions that are possible to achieve from one structural state to another, and creates a statistically correct chemical reaction trajectory. This original algorithm was presented by Daniel T. Gillespie in his 1976 and 1977 papers [29, 30], where he presented the "direct" and "first-reaction" methods of implementing a Monte Carlo scheme for molecular

reactions in a volume $V$. The approach presented here is a variation that resembles the "first-reaction" method, which produces non-uniform continuous time steps, and is developed to suit the needs of the MT system. Essentially, the algorithm randomly samples a waiting time for each possible event, and selects the event with the smallest time to dictate the transition at each step.

Recall that the set $S$ represents the different structural configuration states that a MT can take on, and let $x \in S$ be an arbitrary configuration state. Let $R_i$ be one of the five types of unidirectional reaction events that can transition $x$ into another structural configuration. Note that the integer index $i$ depends on the state $x$, which restricts $R_i$ to one of the following types of reactions events that can occur:

- Polymerization: subject to tubulin concentration levels, independent of $x$
- Depolymerization: subject to the number and types of subunits in PF tips in $x$, and the G-subunits
- Lateral Bond Formation: subject to the types of AG-subunit pairs in $x$
- Lateral Bond Breaking: subject to the types of G-subunit pairs in $x$
- Hydrolysis: subject to the number of eligible GTP-bound subunits in $x$

For each reaction $R_i$, the corresponding reaction rate constants, the parameters $k_i$ have been presented with more detail in Section 3.3.2. To develop the algorithm, it is necessary to make the following fundamental hypothesis, and the "only assumption", which states that the reaction parameter $k_i$ can be defined as:

$$k_i \delta t \equiv \text{average probability, to first order in } \delta t, \text{ that a reaction } R_i \text{ will occur in the next time interval } \delta t. \tag{3.6}$$

Note that the probability of more than one reaction occurring during the interval $\delta t$ is $o(\delta t)$. Since the limit $\delta t \to 0$ will eventually be taken, it is reasonable to restrict

the number of reaction events to occur during the interval $\delta t$ to be no more than one [29].

The Monte Carlo step in this algorithm requires sampling waiting times by using uniformly distributed random numbers between 0 and 1. From the hypothesis in 3.6, and the theory outlined in [29], it follows that the waiting times between reaction events follow a Poisson process, such that the time duration of each possible reaction event can be sampled as:

$$\Delta t_i \ = - \frac{\ln(r_i)}{k_i} \tag{3.7}$$

where each $r_i \sim U[0,1]$ is randomly generated for each possible reaction $R_i$. Note that the time values being sampled will represent continuous-time durations, and do not depend on uniform time steps between reaction events. Once a time has been sampled for all possible reaction events, the reaction $R_i$ corresponding to the smallest waiting time $\Delta t_i$ is chosen to be the transition event to occur at that step in the simulation. Clearly, those reaction events with larger corresponding rates will have a better chance of being selected as the event to occur. Some reaction events (depolymerization and lateral bonds forming and breaking) strongly depend on the structural configuration, so for these reactions, the index $i$ is specific to the particular reactant component. For example, in the 13-PF MT model, if there are only 10 possible locations where lateral bonds can form, then 10 of the waiting times that are sampled would be associated to those 10 specific lateral bonds forming. However, polymerization and hydrolysis only require a single time value to be sampled; there is only one way for the polymerization rate to be calculated using the Michaelis-Menten formula, and the hydrolysis rate is a scalar multiple between the rate constant $k_h$ and the number of hydrolyzable GTP-bound subunits available in a configuration state. Once one of these reactions corresponds to the smallest sampled time duration for the transition

event, then the specific PF or GTP-bound subunit that is to be the reactant can be chosen stochastically by generating a new random integer accordingly.

So, using this method, the following algorithm steps are used in the stochastic simulation for computational MT model presented here:

- Step 0: Choose an initial configuration $x \in S$ at time $t = 0$, usually the MT seed.

- Step 1: Determine the reaction rates $k_i$ for the possible events $R_i$ that the current configuration state $x$ can undergo, at the current time $t$. (For events not possible for configuration $x$, assign $k_i = 0$.)

- Step 2: Generate a random number $r_i$ for each possible event $R_i$ to sample waiting times $\Delta t_i$ using Equation 3.7. (Those impossible events will be assigned a very long waiting time).

- Step 3: Determine the index for the smallest waiting time,

$$i^* = \arg\min_{\forall i}(\Delta t_i)$$

- Step 4: Evolve the MT configuration $x$ with the reaction event $R_{i^*}$ corresponding to smallest waiting time found in Step 3

  - Step 4.1: If $R_{i^*}$ is polymerization, then generate a new random integer from 1 to $N_{PF}$ = (the number of PFs in the MT) to choose which PF receives the new subunit.

  - Step 4.2: If $R_{i^*}$ is hydrolysis, then generate a new random integer from 1 to $N_h$ = (the number of hydrolyzable GTP-bound subunits in $x$) to choose one eligible GTP-bound subunit as the reactant.

- Step 5: Update the current time with the time duration from reaction $R_{i^*}$,

$$t = t + \Delta t_{i^*}$$

- Step 6: Repeat Steps 1-5 until $t$ reaches the desired total simulation time.

Utilizing the Gillespie stochastic simulation algorithm as laid out in Steps 1-6 simulates a trajectory of MT structural states transitions that are chemically possible, with corresponding waiting times that are statistically correct. A computational

model following this algorithm will generate the output desired to study the MT structural features at any given moment in the simulation, particularly at those times where significant dynamic changes to the total MT length are detectable from a macro-level perspective. This model, brings this study one step closer towards the goal of determining the tip structure of a MT during catastrophe and rescue events, and to gain a deeper understanding of the mechanisms that lead to those drastic changes observed in DI behavior.

## 3.6 Model Parameter Values

The model parameters are comprised of the kinetic rate constants for the different reaction events that are possible, and the input parameter representing the tubulin concentration levels in the cytoplasm available for polymerization. The values used for the computational model developed here utilize the kinetic rate constants listed in Table 3.1. They were obtained from "parameter set C" in [56], which is one of three parameter sets tuned to generate reasonable DI behavior that match the experimental observations of MT systems. Some of the kinetic rate constant values were selected within a reasonable range of bio-chemically acceptable values, as well as their relation to other rate constants. The calibration of the entire "parameter set C" values was conducted through a trial and error approach, and is admittedly not an exhaustive method. Thus, there probably exist other sets of values for the kinetic rate constants that simulate behavior resembling DI behavior.

However, "parameter set C" is chosen as preferred set for its characteristics matching the manner in which catastrophe and rescue phase transition occur, using the prior generation of the model implementing the hydrolysis approximation. Since that approximation was a reasonable one, the resulting differences in simulated DI behavior is minimal for the extended computational model being presented here. In other words, the simulated DI behavior in the newer, extended model is very similar to

that created by the older model, especially when taking into consideration the highly stochastic nature of the system being modeled.

Furthermore, the purpose of this study is to focus on the structural details, and the MT state trajectories are more complete with the new extended model. "Parameter set C" combined with a $10\mu M$ tubulin concentration level has been used in the past for simulating MT behavior rich in catastrophe and rescue transition events [34, 47, 55, 56]. This combination of parameters has been satisfactory, therefore pursing a parameter set more closely representing experimental conditions was not attempted. For this reason, the same combination of parameters is used in this study, and more details for measuring the DI behavior from a macro-level perspective are discussed the Chapter 4.

## 3.7 Computational Implementation

As it has been mentioned throughout this chapter, the purpose of extending the detailed level stochastic model of [55, 56] was to unearth the structural details that can be observed during key dynamic moments, like catastrophes and rescues. The approximation treatment used by the prior generation of the stochastic model skips over possible structural configurations by allowing multiple subunits to hydrolyze after one of the other four reaction events take place. Ultimately, this approximation was reasonable, especially when making macro-level measurements on DI behavior. The main benefit offered by removing this approximation and following the Gillespie Algorithm more closely, is that the trajectory of structural configuration states that is bio-chemically exact, and the time durations between reaction events is statistically correct. Thus, the newer model simulations reveal more information on the times spent in each structural state by a MT polymer.

TABLE 3.1

PARAMETER VALUES FOR KINETIC RATE CONSTANTS USED IN THE 13-PF COMPUTATIONAL MODEL

| | | |
|---|---|---|
| Polymerization | $k_{poly}^{GTP} = 1.25$ | |
| Depolymerization | $k_{depoly}^{GTP} = 0.02$<br>$k_{depoly}^{GDP} = 20$ | |
| Lateral Bond Forming | Non-seam Bonds | Seam Bonds |
| | $k_{bond}^{TT} = 100$ | $k_{S-bond}^{TT} = 200$ |
| | $k_{bond}^{TD} = 100$ | $k_{S-bond}^{TD} = 200$ |
| | $k_{bond}^{DT} = 100$ | $k_{S-bond}^{DT} = 200$ |
| | $k_{bond}^{DD} = 100$ | $k_{S-bond}^{DD} = 200$ |
| Lateral Bond Breaking | Non-seam Bonds | Seam Bonds |
| | $k_{break}^{TT} = 70$ | $k_{S-break}^{TT} = 140$ |
| | $k_{break}^{TD} = 90$ | $k_{S-break}^{TD} = 180$ |
| | $k_{break}^{DT} = 90$ | $k_{S-break}^{DT} = 180$ |
| | $k_{break}^{DD} = 400$ | $k_{S-break}^{DD} = 800$ |
| Hydrolysis | $k_h = 0.7$ | |

These rates are equivalent to "parameter set C" in [56]. They have units $sec^{-1}$, and are proportional to probabilities of occurrence per unit time.

### 3.7.1 A Lower Cost Implementation of Hydrolysis Reactions

The hydrolysis approximation in the former generation of the computational model was primarily implemented as a time saving tactic. The Monte Carlo step dealing with hydrolysis was a time consuming task, where each complete PF had to be searched for eligible GTP-bound subunits. The time durations and the order of reaction events (excluding hydrolysis) were sampled and selected in a similar man-

ner to the algorithm described earlier in this chapter. After each reaction event was selected to occur, the probability of an individual GTP-bounds subunit to hydrolyze during each time duration was estimated. As the MT structure was searched, and an eligible GTP-bound subunit was encountered, a separate Monte Carlo step was implemented to stochastically choose multiple subunits for hydrolysis. As the 13-PF MT grows longer, often thousands of subunits long, so does the computational cost for searching for eligible GTP-bound subunits and generating a random number for each one.

In order to ease this costly computational after removing the approximation, a speedup in the code was implemented. The code was first modified to track the number of eligible GTP-bound subunits in each PF, at each step of the simulation. This was not difficult, especially since each step of the simulation would change the MT structure one subunit at a time, except for depolymerization events, which required special treatment to determine how many eligible GTP-bound subunits it dissociated from the MT. This special circumstance particularly benefited from the binary representation of subunits that was used, where a 1 and 0 resembles a GTP- and GDP-bound subunit respectively. By tracking the number of hydrolyzable GTP-bound subunits in each PF, the total hydrolysis rate was easily computed. If a hydrolysis reaction event was chosen, then a new random number was generated to help determine which PF contains the reactant GTP-bound subunit. The likelihood of a PF being selected was determined by the number of hydrolyzable GTP-bound subunits it contained. Once a particular PF was determined, another random number was generated to stochastically choose which eligible GTP-bound subunit within that PF would transition into a GDP-bound state. With this new implementation, only a single PF is involved in the hydrolysis implementation, rather than the entire MT structure in the older model. Additionally, after a hydrolysis event has been selected, only two new random numbers are generated: the first when selecting the PF, and

the second for identifying a GTP-bound subunit within that PF. Recall that the old version generated a new random number for every GTP-bound subunit in the entire MT. So, the speedup in the new version of the code not only reduced the time to search for a GTP-bound subunit, but also reduced the amount of random numbers that need to be generated.

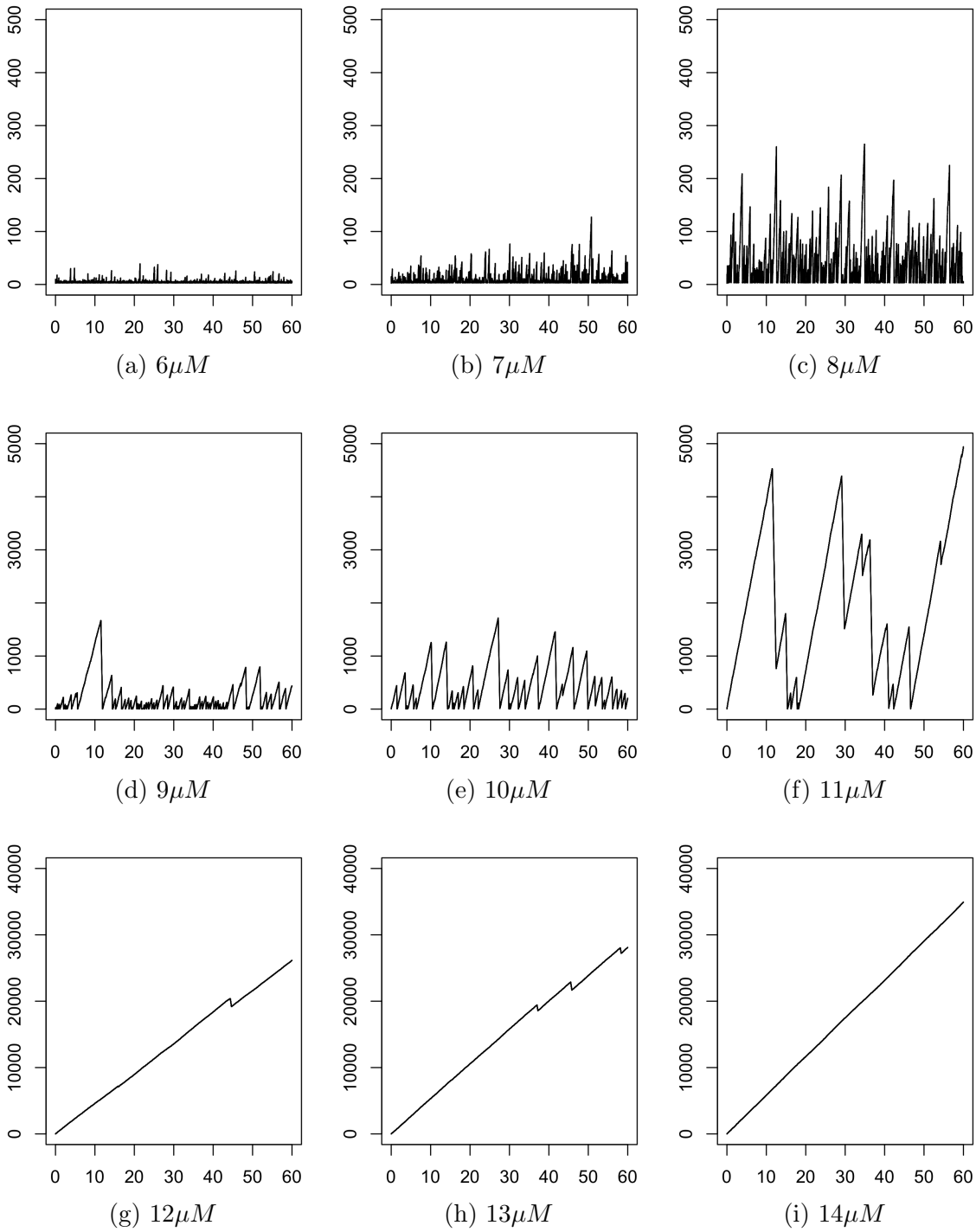
Removing the hydrolysis approximation effectively does add more steps throughout the simulation, since the structure after every hydrolysis event is now realized. This in turn should add to computational costs, and ultimately lengthen the total time to run the new code. However, it should be noted that utilizing the speedup significantly reduced the computational cost for dealing with any hydrolysis events. Generating a one-hour simulation takes $\approx 10min$, which is comparable to the wall-clock time of the older approximated version of the code reported in [55, 56], while considering fluctuations inherent to the stochastic nature of the simulation.

### 3.7.2 New Types of Data for the MT Tip Region

After the new version of the model and code were established, the code had to be re-run to generate new simulation output. In addition to the extra structural states being observed in an exact bio-chemical trajectory of states, the code was modified so that the simulation output included additional data about the MT tip structure that was previously not reported. The previous code created an output file that was mostly limited to the simulation step times and the total MT lengths. The new code output now includes the number of GTP-bound subunits (to help estimate the GTP-cap size), the height of each lateral bond, the nucleotide bound states of the G- and AG-subunits, and the length of each PF. In addition to all of this, an indexing system took advantage of the the binary representation of the subunits to track the actual configurations of PF tips that extend beyond their surrounding lateral bonds. The PF tip configuration indices help in defining the MT tip as a whole, and serves

to be beneficial for extracting additional information pertaining to specific sequences of subunits that appear in the tip region. For this reason, the new code had to be used to generate new information about the 13-PF model that was previously unavailable. By doing so, a more detailed study of the MT tip structural features during significant moments in DI behavior is possible.

Figure 3.2 displays one hour long length history plots of a single MT for various tubulin concentration levels generated from simulations using the parameter values in Table 3.1. The behavior is consistent with the results of the previous approximated version of the detailed 13-PF MT model simulations in [34, 47, 55, 56]. Note that Figures 3.2(a-c) display the near nucleation behavior for lower tubulin concentration levels, where the MT is rarely longer than 100 subunit; Figures 3.2(g-i) display the behavior regime resembling unbounded growth, and the MT rarely encounters catastrophe events; and Figures 3.2(d-f) display DI behavior as it is classically understood, where significantly long MTs are observed, yet frequent catastrophe events return the biopolymer structure to near seed levels. Furthermore, it is interesting to note that the $10\mu M$ tubulin concentration level is the one that generates MT length sufficiently long (i.e. $> 1000$ subunits long), while providing a rich number of catastrophe and rescue events. In contrast, $11\mu M$ tubulin concentration levels generate MT that are much longer, and this results in far fewer catastrophe and rescue events occurring in a given time frame, which provides less of a variety of data surrounding these key events to study. This was also the case in [34, 47, 55, 56], and the reason why those studies in the past, as well as the study conducted here, make use of $10\mu M$ as the preferred tubulin concentration level to simulate data displaying DI behavior. The data analysis administered in Chapters 4 will only make use of simulations for $10\mu M$ tubulin concentration in the 13-PF MT cases.




(a) $6\mu M$ (b) $7\mu M$ (c) $8\mu M$

(d) $9\mu M$ (e) $10\mu M$ (f) $11\mu M$

(g) $12\mu M$ (h) $13\mu M$ (i) $14\mu M$

Figure 3.2. Length history plots for tubulin concentration levels ranging from 6-14$\mu M$ in (a)-(i), from one hour simulations of the extended 13-PF MT model, and the parameter values defined in Table 3.1. The horizontal axis represents time in minutes, and the vertical axis is the length of the MT measured in number of subunits.

## 3.8 The Simplified 2-PF MT Model

The Gillespie stochastic simulation algorithm extension to the latest version of the computaional model was important to extract the detailed level structural trajectory and features of the tip region of a 13-PF MT. However, the 13-PF MT has an inherent caveat dealing with the complexity of the tip region structure. The variety of combinations of PF tips and differing lateral bond heights creates a very large number of possible tip configurations. Without even considering the complexity of where to define the cutoff for the tip region, individual PF tips can extend above the MT structure, or even groups of neighboring PFs structures that resemble sheets can form as different components of the MT tip structure. Additionally, the 13 lateral bonds between PFs can make it difficult to have a uniform row of subunits above which the tip region can be defined. However, the interplay between PFs and lateral bonds is part of the tip structure that requires a better understanding, and the 13-PF MT model is challenging to tackle as is. For this reason, a 2-PF simplification which includes lateral bond interactions is proposed in this section, such that it creates a more suitable situation for studying the tip structure of MT biopolymers.

This 2-PF model can be thought of as representing a generic two-PF polymer, or it can be thought of as two neighboring PFs embedded within a 13-PF MT lattice. The simplification from 13-PFs to 2-PFs provides a feasible setting to demonstrate a novel approach of simulating MT dynamics using mathematical and numerical methods, while analyzing data with statistical tools, including machine learning techniques, connect micro-level structural features to macro-level phases identified from a length history evolution of a MT. The 2-PF model presented in this section will be used in Chapters 4 and 5 to develop the predictive models that bridges the gap between the dynamics occurring at different scales, and potentially inspire the development of methods to study the MT structure at the 13-PF level.

### 3.8.1 Novelty of the 2-PF MT Model

When considering a MT formed of 13-PFs, which would have 13 lateral bonds between them, the tip region definition takes on many variations depending on where one decides to separate the tip from the rest of the MT structure. For example, if the cutoff is chosen as the shortest lateral bonds height position, then the tip may include a substantial section of PFs that are laterally bound to their neighbors, a relatively stagnant section of structure when compared to the laterally unbonded parts that are more inclined to change from the MT dynamics. This classical representation of a 13-PF MT serves too complex a scenario for tackling the tip structure study as is. In the effort to tackled the issues associated with the structural complexities inherent to the 13-PF MT model, it is important to recall the importance of lateral interactions in MTs [4, 5, 8, 12, 27, 34, 47, 55, 56, 76, 77, 82] as the need for a simpler representation of the MT tip configurations.

To this end, a 2-PF MT is presented: 2-PFs with a single lateral bond between the adjacent subunit pairs. Modeling the MT as a 2-PF polymer with one column of lateral bonds between two PFs is the simplest scenario that explicitly models the lateral bond dynamics, and also relaxes the complexities of the MT tip structure found in the full 13-PF case. This representation can be perceived as an arbitrary 2-PF polymer, or as a pair of neighboring PFs embedded within a larger 13-PF MT. Although this offers a less complicated scenario than the 13-PF case, it is worth recognizing that the proposed 2-PF case is relevant to actual two-stranded filaments that exhibit dynamic instability behavior, such as a protein called ParM [28]. However, since the scope of the study is making reference to model parameters associated with prior MT studies, the 2-PF structure and its behavior will be referred to as a special case MT. As is shown later in Section 3.8.5, the proposed 2-PF MT model is successful in simulating DI behavior using model parameters comparable to the 13-PF MT model. In doing so, attention can also be turned to the MT tip region while allowing for

the lateral interactions in dynamic instability to be more tractable. Thus, the 2-PF model provides a more manageable number of tip configurations to consider, and therefore more favorable conditions to study the most dynamic structural features of MT during DI behavior.

Effectively, in the 2-PF case, the single lateral bond allows for a single crack between the two PFs to be considered, and thus renders a strict "laterally unbonded" definition of the tip region to be applicable. Particularly, the tip configurations stored from the computational simulation output will refer to the top sections of both PFs lacking a lateral bond. For example, assuming that any PF-tip can have a maximum of $L = 10$ laterally unbonded subunits in it, and the MT tip configuration is defined as the collection of PF tip configurations (ordered sequence of GTP- and GDP-bounds subunits) protruding above its neighboring lateral bonds, then the number of unique MT tip configurations that are possible for each variations is as follows:

- 2-PF MT Tip Configurations:

$$\left(\sum_{j=0}^{L} 2^j\right)^2 = \left(2^{L+1} - 1\right)^2 = 4,190,209 \text{ for } L = 10 \tag{3.8}$$

- 13-PF MT Tip Configurations:

$$\left(\sum_{j=0}^{L} 2^j\right)^{13} \approx 1.108 \times 10^{43} \text{ for } L = 10 \tag{3.9}$$

Note that the latter value corresponding to the 13-PF MT tip configurations is larger than Avogadro's number, which makes an obvious case for the reduction in complexity offered by the 2-PF MT model. So, the 2-PF MT model allows for this tip region concept to be developed more easily, such that the Gillespie stochastic simulation algorithm can be used to simulate an exact trajectory of MT structural features, particularly focusing on the tip. This will be the model that is used in Chapter 5 to

test how well the structural features of the tip region from a micro-level can predict the significant dynamic changes that are detectable from a macro-level perspective. In other words, the goal of this complete study is to capture the exact structural configurations that lead to interesting DI events like catastrophes and rescues.

### 3.8.2 Model Structure

The simplified structure of the 2-PF MT utilizes a single sequence of lateral bonds between two neighboring PFs as illustrated in Figure 3.3. When compared to the 13-PF model, the fundamental differences that arise in the 2-PF case are embedded in the fact it lacks a helical tube form. Since the lattice of PFs does not wrap around and connect the first PF to the last one, the seam is not present. Without a seam, there does not exist a set of lateral bonds that deserve special treatment in terms of lateral bond formation/breakage dynamics. This is a simplification in the sense that a subset of kinetic rate constants are omitted from the 13-PF case, as well as not having to deal with the shift in neighboring subunit associations that occur there. Additionally, the 2-PF structure without a helical alignment means that the MT seed does not need to accommodate for the shift at the seam. Instead, the 2-PF MT seed is symmetrical, mean that the indestructible subunits in the seed corresponding to each PF, and the lateral bonds between them, have the same height. Otherwise, the seed structure plays the same role in either MT case, meaning that it acts as a typical initial condition, as well as the minimal structure to which a shortening MT can be reduced.

With the 2-PF conformation, the single lateral bond makes it simpler to define some of the other components for the 13-PF MT that have already been introduced earlier in Section 3.2. The most significant simplification, by design, is for the tip region which is most affected by the molecular dynamics. First, having a single sequence of lateral bonds limits the lateral bond height as the sole interface between

the laterally bonded portion of the structure and the cracked portion lacking lateral bonds. This creates a scenario where there is only a single pair for each of the neighboring G- and AG-subunits surrounding this gate or interface. Second, with limited complexity, the tip region can be defined using the pair of PF tips. A polymerization would add a single laterally unbonded GTP-bound subunit to a PF tip, while a depolymerization reaction events would target those subunits that lack a lateral bond already in the PF tips. Even lateral bond formation deals with only those subunits at the bottom of PF tips, the AG-subunits. However, the lateral bond breaking events depend on the subunits surrounding the top-most lateral bonds, the G-subunits. For this reason, the combination of just the PF tips that surround the crack are referred to as the **cracked tip**, while the more complete combination of PF tips and the G-subunits together is referred to as the **gated tip** (see Figure 3.3). As far as hydrolysis changing the structure, the uni-directional reaction that changes GTP-bound subunits to a GDP-bound state still manifests into the formation of a GTP-cap. However, the random selection of subunits to hydrolyze makes it difficult to clearly define the hazy boundary between a GTP-cap and the rest of the MT structure predominantly consisting of GDP-bound subunits. In Chapter 5, an estimate for measuring the GTP-cap is used to make up for this shortcoming.

### 3.8.3 Model Dynamics

Another difference arising from the simplifications of the 2-PF MT affects the dynamics of the lateral bonds. Recall that in the 13-PF MT case, the likelihood of a lateral bond breaking is reduced by a factor of $\pi_{break} = 1000$ when there already exist lateral bonds on the opposite sides of the subunits connected by the bond in question. However, the PFs in the 2-PF MT model only have a single sequence of lateral bonds. In order to utilize model parameters within a similar range of values as those found to create DI behavior in the 13-PF case, a modification is necessary. For

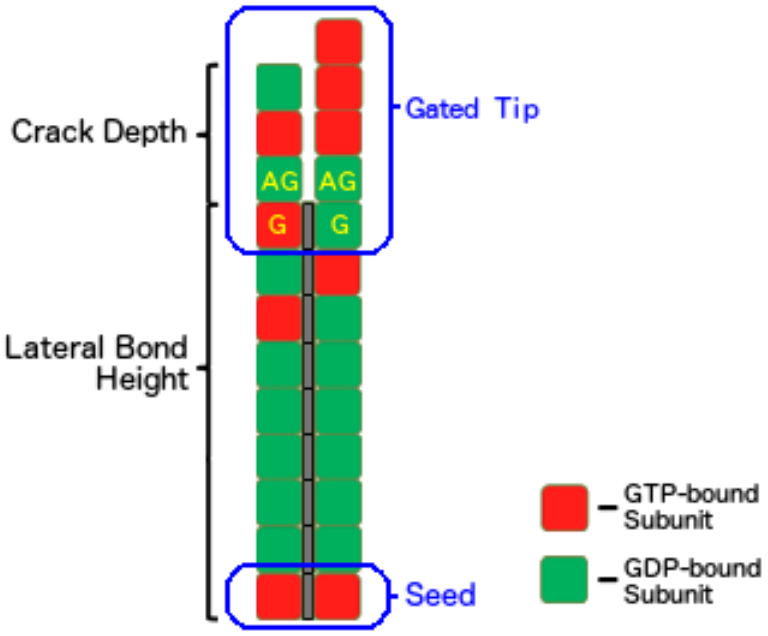



Figure 3.3. An arbitrary example of a 2-PF MT structure and its components. The red and green boxes represent GTP- and GDP-bound subunits respectively. The vertical sequence of subunits create the two parallel PFs in the structure. The gray boxes are the single sequence of lateral bonds allowed to form between PFs. The seed is the indestructible portion at the very bottom of the 2-PF MT structure. The gate and above-gate subunits (G- and AG-subunits respectively) are located near the interface of the top-most lateral bond, and the cracked portion of missing lateral bonds between PFs. The crack depth is measured by the number of subunits in the shorter PF tip. The gated tip is the combination of the individual laterally unbonded PF tips and the G-subunits together.

this, reference is made to effective breaking rate of a randomly picked lateral bond in [55], which is approximated as follows:

$$k_{break,eff} = (1-q)k_{break} + q\frac{k_{break}}{\pi_{break}} \tag{3.10}$$

where $q$ is the probability that a randomly picked lateral bond height will be connecting two PFs, such that the lateral bond heights on their opposite sides are not lower than the later bond height in question. In other words, if $X$, $Y$, and $Z$ were discrete

random variables representing neighboring lateral bond heights in a 13-PF MT such that $Y$ was the lateral bond height in the middle, then $q = P(X \geq Y, Z \geq Y)$. A scenario that satisfies the conditions of probability $q$ represents a lateral bond that has additional support in the PF lattice, it is less likely to break, which in turn reduces its effective rate of breaking by the factor $\pi_{break}$. Furthermore, [55] used $q = 1/3$ in the mean-field approximations assuming that $X$, $Y$, and $Z$ were independently and identically distributed, however knowing that they may be dependent yields that $1/3 \leq q < 1$ in a 13-PF MT with a shifted seam.

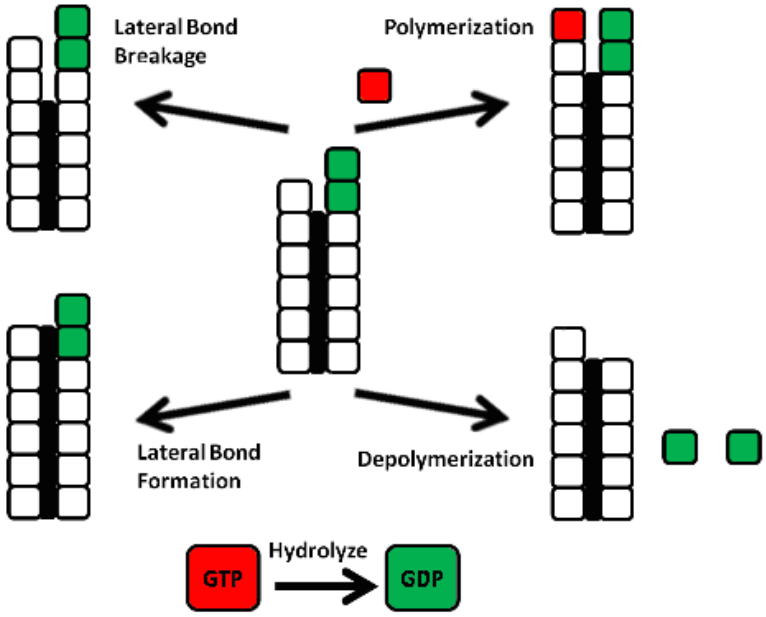



Figure 3.4. The five possible dynamic events that can change the 2-PF structure. Polymerization can lengthen a PF tip by one GTP-subunit. Depolymerization can remove a consecutive sequence of laterally unbonded subunits in a PF tip. The top-most lateral bond can break. A new lateral bond can form immediately above the top-most lateral bond, given that the space between two laterally neighboring subunits exists. Hydrolysis can irreversibly change a GTP-bound subunit into a GDP-bound state (adapted from [46]).

In an effort to use the same values for the remaining kinetic rates of the parameter set that delivered DI behavior with a similar range of tubulin concentration that was comparable to the 13-PF case, a non-exhaustive sweep of appropriate values for $q$ revealed that $q = 2/3$ was a good choice. Separately, it can be shown that $q = 2/3$ is a good choice if it were assumed that the likelihood that the height of a lateral bond was the same height as the neighboring lateral bonds to the left and right of it followed a two-dimensional Gaussian distribution, and the probability formulas reported in [55] were used to calculate $q$. So, the resulting effective breaking rate calculated from Equation 3.10 is utilized in the 2-PF MT simulations, and the lateral bond breaking rate constants are reduced by a factor of $(1 - q) = 1/3$. More specifically, using tubulin concentrations near $10\mu M$ are the preferred levels for delivering DI behavior in the 13-PF MT models, as discussed in the previous section. When using $q$ values below $2/3$, tubulin concentrations well above $10\mu M$ were needed to simulate the DI behavior regime. Conversely, for $q$ values above $2/3$, tubulin concentration levels near $10\mu M$ were already near the persistent growth regime, where polymerization rates easily overtake the simulated dynamics, and MTs grow very long due to catastrophe events being very rare; clearly this is not the desired DI behavior regime. However, when using $q = 2/3$, tubulin concentrations near $11 - 12\mu M$ generate simulated length history plots that certainly display DI behavior, rich with catastrophe and rescue events (see Figure 3.5).

Aside from this modification, the remaining features of the dynamic events that can alter the 2-PF MT structure remain the same as the 13-PF case, without having to make considerations for the seam. As illustrated in Figure 3.4, polymerization, depolymerization, lateral bond formation and breakage, and hydrolysis all obey the same rules as before. The corresponding kinetic rates constants for these events are listed in Table 3.2, where the modification for the effective lateral bond breaking

TABLE 3.2

PARAMETER VALUES FOR KINETIC RATE CONSTANTS USED IN THE 2-PF COMPUTATIONAL MODEL

| | | | |
|---|---|---|---|
| Polymerization | $k_{poly}^{GTP}$ | = | 1.25 |
| Depolymerization | $k_{depoly}^{GTP}$ | = | 0.02 |
| | $k_{depoly}^{GDP}$ | = | 20 |
| Lateral Bond Forming | $k_{bond}^{TT}$ | = | 100 |
| | $k_{bond}^{TD}$ | = | 100 |
| | $k_{bond}^{DT}$ | = | 100 |
| | $k_{bond}^{DD}$ | = | 100 |
| Lateral Bond Breaking | $k_{break}^{TT}$ | = | 23.333 |
| | $k_{break}^{TD}$ | = | 30 |
| | $k_{break}^{DT}$ | = | 30 |
| | $k_{break}^{DD}$ | = | 133.333 |
| Hydrolysis | $k_h$ | = | 0.7 |

Compared to Table 3.1, there are no lateral bond kinetic rates for a seam, and the breaking rates have been modified to allow for the effective breaking rates, which are reduced by a factor of 1/3. All rates have units $sec^{-1}$, and are proportional to probabilities of occurrence per unit time.

rates have already been applied.

### 3.8.4 State Space and Master Equation

Now that the MT model has been simplified, the structural states recognized through the Markov chain have changed along with the model. Let $\tilde{S} = \{$all the 2-PF MT structural configurations$\}$ be the state space for the Markov chain in the simplified 2-PF MT model. The minimum structure in $\tilde{S}$ would still be the MT seed,

and the rest of $\tilde{S}$ would contain any sequence of GTP- or GDP-bound subunits that construct different PFs on top of the MT seed, with a consecutive sequence of lateral bonds between them, given that the lateral bond height has a position at the same height or below the height of the surrounding PFs. This is similar to the general MT case, however the simplified 2-PF MT structure creates a state space $\tilde{S}$ far smaller than the space $S$ for 13-PF MTs. To describe the probability of a MT structure attaining any one configuration, the master equation for this Markov process is also developed.

The possible reaction events that would transition a 2-PF MT from one configuration in $\tilde{S}$ to another, along with their kinetic rates, were listed earlier in this section. By the nature of a Markov process, the possible reactions that can occur depends on the likelihood of being able to transition into that state. Let a random variable $\tilde{X}$ represent any MT configuration in the set $\tilde{S}$. Let $p(\tilde{x}) = p(\tilde{X} = \tilde{x})$ be the occurrence probability of an arbitrary 2-PF MT configuration $\tilde{x} \in \tilde{S}$. Then, the following master equation can be formulated:

$$
\begin{aligned}
\frac{d}{dt}p(\tilde{x}) = & \sum_{\tilde{y}\in\tilde{Y}_{poly}(\tilde{x})} p(\tilde{y})k^{GTP}_{poly}\left(\frac{cc_{1/2}}{c+c_{1/2}}\right) & + \\
& + \sum_{\tilde{y}\in\tilde{Y}^{GTP}_{depoly}(\tilde{x})} p(y)k^{GTP}_{depoly} & + \sum_{\tilde{y}\in\tilde{Y}^{GDP}_{depoly}(\tilde{x})} p(\tilde{y})k^{GDP}_{depoly} & + \\
& + p\left(\tilde{X} = \tilde{y}^{TT}_{break}(\tilde{x})\right)k^{TT}_{break} & + p\left(\tilde{X} = \tilde{y}^{TD}_{break}(\tilde{x})\right)k^{TD}_{break} & + \\
& + p\left(\tilde{X} = \tilde{y}^{DT}_{break}(\tilde{x})\right)k^{DT}_{break} & + p\left(\tilde{X} = \tilde{y}^{DD}_{break}(\tilde{x})\right)k^{DD}_{break} & + \\
& + p\left(\tilde{X} = \tilde{y}^{TT}_{bond}(\tilde{x})\right)k^{TT}_{bond} & + p\left(\tilde{X} = \tilde{y}^{TD}_{bond}(\tilde{x})\right)k^{TD}_{bond}ß & + \\
& + p\left(\tilde{X} = \tilde{y}^{DT}_{bond}(\tilde{x})\right)k^{DT}_{bond} & + p\left(\tilde{X} = \tilde{y}^{DD}_{bond}(\tilde{x})\right)k^{DD}_{bond} & + \\
& + \sum_{\tilde{y}\in\tilde{Y}_{h}(\tilde{x})} p(\tilde{y})k_h & - \quad p(\tilde{x})\sum k_* & (3.11)
\end{aligned}
$$

where the configurations $\tilde{y}$ in each summation term belong to the following subsets of $\tilde{S}$:

$$\begin{aligned}
\tilde{Y}_{poly}(\tilde{x}) = &\quad \{\tilde{y} \in \tilde{S} |\ y \text{ can polymerize a GTP-bound subunit to form } \tilde{x}\} \\
\tilde{Y}^{GTP}_{depoly}(\tilde{x}) = &\quad \{\tilde{y} \in \tilde{S} |\ y \text{ can break a longitudinal bond above} \\
&\qquad \text{a GTP-bound subunit to form } \tilde{x}\} \\
\tilde{Y}^{GDP}_{depoly}(\tilde{x}) = &\quad \{\tilde{y} \in \tilde{S} |\ y \text{ can break a longitudinal bond above} \\
&\qquad \text{a GDP-bound subunit to form } \tilde{x}\} \\
\tilde{Y}_{h}(\tilde{x}) = &\quad \{\tilde{y} \in \tilde{S} |\ y \text{ can hydrolyze a GTP-bound subunit to form } \tilde{x}\}
\end{aligned}$$

Also, each $\tilde{y}^{ij}_{break/bond}(\tilde{x})$ configuration corresponds to the individual configurations in $\tilde{S}$ that can transition into $\tilde{x}$ through a lateral bond breaking/bonding reaction event, subject to the AG/G-subunits pairs being $ij \in \{TT, TD, DT, DD\}$ nucleotide bound states. The kinetic rate constant notations the same as those defined for the 13-PF MT case in Section 3.4.2, except the last term in Equation 3.11, which uses the following interpretation:

$$\sum k_* = \sum \{\text{kinetic rates for all the events that } x \text{ can undergo}\}.$$

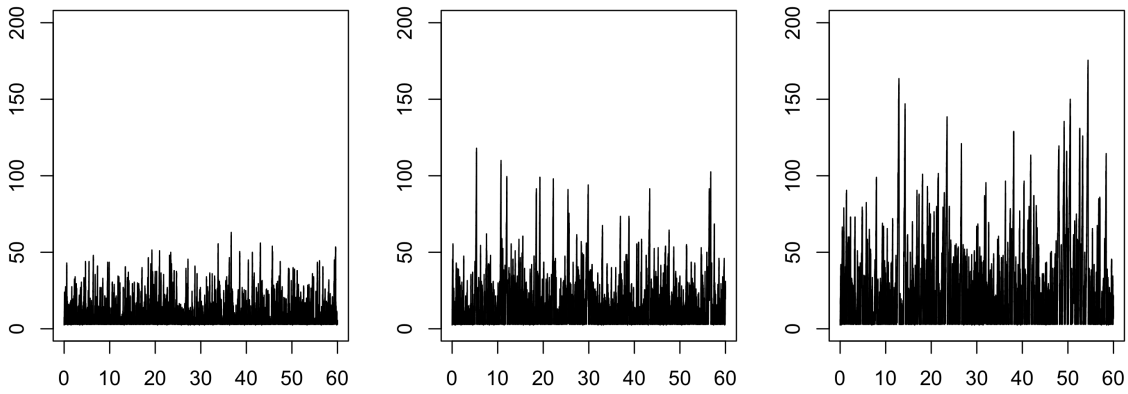
### 3.8.5 Model Simulations

Figure 3.5 displays one hour long length history plots of a single 2-PF MT for various tubulin concentration levels generated from simulations using the parameter values in Table 3.1. The modification from the effective lateral bond breaking rates helped to make the 2-PF MT model output consistent with the results of the 13-PF case, though there are some inevitable differences between the full MT model and the simplified version. Note that Figures 3.5(a-c) display the near nucleation behavior for lower tubulin concentration levels, where the MT is rarely longer than 100 subunit;

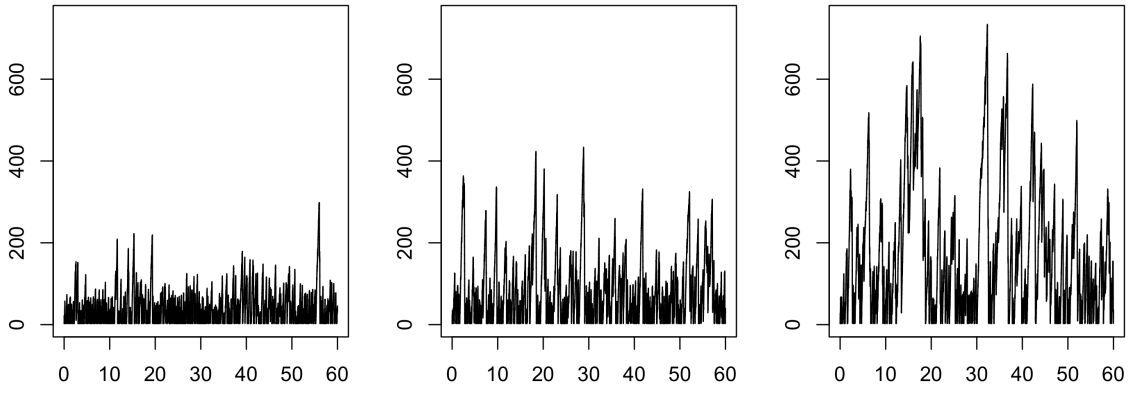
in Figures 3.5(g-i) display the behavior regime resembling unbounded growth, and the MT rarely encounters catastrophe events; and in Figures 3.5(d-f) display DI behavior as it is classically understood, where significantly long MTs are observed, yet frequent catastrophe events unravel the biopolymer structure back to near seed levels. However, in the 2-PF MT case, it is the $12\mu M$ tubulin concentration level the one that generates MT length sufficiently long with a rich number of catastrophe and rescue events, prior to entering the persistent growth regime (seen with the $13\mu M$ tubulin concentration level).

Some minor adjustments to the probability $q$ in Equation 3.10 could help line up the two concentration levels more perfectly, but the stochastic nature, and the inherent differences between the two computational models would make this a difficult task, and doing so is beyond the scope of this study. Recall that the goal is to create a simplified MT model that generates simulations with DI behavior in order to study the structural features near key moments of phase transitions. At this point, the 2-PF meets the qualifications, however it is interesting to note some of the characteristic differences in the length history profiles created by the two models. Specifically, the MT length is more sensitive to changes in the 2-PF MT, because the MT length is measured as an average over the PF lengths. Since fewer reaction events are required to alter the average measurements of the 2-PF MT, the length history profiles in Figure 3.5 tend to be more sporadic, with more fluctuations in the length during a given time duration. In contrast, the 13-PF MT length history plots in Figure 3.2 tend to be more steady, since changes to MT length are averaged over 13 PFs. Particularly the growth to shortening profiles before and after a catastrophe event tend to be broader at the preferred $10\mu M$ tubulin concentration level for the 13-PF MT model, whereas the growth to shortening profiles in the $12\mu M$ counterpart for the 2-PF MT model tend to last shorter periods of time, and tend to have more interruptions. In any case, $12\mu M$ is the preferred choice of tubulin concentration

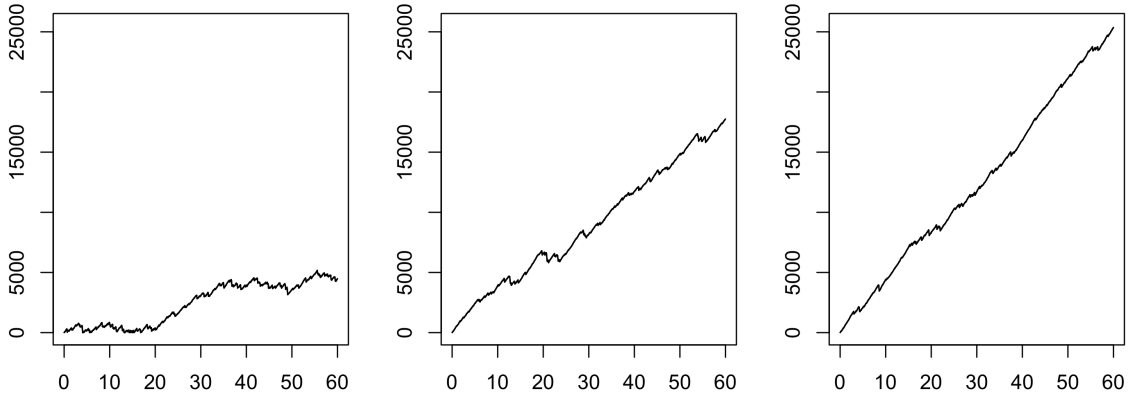
levels for the 2-PF MT simulations displaying DI behavior, and will be used for the data analysis administered in Chapters 4 and 5 for its rich variety in catastrophe and rescue transitions.




(a) $7\mu M$ (b) $8\mu M$ (c) $9\mu M$




(d) $10\mu M$ (e) $11\mu M$ (f) $12\mu M$




(g) $13\mu M$ (h) $14\mu M$ (i) $15\mu M$

Figure 3.5. Length history plots for tubulin concentrations ranging from 6-14$\mu M$ in (a)-(i), from one hour simulations of the simplified 2-PF MT model, and the parameter values defined in Table 3.2. The horizontal axis represents time in minutes, and the vertical axis is the length of the MT measured in number of subunits.

# CHAPTER 4

# DATA ANALYSIS I: A NOVEL METHOD TO IDENTIFY MACRO-LEVEL PHASES AND DETERMINE DYNAMIC INSTABILITY PROPERTIES

Dynamic instability (DI) behavior is observed in the length history plots of MTs, and they are characterized by the sporadic and sudden switches from sustained periods of growth to much more rapid shortening (catastrophe), and the rare switch back to sustained growth (rescues). Length history data is considered to be the macroscopic level measurements in this study since this perspective is concerned with the length of the entire MT, and not the detailed structural configurations. Measurements that describe dynamic instability include growth and shortening rates, and the frequency of catastrophe and rescue occurrences. The appropriateness of these measurement come into question as data acquisition methods have improved with finer resolutions, thanks to advancements in lab experimental conditions and detailed-level simulations. The unpredictable and high-frequency nature of recent MT length data now challenges the accuracy of pinpointing the location of dynamic changes of prior methods that have been used to make dynamic instability measurements. To accommodate this recent data, an improved approach is desired to help characterize the dynamic instability behavior, and in doing so, expose any new features that had previously been overlooked in the coarser resolution data from the past. In this chapter, a computational tool is presented to apply an automated approach to segment periods of consistent dynamic behavior with a continuous piece-wise linear approximation, and to classify these segments into dynamic phases in order to facilitate the respective DI measurements.

The development of this tool took into consideration any data displaying dynamical instability, regardless of its source being lab experiments or simulations. An adaptive approach accommodates the stochastic data by first identifying periods of consistent behavior, and then classifying the possible phases using an unsupervised method, $K$-means clustering. Using this tool on simulated data from the detailed MT model developed in Chapter 3 revealed the existence of a phase class consisting of segments with attenuated dynamic behavior, called "stutters", which were different from and intermediate to the classically recognized growth and shortening phases. The term stutter is used, because the MT structure changes that occur during these periods create short fluctuations to MT length, but the total affect on MT length throughout this period is small, especially compared to the the total change in length made during growth and shortening periods. Finally, the possible dynamic phase change patterns were analyzed considering all possible combinations of these classes. The results showed that a significant number of stutter phases occur between the switch from growth to shortening, hence characterizing their transitional role during the catastrophe phenomena. In past experimental studies, "slow-down" periods were detected before the onset of shortening periods, however they were not separated and quantified as a different class of behavior [21]. Instead, what are referred to as stutters were lumped together with the growth periods, leading to not only overlook a possible third phase of DI behavior, but also introducing errors to measuring growth rates. Thus, the DI phase classification method introduced in this chapter not only improves upon the accuracy for measuring DI rates, but also allows for the separate treatment of a third phase of DI.

It should be noted that these stutter phases are a different type of third phase class in MT dynamics than the "pause" periods identified in previous biological studies [21, 70]. In fact, those "pauses" have been attributed to different circumstances, including affects from MT binding proteins (MTBPs) in both *in vivo* and *in vitro*

experiments, as well as situations where the MT becomes stuck to the glass surface substrate in *in vitro* experiments. These situations give rises to a stall or pause in MT dynamics where little to no structural changes take place. They can last on the order of 30 seconds to several minutes, and are rarely observed in experiments with pure tubulin, which is the scenario being considered in this study. Pauses have been studied, and their role as a third phase has been identified more as a period of no MT structural change, typically occurring after the MT has depolymerized to nucleation levels and cannot regain growth easily[72, 83, 84]. In contrast, the stutter phases are more limited in their duration, lasting a few seconds at most, and they capture periods of MT dynamics where the MT structure is actively changing without contributing significant changes to the MT length as a whole. The stutters identified in this study were all detected from MT lengths well above nucleation levels. Dilution induced experiments do show stutter-like delay periods of attenuated dynamics emerging as in the length history plots prior to the onset of a rapid depolymerization period, however they were not studied a separate phase [56, 79, 81]. Additionally, they have been overlooked completely in other studies by only assuming that growth and shortening phases make up the DI behavior regimes [51].

The term "stutter" being used in this dissertation should be distinguished from the terminology used in studies associated with protein structure sequences [7, 13]. The nano-scale studies of protein structure sequences utilized the terms skip, stutter, and stammer to describe different patterns of discontinuities that occur in sequences of protein structures, where stutters specifically refer to the deletion of three residues in the heptad repeats of $\alpha$-helical coiled-coil sequences [7]. This is different than the terminology used in this dissertation, where the term stutters describe the new macro-level phase in DI behavior found in the MT length history plots.

The computational tool presented in this chapter does not make any assumptions that restrict the number of possible phases that can exist in DI. Thus, this approach

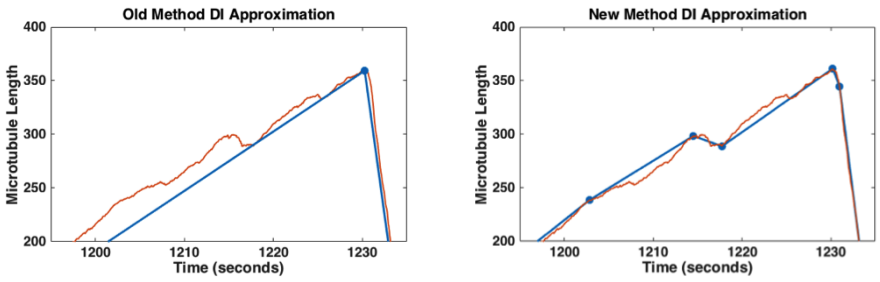
not only makes it possible to detect and accurately measure DI parameters, but also consistently provides the precise moments for when major dynamic changes occur, a key feature that opens up future work dealing with the detailed level MT structures associated with catastrophe and rescue events. Being able to conduct the macro-level data analysis provided by this DI phase classification tool is a necessary step to study the micro-level mechanisms that lead to DI phase transitions.

## 4.1 Motivation for an Improved Phase Classification Method

Dynamic instability in MTs has classically been studied from a macro-level perspective with regards to the MT structure by utilizing the length history profile, or the evolution of a single biopolymer's length in time. After all, it is in the length history data plots where catastrophe and rescue events are labeled. Recent technological advancements in the lab have provided more detail with respect to finer time and space resolution in data collected for MT lengths, and this has generated the need for developing automated tools to measure this data more accurately [15]. Additionally, novel computational models simulate individual molecular level reactions involved in the polymerization dynamics, and thus provide a very rich set of detailed level data on the evolution of the MT length, including the models in [34, 47, 56, 85], and the computational models presented in Chapter 3. However, when it comes to measuring the dynamic instability parameters that describe the characteristic behaviors seen in the data, it becomes increasingly challenging to sort periods of consistent behavior into just growth or shortening phases. Prior methods were admittedly inaccurate in measuring shortening rates. They further assumed that all behavior followed a near constant positive value for growth, and a near constant negative value for shortening. This assumption is especially invalid when considering different tubulin concentration levels. With increasing detailed level data comes the revelation of another class of behavioral periods, or consistent behavior with dynamics with smaller magnitude

rates of change to MT length than what is expected from the classical growth and shortening phases [31]. These periods of attenuated dynamics may indicate an unstable steady-state, during which relevant structural changes may make the MT more prone to either grow or shorten [31, 55]. Also, the stochastic nature of dynamic instability makes it difficult to pinpoint an exact moment in the detailed level data when the irregular changes between phases occurs. As a result, current methods tend to underestimate shortening rates, as well as overlook the role of periods with subtle changes to MT length during transitions between growth and shortening phases (see Figure 4.1a). These periods are observed in high frequency data show that they are indeed dynamically active, and that structural changes to the MT do not stop. These segments that are not considered either growth or shortening phases, have small net height change, but they still capture a significant sequence of molecular level reactions that alter the MT structure. For this reason, this study refers to these periods with attenuated length changes as stutters, especially since a large part of the data segments different from growth or shortening have some non-zero net height change, though minor fluctuations in MT length persistent throughout their duration. Any behavior in the data displaying halted dynamics to structural changes would still be captured first as a stutter by the proposed methodology here, though further analysis would be necessary to segregate them as periods that exhibit a true pause in dynamic behavior.

In this chapter, a semi-automated procedure is presented in order to address the need for making more accurate parameter measurements for data resembling dynamic instability behavior. The corresponding algorithm accepts the input data from any source, simulation or experimental based, and calculates rates and frequencies without any *a priori* assumptions on the possible number of phases present in the data. The procedure operates in three stages: segmentation, classification, and pattern analysis. The segmentation adaptively creates a continuous piece-wise linear approx-




(a) (b)

Figure 4.1. A comparison between the (a) old approximation methods, which identifies strictly growth or shortening periods by seeking changes in dynamic directionality (i.e. positive to negative slope, or vice versa) and (b) the new proposed approximation method, which seeks any significant changes in dynamic rates, and thus captures more subtle behaviors regardless of the prior segments directionality.

imation to the given time dependent data. The classification procedure categorizes linear segments into phases using an unsupervised clustering method based on the measurable features of those segments. The pattern analysis considers the possible changes between phases, which include but are not limited to catastrophe and rescue events. This procedure is applicable to any scenario that can benefit from accurate measurements of stochastic dynamic behavior, including MT length history from *in vivo*, *in vitro*, or *in silico* sources. The demonstrations here uses simulated data representing the classically understood dynamic stability scenario of MTs. In particular, to extract any possible phases or patterns that can exists in the data, the 13-PF MT model simulations using $10\mu M$ tubulin concentrations were utilized for its ability to generate DI behavior rich with catastrophe and rescue events. Additionally, the simulated output represents 10 hours of MT activity, a long time run to ensure a large number of possible outcomes in the stochastic process being modeled. The method is executed in MATLAB with minimal user input and human intervention in order to facilitate ease of use considering the various scientific backgrounds that would benefit from this method. The user-defined values are as follows:

- Nucleation height threshold: this separates the shorter MT lengths that are too close to the seed structure to ensure undistorted information on the dynamic behavior.

- Minimum time duration of a linear segment: this limits the time duration of a segment so that a macroscopic measurement with a large number of molecular reaction events is ensured.

- Maximum height error tolerance: this value dictates the accuracy of the linear segmentation by restricting the largest allowable point-wise error with respect to the height measurements to the given data.

- Maximum slope magnitude for near-zero slope segments: this helps identify the stutter segments with net height change near zero, which separate positive slope segments form negative ones.

- The $K$-value (i.e. the number of centroids in the collection of segments) for positive and negative slopes, respectively: these classification relevant values are tuned as part of a diagnostic mode. The ideal number is suggested to the user, who then chooses it for the remainder of the classification stage process.

The remaining sections of this chapter describe the details of the segmentation, classification, and analysis procedures, and how these user-defined values are used during each stage of the process.

## 4.2 Segmentation

The purpose of this stage is to identify the starting and ending points of periods with consistent behavior in the MT length history data. Ideally, each linear segment would accurately identify the exact moments a MT switches from one phase into another, effectively marking the instance of a significant dynamic change. Furthermore, the slope of a well-approximated linear segment would measure the rate of change in MT length for that period, which is already relevant for measuring dynamic instability parameters. Previously, a bi-phase assumption only considered periods that were strictly growth or shortening, along with a minimum height change threshold to help identify switches between the two phases. Additionally, the growth

and shortening rates were assumed to be near constant, thus justifying of a linear approximation. However, the challenge has been to identify the correct moment that separates growth and shortening phases from each other, especially in the presence of attenuated dynamic periods that stray from the bi-phase assumption. These periods create errors in the DI parameter measurements, as well as challenges the constant linear rate assumption. So, the method developed here does away with the bi-phase assumption in order to capture all periods of consistent behavior that may be present in the stochastic data. Thus, an adaptive approach is sought, such that points of significant dynamic changes are iteratively included to create a continuous piece-wise linear approximation, i.e. a sequence of line segments accurately resembling the input data.

It should be noted that there may be other methods of approximating the length history plot, but the nature of dynamic instability data renders those approaches inappropriate. For example, higher order spline methods would reduce the approximation error, but would not help in identifying an exact instance of dynamic change, while adding complexity to measuring dynamic instability parameters. For this reason, the linear approximation is preferred. Another option would attempt to fit a sequence of linear regressions, which would maintain the consistency with the rate parameters, but that would already require the number and location of dynamic phase changes that is unavailable due to the stochastic nature of the data. Rather, an approach that is adaptive to the sporadically varying rates, time durations, and height changes of these consistent periods is required, while delivering a continuous piece-wise approximation.

At the heart of the segmentation process is the search for the points in the data where significant dynamic changes occur. Each of these points can be considered a vertex between line segments, which resemble "elbow points", such that if one where to connect all of them, an approximation to the length history data is created. In

other words, these vertices and the line segments between them collectively create the continuous piece-wise approximation that is desired. The following steps summarize the algorithm that finds these vertices in data resembling dynamic instability:

- Step 0: Read-in raw length history data of MT length vs. time. Include the first and last data points into the vertex list.
- Step 1: Identify local extrema (peaks and valleys) in the data above the nucleation range, and add them to the vertex list.
- Step 2: Identify data points where the MT length enters/exits the nucleation height, and add them to the vertex list.
- Step 3: Use each consecutive pair of vertices as endpoints to create a linear segment that approximates the corresponding portion of the data.
- Step 4: Identify the data point where the maximum point-wise error for the MT length occurs, and add it to the vertex list (adjust as needed to obey minimum threshold for time duration between vertices).
- Step 5: Repeat Steps 3-4 until the maximum point-wise error of each linear segment approximation in Step 4 is not greater than the user-defined maximum height error tolerance.

Note that if the data resembles a persistent growth regime illustrated in Figure3.2 (i), then only the starting and ending points included in the vertex list from Step 0 would be necessary to identify the sole growth phase. Similarly, if the data was generated from conditions with too little tubulin concentration that does not allow for MT lengths above a nucleation threshold (see Figure 3.2(a) ), then only the starting and ending points included in the vertex list from Step 0 would be necessary to identify the sole nucleation phase. The rest of the steps in the algorithm are good for identifying the vertex points in DI data that displays dynamic changes in MT lengths above the user-defined nucleation threshold. Step 1 in the segmentation procedure begins with first identifying local extrema that stand out with a certain prominence with respect to the surround data values, which are in effect instances of very signifi-

cant dynamic change. Using the *findpeaks* function in MATLAB (a common tool for initial signal processing procedures) on the raw data identifies the local peaks (local maxima), and using it on the negative values of the raw data identifies the local valleys (local minima). The prominence of these extrema is defined by the user-defined maximum error tolerance, meaning how high a peak stands up with respect to the nearest valleys on either side of the peak. Additionally, an option in *findpeaks* keeps only the local maxima that occur above the user-defined nucleation height threshold. At this time, the user-defined minimum time duration criteria is used to omit any peaks or valleys occurring too close to each other, as well as removing redundant peaks or valleys that were identified in excess by *findpeaks*. What remain are peaks (purple diamond) and valleys (gold squares) illustrated in Figure 4.2. In Step 2, the maximum nucleation height threshold is used to identify points entering and exiting the nucleation region in order to prevent spending computational resources on moments of dynamic behavior bearing no relevance to the study. The demonstrative example uses 75 subunits for its nucleation threshold, and can be seen clearly as the sequence of blue points at the same height in Figure 4.2. This treatment is synonymous to the range of short MT lengths too difficult to measure accurately in experimental conditions. In the event that behavior in the nucleation region is relevant, the user-defined value can be set to zero (or some small number) to effectively include all the behavior available from the given data into the approximation.

The continuous piece-wise linear approximation defined only by the points included in the vertex list so far provide a good starting point, although the point-wise errors prevent this segmentation from being satisfactory yet. Steps 3-4 improve this approximation, by putting each linear segment is through an iterative process that identifies instances of significant dynamic change during the corresponding time duration. For two consecutive peak or valley points in the vertex list, say points $A$ and $B$, the data point where the maximum point-wise error made by the line segment

approximation $\overline{AB}$ occurs, call it point $C$, is identified and included into the vertex list as a refinement. Now, the data point where the maximum point-wise error made by the line segment approximation $\overline{AC}$ occurs is identified and included into the vertex list. This iterative step is repeated until the maximum point-wise error of any of the line segment approximations formed by newly incorporated vertex points are less than or equal to the user-defined maximum error tolerance. Once the error criteria is satisfied, the process moves on to the segment defined by next consecutive pair of vertices between which the iterative process has not yet been applied. This is repeated for until the maximum error criteria is satisfied for the entire approximation formed of all the segments with consecutive vertices as endpoints. The collection of linear segments created between consecutive points in the final vertex list provides the segmentation that defines the continuous piece-wise linear approximation, such that each linear segment represents consistent behavior, and the segment endpoints indicate significant changes in the MT's dynamic behavior.

The accuracy of the approximation created during the segmentation procedure is especially sensitive to two of the user-defined values: the minimum time duration of a linear segment, and the maximum height error tolerance. The latter of the two has an obviously direct impact in dictating the point-wise errors used for the stopping criteria of finding vertex points. However the minimum time duration of each segment is less obvious, in that it can prevent in the incorporation of a data point into the list of vertices. More specifically, in the event that the highest point-wise error is identified near an existing vertex, and the time difference between them is less than the user-defined minimum, then another data point is chosen instead, close to the desired location, but satisfying the minimum time criteria. This means that if a user chooses minimum time durations that are too large for a given error tolerance, there may be unintended errors created in the approximation that cannot be reconciled due to the conflicting criterion. This is demonstrated in Table 4.1, which displays

the the number of irreconcilable errors, and the maximum errors that occur (above the nucleation threshold), and how many of them were greater than 50 subunits in the approximation created when using 25 subunits as the maximum height error tolerance for different minimum time duration values.

TABLE 4.1

LENGTH HISTORY APPROXIMATION ERRORS

| Min Time Step: | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| # of Irreconcilable Pts: | 0 | 0 | 0 | 0 | 3 | 4 | 12 | 22 | 45 | 77 |
| Max Error: | 25 | 25 | 25 | 25 | 26 | 27 | 36 | 56 | 56 | 72 |
| # of Errors > 50: | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 13 |

These approximation results are from using a tolerance of 25 subunits for different minimum time steps to process the data. The irreconcilable points refer to those errors greater than the minimum error tolerance and that could not be resolved due to the conflict with the minimum time step criteria. These were generated from the segmentation procedure for the 10 hour 13-PF MT model simulation data. Time units are in seconds, and the error values are in number of subunits. Using minimum time steps much larger than 1.0 seconds did not produce meaningful approximations, so those results were not included here.

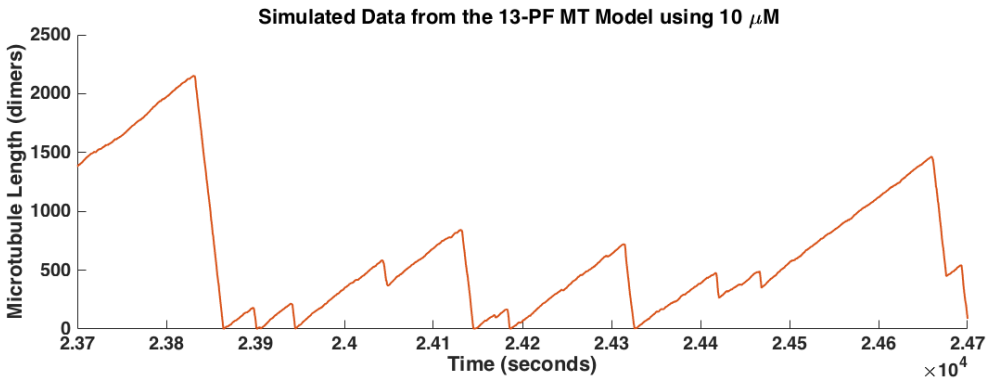
Note that using smaller time duration criteria makes it easier to avoid irreconcilable error points, whereas the larger time steps would require raising the height change error tolerance to resolve any conflicts. For this reason, the demonstrative example in this chapter uses 25 subunits for the maximum height change error tolerance, and $100ms$ for the maximum time duration of each segment. The time duration criteria does deliver at least 70 micro-level reaction events during each segment period, and

the 25 subunit error is in agreement with spatial resolutions available in experimental data. The performance of these criterion are demonstrated in with approximation in Figure 4.2 using the length history plot from a 13-PF model simulation, where the blue dots are the vertices separating segments of consistent behavior. Setting a lower minimum height error threshold would have identified additional points with smaller differences between the actual MT length (red line) and the approximation (blue line). However, since macro-level behavior is being measured, this level of approximation is satisfactory.

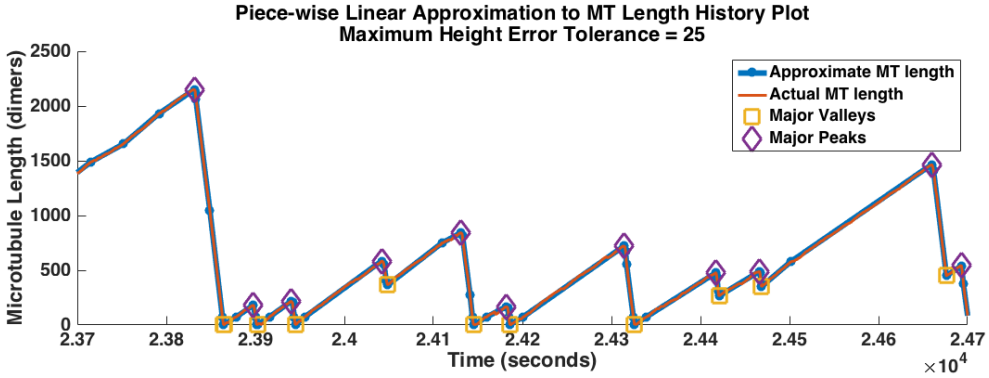
## 4.3 Classification

The linear segments provided by the accepted continuous piece-wise linear approximation represent periods of consistent behavior, and thus is appropriate to label each as being in a particular phase corresponding to the segment properties that can be measured. The three measurements available for each linear segment are the time duration (run), the height change (rise), and the rate of change (slope = rise/run). All of the points reside on the surface manifold defined by $z(x, y) = y/x$ as illustrated in Figure 4.3. It can be argued that only two of these measurements are needed, however, as evident in the characteristics of $z(x, y)$, the combinations of the run, rise, and slope values that would capture the greatest variance in data points on this curved surface depends on the region of the surface being considered. At this time, no assumptions are being made with regards to which measurement contributes more to separate particular classes. For completeness, any combination of the three measurements is considered, which allows for the more complex case where each measurement has similar weights determining the phase class separation, as well as simpler cases where only one or two of the measurements contribute nothing.

Initial attempts to use these measurements separately to establish rules for classification were fruitless, and indicated that indeed combinations of the three mea-




(a)




(b)

Figure 4.2. A 1,000 second excerpt from (a) the simulated length history plot, and (b) the resulting piece-wise linear approximation for a portion of MT length history output from a 10hr simulation of the 13-PF MT model, using a minimum height error threshold of 25 subunits. The red plot is the raw output, and tends to be very noisy at a finer scale. The purple diamonds and gold squares are the significant local maxima (peaks) and local minima (valleys) respectively, used to initiate the iterative process. The blue line segments are the resulting piece-wise linear approximation segments. The blue dots represent the segment endpoint vertices, which are separated by at least the user-defined minimum time duration for each segment. Note the additional points at height = 75 that identify moments of nucleation phase entry/exit.

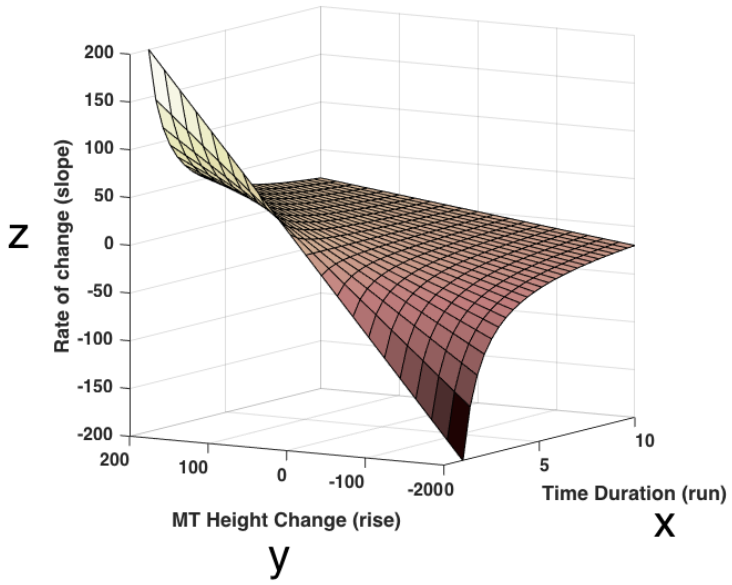



Figure 4.3. The $z(x, y) = y/x$ surface manifold on which the points representing the linear segment characteristics reside.

surements may be necessary to characterize each segment into different phases. To consider all three measurements together, each segment is assigned a 3D-point, such that the relationships between different segment characteristics can be observed. Each point plotted in Figure 4.4 represents the individual segments identified from the same 13-PF MT used for the segmentation demonstration in the previous section. Doing so revealed that the data points organized into more complex subgroups than the previously used bi-phase growth and shortening phases had assumed. In order to separate these data clusters, an unsupervised clustering method is utilized such that no *a priori* assumptions about the number of possible phases are made, and each segment can be classified into a phase according only to its relation to other segments with similar characteristics.

To assist the unsupervised classification procedure, there is enough information at this time to separate the nucleation phases based on the user-defined criteria for the

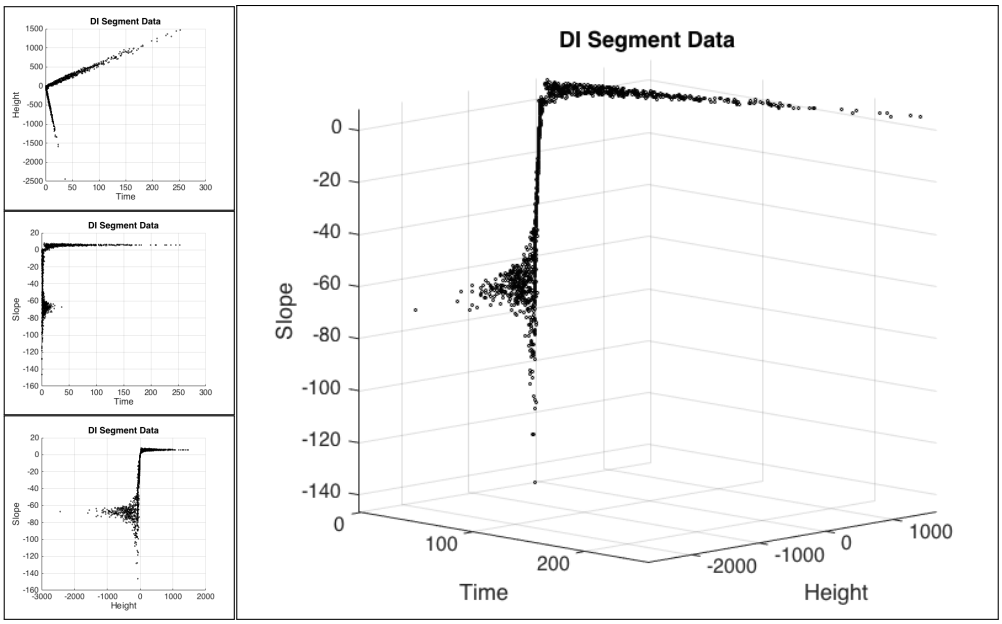



Figure 4.4. The data points for each segment obtained from the linear segmentation on the MT length history for a 10hr simulation of a 13-PF MT, such as those segments identified in Figure 4.2(b). Different perspectives are shown here to aid the visualization of the 3D plot.

maximum nucleation height threshold. Also, those stutter segments with near zero slope or height change, which are called flat stutters, can also be classified according to the user-defined maximum values for height change and slope. By removing these instances from the data that needs to be classified, the region now separating the points for segments with positive slopes and those with negative slopes becomes a clearer and better separated. Furthermore, for the data set being used in this demonstration, the points for positive sloped segments lie on a relatively flat portion of the $z = y/x$ surface, which is not parallel to the relatively flat region corresponding to the negative slope points. This provides the first indication that the developed classification procedure should treat the segments with positive and negative slopes separately.

### 4.3.1 $K$-means Clustering

At the heart of this stage lies the $K$-means clustering method, an unsupervised classification method that will separate the given data into $K$-many clusters, which are defined by Voronoi cells, and can be representative of different classes present in the data [49, 50]. The algorithm for $K$-means begins by randomly selecting $K$-many points as cluster centroids in the data set, and defining a cluster including the data points closest to each respective cluster. Then, the true centroid of each cluster is measured, and the cluster is refined by including the data points closest to these refined centroids. This is centroid selection is recursively repeated until the cluster composition converges, and the centroids after convergences are stored. For irregularly shaped data sets, the resulting centroids after convergence may depend on the random points in the initialization step. To overcome this, the entire process from the random initialization to cluster convergence is repeated many times, and the most popular centroids are used to define the final Voronoi cells that define the separations between clusters.

The desired methodology being developed here needs to consider the possibility of any number of clusters, since we are ultimately looking to identify those segments that are clearly growth and shortening, as well as any other unknown phases. However, this approach works best when the $K$-value is known, but it can still be used to investigate the number of centers that the data naturally clusters around [36, 37, 42, 49]. To address this, the search for the $K$-value that best separates the data refers to the gap statistic measured from the resulting clusters for different $K$-values. The gap statistic considers the changes in the dispersion within clusters compared to a randomized distribution for reference [74]. In practice, the $K$-means clustering procedures are repeated for a range of $K$-values, the gap statistic in each clustering result is measured, and the first local maximum is sought [74]. The corresponding $K$-value of the first local maximum gap statistic corresponds to the lowest number

of values that shows an improvement in separating the data compared to nearby $K$-values. For purposes of identifying the $K$-value that best separates the data in the method being developed here, only 100 random starts are used to compare the results from different $K$-values and measure the corresponding gap statistics, but 500 starts are used when creating the final clusters in the classification procedures.

Finally, the large range and variance in the measured linear segment values make it challenging to classify the raw data points as is. Since $K$-means is a distance based metric, and the linear segment data set has various elongated features, it is important to first pre-condition the data points so that the resulting shape of the data set is more favorable for classification. Therefore, the method begins by applying a natural logarithmic transformation to the data, and then standardizing the points so that the measured differences would be more comparable. This helps make the separations between resulting clusters more apparent, and also helps the classification procedure being utilized, since $K$-means has a tendency to identify similarly sized clusters regardless of the different shapes possible.

### 4.3.2 Clustering the Entire Data Set at Once

Attempting to classify the complete collection of the segment points together returns the gap statistic plot displayed in Figure 4.5. In this case, the gap statistic values are monotonically increasing with the $K$-values. The increasing gap values suggests that there may exist more substructures of the data that create better clustering results [74]. If there was a clear separation established between positive and negative slope segment points, it is expected that $K$=2 should stand out at least as a possibility for a good result. This leads to the conclusion that separating the data into just two classes, such as only growth and shortening, does not divide the data points along the natural separations. Additionally, the chosen boundaries between clusters are unsatisfactory, particular on the positive slope side. The poor results from the

gap statistic analysis here is the second indication that the data with positive slopes should undergo classification separately from the negative slope data.

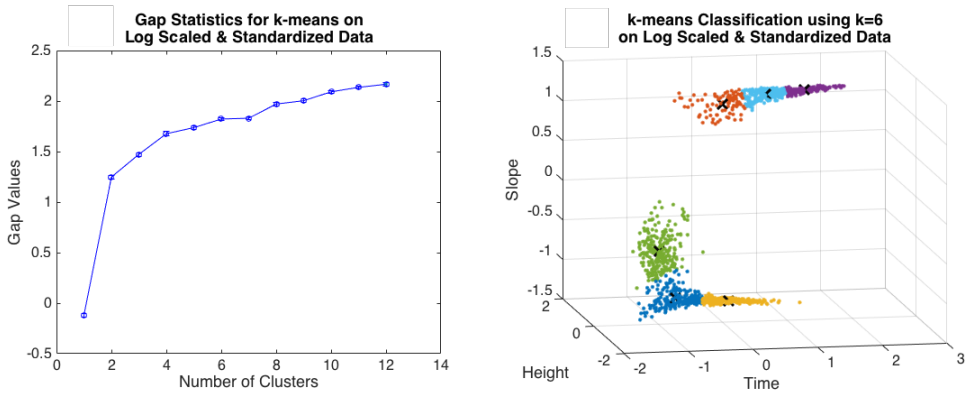
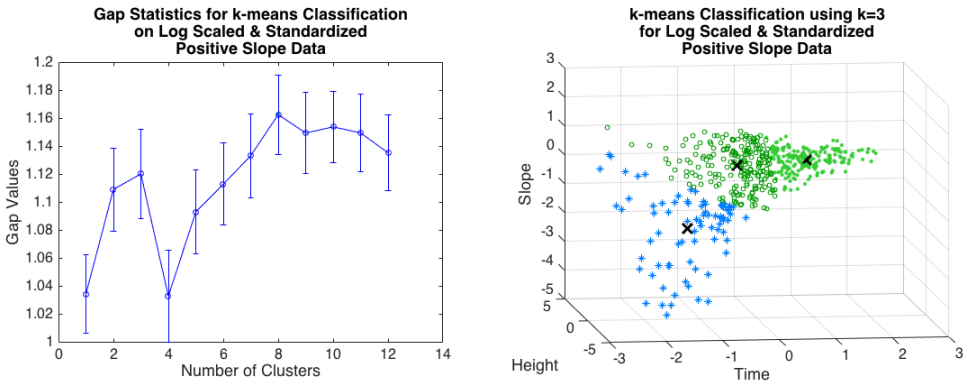


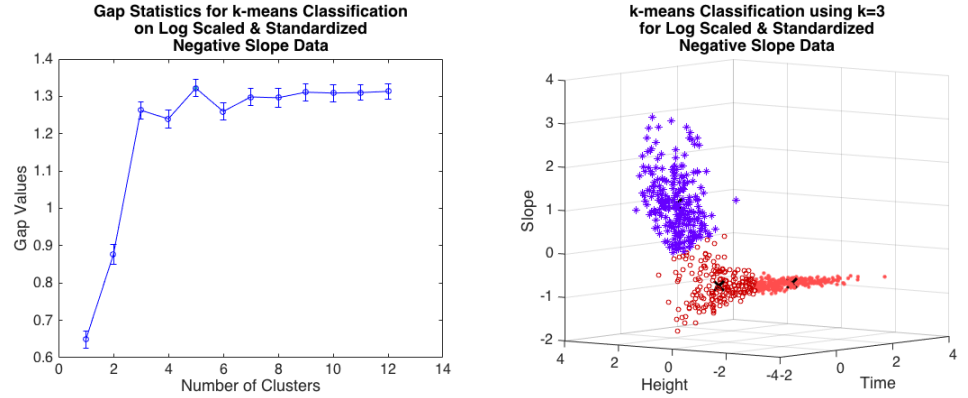

Figure 4.5. (Left) Gap statistic plot for data representing segments with both positive and negative slopes. The monotonically increasing plot indicates no good clustering results for the given data set. (Right) Clustering results using $K = 6$ for the data representing segments with both positive and negative slopes, which do not create satisfactory boundaries to separate the anticipated substructures in the data.

### 4.3.3 Clustering Positive- and Negative-Slope Data Separately

In the data used for the demonstration, the previously mentioned flat regions for the positive and negative sloped segments display the respective data as a cloud of points in 2D separately, though it may depend on some combination of the 3D dimensions derived from line segment properties. This condition helps the applications of $K$-means, which work best with data resembling a Gaussian distribution [42]. Although the flat region is embedded in three dimensions, the Euclidean distances used in $K$-means will be unaffected, and therefore reducing the dimensionality is unnec-




(a)




(b)

Figure 4.6. The classification results for (a) positive slope segment data, and (b) negative slope segment data generated from 10hr simulation of the 13-PF MT model. (Left) Gap statistics when clustering for different $K$-values. In both cases, the first local maximum appears at $K = 3$. (Right) The $K$-means clustering results on the log scaled and standardized positive slope data points for $K = 3$, displayed with different colors and markers. A black $\times$ marks the center of each cluster.

essary since it would not change the classification results. Therefore, the $K$-means procedure is implemented twice; the positive and negative slope segment data sets are each processed separately. In each case, the raw data values are first scaled using a logarithmic transformation. The negative slope data are multiplied by $-1$ to distinguish them from the positive slope data. Then, each group of points are standardized according to the mean and standard deviation of their respective transformed data values. The respective mean and standard deviation for the positive and negative slope segment points are stored for future use. This is one beneficial aspect of this method that can help classify different length history data in the future without having to repeat this process. Upon transforming and standardizing the data, $K$-means and the corresponding gap statistics are measured to test the separation of the data points into relevant phases. Executing this procedure for the positive and negative slope segment data separately delivers gap statistic plots that are more easily interpretable when compared to the trying to process the entire data set all together. As a result the separations between clusters that are more satisfactory, as seen in Figures 4.6a and 4.6b.

### 4.3.4 The Classification Algorithm

The latter process that treated positive and negative slope segments separately is the preferred approach for classification, since the boundaries between data clusters was more satisfactory, and because the corresponding gap statistics revealed a clear indication for confidently choosing a $K$-value. So, the classification algorithm can be summarized using the following steps:

- Step 0: Retrieve the segmentation results for a length history data set.
- Step 1: For each segment, assign a point in 3D, such that each dimension

represents the following line segment features:

$$x = \text{Time Duration (run)}$$
$$y = \text{Height Change (rise)}$$
$$z = \text{Rate of Length Change (slope)}$$

- Step 2: Identify, remove, and set aside those points with near zero height change and slope values according to the user-defined thresholds.

- Step 3: Scale the data using a logarithmic transformation for each point $\vec{X}_{raw} = (x_{raw}, y_{raw}, z_{raw})$ using the following formula that preserves zeros:

$$\vec{X}_{transformed} = \text{sign}(z_{raw}) \cdot log(|\vec{X}_{raw}| + 1)$$

- Step 4: Perform $K$-means clustering on the positive slope data points and compute the corresponding gap statistics for different $K$-values

- Step 5: Identify the $K$-value corresponding to first local maximum in the gap statistic values

- Step 6: Repeat the $K$-means clustering results using the $K$-value from Step 5, and using more starts to achieve better results.

- Step 7: Repeat Steps 4-6 for the negative slope data points.

- Step 8: Label the raw data points according to the classes identified.

### 4.3.5 Diagnostic and Fully Automated Modes

Until this point, the methodology being implemented in MATLAB allows for the user to modify the input parameters, and to test for various outcomes. In other words, the modification of the user-defined thresholds are part of a diagnostic process, which delivers the gap statistic results and a suggestion for the best choice for the $K$-value to be used for the positive- and negative-slope data sets. After receiving the suggestion, the user can decide the $K$-values to be used for the remainder of the classification and resulting pattern analysis. Once the number of centroids is decided upon, the remainder of the method presented in this chapter is completed

automatically. The thresholds leading to the segmentation results earlier in this section, and the criteria needed to select out the near-zero slope segments have already been discussed. All that remains to continue to the next step are the $K$-values for the positive and negative slope segments, which have clearly indications to be 3 for both cases. The complete list of the user-defined threshold values utilized in the demonstrated procedures are listed in Table 4.2, unless otherwise specified.

TABLE 4.2

USER-DEFINED THRESHOLDS AND VALUES FOR THE DEMONSTRATED DI PHASE SEGMENTATION, CLASSIFICATION, AND PATTERN ANALYSIS METHOD APPLIED TO THE 13-PF MT MODEL LENGTH HISTORY PLOT

| | |
|---|---|
| Nucleation height threshold: | 75 subunits |
| Minimum time duration of a linear segment: | 100 ms |
| Maximum height error tolerance: | 25 subunits |
| Maximum height change for near-zero slope segments: | 3 subunits |
| Maximum slope magnitude for near-zero slope segments: | 2 subunits/sec |
| Number of centroids for positive slope segments: | 3 |
| Number of centroids for negative slope segments: | 3 |

### 4.3.6 Distinguishing Stutters from Growth and Shortening

In both the positive and negative slope data subsets, closer observation of Figures 4.6a and 4.6b reveals that two of the cluster centers (those surrounded by filled and hollow circles) share similar slope values. Additionally, their respective cluster points seem to be distributed around the line connecting the cluster centers. However, the third cluster center (surrounded by asterisk) has a slope certainly smaller in magnitude than the other two cluster centers, and the respective cluster points are closer to the origin than the other points (labeled with circles). For this reason, the points corresponding to cluster centers with a smaller magnitude slope value (labeled with asterisk) are referred to as stutters, since they represent dynamic instability segments with smaller MT length changes than their counterparts. This distinguishes them from the other two clusters that display more classically expected behavior of growth or shortening, when they have positive or negative slopes respectively. In fact, in both the positive and negative slope cases, the major factor separating the non-stutter clusters is the time duration component. This means that using the $K$-means clustering has helped identify long growth (light green), brief growth (dark green), and up stutter (light blue) phases within the positive slope segments, and long shortening (light red), brief shortening (dark red), and down stutter (purple) phases within the negative slope segments (see the color legend in Figure 4.7).

By combining these results for the running demonstrative example, and including the near-zero slope segments separated before the classification process as flat stutters (dark blue), a more complete classification is presented of the dynamic instability segments into appropriate phases. In other words, the points in Figure 4.4 have now been classified into appropriate DI phases, and they they are labeled using color legend listed in Figure 4.7. The resulting the labeled length history plot is presented in Figure 4.8.

With each segment labeled as being within a particular phase, the overall phase

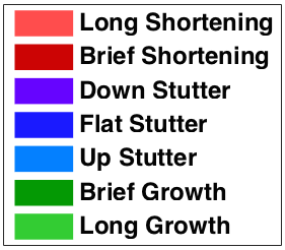



Figure 4.7. The color legend used to identify the different DI phases that have been classified.

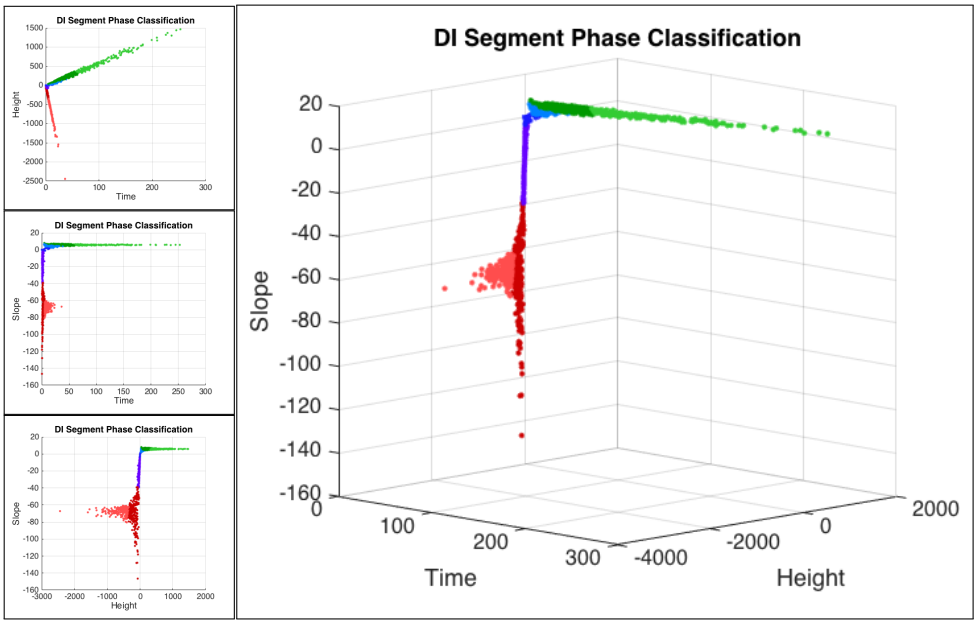



Figure 4.8. Phase classification results on the data set displayed in Figure 4.4. Different classes are labeled according to the legend in Figure 4.7.

classes can be compared using average measurements and corresponding box plots, which are calculated and displayed in Figure 4.9. These measured values reveal the similarity within the slope values shared by both long and brief growth classes, and by both long and brief shortening classes. Meanwhile, the three stutter classes have slope values certainly smaller than their counterparts. The case is more drastic for

the negative slope segments, mostly because events corresponding to the shortening of a MT's length can involve the loss of multiple subunits at a time, compared to the single subunit addition during polymerization. Details of this were also discussed in the rate of subunit addition and loss in Section 3.3.3. The differences between the stutter phases and the other two classes is further evidence that indeed macro-level behavior detected during these periods are different, and that the bi-phase assumption cannot be valid one, especially at this level of detail for MT activity. Also, the boxplots in Figure 4.9 illustrate how there is an overlap of measurements between segment classes, thus justifying the need to consider all three variables when conducting the classification procedures, rather than attempting to separate the data by using any one variable independently.

**Segment Measurements for each DI Phase**

| | # of Segments | Time Mean | Time StDev | Height Mean | Height StDev | Slope Mean | Slope StDev |
|---|---|---|---|---|---|---|---|
| LngGr | 195 | 100.7 | 40.8 | 583.2 | 238.2 | 5.8 | 0.4 |
| BrfGr | 214 | 37.2 | 12.1 | 214.6 | 70.4 | 5.8 | 0.8 |
| UpSt | 79 | 18.2 | 8.3 | 72.5 | 37.3 | 4.0 | 1.0 |
| FltSt | 22 | 3.8 | 4.2 | 2.0 | 7.8 | -0.3 | 1.3 |
| DnSt | 255 | 1.6 | 0.6 | -24.8 | 14.7 | -16.7 | 8.8 |
| BrfSh | 240 | 2.7 | 1.3 | -185.6 | 89.1 | -70.5 | 16.5 |
| LngSh | 177 | 9.6 | 4.3 | -649.3 | 289.2 | -67.9 | 3.9 |

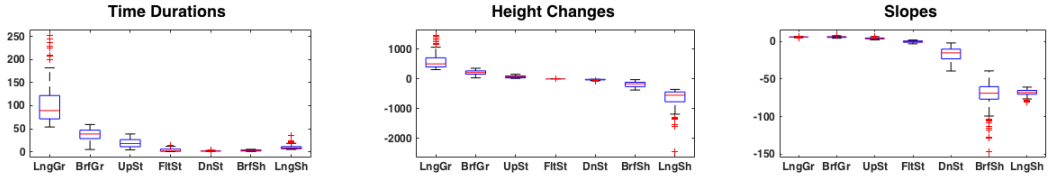
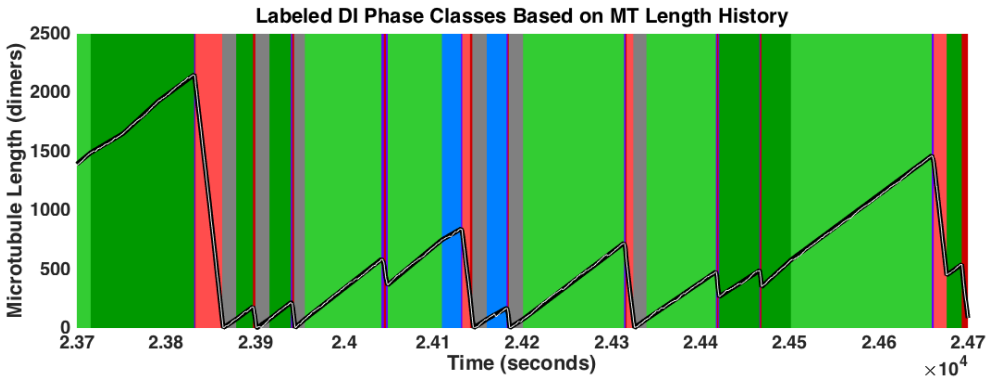



Figure 4.9. (Top) Mean and standard deviation for the time duration, height change, and slope measurements for each dynamic instability segment phase identified in the 13-PF MT model simulations. (Bottom) Box plots of the time duration, height change, and slope measurements for each phase class. The red crosses represent line segment data that are outliers in their respective phase class.

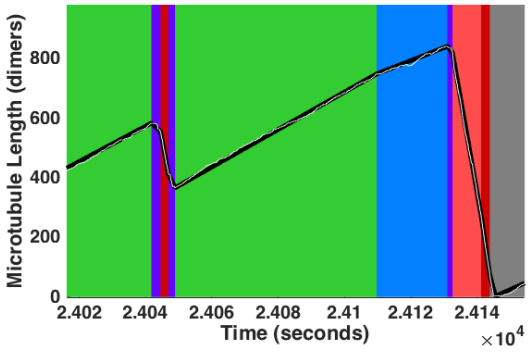
## 4.4 Phase and Pattern Analysis

At this time, the linear segments identified from the segmentation stage are now classified as one of the following 7 phases: Long Growth, Brief Growth, Up Stutters, Flat Stutters, Down Stutters, Brief Shortening, or Long Shortening. In addition, the segmentation procedure identified the periods with shorter MT lengths as nucleation phases. To help visualize these phases in terms of the MT dynamics, the line segments illustrated in Figure 4.2 can be labeled using the color code defined in Figure 4.8. This provides the color labeled plot in Figure 4.10, where the grey segments represent nucleation phases. Now that each segment has been allocated to a particular phase class, more detailed measurements can be taken to compare the phases to each other, as well analyzing the chronological orders in which phases appear to study the transitions between phases. The remainder of this section deals with these type of measurements and the associated plots that help visualize them.

Using this classification, properties of the phases can be quantified for the entire length of the simulated data. Figure 4.11 shows these measurements made for the 10hr simulated data, which includes the following measurements associated with each phase: frequency (total number of occurrences), percent of simulation time spent, percent of total height changes, and various slope averages (the mean (grey yellow) and medians (dull yellow) of the segment slopes for each phase, and the weighted slope $= \frac{(total\ height\ changes\ of\ a\ phase)}{(total\ time\ spent\ in\ a\ phase)}$ (bright yellow).) Note that the slope averages are quite similar for all three methods of measurement, suggesting an agreement in measuring the average rates using either formula. It is worth noting the number of nucleation segments encountered, and this is indicative of choosing a good tubulin concentration level that delivers MT behavior rich in dynamic instability, because of the balance between the large MT heights achieved along with plenty of catastrophe events that cause the MT to return to nucleation levels. A tubulin concentration level that is too low would create a simulation that spends far too much time in




(a)




(b)

Figure 4.10. (a) Color labeled representation of line segments that have been classified from a portion of length history data from the 13-PF MT model 10hr simulation. Each segment is labeled using the color legend in Figure 4.7, in addition to periods of Nucleation (gray). (b) A zoomed in excerpt from the same plot.

nucleation, and a level that is too high would create infrequent catastrophe events and rarely return to nucleation levels, if at all. This is additional indication that the $10\mu M$ tubulin concentration level delivered an appropriate amount of DI behavior to conduct this study.

Further pattern analysis can be conducted on the chronological sequences in which the phases occur. For simplicity, the phases that are strongly related to each other are bundled together, so that the need for sub-classes is omitted, and effectively reduced the sheer number of different types of phase transitions that can occur (see

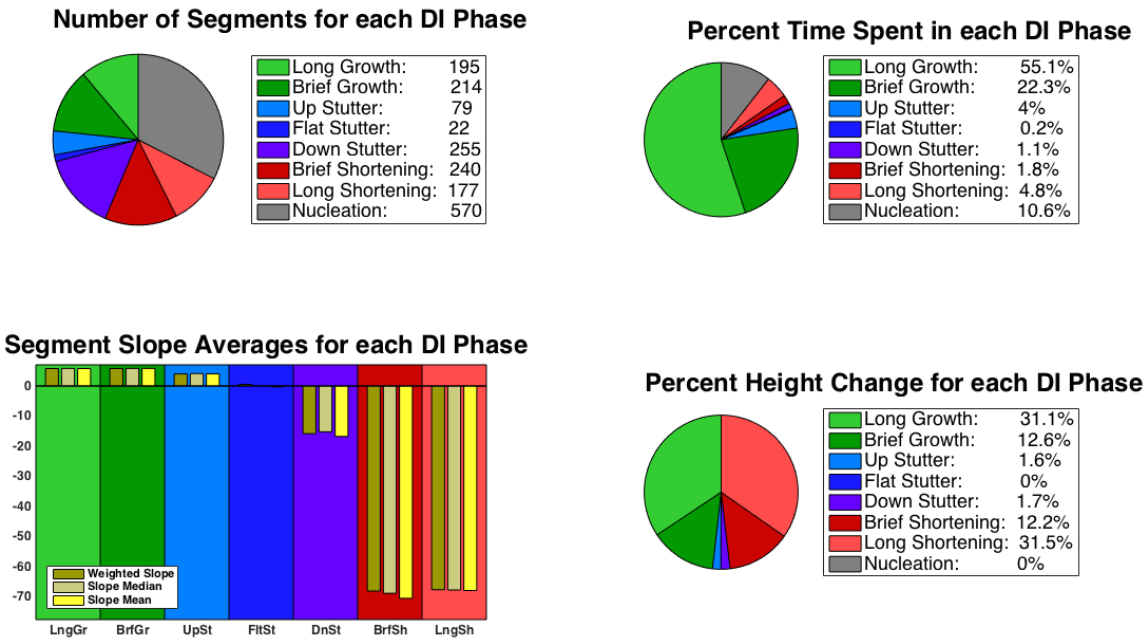



Figure 4.11. Different properties measured for each DI phase identified in the 13-PF MT model's 10hr simulation.

Figure 4.12). This is strongly motivated and supported by the fact that many of the slope measurements for Long and Brief Growth phases are close in value, especially when compared to other phases. Similarly, the Long and Brief Shortening slope measurements stand apart from the other phases also. For this reason, a simpler set of only four bundled phase classes are introduced, and used for the remainder of this study:

a Growth Phases: long and brief in time duration, steeper positive slopes and higher height gains than other segments

b Shortening Phases: long and brief in time duration, steeper negative slopes and higher height losses than other segments

c Stutter Phases: brief in time duration, smaller magnitude of slopes and height changes than other segments

d Nucleation Phases: segments with MT lengths too short for dynamic relevance,

and thus ignored from analysis

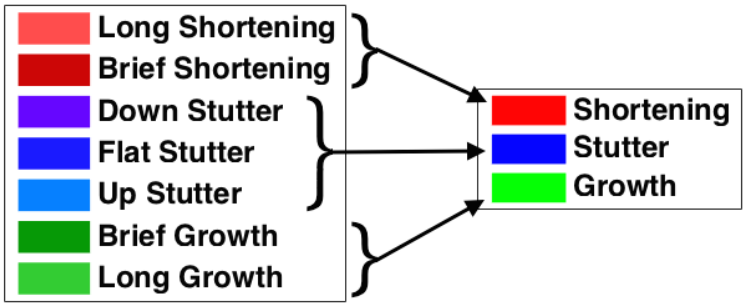



Figure 4.12. The color legend used to identify the different DI phases that have been classified.

Using these bundled class labels, the transitions between growth, shortening, stutter, and nucleation phases can be studied. Additionally, since the nucleation phase by design contains MT structures that are relatively close to the seed, they are omitted from dynamic analysis with respect to the possible permutations in phase transitions. In doing so, it should be noted that the stutter phase class creates a new variety of dynamic transition events when switching the bundled phase classes.

- Catastrophes: possible ways of changing from growth to shortening
  - Abrupt Catastrophe: Growth-Shortening 
  - Transitional Catastrophe: Growth-Stutter-Shortening 
- Rescues: possible ways of changing from shortening to growth
  - Abrupt Rescue: Shortening-Growth 

  - Transitional Rescue: Shortening-Stutter-Growth 

- Interruptions: involving stutters without changing dynamic directionality

  - Interrupted Growth: Growth-Stutter-Growth 

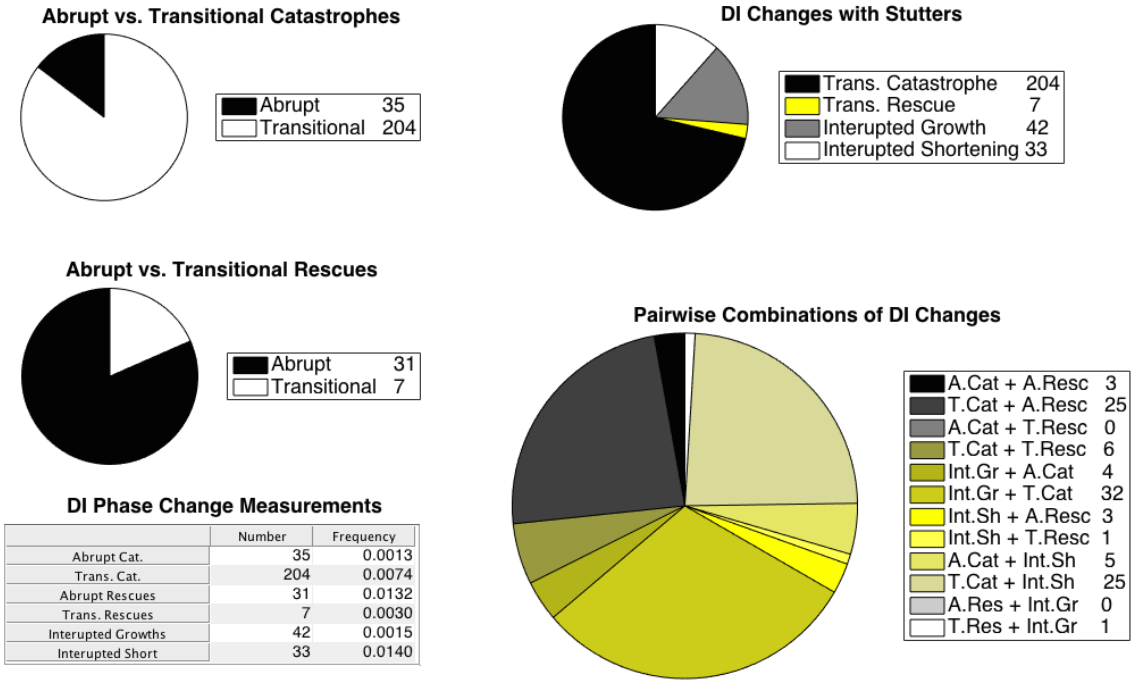
  - Interrupted Shortening: Shortening-Stutter-Shortening 

The abrupt catastrophe and abrupt rescue transitions are the classically understood catastrophe and rescue events respectively. The bi-phase assumption resulted in those definitions, and the abrupt variety of these phase transitions allows for the continuity of that understanding. Additionally, the role of nucleation segments present in the form of phase transitions, but is utilized in preventing moments of MT dynamics being categorized as a rescue if the MT length reaches nucleation levels during that time period. In those cases, the MT length reaches very short levels, where a significant amount of the MT structure has been lost, and thus is disqualified from being considered a rescue. Frequencies of these phase transitions can be calculated using the following:

- Catastrophe Frequency (Abrupt or Transitional) = $\frac{(\#\ of\ catastrophes)}{(total\ time\ in\ growth)}$
- Rescue Frequency (Abrupt or Transitional) = $\frac{(\#\ of\ rescues)}{(total\ time\ in\ shortening)}$
- Interruption Frequency (Growth or Shortening) = $\frac{(\#\ of\ interruptions)}{(total\ time\ in\ interrupted\ phase)}$

This proposed formulation is an improvement to prior methods of DI parameter measurements, because the segmentation process was based on identifying vertex points that marked moments of significant dynamic change better than the course-grained bi-phase approach used in the past. Additionally, the methods presented here takes into consideration the variety of dynamic change occurring. If a direct comparison between older methods and the formulation presented here is desired,

then a simple summation of the abrupt and transitional varieties can be performed to combine frequency values together, a valid approach since they share a denominator.




| | Number | Frequency |
|---|---|---|
| Abrupt Cat. | 35 | 0.0013 |
| Trans. Cat. | 204 | 0.0074 |
| Abrupt Rescues | 31 | 0.0132 |
| Trans. Rescues | 7 | 0.0030 |
| Interupted Growths | 42 | 0.0015 |
| Interupted Short | 33 | 0.0140 |

Figure 4.13. Measurements of possible dynamic transition events generated by the perturbations created from growth, shortening, stutter, and nucleation phases found in the 10 hour long 13-PF MT model simulations.

Figure 4.13 shows the measurements made considering these dynamic transitions identified from the length history plot for the 13-PF MT model 10 hour simulation, including the measured phase change occurrence and frequency values listed in the bottom left table. It is interesting to note that a large fraction of catastrophes occur in the transitional manner, where the MT first enters a stutter phase prior to rapidly losing its polymer structure during shortening. Conversely, rescues tend

to occur in an abrupt fashion, where the switch from shortening to growth is more sudden. Additionally, a significant number of stutters occur during interruptions of growth and shortening, but far more of them appear during catastrophes. This further supports the idea that stutter phases predominantly play a transitional role during catastrophes, during which a MT undergoes structural changes involving little subunit exchange, but leading to a more unstable structure more prone to enter a shortening phase. Finally, the bottom right pie chart in Figure 4.13 considers pairwise combinations dynamic events. The 10 hour long 13-PF MT simulation that was analyzed indicates that the most popular dynamic patterns (without entering nucleation) include the following three combinations of phase transitions:

1. Interrupted Growth + Transitional Catastrophe: 32 instances
2. Transitional Catastrophe + Interrupted Shortening: 25 instances
3. Transitional Catastrophe + Abrupt Rescue: 25 instances

Two out of these three patterns of dynamic change involve the MT entering a stutter phase twice. It's interesting to see the active presence that stutter phases have in common MT dynamics being simulated, especially considering the fact that they were overlooked in course-grained methods used in the past. However, the number of occurrences and the amount of time spent during stutter phases is on the same order of shortening phases, as indicated in the values listed in Figure 4.11. Combine with the common role that stutters have during phase transitions, it should be clear that stutter phases are an important facet of MT dynamics that should be focused on.

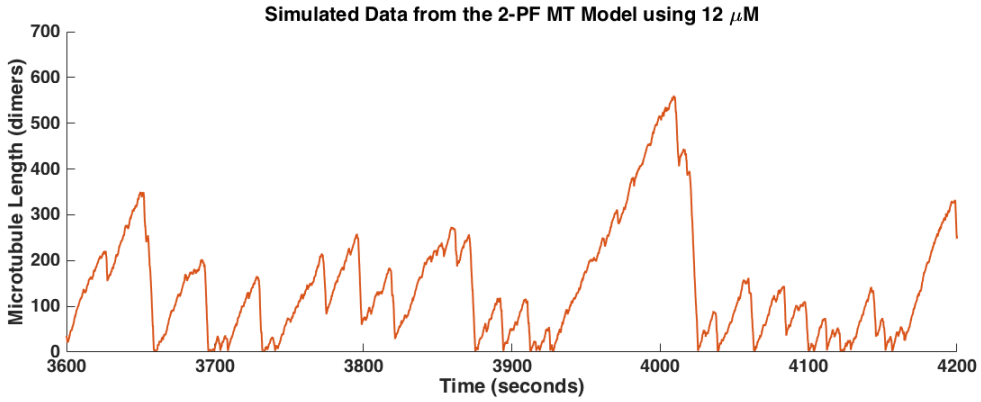
## 4.5 Phase Classification for the Simplified 2-PF MT Model

In Section 3.8, the simplified 2-PF MT model was introduced in an effort to to study the micro-level MT tip structures during significant events in DI behavior. However, prior to zooming into arbitrary micro-level tip structures, a measured ap-

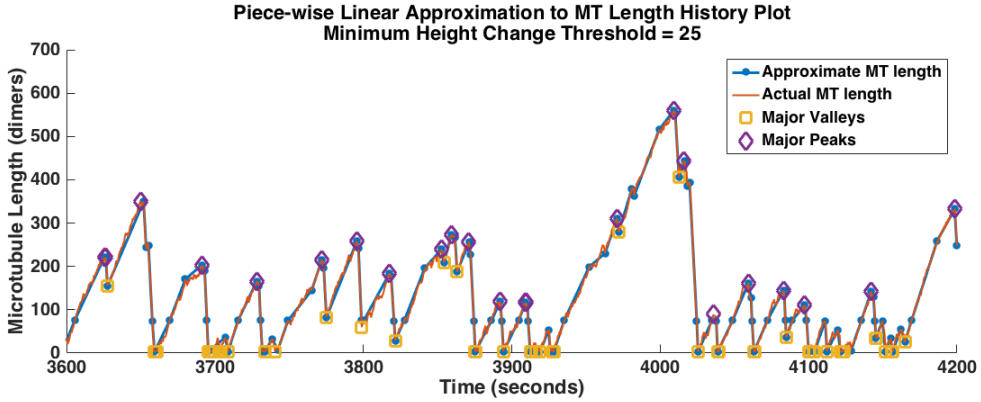
proach requires knowing where the significant changes in the DI behavior are located. For this, the method laid out in this chapter can be utilized to conduct the DI phase segmentation, classification, and analysis in simulated data generated from the 2-PF MT model. Similar to the 13-PF MT model example that was demonstrated, a long time simulation displaying a rich variety of DI behavior is desired in order to capture a robust range of the possible dynamics that can be observed in this system. The $12\mu M$ tubulin concentration level is used to create a 10 hour long simulated length history data of the 2-PF MT model, which is analogous to the $10\mu M$ tubulin concentration levels that provided the rich variety of DI behavior measured in the 13-PF MT case. This section goes through the process and reports the results when segmenting, classifying, and analyzing the DI phases identified in the length history data for the 2-PF MT case.

The process begins with the length history data representing 10 hours of DI behavior from the 2-PF MT model using a tubulin concentration of $12\mu M$ plotted in Figure 4.14a. When compared to the 13-PF MT model, the 2-PF MT model generates length history data with far more fluctuations on a finer scale, possible due to the fact that structural changes to only two PFs can more easily change the length of the MT structure. This scenario actually creates more of the periods that challenge the bi-phase assumption, where more instances of intermediate dynamics are visible. However, the segmentation step using the same minimum time duration threshold as the 13-PF MT model case is small enough to satisfy the same maximum height change error tolerance of 25 subunits, which creates the continuous linear piece-wise approximation displayed in Figure 4.14b.

The next step in the process is to gather the 3D data from the time duration, height change, and slope of each linear segment identified from the segmentation step. The segment data from the 2-PF MT model simulation is displayed in Figure 4.15. The structures in the scatter plot are qualitatively similar to the 13-PF MT segment




(a)




(b)

Figure 4.14. A 800 second excerpt from (a) the simulated length history plot, and (b) the resulting piece-wise linear approximation for a portion of MT length history output from a 10hr simulation of the 2-PF MT model, using a minimum height error threshold of 25 subunits. The red plot is the raw output, and tends to be very noisy at a finer scale. The purple diamonds and gold squares are the significant local maxima (peaks) and local minima (valleys) respectively, used to initiate the iterative process. The blue line segments are the resulting piece-wise linear approximation segments. The blue dots represent the segment endpoint vertices, which are separated by at least the user-defined minimum time duration for each segment. Note the additional points at height = 75 that identify moments of nucleation phase entry/exit.

data. In contrast, the 2-PF MT model segments have a wider variance in values, though the time durations do not extend as far out as the positive slope segments found in the 13-PF MT model simulations.

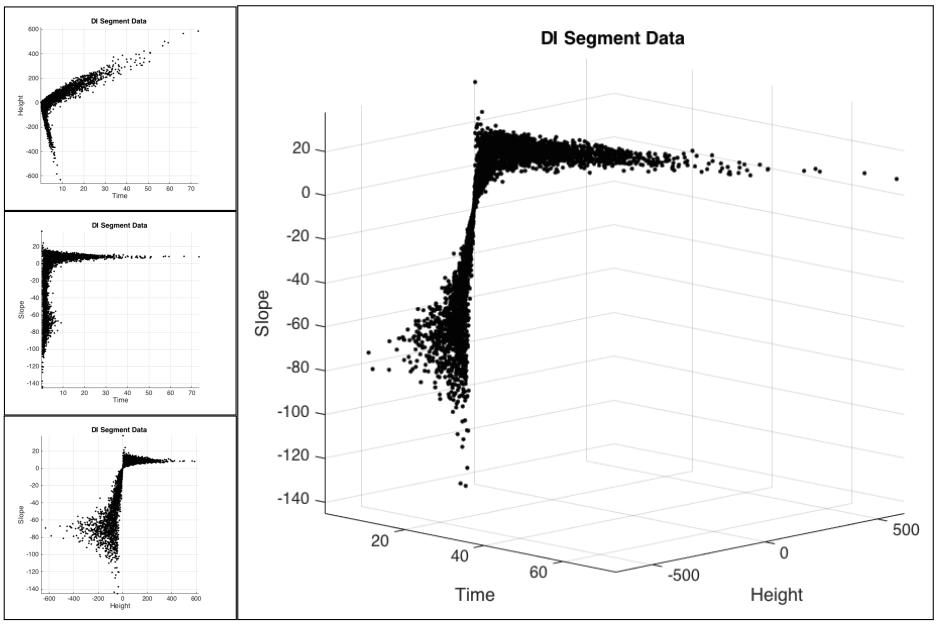


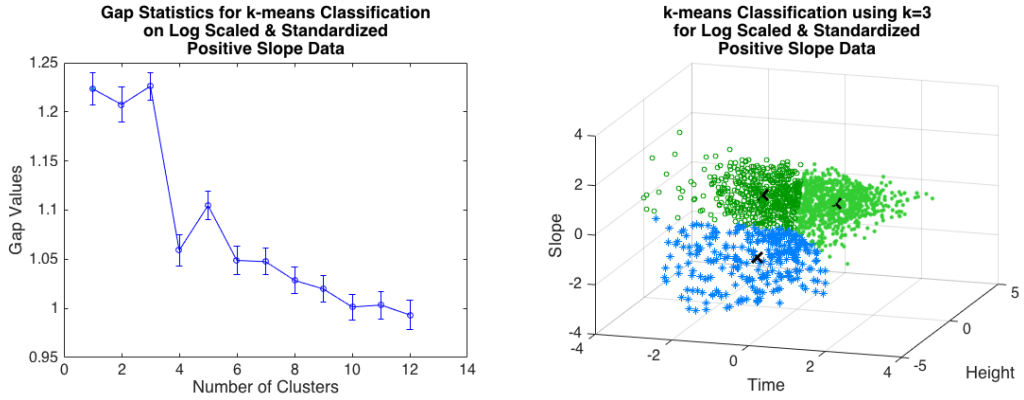

Figure 4.15. The data points for each segment obtained from the linear segmentation on the MT length history for a 10hr simulation of a 2-PF MT, such as those segments identified in Figure 4.14(b). Different perspectives are shown here to aid the visualization of the 3D plot.

Once the 3D segment data is available, the nucleation phase segments can be separated using the near zero height change and slope criteria. The same near-zero segment criteria is used to separate the positive- and negative-slope segment for both the 2- and 13-PF MT model segment data. After removing the near-zero slope segments, the $K$-means clustering step can be performed on the positive- and negative-slope segments separately. Figure 4.16 shows that the gap statistics clearly

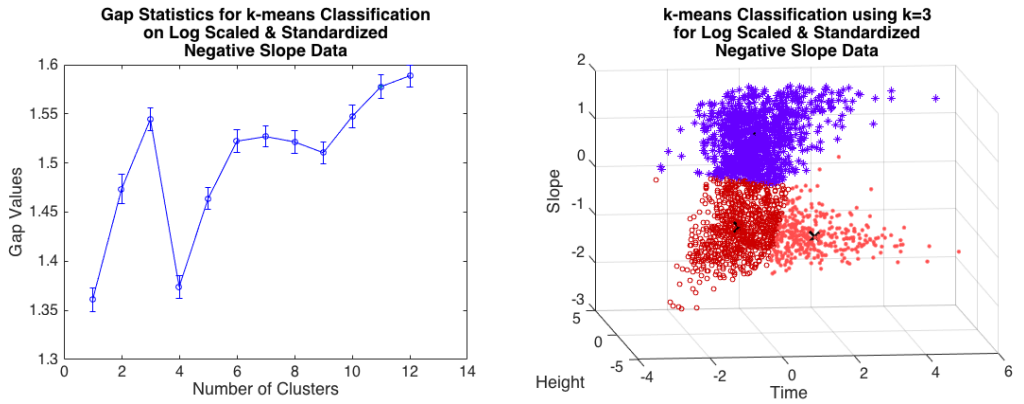
indicate $K = 3$ to be the best number of clusters to used in each portion of the data. In fact, the first local maxima are more drastic than the 13-PF MT clustering results. The better separation measured can be a result of the larger variance in values in the 2-PF MT segment data.

The gap statistic results verifying the choice for $K = 3$ completes the diagnostic portion of the classification process, as well as the list of the threshold and criteria values needed to finish the classification step. The diagnostic portion results indicated the same values listed in 4.2 to work for the 2-PF MT segment data as well. Moving forward with the classification procedure results in the labeled segment plot displayed in Figure 4.17. With the phase classes separated, the average measurements can be made, which are displayed in the table and box plots in Figure 4.18. Compared to the 13-PF MT model phase segment measurements, the 2-PF MT model phases are certainly shorter with respect to the time duration and height changes, especially for growth and shortening segments. However the growth segments tend to have steeper slopes in the 2-PF MT case, and nearly the same slope for the other phases with a wider variance in values. Additionally, the brief growth phase achieves average slope values higher than the long growth phases, mostly because the Voronoi cell corresponding to the brief growth phase covers a portion of the $z(x, y) = y/x$ manifold that protrudes higher in the $z$-direction than the rest of the surface on which the data segment points lie.

With each segment classified as one of the seven phases, it is possible to visualize each period of consistent behavior with phase labels, as displayed in Figure 4.19. Further analysis on the phases with respect to their contribution to the simulation is measured and shows in the plots of Figure 4.20. Due the higher rate of fluctuations that appear in the 2-PF length history data, the shear number of segments identified is larger than the 13-PF MT model case. However, the relative number for most phases, especially nucleation, is similar in both of the MT models. Additionally, the




(a)




(b)

Figure 4.16. The classification results for (a) positive slope segment data, and (b)negative slope segment data generated from 10hr simulation of the 2-PF MT model. (Left) Gap statistics when clustering for different $K$-values. In both cases, the first local maximum appears at $K = 3$. (Right) The $K$-means clustering results on the log scaled and standardized positive slope data points for $K = 3$, displayed with different colors and markers. A black $\times$ marks the center of each cluster.

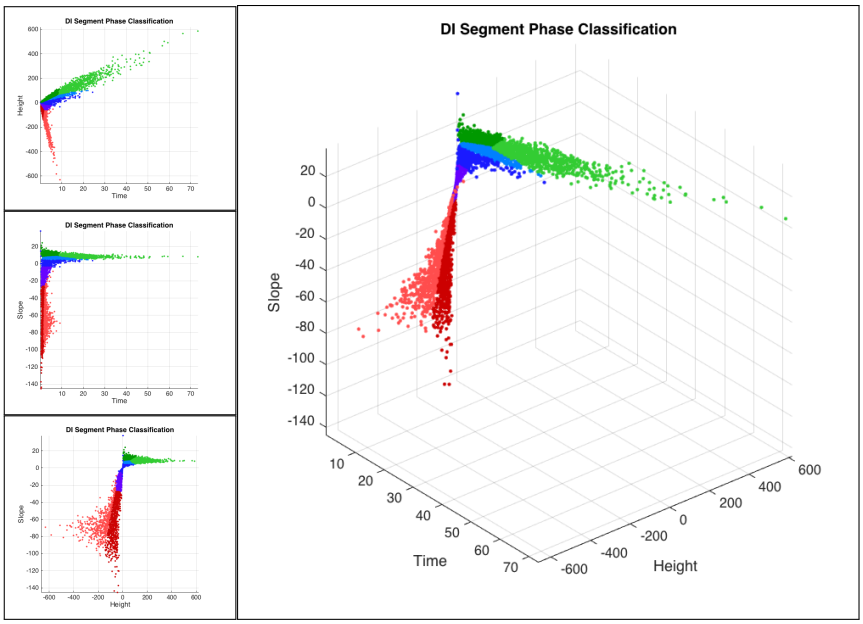



Figure 4.17. Phase classification results on the data set displayed in Figure 4.15. Different classes are labeled according to the legend in Figure 4.7.

**Segment Measurements for each DI Phase**

| | # of Segments | Time Mean | Time StDev | Height Mean | Height StDev | Slope Mean | Slope StDev |
|---|---|---|---|---|---|---|---|
| LngGr | 806 | 18.4 | 8.7 | 155.8 | 69.2 | 8.7 | 1.5 |
| BrfGr | 445 | 4.5 | 2.2 | 48.1 | 24.0 | 11.0 | 2.2 |
| UpSt | 396 | 7.8 | 4.0 | 45.9 | 21.7 | 6.0 | 1.2 |
| FltSt | 448 | 3.5 | 3.7 | 4.3 | 13.7 | 0.3 | 3.0 |
| DnSt | 837 | 1.7 | 0.7 | -22.5 | 11.3 | -13.9 | 6.1 |
| BrfSh | 873 | 1.2 | 0.4 | -65.7 | 26.5 | -59.1 | 23.4 |
| LngSh | 509 | 2.9 | 1.1 | -180.5 | 83.3 | -63.4 | 15.6 |

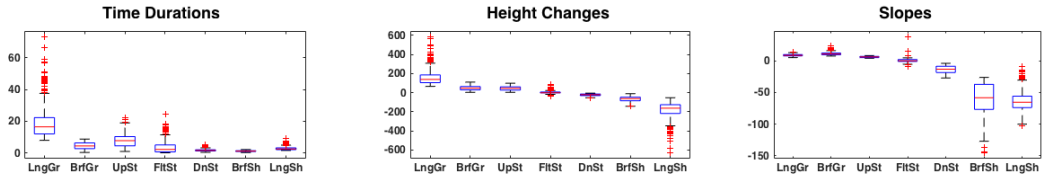



Figure 4.18. (Top) Mean and standard deviation for the time duration, height change, and slope measurements for each dynamic instability segment phase identified in the 2-PF MT model simulations. (Bottom) Box plots of the time duration, height change, and slope measurements for each phase class. The red crosses represent line segment data that are outliers in their respective phase class.

higher fluctuation rate meant that the 2-PF MT spent less time in growth phases, leading to shorter MTs, and more time spent closer to the MT seed, leading to more time spent in nucleation phases. Despite a disagreement in the percent of time spent in each phase between the 13- and 2-PF MT model simulations, in should be noted that the level of tubulin concentration in both cases indicates a balance of phase time durations for displaying a variety of DI behaviors. Similar to the $10\mu M$ levels used in the 13-PF MT model simulations, using less than $12\mu M$ tubulin concentration levels would have produce length history data that spent too much time in nucleation, and any more than $12\mu M$ would bias the data to have have much more growth. So, these plots are also useful for verifying that indeed an appropriate tubulin concentration level is being used to generate the desired DI behavior.

Additionally, a comparison of the different slope measurements shows a good separation between phases. The percent height change for each phase chart also provides support of how the stutter phases truly make up for a more subtle portion of length changes that occur in the simulated data. Qualitatively, the overall phase measurements shown in Figure 4.20 are in agreement with the results for the 13-PF MT model simulations in Figure 4.11, and the fact that the stutter phases cannot be ignored remains to be the case.

The next step is to consider the chronological order in which the phases appear, and analyze the emerging patterns in the phase transition. Recall that, in order to ease the number of permutations of phase transitions possible, only bundled phases are considered for this part of the analysis. This means that phases lose their sub-phase labels, and are only part of either growth, stutter, or shortening phases. For the 2-PF MT model case, the results displayed in Figure 4.21 again indicate that the onset of a shortening phases is more common to be a transitional catastrophe, such that it first goes through a stutter phase before rapidly losing its polymer mass. In contrast, abrupt rescues are more common when switching from a shortening to

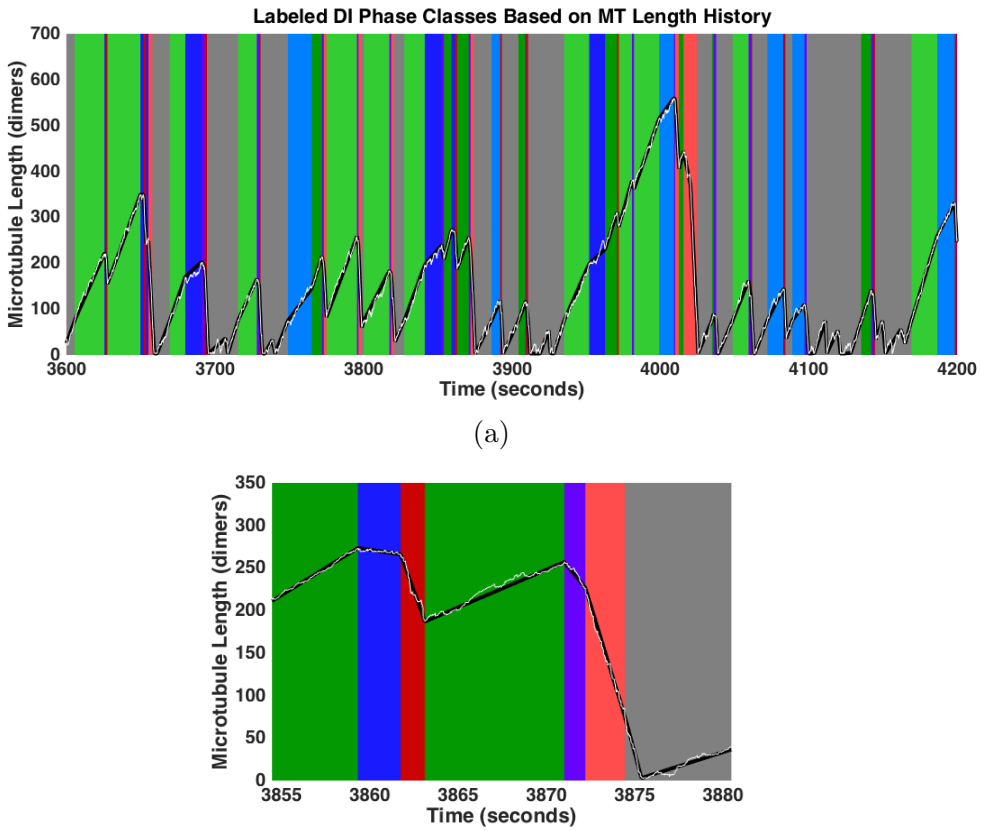



Figure 4.19. (a) Color labeled representation of line segments that have been classified from a portion of length history data from the 2-PF MT model 10hr simulation. Each segment is labeled using the color legend in Figure 4.7, in addition to periods of Nucleation (gray). (b) A zoomed in excerpt from the same plot.

a growth phase, indicating that less structural changes are necessary for a MT to suddenly change back to polymerization after a period of sustained subunit loss. Also, related to the more frequent fluctuations in the 2-PF MT model length history data, the frequencies and shear number of phase transitions is more than for the 13-PF MT case. Even so, the role of stutter phases remains to predominantly be more during transitional catastrophes, and certainly less for rescues. Finally, when considering pair-wise combinations of phase transitions, the 2-PF MT model shows somewhat different patterns than those observed in the 13-PF MT model case. The

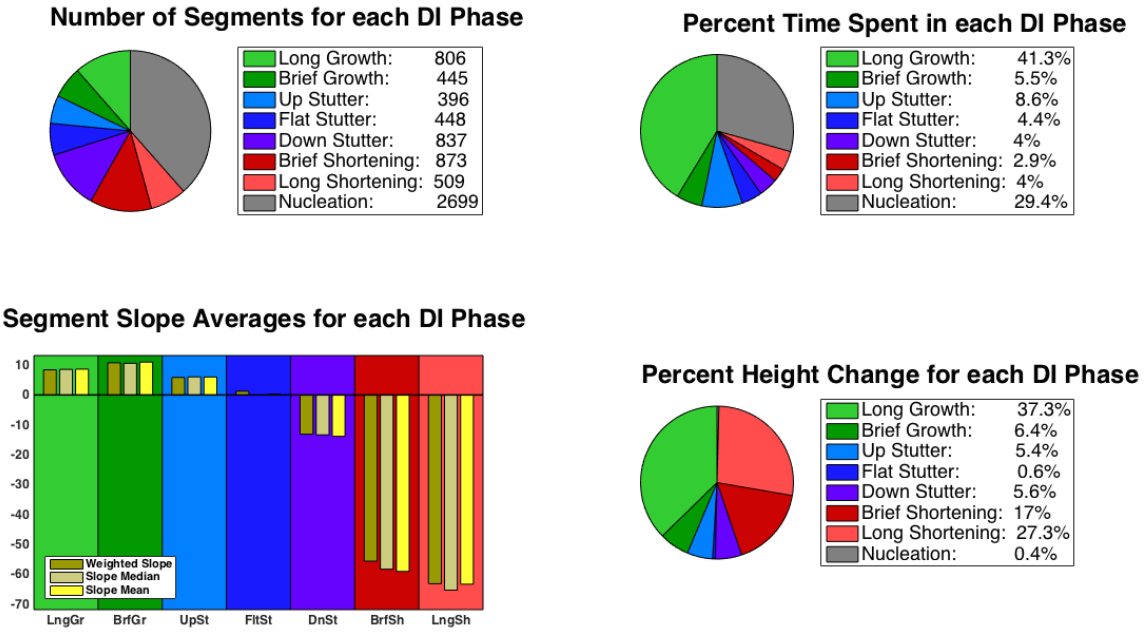


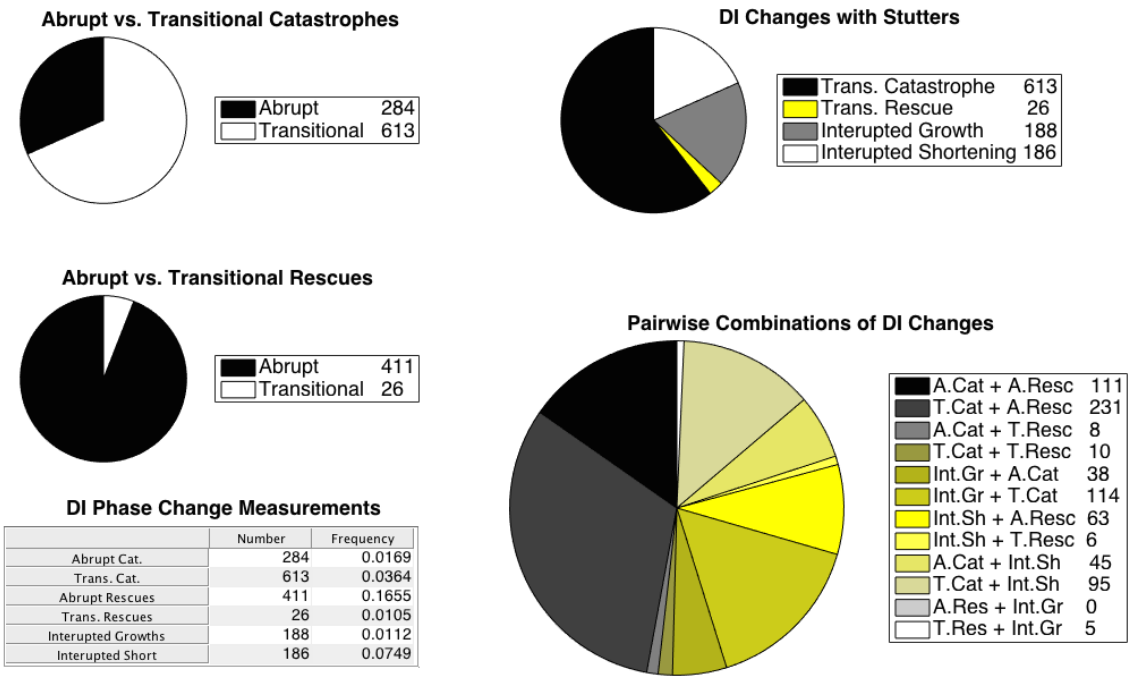

Figure 4.20. Different properties measured for each DI phase identified in the 2-PF MT model 10hr simulation.

most commonly observed phase transition combinations are the following:

1. Transitional Catastrophe + Abrupt Rescue: 231 instances
2. Interrupted Growth + Transitional Catastrophe: 114 instances
3. Abrupt Catastrophe + Abrupt Rescue: 111 instances
4. Transitional Catastrophe + Interrupted Shortening: 95 instances

Aside from having many more transitions to consider, it is interesting to see that the most frequent pattern is the Transitional Catastrophe + Abrupt Rescue combination. Additionally, the Abrupt Catastrophe + Abrupt Rescue combination is a frequent pattern that was not commonly observed in the 13-PF MT model case.

It is important to note that a large number of different patterns of phase transitions are successfully captured in the 10 hour long simulations of the 2-PF MT model, at the $12\mu M$ tubulin concentration level, satisfying the intended result of using the




**DI Phase Change Measurements**

| | Number | Frequency |
|---|---|---|
| Abrupt Cat. | 284 | 0.0169 |
| Trans. Cat. | 613 | 0.0364 |
| Abrupt Rescues | 411 | 0.1655 |
| Trans. Rescues | 26 | 0.0105 |
| Interupted Growths | 188 | 0.0112 |
| Interupted Short | 186 | 0.0749 |

Figure 4.21. Measurements of possible dynamic transition events generated by the perturbations created from growth, shortening, stutter, and nucleation phases found in the 10 hour long 2-PF MT model simulations.

model parameters that delivered this variety of DI behaviors, which is part of the goal of this study. Also, the aim of the 2-PF MT model by design was to simplify the tip region structure, so that focus can be placed on the most dynamic part of the MT structure. The results of this section provided the phase classes identified from periods of consistent behavior, separated by moments of significant dynamic changes in MT length. This information will be used in the next chapter, which makes connections between the structural features of the MT tip region, and the DI phases during which they occur.

# CHAPTER 5

# DATA ANALYSIS II: PREDICTING DI PHASES FROM THE TIP STRUCTURES IN THE 2-PF MT MODEL

In Chapter 2, the biological importance of the MT tip region was discussed, since that is the portion of the MT structure that contains the subunits and bonds targeted by the molecular dynamic events. In Chapter 3, a computational model representing the MT structure and dynamics was presented, such that it simulated an exact trajectory of bio-chemical exact MT structural states. The simplified 2-PF MT model was also presented to ease the study of the tip region, while maintaining the DI behavior that is characteristic to MTs. In Chapter 4, a method was presented to segment, classify, and analyze any possible macro-level phases that exist in length history data resembling DI behavior. The contents of this chapter combine the focus and results from all of the previous chapters together in order to connect the micro-level MT tip structures from simulations to the macro-level DI phases and phase transitions during which they are observed. The Random Forest machine learning method is utilized to treat the different measurements of individual 2-PF MT tip features as predictor variables in order to predict the response variables, represented by the DI phase labels identified in Chapter 4. Different arrangements of this data set allows the learning approach to predict phases by only looking at tip structures, and to forecast upcoming phase transitions. An added benefit is the ability of these approaches to rank the importance of the MT tip features used in making these predictions and forecasts. This chapter presents the development of these predictive models, and the resulting knowledge providing deeper insight into the mechanisms

that lead to the sporadically occurring changes in MT dynamics.

## 5.1 Motivation for Developing Predictive Models

The ultimate goal of this study is to develop a deeper insight into the mechanisms of MT dynamics that lead to significant phase transitions in DI. The knowledge acquired in the past provided information about general expectations of the MT structure during growth and shortening phases exclusively. This included generalizations of a healthy GTP-cap being present during growth, and the lack of one during shortening, which exposed less stable MT structures vulnerable to rapid loss of subunits. Verifying these conclusions is possible through experimental observations, however laboratory conditions limit these confirmations from reaching the smaller tip structures known to be most dynamic portions of the MT. Additionally, the time scale at which dynamic changes alter the MT tip region are very quick, and the speed at which experimental data can be observed is also limited. For this reason, the computational models representing MT structures and dynamics are an invaluable tool, such that simulated data provides a scope that is unattainable in current laboratory settings.

The exact method implementation of the detailed MT computational model presented in Chapter 3 simulates the trajectory of MT structure states that are bio-chemically realistic. Doing so exposes the information about the MT structure at a very fine time resolution. The simplified 2-PF MT model reduces the number of various tip structures possible to create a feasible scenario for studying the most dynamic part of the MT. The simulations provide micro-level structures that can be calculated with respect to various properties that are bio-chemically relevant, or to distinguish similar tip configurations from each other. The same 2-PF MT model simulations also generates length history data, for which macro-level DI phase information is extracted, as demonstrated in Chapter 4. These DI phase assignments are applied as

labels to each observation in the simulation that occurs during the time duration of the corresponding phase segment. To connect these two forms of data extracted from different time scales, a machine learning method is sought to associate the structural properties of the MT tip structures to the DI phases during which they occur.

Considering the large number of possible tip configurations, the numerous features that describe a tip structure configuration, and the unknown micro-level characteristics that distinguish stutters from growth and shortening phases, a simple and effective approach is provided by the Random Forest learning method. This machine learning approach helps to test the ability for tip structure features to predict DI phases, and thus delivers a deeper understanding of how the tip structures relate to the macro-level dynamics. Also, reconfiguring the data to consider each phase separately can test the ability of structural features to forecast future changes out of DI phases. The remainder of this chapter covers the organization of the simulated 2-PF MT model data, and how the Random Forest approach is used to gain a deeper understanding on the role MT tip structures play in regards to the macro-level dynamics seen in DI behavior.

## 5.2 2-PF MT Tip Data

The 2-PF MT model was introduced to help simplify the tip configurations, so that the micro-level structural information about the MT is created at a level unavailable in laboratory settings. Using the DI phase classification methods presented in Chapter 3, each micro-level observation is now matched with a macro-level phase. In this section, the measurements that describe the individual micro-level structural features are presented. Additionally, since the DI phase labels are also available, an initial exploration into how the individual feature measurements relate to DI phase is provided with visualized comparisons.

### 5.2.1 Simulating Tip Data from the 2-PF MT Model

The output from these 2-PF MT model simulations supplied provided over 4.6 million observations, such that the data is relevant to the distinguishing different features of individual structural configurations. More specifically, the simulation output included the actual configuration of the 2-PF MT's gated tip, such that the exact ordering of GTP- and GDP-bound subunits was known for the G-subunits, and the PF tips above them. Early data explorations revealed that the information about the gated tip region alone, however relevant, was not sufficient to distinguish DI phases. It turns out that similar tip configurations were commonly observed amongst all of the DI phases. For this reason, additional information about the number of GTP-bound subunits in the entire MT was utilized to approximate the size of a GTP-cap near the MT tip region.

Though previous studies have indicated some expectations on the size of a GTP-cap, tracking the actual configurations of subunits that would guarantee containing the entire GTP-cap structure would require far more configurations than the 4 million gated tip configurations possible in the 2-PF model (see Equation 3.8). As a first attempt for studying the specific tip configurations, estimates for the GTP-cap size are sufficient at this time. Additionally, it should be noted that many of the gated tip configurations were tracked on the fly during the simulation run, and any measurements on the features were calculated in post-processing. Trying to make these measurements beyond the gated tip region are possible during the simulation, however this would greatly add to the computational cost of making these calculations for each MT state encountered during each simulation step.

Using the DI phase classification, over 7,500 phases were identified, and each observation was additionally labeled with the corresponding bundled DI phase (growth, stutters, and shortening). Measurements of 2-PF MT tip structural features were made for each observation independently, and some insight to their relation between

individual features and DI phases was provided in the form of box plots and histogram plots in Section 5.2.2. The rates of reaction events were also included in these measurements, because they too depend on the tips structure, and provide some additional information by weighting GTP- and GDP-bound subunits differently. Though many of these measurements have a poor separation in regards to DI phases, these tip structure measurements still carry a bio-chemical relevance to the possible molecular dynamics, as well as help uniquely describe individual tip configurations. Even so, many of these measurements do provide evidence that stutter phases behave as a transition between growth and shortening phases. These tip structure features will be used as the predictor variables in the predictive models developed later in this chapter. The Random Forest method offers variable importance rankings that will help determine which of these features are most relevant to identifying the DI phases to which they correspond.

### 5.2.2 Calculating Tip Structure Features

Below is a description of the measurements for simulated 2-PF MT tip structures and the corresponding box plots and histogram plots split by the three bundled DI phases being considered (shortening phases are in red, stutter phases are in blue, and growth phases are in green). Histogram plots also display a comparison for the distribution of the measured values across the entire data set of simulated MT tip structures (shown in black). It should be noted that in any tip configuration, left and right PF assignments are not used, and instead, to take advantage of the simplification offered by symmetry, measurements are made for the longer or shorter PFs accordingly. Some measurements deal exclusively with individual PFs, while others consolidate measurements across both PFs to reference the entire MT tip. Recall that the term "gated tip" refers to the cracked (or laterally unbonded section) of the either PF or MT tip plus the G-subunits (the top-most laterally bonded subunits) (see

Figure 3.3 for details). This simple gated tip definition is unique to the simplification offered by the 2-PF MT model. For future work, similar types of measurements could inspire means for measuring 13-PF MT tip structures.

1. Length of the Longer PF Tip: measures the number of laterally unbounded subunits in the longer PF (see Figure 5.1).

2. Length of Shorter PF Tip / Equivalent to Crack Depth: measures the number of laterally unbounded subunits located in the shorter PF, which also defines the depth of the crack between the two PFs (see Figure 5.2).

3. Ratio of the Shorter PF tip Length to the Longer PF tip Length: compares the lengths of the two PF tips, such that a value of 1 indicates the same length, and a value of 0 indicates that shorter PF tip has no laterally unbonded subunits (i.e. it has the null configuration) (see Figure 5.3).

4. Total number of GTP-bound subunits in the entire MT: this measures all of the GTP-bound subunits throughout the entire MT structure, including those in the seed (see Figure 5.4).

5. Total number of GTP-bound subunits in the gated MT tip: considers the number of GTP-bound subunits in the gated tip region (see Figure 5.5).

6. Percentage of all the GTP-bound subunits located in the gated MT tip region: compares measurements 4 and 5, as the percentage of the total GTP-bound content found only within the gated tip region (see Figure 5.6).

7. Percentage of the Longer PF gated tip comprised of GTP-bound subunits: measures the percentage of the subunits in the longer gated PF tip that are GTP-bound (see Figure 5.7).

8. Percentage of the shorter PF gated tip comprised of GTP-bound subunits: measures the percentage of the subunits in the shorter gated PF tip that are GTP-bound (see Figure 5.8).

9. Percentage of the MT gated tip comprised of GTP-bound subunits: measures the percentage of only the subunits in the gated MT tip that are GTP-bound (see Figure 5.9).

10. Number of neighboring subunit pairs in the gated crack containing at least a GTP-bound subunit: indicates the health of the cracked region, since neigh-

boring subunit pairs across the crack dictate the likelihood of a lateral bond breaking after future reaction events form bonds there (see Figure 5.10).

11. Number of G-subunits that are GTP-bound: helps determine how strong the top most lateral bond is, since the state of those subunits determines the rate of a lateral bond breaking (see Figure 5.11).

12. Number of AG-subunits available: the AG-subunits are located above the G-subunits, the lowest subunits without a lateral bond at the bottom of the crack. If they don't exist, a new lateral bond can't form. It's interesting that there is little difference across DI phases. (see Figure 5.12).

13. Number of AG-subunits that are GTP-bound: helps determine the likelihood of a lateral bond forming, if there is the appropriate space for one to form (see Figure 5.13).

14. The estimated GTP-cap size: measures an estimate for the GTP-cap assuming all the GTP-bound subunits are adjacent and located in the MT tip (see Figure 5.14).

15. The estimate of the GTP-cap size below the crack depth: considers the estimate for the GTP-cap size in measurement 14, and measures how far below the crack it extends. Since this is a computed difference, negative values are allowed, and they indicate that the crack penetrates deeper down the MT lattice than the estimated GTP-cap (see Figure 5.15).

16. The ratio of the estimated GTP-cap size to the average PF Tip Lengths: compares the size of the estimated GTP-cap size (measurement 14) to the gated MT tip length (average of measurements 1 and 2). A value of zero indicates no GTP-bound subunits are in the gated tip region (see Figure 5.16).

17. Dispersion of GTP-bound subunits in the Longer gated PF tip: measures the dispersion of GTP-bound subunits in the longer PF's gated tip region by considering the mean longitudinal distance between GTP-bound subunits. The measurement includes "ghost" GTP-bound subunits above and below the gated PF tip to take into account the sequence of GDP-bound subunits that would otherwise go undetected. If no GTP-bound subunits are available, the value of 25 is assigned since the longest MT tip lengths are near 20 subunits long. Larger values indicate sparsely spaced GTP-bound subunits, and smaller values near 0 indicate nearly adjacent GTP-bound subunits (see Figure 5.17).

18. Dispersion of GTP-bound subunits in the Shorter gated PF tip: measures the dispersion of GTP-bound subunits in the shorter PF's gated tip region by con-

sidering the mean longitudinal distance between GTP-bound subunits. Additionally, the measurement includes "ghost" GTP-bound subunits above and below the gated PF tip to take into account the sequence of GDP-bound subunits that would otherwise go undetected in the measurement. If no GTP-bound subunits are available, the value of 25 is assigned since the longest MT tip length being observed above 20 subunits long. Larger values indicate sparsely spaced GTP-bound subunits, and smaller values near 0 indicate nearly adjacent GTP-bound subunits (see Figure 5.18).

19. Dispersion of GTP-bound subunits in the gated MT tip: measures the dispersion of GTP-bound subunits in the gated MT tip region by averaging all of the individual distances identified in measurements 17 and 18 (see Figure 5.19).

20. Standard deviation of GTP-bound subunit positions in the gated MT tip: a second measurement of dispersion of GTP-bound subunits. This was included in case it provided more information that the mean distance between GTP-bound subunits, but as the box plots indicate, there is little difference between DI phases (see Figure 5.20).

21. Rate of Hydrolysis: a scalar multiple of the rate constant for hydrolysis and the number of hydrolyzable GTP-bound subunits in the entire MT structure, which does not include GTP-bound subunits in the MT seed or the top-most position of PF-tips. This number is calculated at each step in the simulation to determine the likelihood of a hydrolysis event to be the next one to occur in the Markov process (see Figure 5.21).

22. Expected Rate of Subunit Loss as a function of Tip Configuration: this is a weighted sum taking into consideration the possible rate of a particular longitudinal bond breaking and the number of subunits that would be lost as a result of that event taking place. This value strongly depends on the tip configuration, i.e. the specific sequence of GTP- and GDP-bound subunits that define the tip structure. The formula for this measurement is provided in Equation 3.4 (see Figure 5.22).

23. Rate of Lateral Bond Breaking: These values are calculated by taking a scalar product of the rate constants for lateral bond breakage and the bounded state of the G-subunits. It strongly depends on the number of G-subunits that are GTP-bound measured earlier (see Figure 5.23).

24. Rate of Lateral Bond Forming: These values depend solely on the availability of the AG-subunits from measurement 12. The consistency in the plots shown here is an artifact of the parameters choices used in the simulation, and verifies an expected outcome (the rate constants for lateral bond formation used to simulate this data are uniform for all AG-subunit states) (see Figure 5.24).

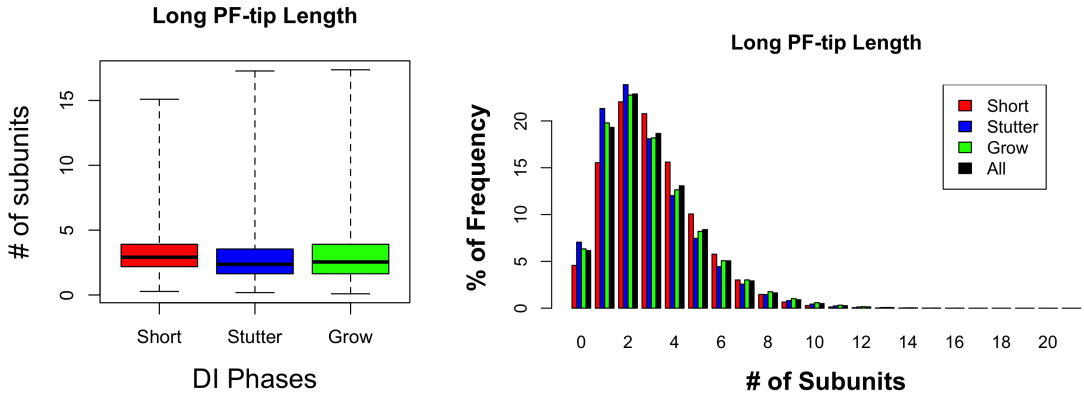



Figure 5.1. Comparison between different DI phases for the longer PF tip length.

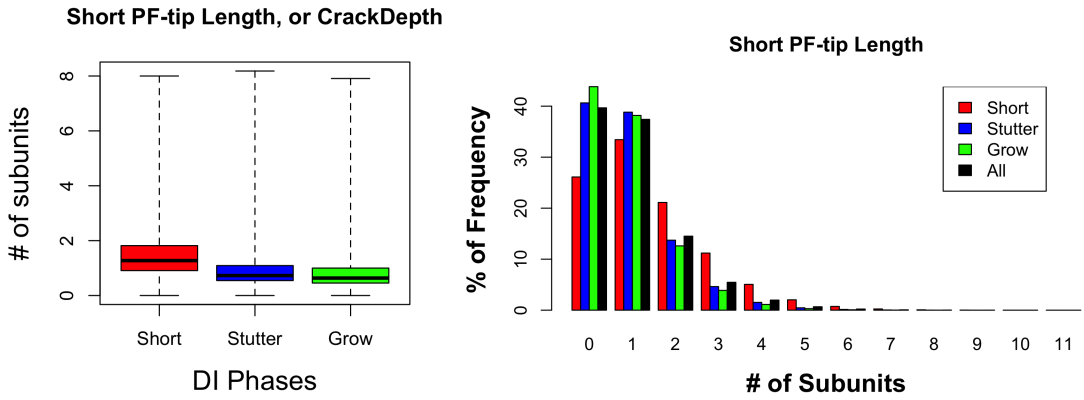



Figure 5.2. Comparison between different DI phases for the shorter PF tip length.

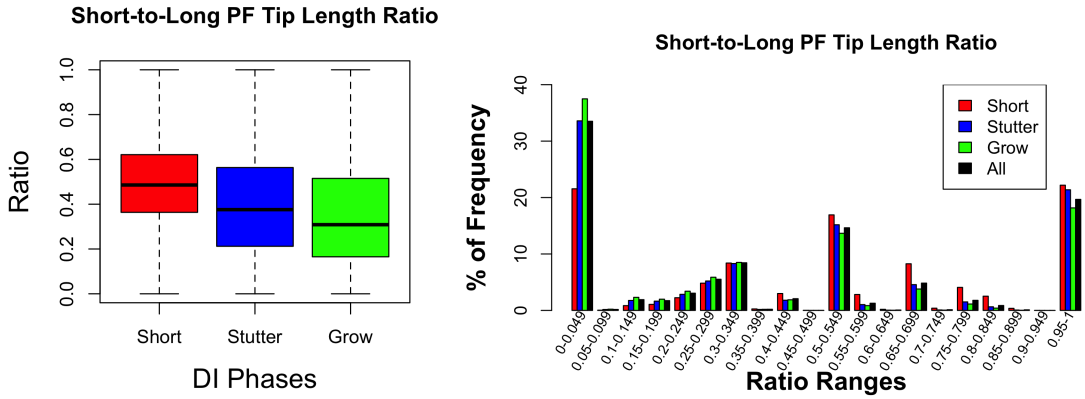



Figure 5.3. Comparison between different DI phases for the ratio of shorter to longer PF tip lengths.

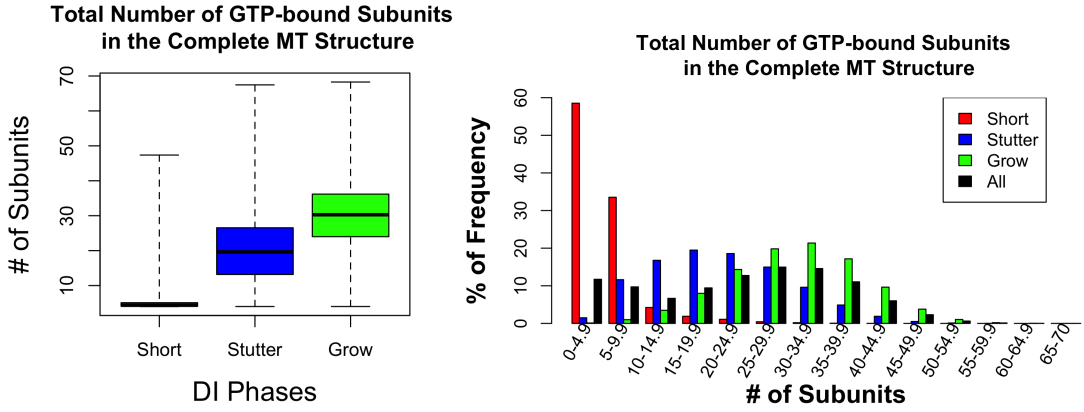



Figure 5.4. Comparison between different DI phases for the total number of GTP-bound subunits in the entire MT.

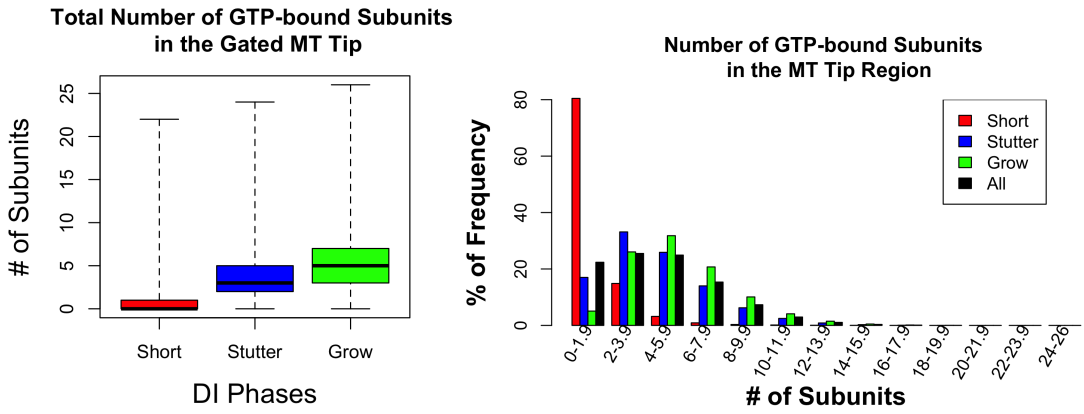



Figure 5.5. Comparison of the total number of GTP-bound subunits in the gated MT tip between different DI phases.

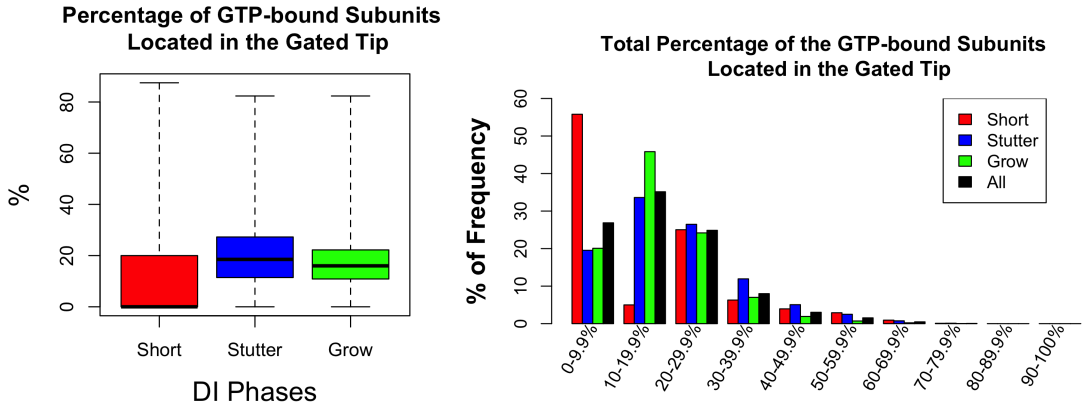



Figure 5.6. Comparison between different DI phases for the percentage of the GTP-bound subunits located in the gated MT tip.

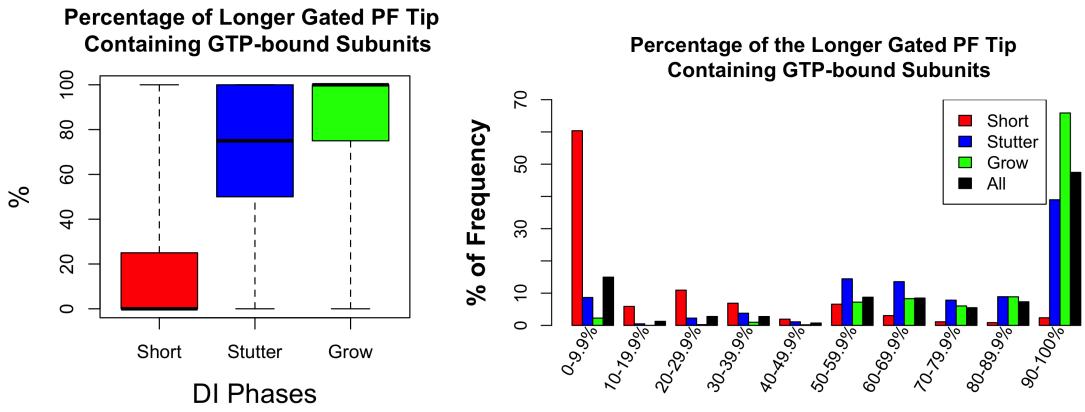



Figure 5.7. Comparison between different DI phases for the percentage of the longer gated PF tip being comprised of GTP-bound subunits.

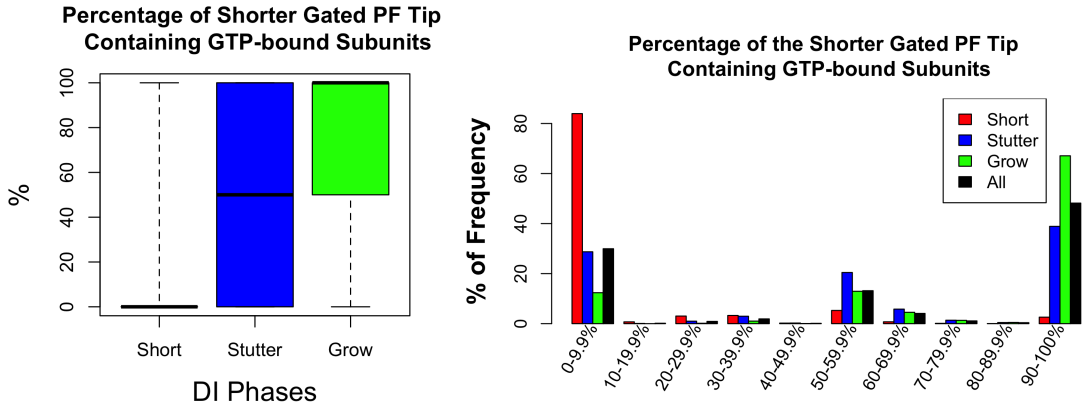



Figure 5.8. Comparison between different DI phases for the percentage of the shorter gated PF tip being comprised of GTP-bound subunits.

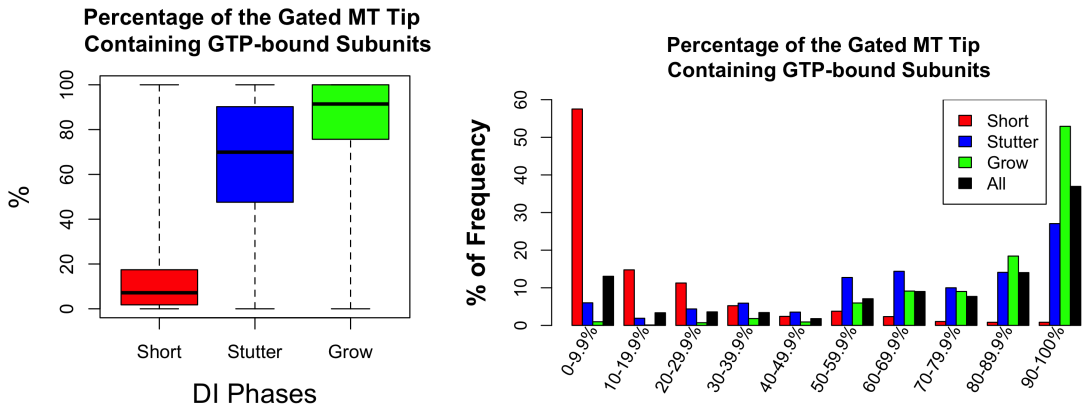



Figure 5.9. Comparison between different DI phases for the percentage of the gated MT tip being comprised of GTP-bound subunits.

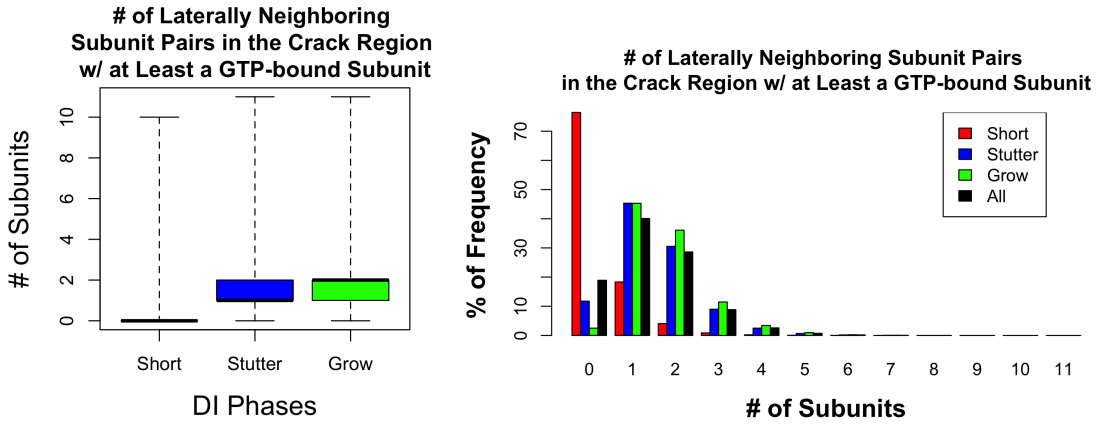



Figure 5.10. Comparison between different DI phases for the number of subunit pairs in the gated crack containing at least a GTP-bound subunit.

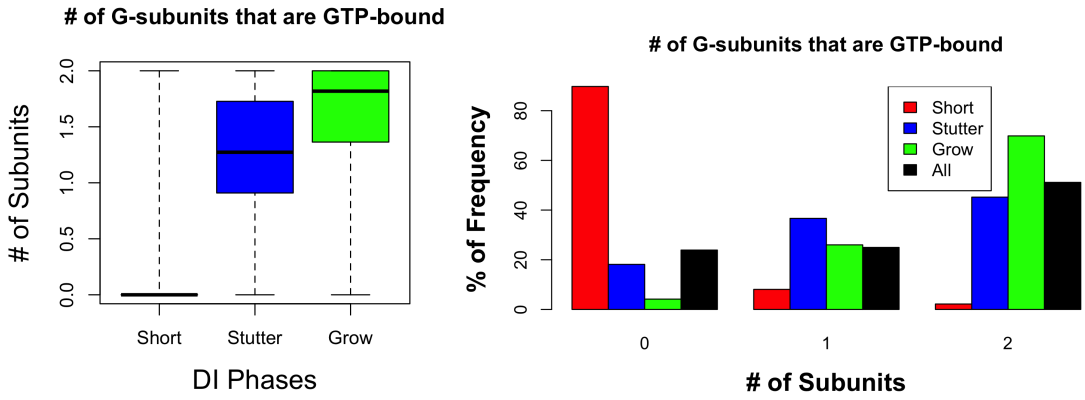



Figure 5.11. Comparison between different DI phases for the number of G-subunits that are GTP-bound.

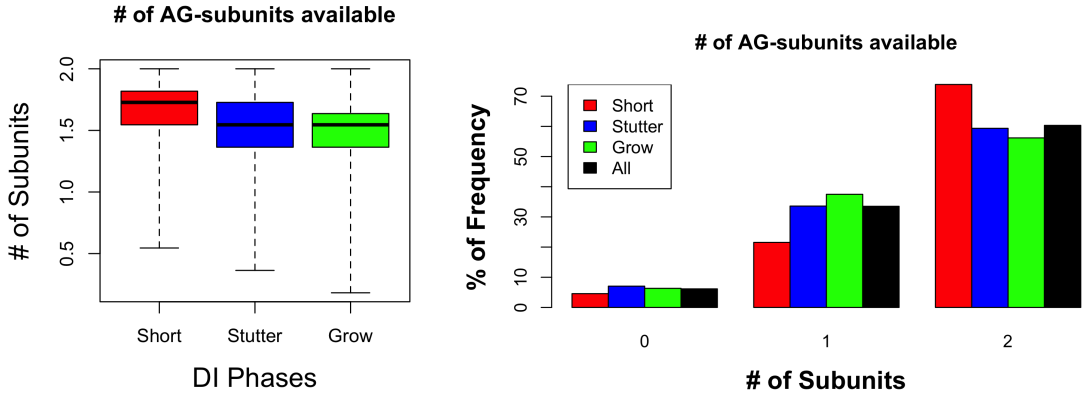



Figure 5.12. Comparison between different DI phases for the number of AG-subunits available.

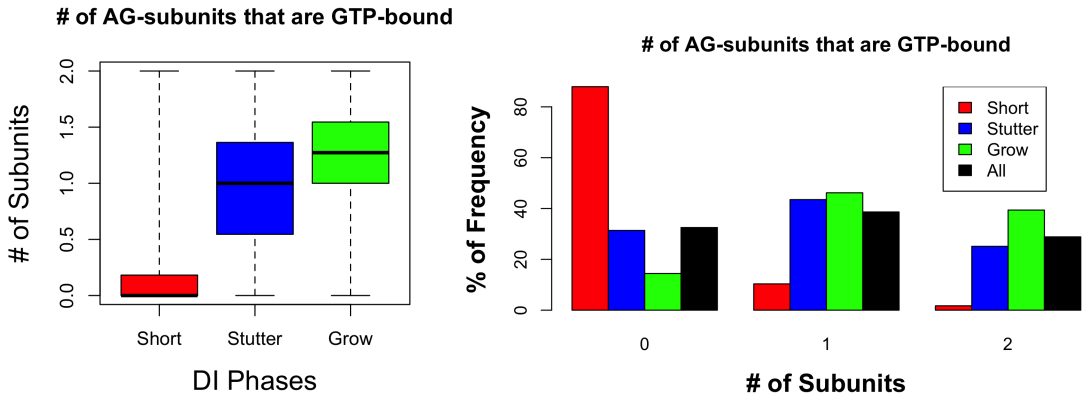



Figure 5.13. Comparison between different DI phases for the number of AG-subunits that are GTP-bound.

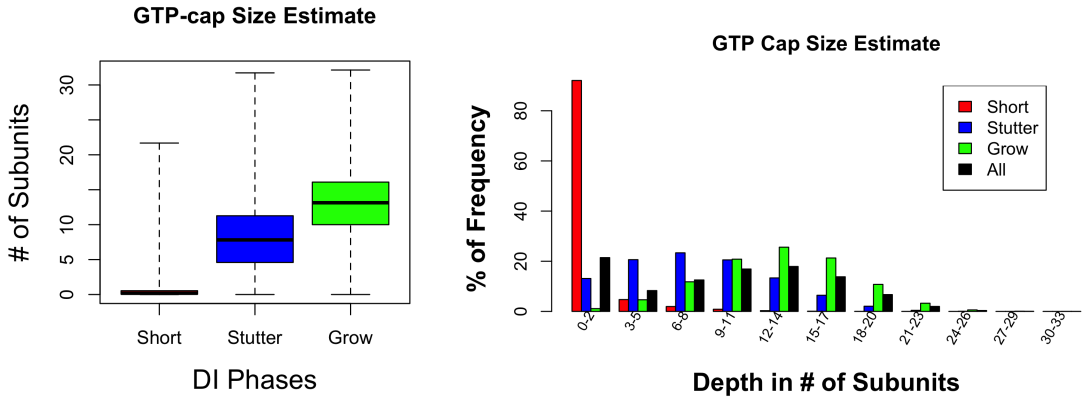



Figure 5.14. Comparison between different DI phases for the estimated GTP-cap size.

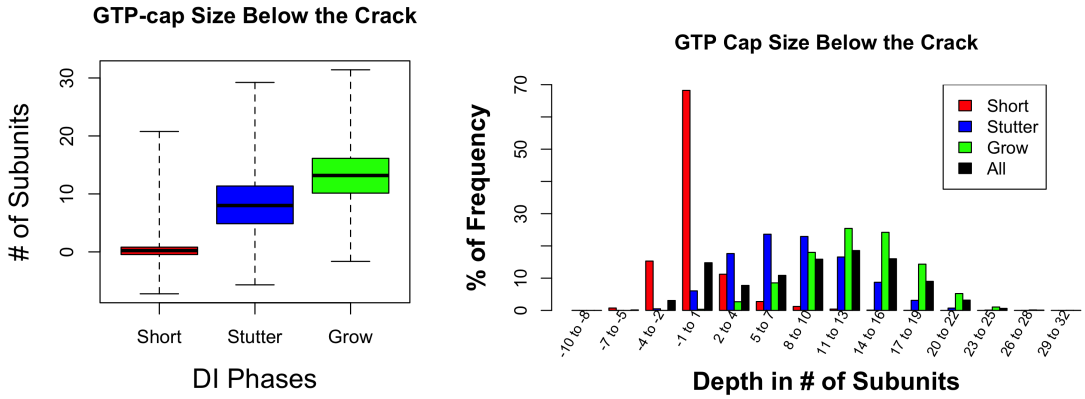



Figure 5.15. Comparison between different DI phases for the estimate of how far below the GTP-cap is from the crack depth.

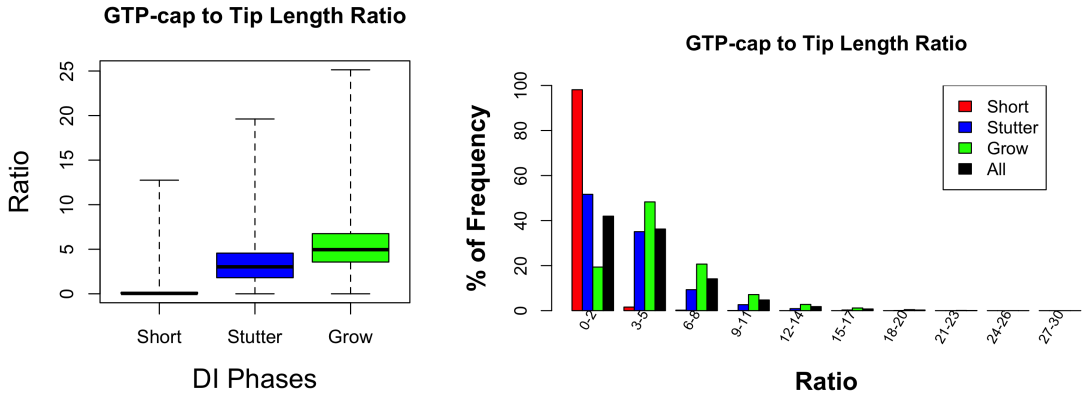



Figure 5.16. Comparison between different DI phases for the ratio of the estimated GTP-cap size to the average PF tip lengths.

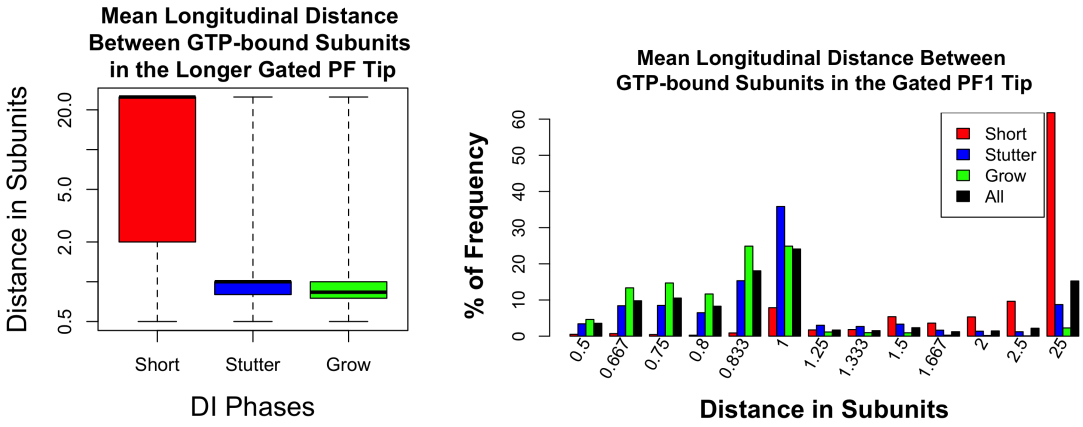



Figure 5.17. Comparison between different DI phases for the mean longitudinal distance between GTP-bound subunits in the longer gated PF tip.

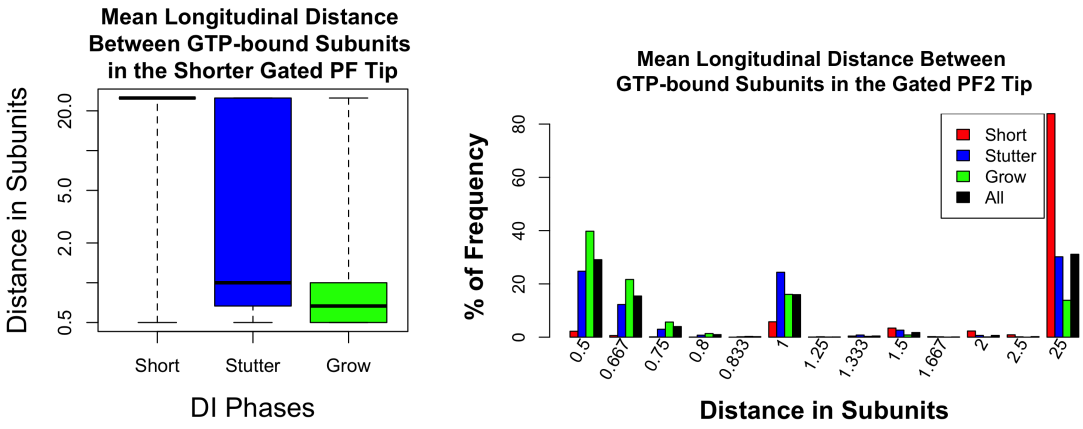



Figure 5.18. Comparison between different DI phases for the mean longitudinal distance between GTP-bound subunits in the shorter gated PF tip.

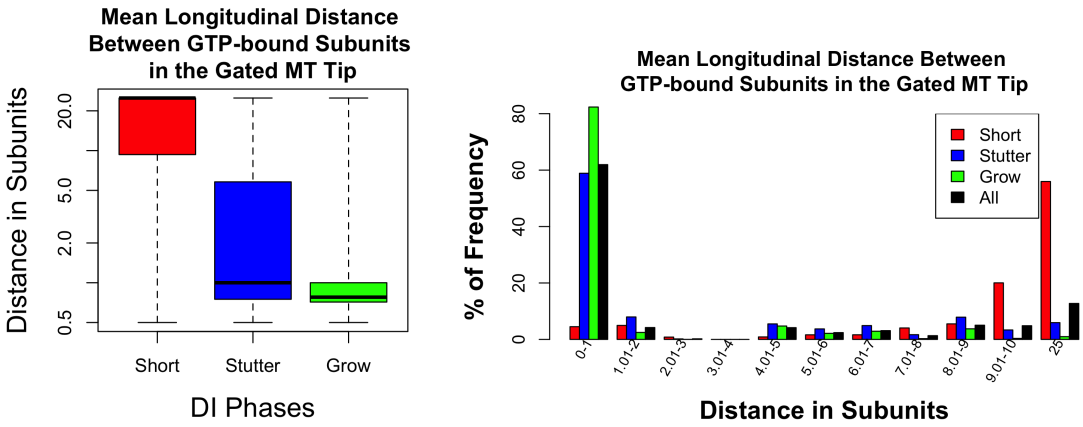



Figure 5.19. Comparison between different DI phases for the mean longitudinal distance between GTP-bound subunits in the gated MT tip.

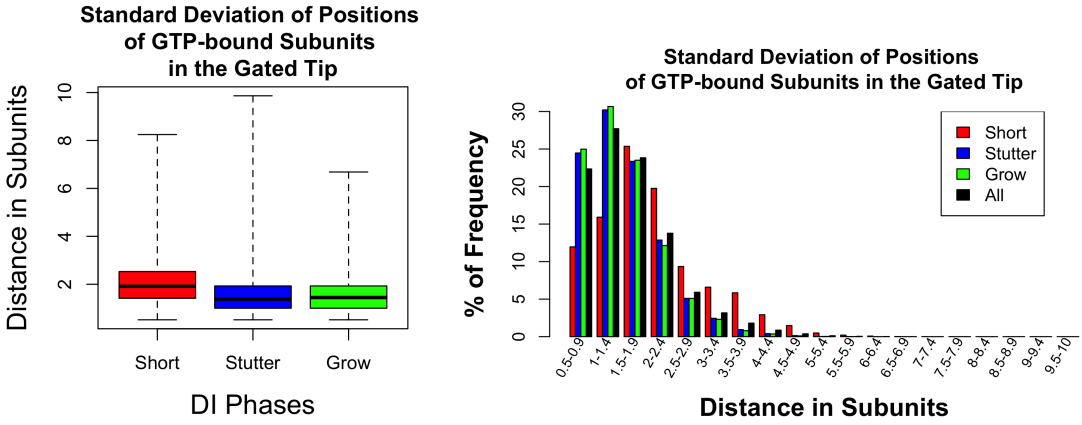



Figure 5.20. Comparison between different DI phases for the standard deviation of longitudinal positions of GTP-bound subunits in the gated MT tip.

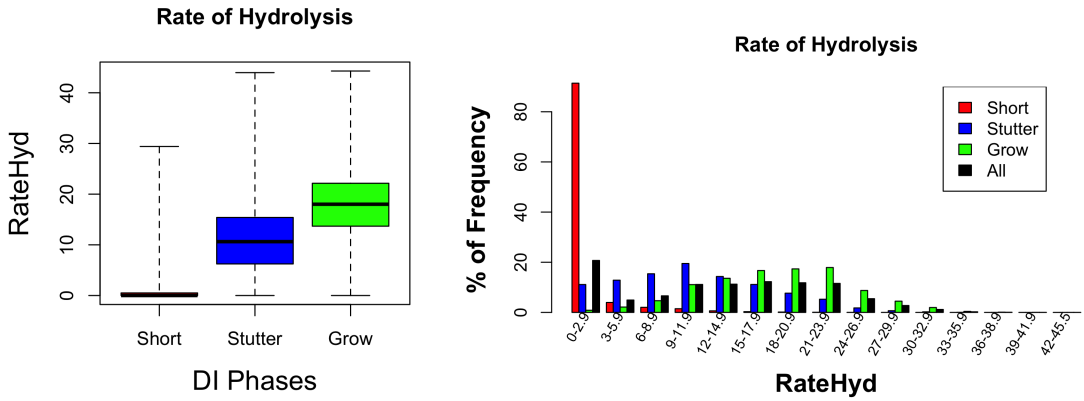



Figure 5.21. Comparison between different DI phases for the expected rate of a hydrolysis event.

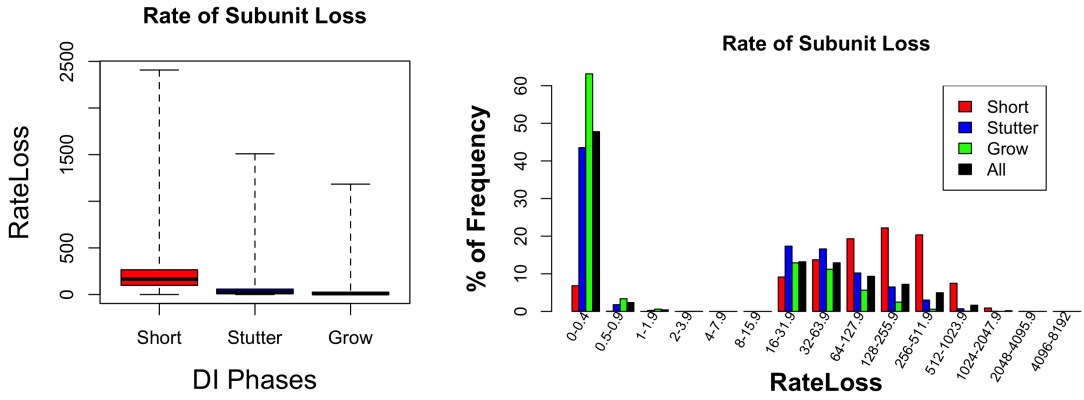



Figure 5.22. Comparison between different DI phases for the expected rate of subunit loss.

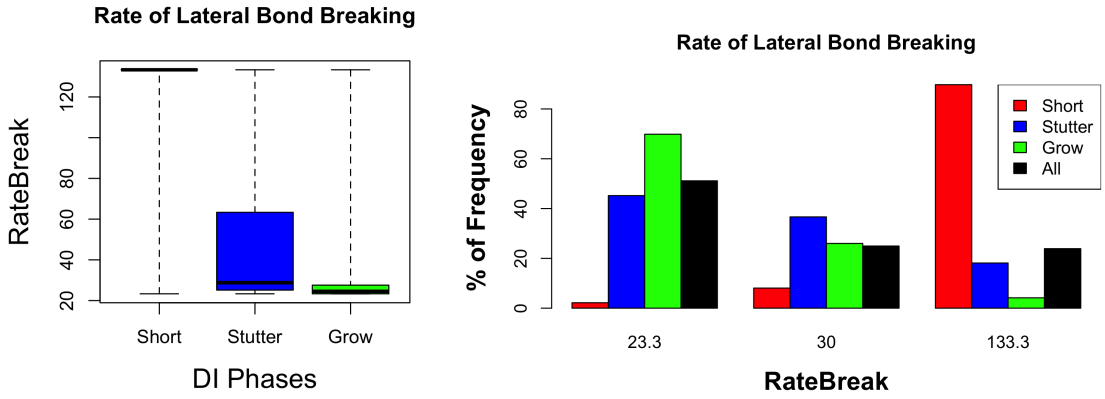



Figure 5.23. Comparison between different DI phases for the expected rate of breaking a lateral bond.

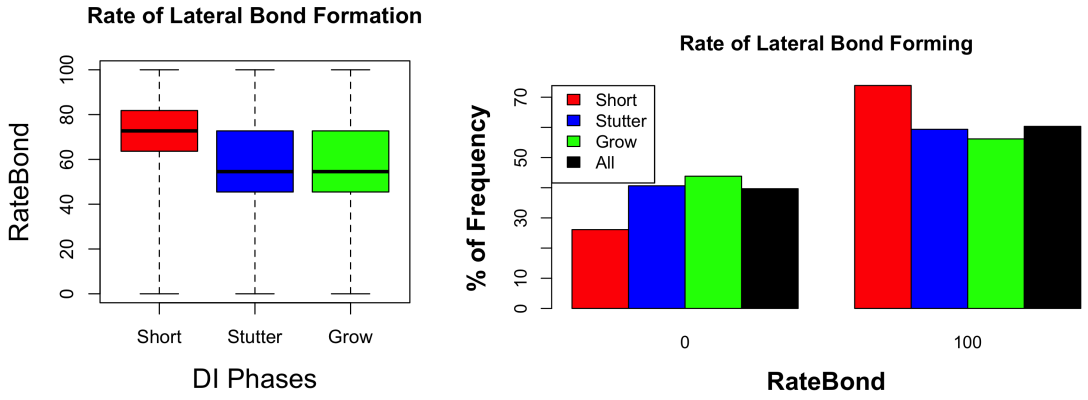



Figure 5.24. Comparison between different DI phases for the expected rate of forming a lateral bond.

## 5.3 Predictive Modeling Methodology

In order to connect the micro-level tip data calculated from individual MT structural observations to the macro-level DI phases identified from the classification procedures applied to the simulated data from the 2-PF MT model, a supervised machine learning method to create a predictive classification model is desired. This data set to be analyzed consists of a large number of observations with far fewer feature variables describing a tip structure. This eases the difficulty of the theoretical problem, however the large data size may be a consideration for computational costs in training the predictive model. This section discusses the reasons for using a the Random Forest approach for classification, and the different perspectives of insight it can offer for various arrangements of the data.

### 5.3.1 Description of Random Forest Classification Algorithm

The micro-level tip structure data and the macro-level DI phase information provided good conditions for a supervised machine learning method to study the relationship within this data set. The Random Forest model is one such method that can be used for classification, where the micro-level tip features can be used as the predictor variables, and the macro-level phase information can be used as the response variable. The Random Forest is an ensemble method consisting of many decision trees that can collectively form a classification model [6, 48]. It is easy to implement, and appropriate for large data sets, such as the $24 \times 3.3$ million tip feature data and the corresponding DI phase labels collected from the 2-PF MT model simulations. Out-of-Bag (OOB) errors are used to tune the predictive model, and are used to calculate the misclassification rates when training the model [6, 48]. OOB refers to the separation created in the existing data between training and testing subsets, similar to cross-validation. Each tree in the Random Forest model sets 36% of the observations aside to test the performance of the classification model being devel-

oped. Bootstrap aggregation (BAG), or random sampling with replacement from the remaining tip structure observations, is used to compensate for the testing data samples, and to create a training data with the equivalent size as the original data set. After the Random Forest model is trained, the respective testing data (the "out of BAG" observations) for each tree is used as the input for the model, and to calculate OOB errors, which provide collective misclassification error rates that help assess how well the different DI phases have been classified.

To create each branch of an individual decision tree in the Random Forest model during the training process, a random sample of the tip structure feature variables are selected, and the most important one of those is used to make the decision at that branch [6, 48]. This process is repeated until each branch leads to a terminal node where one of the classes of the response variable, the DI phase classes, is selected. Since each decision trees to reaches this point, the Random Forest model classification does not require pruning, and thus avoids further estimates based on a probabilistic outcome. The tree training step is repeated until the desired number of trees are created. When the entire "forest" is used, an input observation is put through each decision tree separately, and a majority vote from all the predicted classes is used to decide the resulting class from the entire model [6, 48].

Another benefit of the Random Forest model is that it offers information on variable importance. During the training process, the mean decrease in the Gini index compares the improvement to the data dispersion before and after that variable is used, at each branch where that variable is used throughout the forest of decision trees [6, 48]. Those variables that are attributed a larger mean decrease in Gini index values are considered to be more important, since they do a better job at reducing the dispersion of the data, which indicates a better separation between points associated with different classes.

### 5.3.2 Predictive Modeling with Random Forest

The data derived from the 10-hour long 2-PF MT model simulations are used as the training data for developing the Random Forest model. In this study, the 2-PF MT tip feature data is the input, and associate DI phase classes are the desired output. The Random Forest model parameters that can be adjusted include the number of trees in the forest, and $mtry$, the number of variables sampled when constructing each branch of a decision tree. For training the predictive models, 1000 tree and $mtry = 8$ were used.

The predictive models developed in this study fall into two categories. The first is for predicting DI phases from tip structure features, which tests the relationship between the micro-level information on the MT structure and the macro-level polymer length behavior. The second is for forecasting upcoming phase transitions by only observing tip structures, which tests the ability to sense changes in macro-level dynamics by only observing a short period of tip structures. In order to properly test each Random Forest model, the tip structures information from a new 1 hour long simulation of the 2-PF MT model using $12\mu M$ tubulin concentration levels will be used as a testing data set. The resulting classes will be compared to DI phases identified in this new data set using the same cluster centers for the DI phases identified in the training data from the 10 hour long simulations.

Another issue that needed to be addressed was the large difference in observations for each class. For example, in the tip-to-phase predictive models, many more observations come from growth phases, because their typical time duration is so long compared to stutter or shortening phases. A second example, in the phase transition forecasting models, far fewer observations come from the transition range at the end of each phase, which are only 5, 10, or 20 observations out of the 75+ data points available per phase period. For this reason, observations are randomly sampled from larger portions of the data such that the number of data points associated to each

class in the training data is roughly the same size. Since this does not use the entire data available, the training results were repeated for the other fractions of the large classes, and confirmed that the resulting misclassification rates and OOB errors were in agreement. In fact, no issues were detected in any of the model training results, so for simplicity, the results presented in the remainder of this chapter only showcase the outcomes for one sub-sample that even out the number of observations per class for each model case.

## 5.4 Prediction and Forecasting Model Results

In this section, the results are presented for the different predictive models. The OOB samples from the data test the performance of how each model classifies the different DI phases, and lists the results into confusion matrices, which reports the number of actual observations that were predicted into each of the possible DI phase classes. The OOB errors and the misclassification rates are calculated from the corresponding confusion matrices, and help determine the success of any one model. The confusion tables and the misclassification rates are particularly helpful in segregating the success for the different classes separately.

### 5.4.1 Training Tip-to-Phase Predictive Models

Results indicating the relationship between MT tip structure and their corresponding DI phase is listed here. For these data sets, the growth phase observations were about three times more than observations from stutter or shortening phases. Random Forest has a tendency to create bias towards classes that have a much larger number of observations, so in practice its best to train these predictive models such that the training data has similar sizes for each class. Since the number of observations in stutters and shortening were relatively close, random samples were taken only from the growth phases, such that the tip-to-phase predictive models were trained

using at least 635,000 observations per phase.

The confusion matrix when the raw data from the 10 hour long 2-PF MT model simulations was used to train the model is presented in Table 5.1(a). For this predictive model, the raw data was doing a fine job of predicting shortening phases, where about 90% of the tip structures that came from shortening phases were classified correctly. The incorrectly classified tip structures were mostly predicted as stutters, indicating a clear separation of 2-PF MT tip characteristics for shortening and growth phases. The misclassification rates for stutters and growth were 32.75% and 30.98% respectively. Having misclassification rates less than 40% indicates that there exists some separation in the data between those two phases [6, 48]. However, the near 30% rates are not as satisfactory as hoped, and does imply that there the 2-PF MT tips that come from stutters and growth share a great deal of structural properties. The stutters sharing some features with the other two phases was anticipated, since the DI phase analysis revealed the transitional role of stutter phases. The Random Forest model results here shows that stutters are more similar to growth than shortening phases.

In an attempt to improve the prediction error rates, the data set was restructured to use different variations of the simulated data from the 2-PF MT model. First, trailing moving average data was created by averaging each of the calculated tip features over a sequence of observations. This newly creating data can be thought of as "remembering" the last several tip structures. The moving average window sizes that were considered were for 11 and 21 observations. The corresponding DI phase labels were unchanged. Tables 5.1(b) and (c) display the confusion matrices for the trailing moving average data using a window size of 11 and 21 respectively. It is clear that using a wider averaging window size improves the prediction results for the stutters and growth phases, and make continued improvements to predicting

TABLE 5.1

CONFUSION MATRICES FOR TIP-TO-PHASE PREDICTIONS

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Shortening | Stutter | Growth | Rate |
| (a) | *Actual* | Shortening | 572949 | 64824 | 2826 | 10.56% |
| | | Stutter | 48621 | 508289 | 198885 | 32.75% |
| | | Growth | 3681 | 194336 | 441127 | 30.98% |
| | | | *Predicted* | | | Misclass. |
| | | | Shortening | Stutter | Growth | Rate |
| (b) | *Actual* | Shortening | 590881 | 46450 | 3268 | 7.76% |
| | | Stutter | 34264 | 578310 | 143216 | 23.48% |
| | | Growth | 3484 | 161435 | 474225 | 25.80% |
| | | | *Predicted* | | | Misclass. |
| | | | Shortening | Stutter | Growth | Rate |
| (c) | *Actual* | Shortening | 612680 | 25458 | 2461 | 4.36% |
| | | Stutter | 22174 | 651662 | 81949 | 13.78% |
| | | Growth | 3452 | 107412 | 528280 | 17.35% |
| | | | *Predicted* | | | Misclass. |
| | | | Shortening | Stutter | Growth | Rate |
| (d) | *Actual* | Shortening | 603355 | 35473 | 1771 | 5.81% |
| | | Stutter | 26009 | 627953 | 101814 | 16.91% |
| | | Growth | 1751 | 125592 | 511801 | 19.92% |

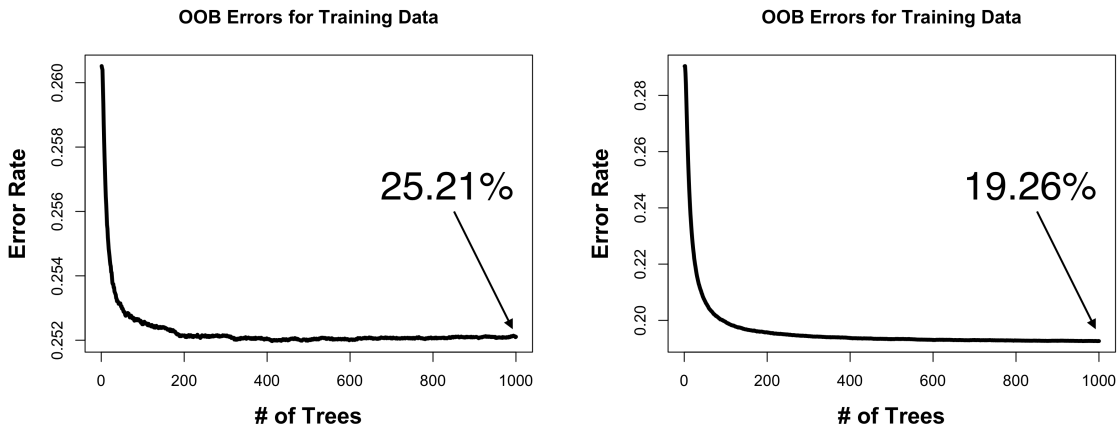
These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

shortening phases. In fact, using a window size of 21 observations has misclassification rates well below 20%. The results of these predictive models indicate that increasing the temporal range of which tip structures are observed may be beneficial to detecting the relationship between 2-PF MT tip structures and macro-level DI phases.

The final model trained for tip-to-phase predictions makes use of the multi-resolution data that is comprised of all the previous three data sets together: the raw data, and both 11 and 21 observations window sizes of the trailing moving average data sets. By combing the columns of these data sets, 72 tip structure features are considered simultaneously as the predictor variables. This requires no additional changes for utilizing the Random Forest methodology, which is another benefit offered by this approach. The results showing in Table 5.1(d) are close to, but not better than, the 21 observation window size trailing average data case.

Comparing the different tip-to-phase prediction models to each other, the first and most obvious result is the OOB estimate error rates. These overall error rates improved along with the misclassification rates as larger average window sizes were used. Also, the multi-resolution OOB error rate was good but not better than using just the trailing moving average data with a window size of 21 observations. Figure 5.25 plots the OOB errors as more trees are added into each predictive model, and demonstrates a convergence of error rates well before the number of trees reaches 1000. It is also worth noting that the models sets using moving average data have a smoother convergence compared to the raw data model.

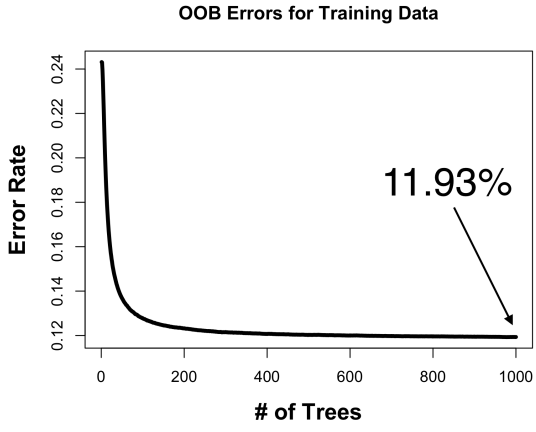
Another comparison amongst the tip-to-phase predictive models is by variable importance. Figure 5.26 displays the mean decrease in Gini index values for each tip feature variable. The primary list of important variables, consisting of the five most important variables shared by the predictive models using training data from raw and trailing moving average tip features, are a follows:



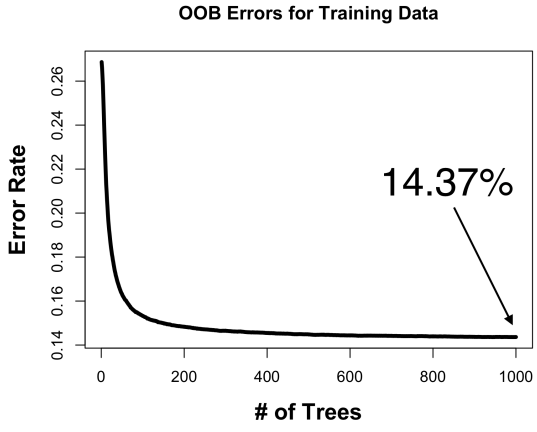

(a)




(b)




(c)




(d)

Figure 5.25. OOB errors as trees are added to the Random Forest model for predictive models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

- The ratio of the estimated GTP-cap size to the average PF Tip Lengths (Variable #16)
- Rate of Hydrolysis (Variable #21)
- The estimate for how far below the GTP-cap is from the crack depth (Variable #15)
- Total number of GTP-bound subunits in the entire MT (Variable #4)
- The estimated GTP-cap size (Variable #14)

These variables have corresponding Gini index values that are clearly stand out from the remaining variables. These all relate to the GTP-bound subunit content available in the MT structure, which is used to estimate the size of the GTP-cap size. In fact, Variable #21, #4, and #14 are close to being scalar multiples of each other. The next five important variables, which are more closely related to the features of the tip region, the secondary list of important variables, are the following:

- Expected Rate of Subunit Loss as a function of Tip Configuration (Variable #22)
- Percentage of the MT gated tip comprised of GTP-bound subunits (Variable #9)
- Dispersion of GTP-bound subunits in the gated MT tip (Variable #19)
- Number of G-subunits that are GTP-bound (Variable #11)
- Percentage of all the GTP-bound subunits located in the gated MT tip region (Variable #6)

Compared to the GTP-cap estimates, this secondary list of variables has smaller mean decrease of Gini index values. Although, in the trailing moving average data cases, the importance of the primary list is less drastic in the calculated importance.

Also, when comparing all of the tip feature variables together in the multi-resolution data set, the primary list of variables from the trailing moving average

data sets dominate the top ten mean decrease in Gini index values. It is not until the $11^{th}$ and $12^{th}$ positions that the raw data variables contribute to the important variables list, and those again are from the primary variable list.

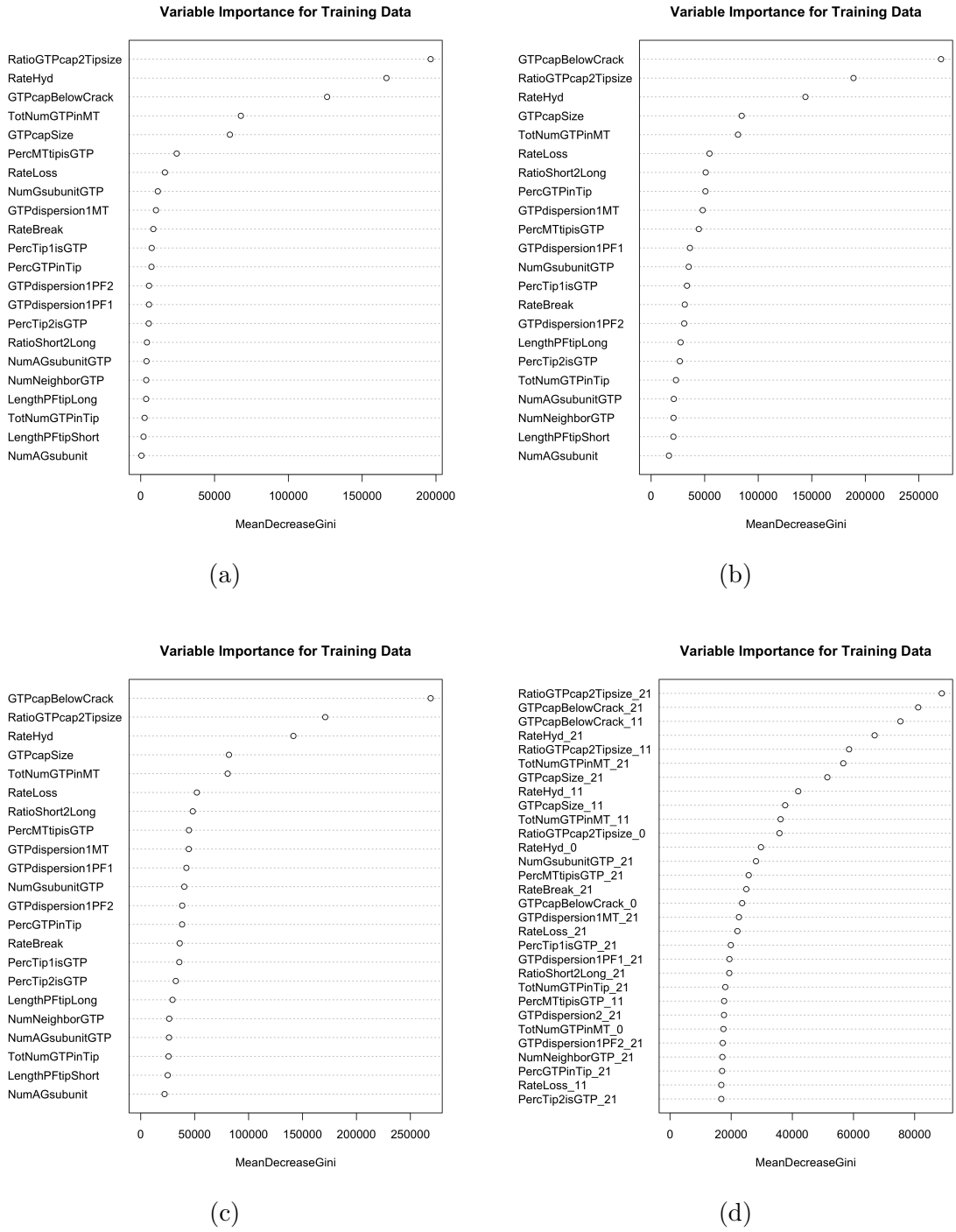



Figure 5.26. Variable importance via the mean decrease in the Gini index for each tip feature for predictive models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

### 5.4.2 Training Phase Transition Forecasting Models

The forecasting models developed here refer to attempting to predict future dynamics. In particular, these models test the ability for the MT tip structures to detect those configurations associated with specific phase transitions. This is done by creating a new set of class labels, such that the majority of the observations in a phase keep the same phase label, however a small set of observations in a pre-transition range are labeled with the phase they are in, and the phase they are headed to. Several pre-transition ranges sizes are used to compare how quickly structural changes can manifest into phase transitions. For example, if a stutter phase followed by a shortening phase is considered, all but the last $N$-many observations would keep their "Stutter" label, and the last $N$-many observations in the stutter phase would be labeled as "Stutter-to-Shortening". By doing so, the tip features are used as the predictor variables, and these new class labels are the response variables.

These forecasting models are repeated for $N = 5,\ 10,$ and 20 observations, and trained using the same data sets from the tip-to-phase predictive models (raw data, trailing moving average over 11 observations, trailing moving average over 21 observations, and multi-resolution data). This creates 12 models that forecast if the MT will continue in a specific phase, or if will transition into one of the other two DI phases. Again, the Random Forest approach faces a challenging scenario where the main phase classes and the pre-transition classes all have different sample sizes. For this reason, the smallest class is used to determine the number of samples to draw from each class when training the forecasting model.

#### 5.4.2.1 Forecasting Transitions out of Shortening Phases

Tables 5.2, 5.3, and 5.4 display the confusion tables when training forecasting models for shortening phases using a pre-transition range 5, 10, and 20 observations respectively. The most successful misclassification rate is attributed to "Shortening-

to-Stutter"classes when using a pre-transition range of 5 observations, and the 21 observation trailing average data set to train the model. "Shortening-to-Growth" classes when trained using moving average data are also worth mentioning, since these were the only other cases to have mislassification rates below 30%.

Figures 5.27, 5.28, and 5.29 display the corresponding OOB error plots as more trees are added into the Random Forest model. These plots display a bit of instability in converging OOB error values, and the most stable results come from the 21 observation trailing moving average data. Though the 5 observation long pre-transition region delivers the lowest OOB errors when using 1000 trees, using 10 and 20 observation long pre-transition regions also deliver satisfactory, and somewhat smooth convergence results. The multi-resolution data surprisingly delivers poor results, seeming oblivious to the advantages of the moving average data it has at is disposal.

When considered the variable importance plots in Figures 5.30, 5.31, and 5.32, the expected tip features involving GTP-cap estimates are at the top of the list. Additionally, in forecasting pre-transition shortening regions, the PF tip lengths and the ratio between them appear as important gated tip features that were not as prevalent in predictive DI phases, especially with the moving average data. The models trained with multi-resolution data were peculiar, in that they focused on mostly raw data information, which degraded the forecasting success rates.

TABLE 5.2

CONFUSION MATRICES FOR FORECASTING 5 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM SHORTENING

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (a) | *Actual* | Short. | 457 | 521 | 318 | 64.74% |
| | | Short. to Gr. | 262 | 703 | 328 | 45.63% |
| | | Short. to St. | 234 | 604 | 458 | 64.66% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (b) | *Actual* | Short. | 652 | 364 | 280 | 49.69% |
| | | Short. to Gr. | 252 | 778 | 263 | 39.82% |
| | | Short. to St. | 183 | 209 | 904 | 30.27% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (c) | *Actual* | Short. | 772 | 308 | 216 | 40.43% |
| | | Short. to Gr. | 172 | 1014 | 107 | 21.58% |
| | | Short. to St. | 85 | 87 | 1124 | 13.27% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (d) | *Actual* | Short. | 568 | 387 | 341 | 56.17% |
| | | Short. to Gr. | 259 | 627 | 407 | 51.51% |
| | | Short. to St. | 270 | 513 | 513 | 60.42% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

TABLE 5.3

CONFUSION MATRICES FOR FORECASTING 10 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM SHORTENING

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (a) | *Actual* | Short. | 736 | 873 | 797 | 69.41% |
| | | Short. to Gr. | 457 | 1118 | 825 | 53.42% |
| | | Short. to St. | 427 | 904 | 1045 | 56.02% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (b) | *Actual* | Short. | 1204 | 646 | 556 | 49.96% |
| | | Short. to Gr. | 471 | 1474 | 455 | 38.58% |
| | | Short. to St. | 341 | 353 | 1682 | 29.21% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (c) | *Actual* | Short. | 1384 | 580 | 442 | 42.48% |
| | | Short. to Gr. | 341 | 1892 | 167 | 21.17% |
| | | Short. to St. | 191 | 155 | 2030 | 14.56% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (d) | *Actual* | Short. | 832 | 861 | 713 | 65.42% |
| | | Short. to Gr. | 560 | 1061 | 779 | 55.79% |
| | | Short. to St. | 468 | 871 | 1037 | 56.36% |

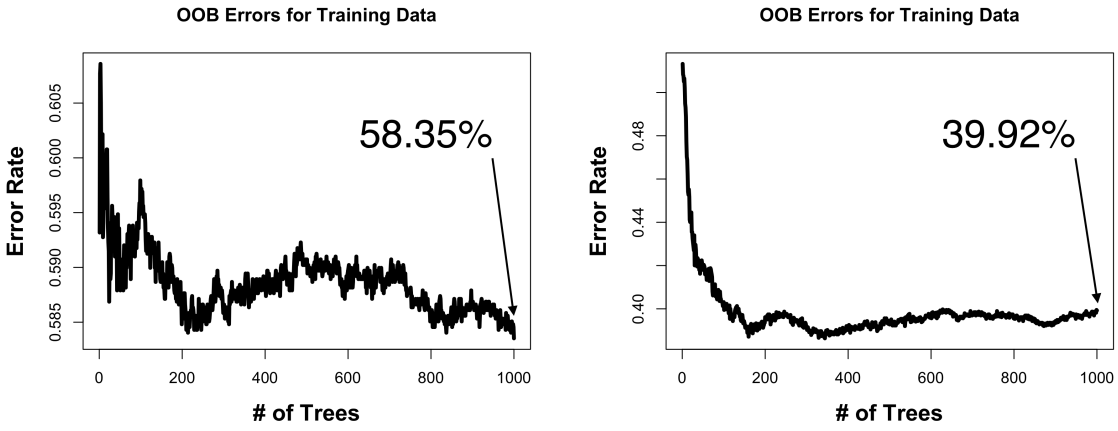
These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

TABLE 5.4

CONFUSION MATRICES FOR FORECASTING 20 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM SHORTENING

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (a) | *Actual* | Short. | 1233 | 1469 | 1851 | 72.92% |
| | | Short. to Gr. | 759 | 1828 | 1966 | 59.85% |
| | | Short. to St. | 706 | 1419 | 2411 | 46.85% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (b) | *Actual* | Short. | 2120 | 1322 | 1111 | 53.44% |
| | | Short. to Gr. | 972 | 2660 | 921 | 41.58% |
| | | Short. to St. | 707 | 747 | 3082 | 32.05% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (c) | *Actual* | Short. | 2723 | 1086 | 744 | 40.19% |
| | | Short. to Gr. | 676 | 3520 | 357 | 22.69% |
| | | Short. to St. | 414 | 283 | 3839 | 15.37% |
| | | | *Predicted* | | | Misclass. |
| | | | Short. | Short. to Gr. | Short. to St. | Rate |
| (d) | *Actual* | Short. | 1341 | 1543 | 1669 | 70.55% |
| | | Short. to Gr. | 853 | 1866 | 1834 | 59.02% |
| | | Short. to St. | 727 | 1601 | 2208 | 51.32% |

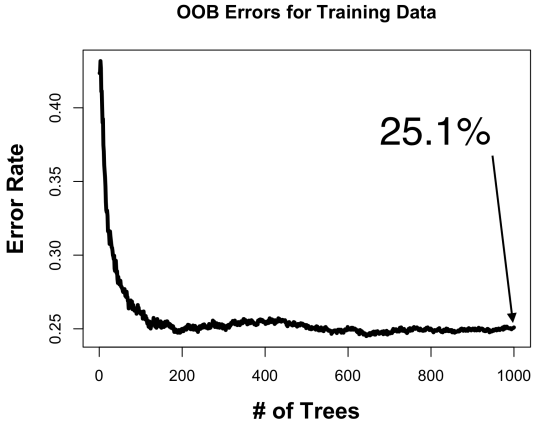
These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.



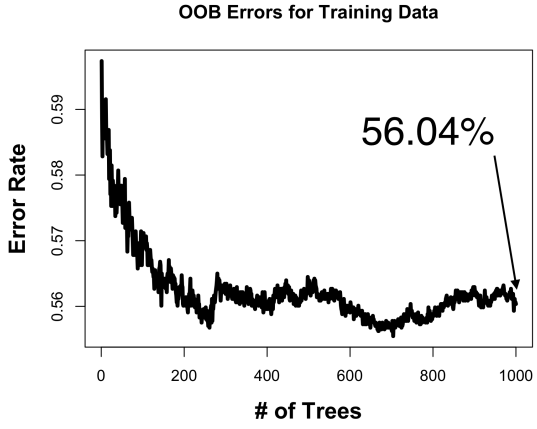

(a)




(b)




(c)




(d)

Figure 5.27. OOB errors as trees are added to the Random Forest model for forecasting 5 observation long regions of pre-transition shortening, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

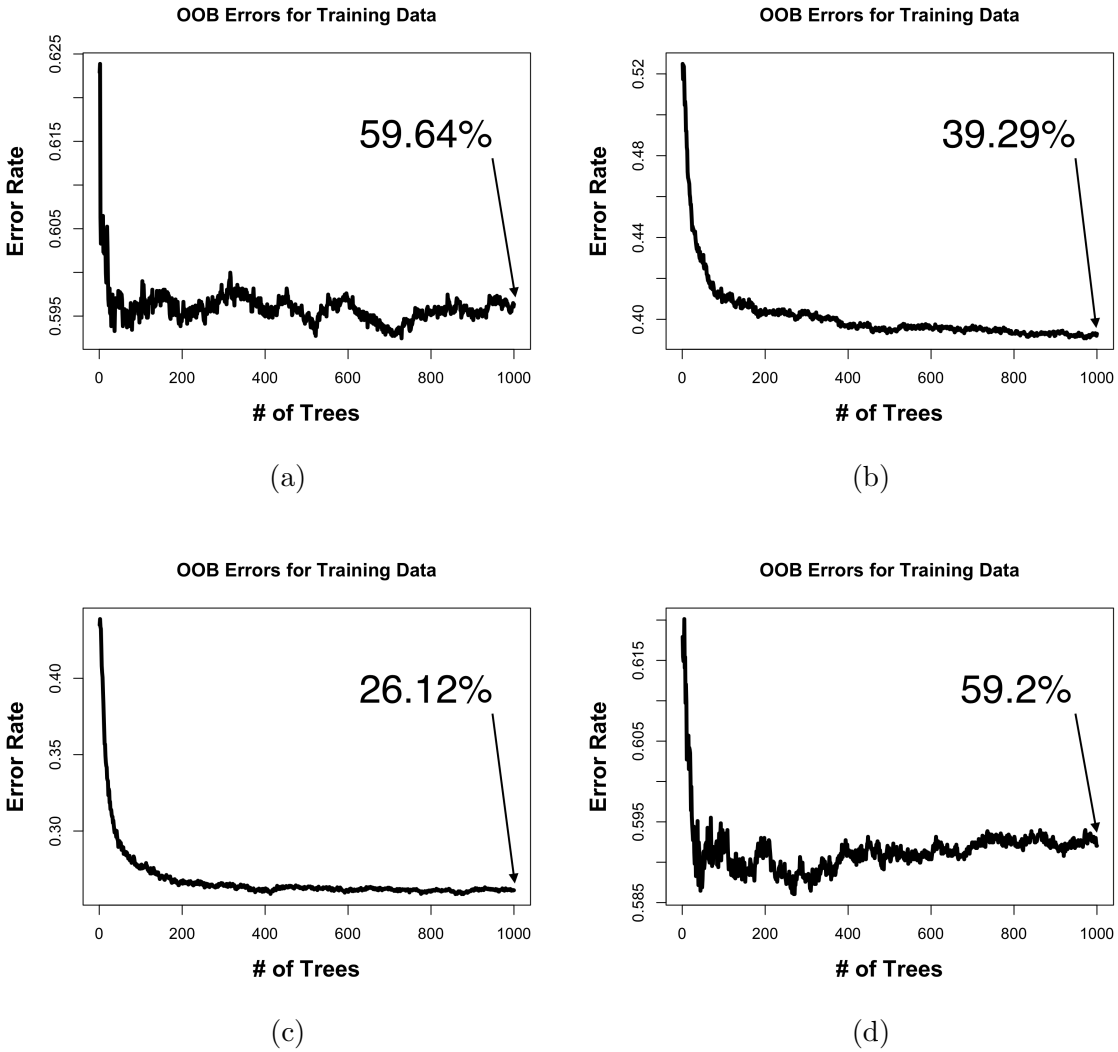
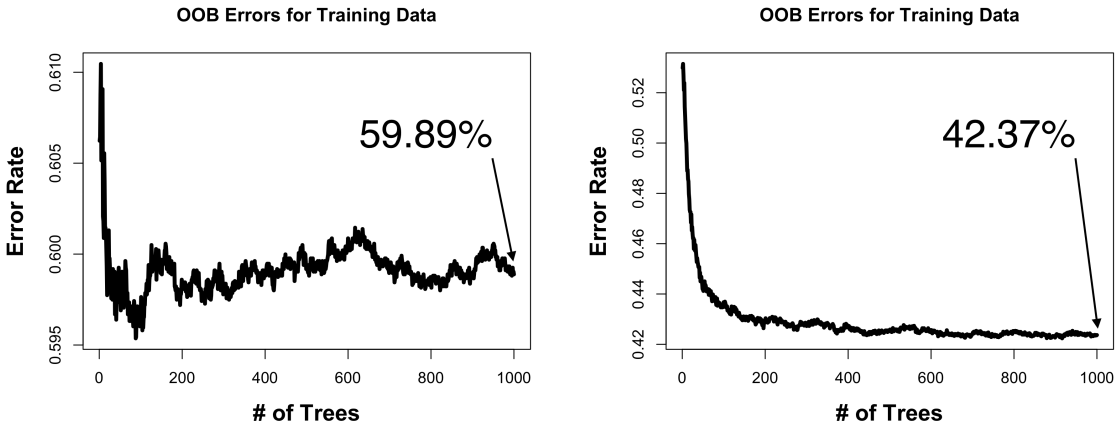
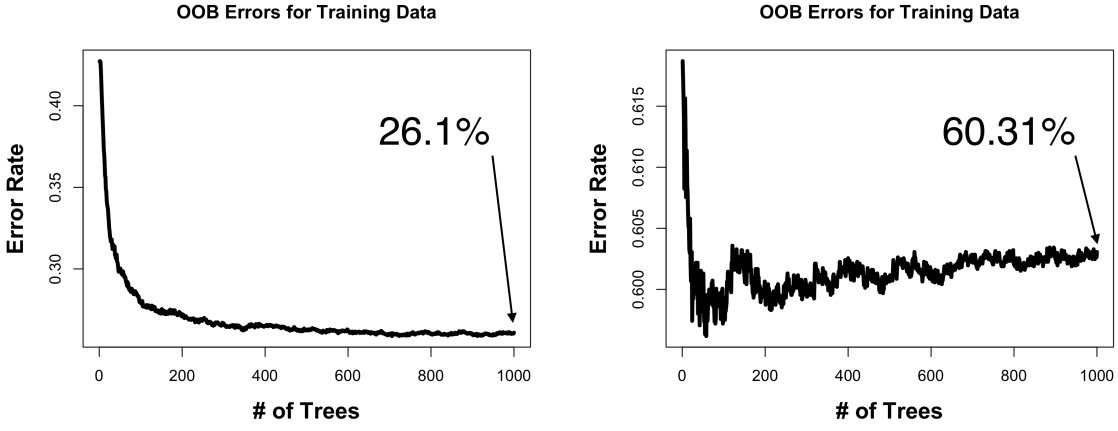



Figure 5.28. OOB errors as trees are added to the Random Forest model for forecasting 10 observation long regions of pre-transition shortening, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.




(a) (b)




(c) (d)

Figure 5.29. OOB errors as trees are added to the Random Forest model for forecasting 20 observation long regions of pre-transition shortening, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.


Figure 5.30. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 5 observation long regions before transitioning out of shortening, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.


Figure 5.31. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 10 observation long regions before transitioning out of shortening, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

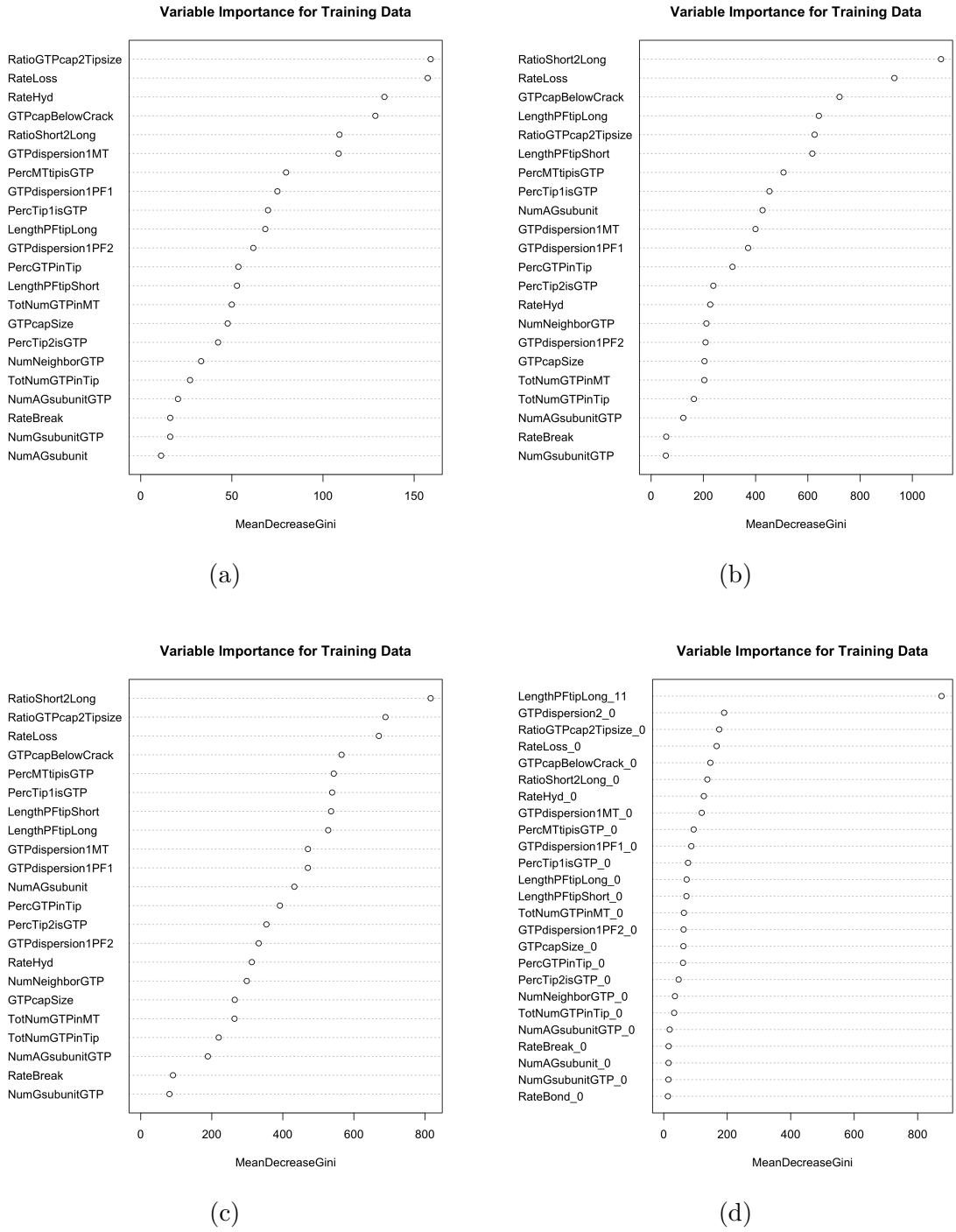



Figure 5.32. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 20 observation long regions before transitioning out of shortening, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

#### 5.4.2.2 Forecasting Transitions out of Stutter Phases

Tables 5.5, 5.6, and 5.7 display the confusion tables when training forecasting models for stutter phases using a pre-transition range 5, 10, and 20 observations respectively. Forecasting transitions out of stutter phases was the most successful out of the three DI phases. The most successful misclassification rates are attributed to "Stutter-to-Growth" classes when using a pre-transition range of 5 observations, and the 21 observation trailing average data set to train the model, this phase transition had misclassification rates less than 10%. "Stutter-to-Shortening" classes also had good misclassification rates, especially when using the 21 observation trailing average data set to train the model.

Figures 5.33, 5.34, and 5.35 display the corresponding OOB error plots as more trees are added into the Random Forest model. These plots also display a bit of instability in converging OOB error values, though much less than the other two DI phases. Though the 20 observation long pre-transition region delivers the lowest OOB errors when using 1000 trees, using 5 and 10 observation long pre-transition regions also deliver satisfactory results with less than 20% error rates. The models trained with 11 observation long average data also deliver good results, with error rates near the 20% range.

When considered the variable importance plots in Figures 5.36, 5.37, and 5.38, the tip features involving GTP-cap estimates dominate top of the list. The ratio between PF tip lengths is present again, but the gated-tip features are far less present here than when compared to the forecasting phase transitions out of shortening case. Once again, the models trained with multi-resolution data was caught up focusing on raw data information, which did not help the success rates.

TABLE 5.5

CONFUSION MATRICES FOR FORECASTING 5 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM STUTTERS

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (a) | *Actual* | Stut. | 975 | 309 | 262 | 36.93% |
| | | Stut. to Gr. | 186 | 1107 | 243 | 27.93% |
| | | Stut. to Sh. | 241 | 285 | 955 | 35.52% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (b) | *Actual* | Stut. | 1100 | 209 | 238 | 28.89% |
| | | Stut. to Gr. | 32 | 1425 | 79 | 7.23% |
| | | Stut. to Sh. | 192 | 179 | 1111 | 25.03% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (c) | *Actual* | Stut. | 1139 | 164 | 245 | 26.42% |
| | | Stut. to Gr. | 12 | 1499 | 25 | 2.41% |
| | | Stut. to Sh. | 168 | 122 | 1190 | 19.59% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (d) | *Actual* | Stut. | 1025 | 291 | 232 | 33.79% |
| | | Stut. to Gr. | 187 | 1153 | 196 | 24.93% |
| | | Stut. to Sh. | 229 | 222 | 1026 | 30.53% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

TABLE 5.6

CONFUSION MATRICES FOR FORECASTING 10 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM STUTTERS

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (a) | *Actual* | Stut. | 1749 | 504 | 536 | 37.29% |
| | | Stut. to Gr. | 314 | 1935 | 567 | 31.29% |
| | | Stut. to Sh. | 512 | 688 | 1578 | 43.20% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (b) | *Actual* | Stut. | 2010 | 357 | 422 | 27.93% |
| | | Stut. to Gr. | 67 | 2593 | 156 | 7.92% |
| | | Stut. to Sh. | 410 | 339 | 2029 | 26.97% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (c) | *Actual* | Stut. | 2092 | 263 | 434 | 24.99% |
| | | Stut. to Gr. | 21 | 2734 | 61 | 2.91% |
| | | Stut. to Sh. | 305 | 185 | 2287 | 17.64% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (d) | *Actual* | Stut. | 1818 | 474 | 497 | 34.82% |
| | | Stut. to Gr. | 308 | 2017 | 491 | 28.37% |
| | | Stut. to Sh. | 463 | 599 | 1714 | 38.26% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

TABLE 5.7

CONFUSION MATRICES FOR FORECASTING 20 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM STUTTERS

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (a) | *Actual* | Stut. | 3436 | 962 | 1004 | 36.39% |
| | | Stut. to Gr. | 618 | 3620 | 1138 | 32.66% |
| | | Stut. to Sh. | 1087 | 1426 | 2803 | 47.27% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (b) | *Actual* | Stut. | 3899 | 610 | 893 | 27.82% |
| | | Stut. to Gr. | 161 | 4834 | 381 | 10.08% |
| | | Stut. to Sh. | 775 | 700 | 3843 | 27.74% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (c) | *Actual* | Stut. | 4109 | 434 | 859 | 23.94% |
| | | Stut. to Gr. | 63 | 5206 | 107 | 3.16% |
| | | Stut. to Sh. | 582 | 323 | 4412 | 17.03% |
| | | | *Predicted* | | | Misclass. |
| | | | Stut. | Stut. to Gr. | Stut. to Sh. | Rate |
| (d) | *Actual* | Stut. | 3447 | 940 | 1015 | 36.19% |
| | | Stut. to Gr. | 651 | 3620 | 1105 | 32.66% |
| | | Stut. to Sh. | 993 | 1391 | 2933 | 44.84% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

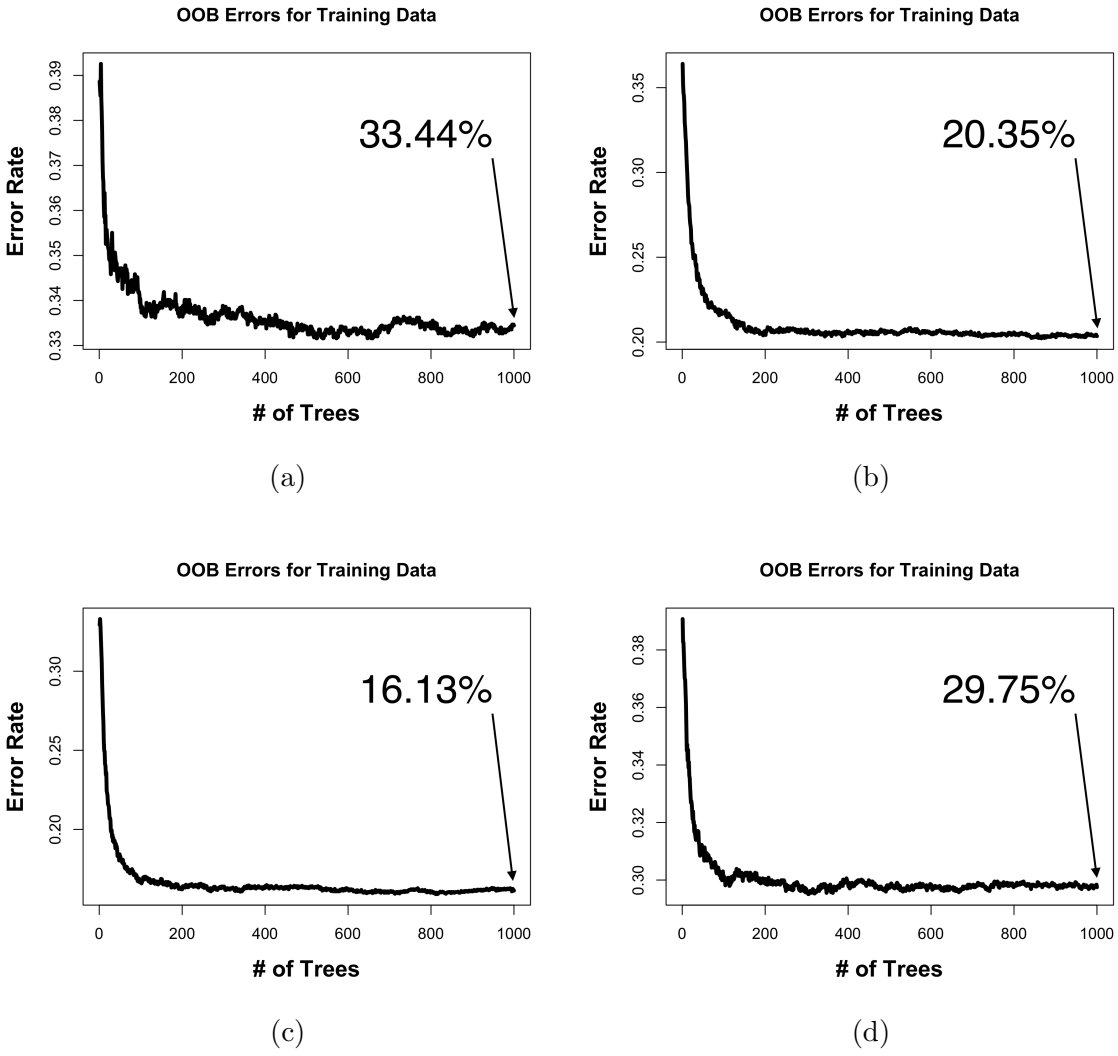
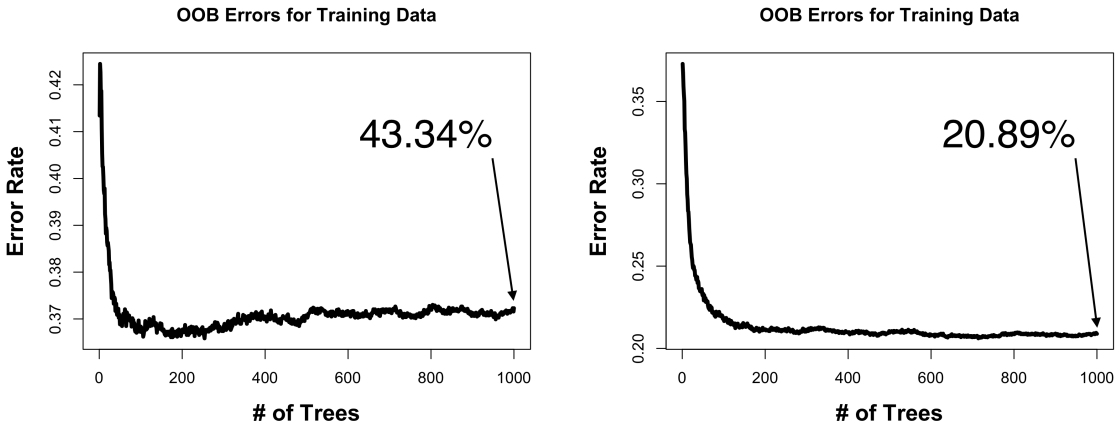


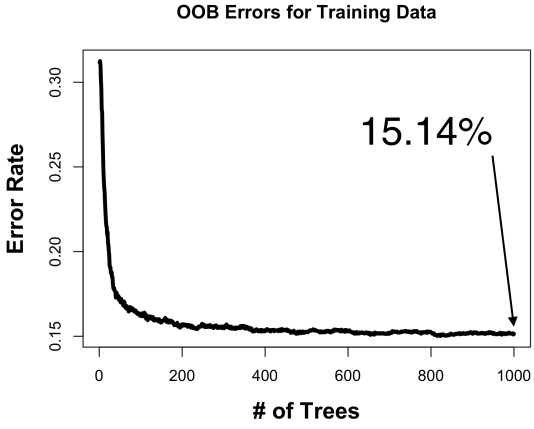

Figure 5.33. OOB errors as trees are added to the Random Forest model for forecasting 5 observation long regions of pre-transition stutters, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.



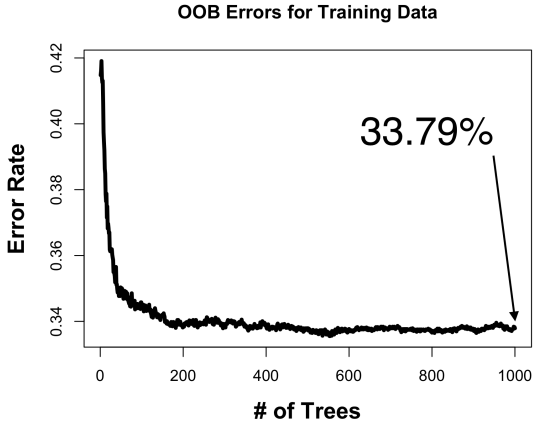

(a)




(b)

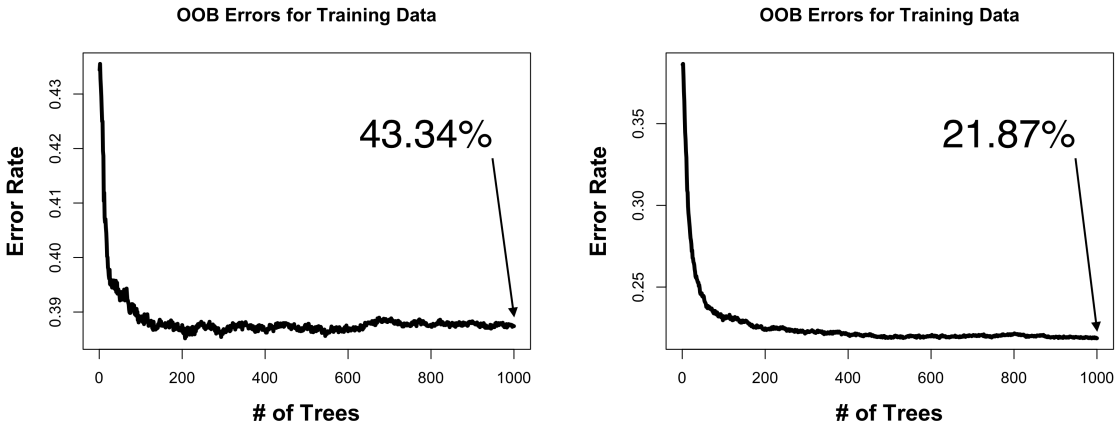



(c)




(d)

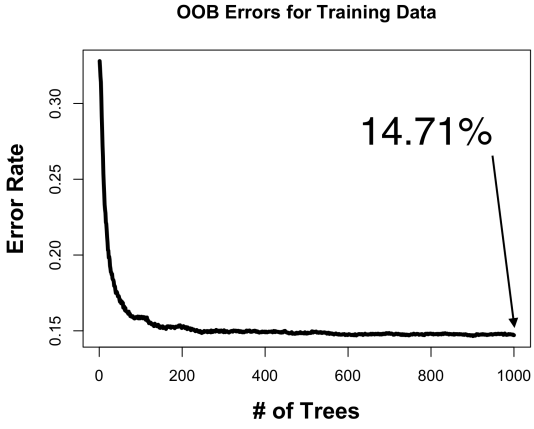
Figure 5.34. OOB errors as trees are added to the Random Forest model for forecasting 10 observation long regions of pre-transition stutters, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.



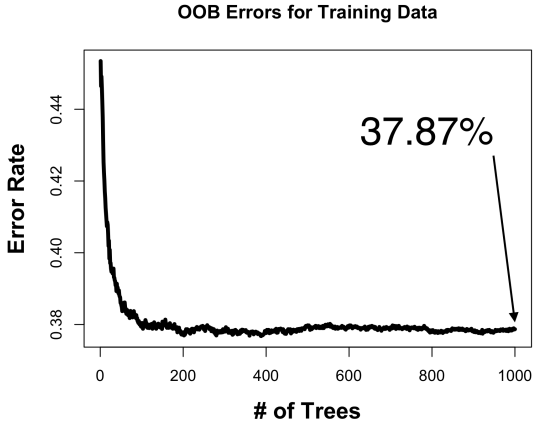

(a)




(b)




(c)




(d)

Figure 5.35. OOB errors as trees are added to the Random Forest model for forecasting 20 observation long regions of pre-transition stutters, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

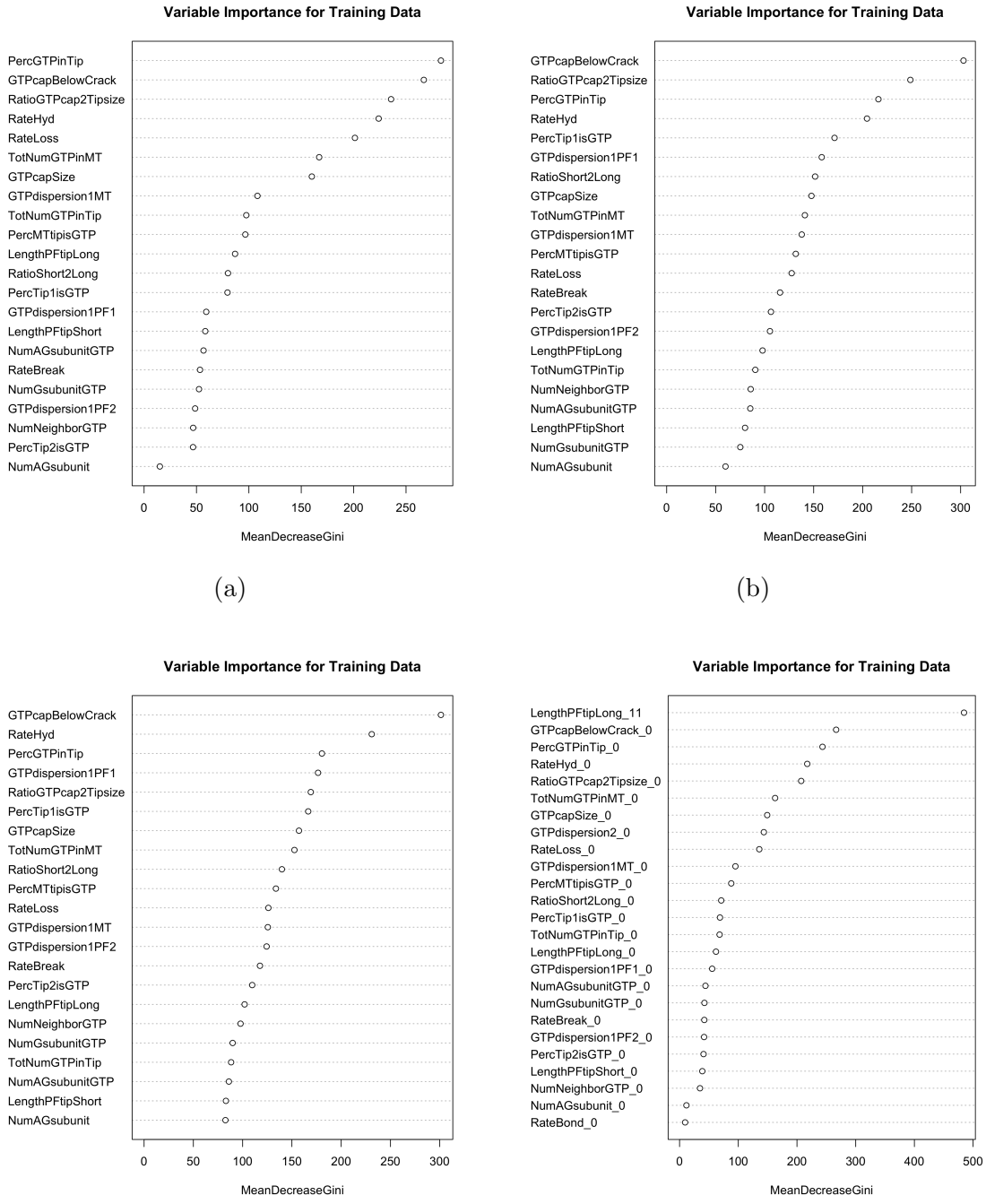



Figure 5.36. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 5 observation long regions before transitioning out of stutters, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from 10 hour 2-PF MT model simulations.

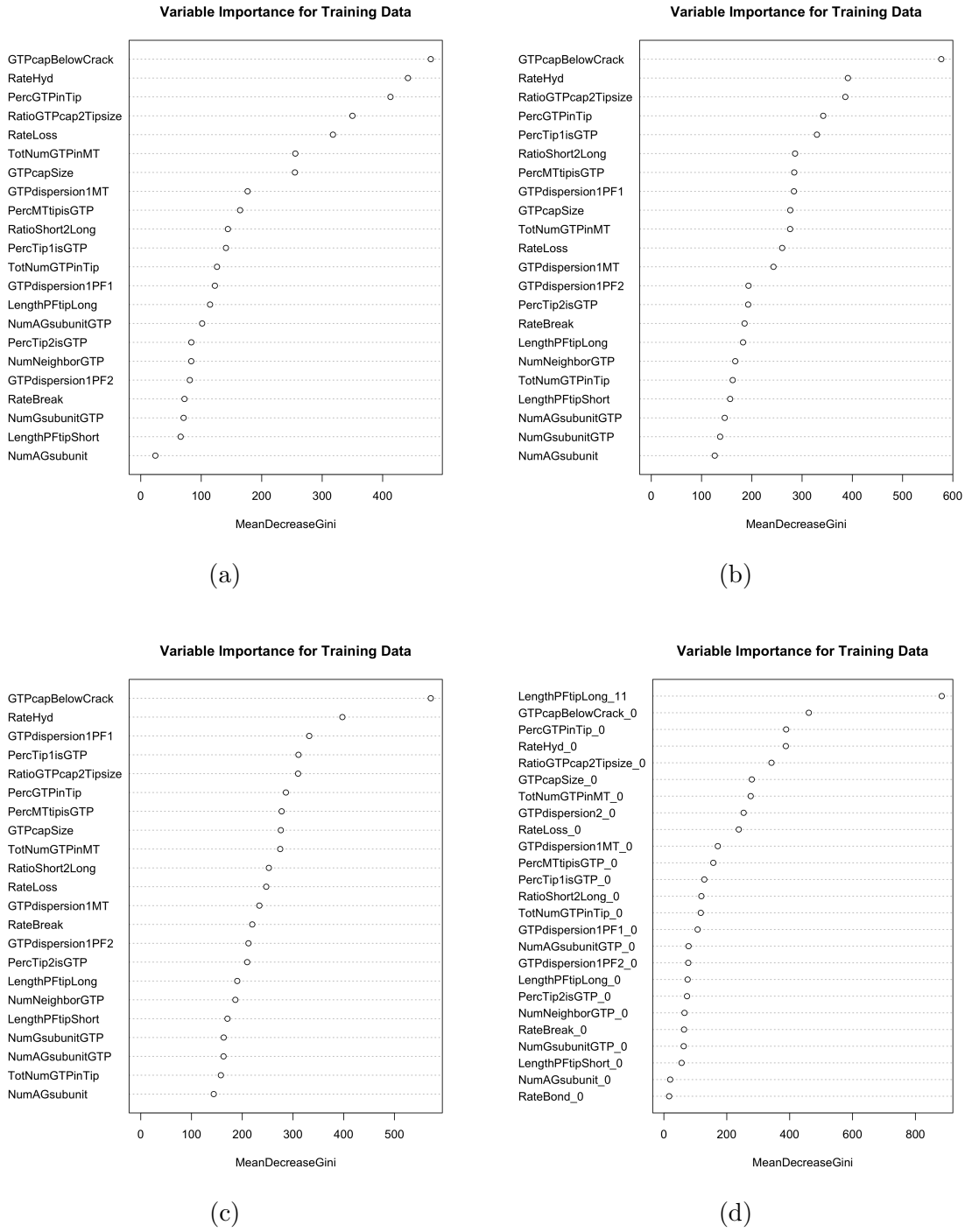



Figure 5.37. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 10 observation long regions before transitioning out of stutters, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from 10 hour 2-PF MT model simulations.


(a) (b)


(c) (d)

Figure 5.38. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 20 observation long regions before transitioning out of stutters, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from 10 hour 2-PF MT model simulations.

#### 5.4.2.3 Forecasting Transitions out of Growth Phases

Tables 5.8, 5.9, and 5.10 display the confusion tables when training forecasting models for growth phases using a pre-transition range 5, 10, and 20 observations respectively. Forecasting transitions out of growth phases was better than the shortening cases. The most successful misclassification rates are attributed to "Growth-to-Shortening" classes when using a pre-transition range of 10 observations, and the 21 observation trailing average data set to train the model. Even so, all cases trained with averaged data had misclassification rates not much more than 10%. "Growth-to-Stutter" classes also had good results when using the 21 observation trailing average data set to train the model.

Figures 5.39, 5.40, and 5.41 display the corresponding OOB error plots as more trees are added into the Random Forest model. These plots also display a bit of instability in converging OOB error values, much less than the shortening case, but not as stable as the stutters case. The 10 and 20 observation long pre-transition region deliver similarly low OOB errors when using 1000 trees, though using 5 observation long pre-transition regions also delivers satisfactory results with about 16% error rates. The models trained with average data over 11 observations also deliver good results, with error rates near 20%.

When considered the variable importance plots in Figures 5.42, 5.43, and 5.44, the tip features involving GTP-cap estimates dominate top of the list again. The ratio between PF tip lengths and the rate of subunit loss appear to be important for forecasting transitions out of growth phases, and so do some details about the longer PF tip, such as the dispersion of GTP-bound subunits and the percentage of subunits in the PF tip being GTP-bound. Again, the models trained with multi-resolution data did not have satisfactory results.

TABLE 5.8

CONFUSION MATRICES FOR FORECASTING 5 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM GROWTH

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (a) | *Actual* | Grow. | 1010 | 289 | 374 | 39.63% |
| | | Grow. to Sh. | 341 | 1061 | 302 | 37.73% |
| | | Grow. to St. | 496 | 382 | 784 | 52.83% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (b) | *Actual* | Grow. | 1011 | 241 | 421 | 39.57% |
| | | Grow. to Sh. | 87 | 1534 | 83 | 9.98% |
| | | Grow. to St. | 313 | 219 | 1130 | 32.01% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (c) | *Actual* | Grow. | 1075 | 217 | 381 | 35.74% |
| | | Grow. to Sh. | 54 | 1625 | 25 | 4.64% |
| | | Grow. to St. | 263 | 131 | 1268 | 23.71% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (d) | *Actual* | Grow. | 973 | 289 | 411 | 41.84% |
| | | Grow. to Sh. | 281 | 1133 | 290 | 33.51% |
| | | Grow. to St. | 439 | 374 | 849 | 48.92% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

TABLE 5.9

CONFUSION MATRICES FOR FORECASTING 10 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM GROWTH

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (a) | *Actual* | Grow. | 1441 | 816 | 873 | 53.96% |
| | | Grow. to Sh. | 527 | 2080 | 517 | 33.42% |
| | | Grow. to St. | 762 | 784 | 1604 | 49.08% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (b) | *Actual* | Grow. | 1817 | 476 | 837 | 41.95% |
| | | Grow. to Sh. | 169 | 2811 | 144 | 10.02% |
| | | Grow. to St. | 589 | 315 | 2246 | 28.70% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (c) | *Actual* | Grow. | 1924 | 387 | 819 | 38.53% |
| | | Grow. to Sh. | 69 | 3000 | 55 | 3.97% |
| | | Grow. to St. | 447 | 160 | 2543 | 19.27% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (d) | *Actual* | Grow. | 1528 | 726 | 876 | 51.18% |
| | | Grow. to Sh. | 514 | 2070 | 540 | 33.74% |
| | | Grow. to St. | 784 | 728 | 1638 | 48.00% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

TABLE 5.10

CONFUSION MATRICES FOR FORECASTING 20 OBSERVATION LONG REGIONS BEFORE TRANSITIONING FROM GROWTH

| | | | *Predicted* | | | Misclass. |
|---|---|---|---|---|---|---|
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (a) | *Actual* | Grow. | 2549 | 1736 | 1620 | 56.83% |
| | | Grow. to Sh. | 1157 | 3867 | 940 | 35.16% |
| | | Grow. to St. | 1595 | 1549 | 2830 | 52.63% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (b) | *Actual* | Grow. | 3299 | 972 | 1634 | 44.13% |
| | | Grow. to Sh. | 390 | 5302 | 272 | 11.10% |
| | | Grow. to St. | 1116 | 559 | 4299 | 28.04% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (c) | *Actual* | Grow. | 3651 | 744 | 1510 | 38.17% |
| | | Grow. to Sh. | 193 | 5680 | 91 | 4.76% |
| | | Grow. to St. | 800 | 300 | 4874 | 18.41% |
| | | | *Predicted* | | | Misclass. |
| | | | Grow. | Grow. to Sh. | Grow. to St. | Rate |
| (d) | *Actual* | Grow. | 2611 | 1575 | 1719 | 55.78% |
| | | Grow. to Sh. | 1121 | 3835 | 1008 | 35.70% |
| | | Grow. to St. | 1553 | 1493 | 2928 | 50.99% |

These results were obtained from models trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

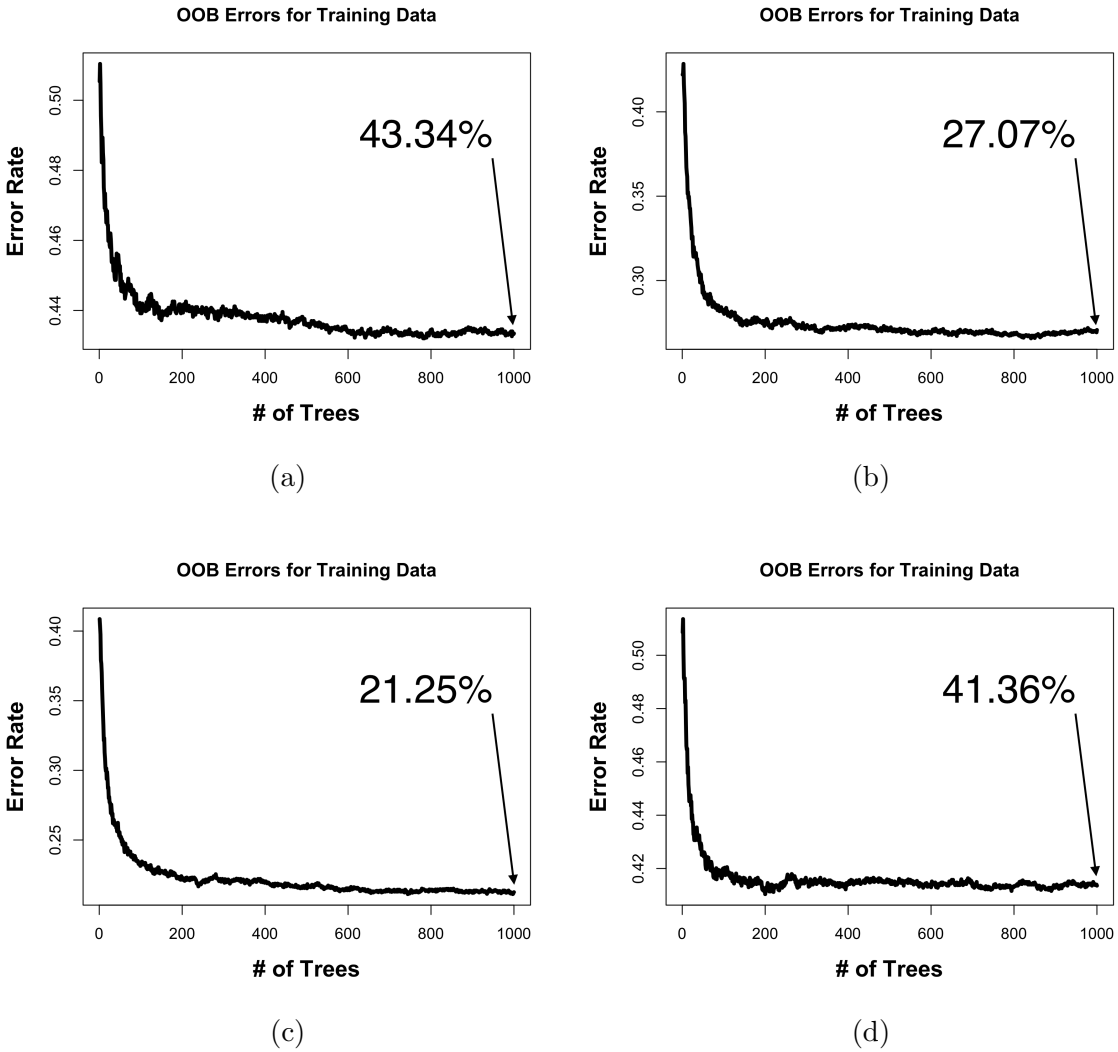



Figure 5.39. OOB errors as trees are added to the Random Forest model for forecasting 5 observation long regions of pre-transition growth, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

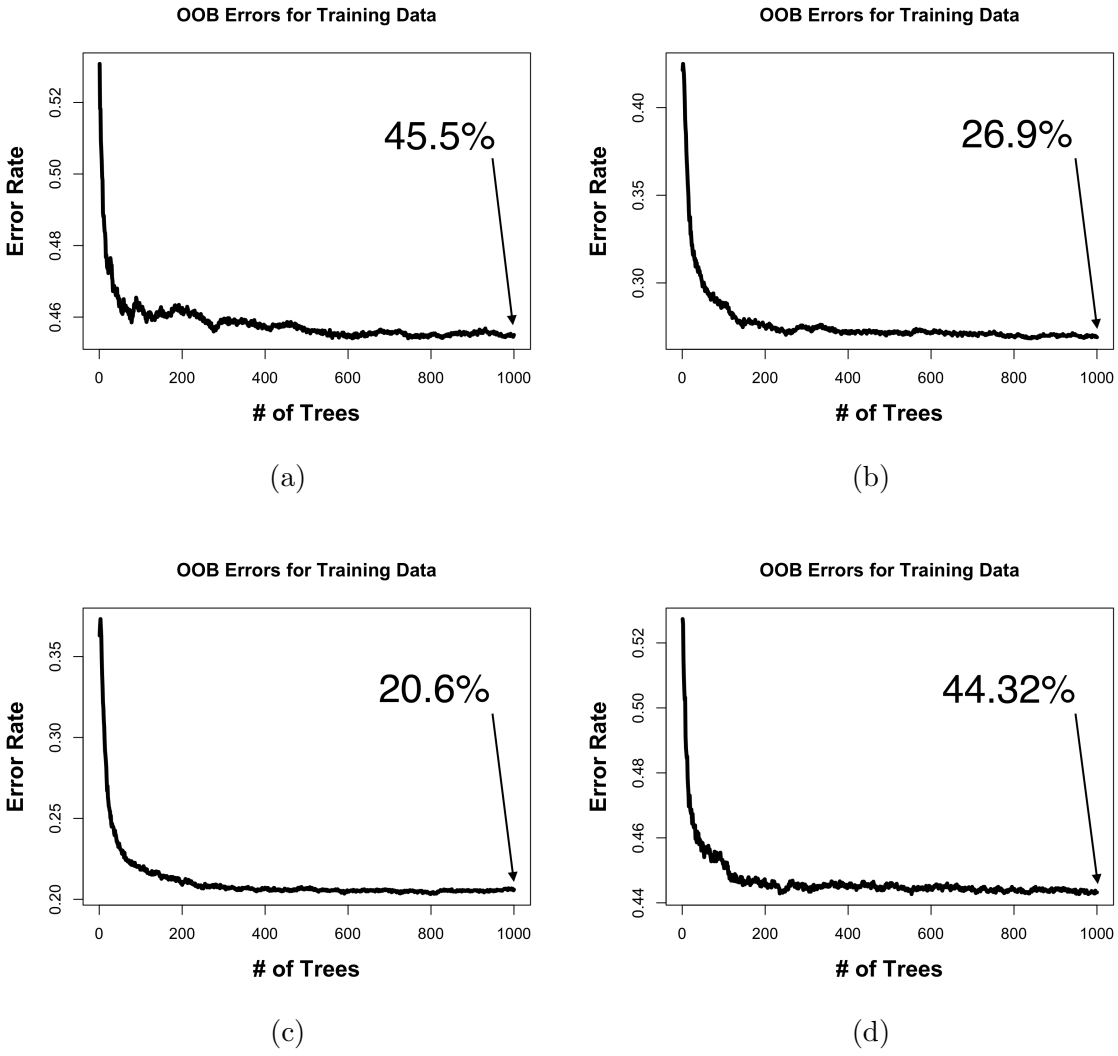



Figure 5.40. OOB errors as trees are added to the Random Forest model for forecasting 10 observation long regions of pre-transition growth, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

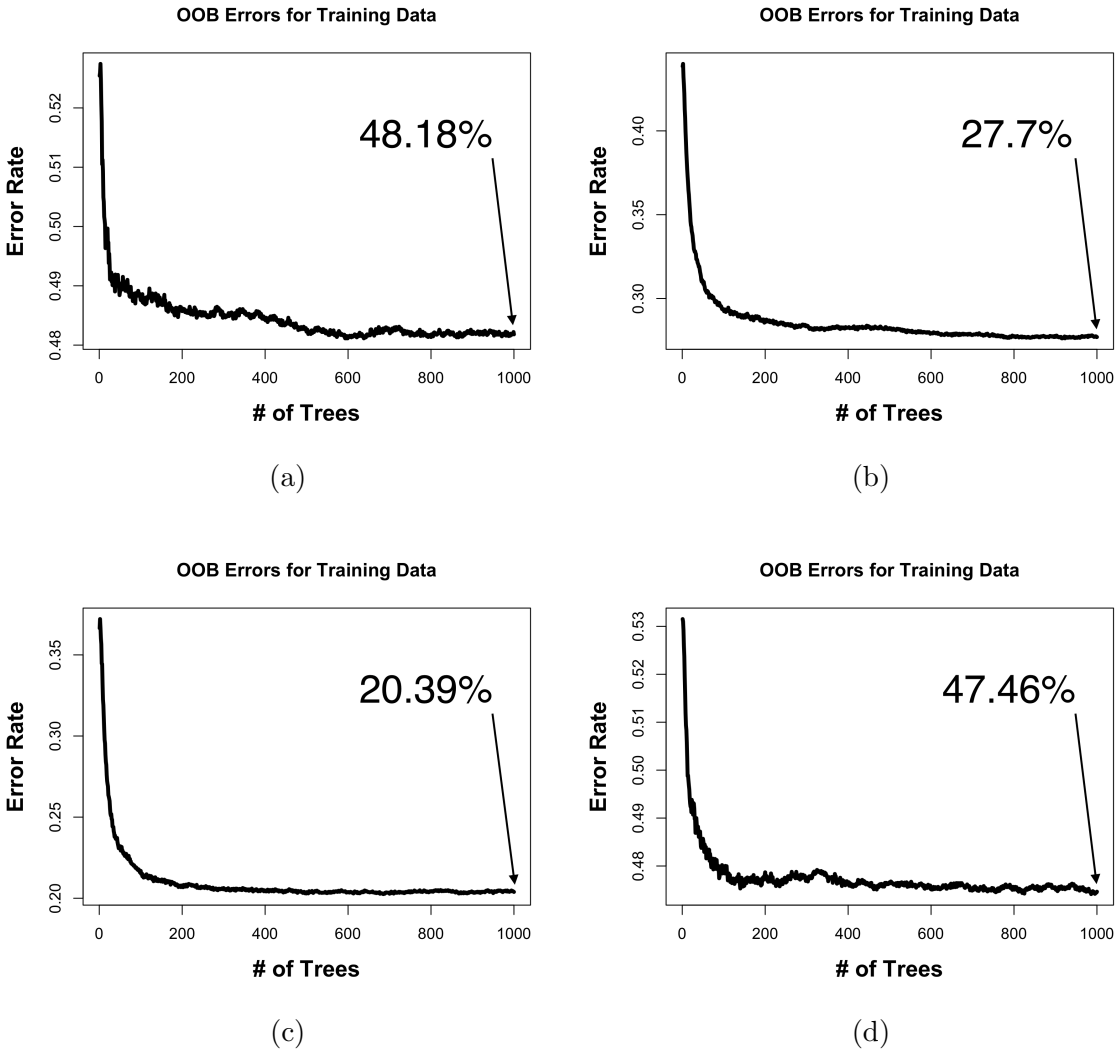



Figure 5.41. OOB errors as trees are added to the Random Forest model for forecasting 20 observation long regions of pre-transition growth, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from the 10 hour 2-PF MT model simulations.

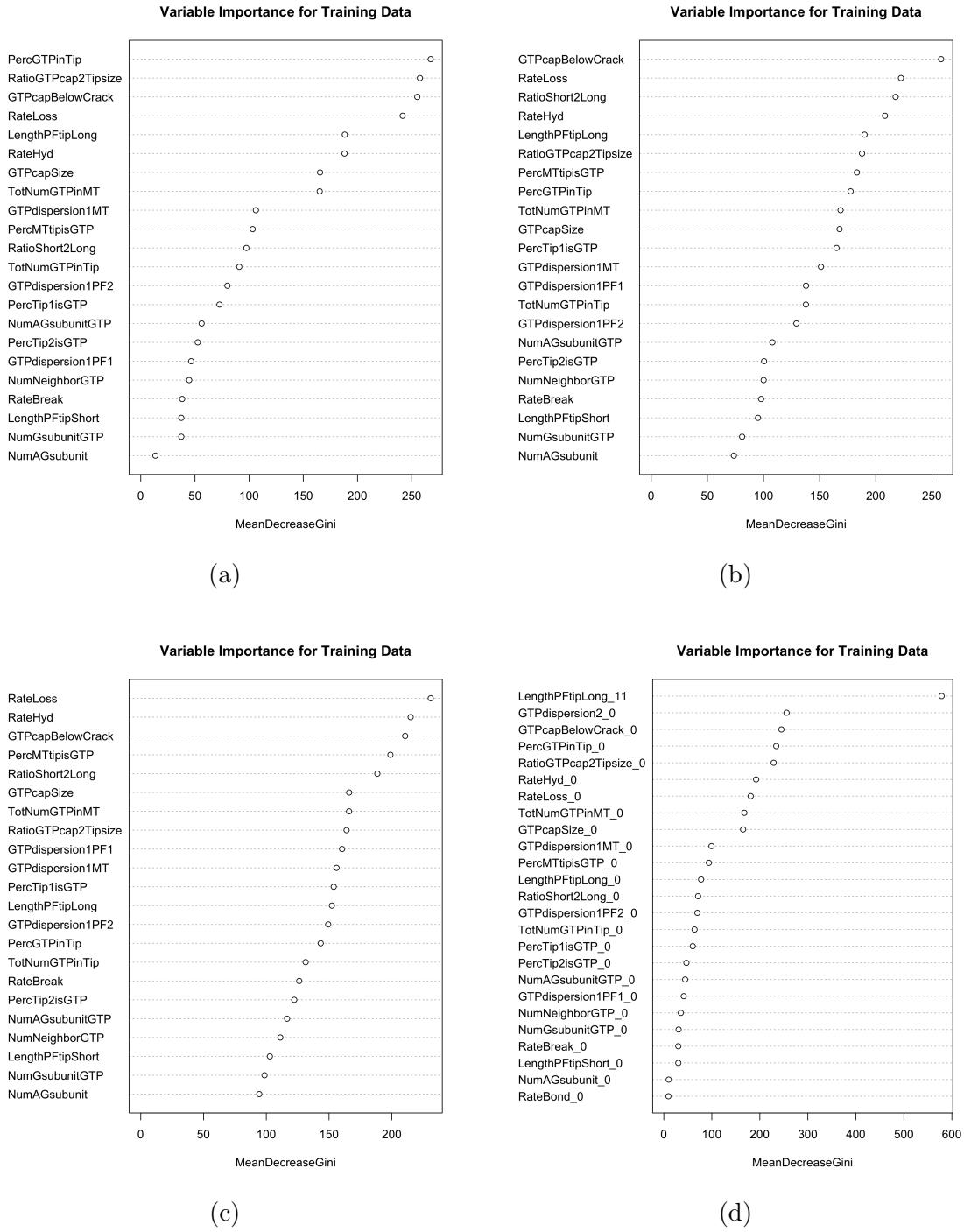



Figure 5.42. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 5 observation long regions before transitioning out of growth, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from 10 hour 2-PF MT model simulations.

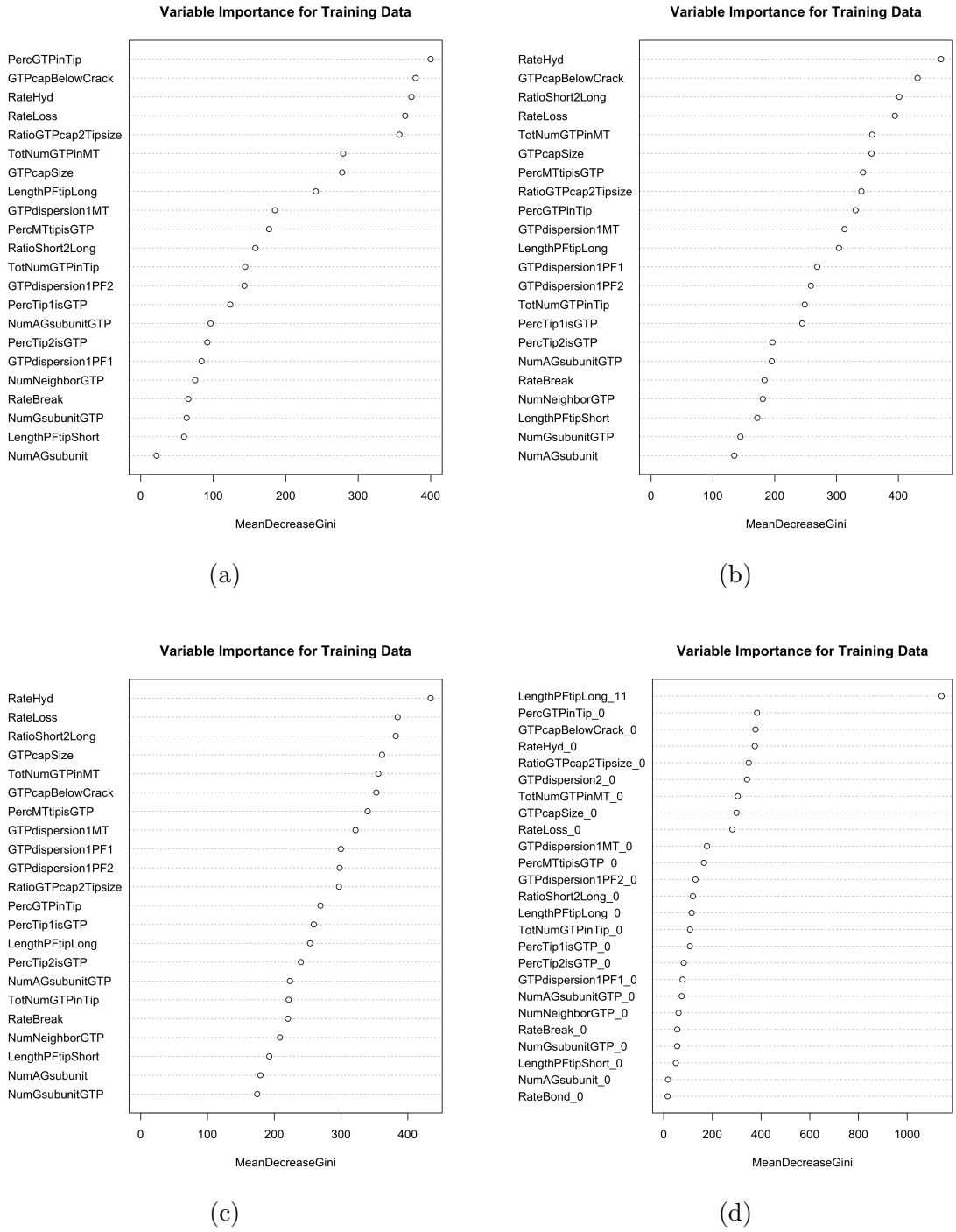



Figure 5.43. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 10 observation long regions before transitioning out of growth, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from 10 hour 2-PF MT model simulations.

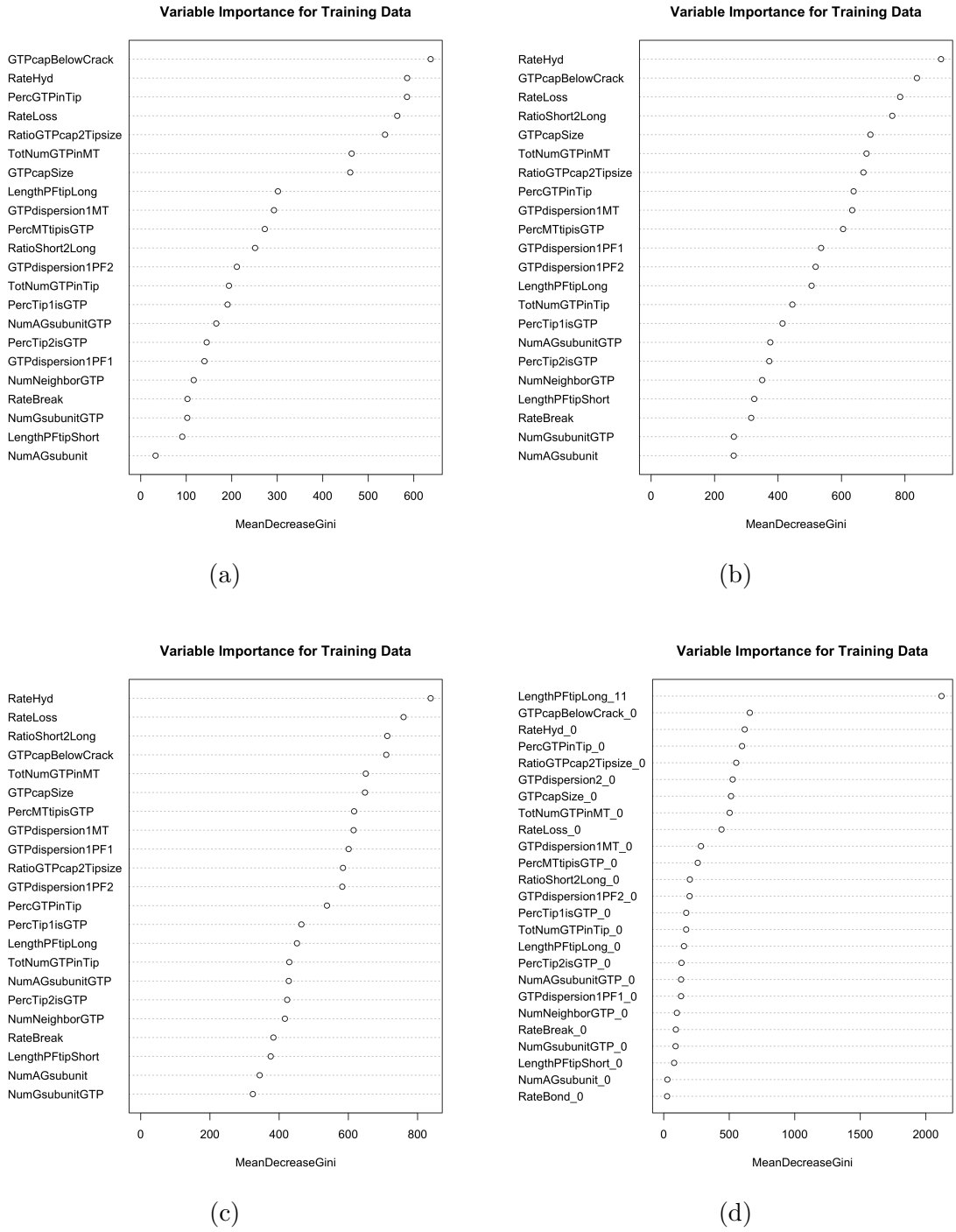



Figure 5.44. Variable importance via the mean decrease in the Gini index for each tip feature when forecasting 20 observation long regions before transitioning out of growth, trained using (a) raw tip structure data, (b) 11 observation trailing average, (c) 21 observation trailing average, and (d) multi-resolution data from 10 hour 2-PF MT model simulations.

# CHAPTER 6

# CONCLUSIONS AND DISCUSSION

A computational modeling approach was utilized in order to investigate how the MT tip structure contributed to DI phase changes observed in length history data. First, the previous version of the detailed 13-PF MT model was extended in order to simulate all the MT structural states possible resulting from a realistic sequence of molecular reaction events. The older detailed 13-PF MT model implemented an approximation that allowed multiple hydrolysis events to occur at a given time, which effectively skipped over MT structure states. This approximated implementation searched through the entire MT structure to find hydrolyzable GTP-bound subunits, and then sampled a different random number for each one to stochastically determine if that subunit underwent hydrolysis. Simultaneous hydrolysis reactions was utilized due to the high computational cost associated with finding hydrolyzable GTP-bound subunits.

The extended 13-PF MT model presented in this dissertation removes this approximation and executes the Gillespie algorithm, which treats hydrolysis the same as every other reaction event, only allows for one reaction to occur at a time, and therefore delivers a bio-chemically exact trajectory of MT structural states. To reduce computational costs, the extended model tracks the hydroyzable GTP subunits available in each PF at a given time, thus avoiding a search through the entire MT structure. This allows for a faster computation of the hydrolysis rate used in the Gillespie algorithm, and only requires two random numbers to be sampled: one to select a PF, and another to select which GTP-bound subunit within that PF that will

undergo hydrolysis. So, the extended 13-PF MT model is able to simulate the same DI behavior as the previous detailed 13-PF MT model, without an approximation requiring simultaneous hydrolysis reactions, and without additional computational cost.

Additionally, different features describing the MT tip region were calculated from the simulated MT structural states, and included as part of the simulation output in order to take advantage of the perspective offered by computational models, which are not currently available in the laboratory. However, the 13-PF MT model has complex tip structures with a large number of possible tip configurations, which make it difficult to conduct this study at this time. So, the 2-PF MT model was developed as a simplified version of the extended 13-PF MT model to help make studying the tip structure more tractable. The 2-PF MT structure uses only two PFs, and allows for a single sequence of lateral bonds between them, which is the minimum structure needed while still considering lateral bonds. Furthermore, the 2-PF MT structure does not have a seam, which removes the need for seam lateral bond forming and breaking rates in the model parameter list.

Despite this simplification, the 2-PF MT model used the relevant model parameter values similar to that of the extended 13-PF MT model to successfully simulate DI behavior. Moreover, the critical concentration of tubulin levels that deliver the unbounded growth regime was in a similar range for both the 13- and 2-PF MT models. So, a 10-hour long simulation of the 2-PF MT model using a tubulin concentration level of $12\mu M$ provided a statistically significant amount of length history data with DI behavior. In addition the MT lengths, over 4.6 million reaction events provided micro-level structural information, and from this 24 features of the tip region structure were calculated for each observation. This 4.6 million$\times$24 matrix was the micro-level MT tip structure data used for the analysis portions of this study.

A novel computational tool to identify, classify, and analyze DI phases was devel-

oped to address the high frequency and low amplitude fluctuations in length history data. This tool is applicable to data obtained by using the detailed leveled computational model simulations in this study, as well as recent experimental data sets that have been collected by using microscopy equipment with improved temporal and spatial resolutions. This DI phase classification tool is an important part in this study, since the ultimate goal is to establish the relationship between micro-level structures and macro-level DI phases and transitions. Previous methods for identifying DI phases cannot be applied for determining the exact moments when phase transitions occur. Additionally, the resulting approximations of the rates of change of the MT length are not good enough for our study, since periods with smaller rates of change were observed at finer time resolutions, and they were not accounted for as separate periods.

An unsupervised machine learning approach was used to develop an automated method to identify, classify, and analyze macro-level phases using length history data with DI behavior. This novel approach eliminated much of the inaccuracies and inconsistencies of older, non-automated methods that were tainted with human error. Application of the novel unsupervised method resulted in the discovery of periods with rates of change of the MT length that were smaller in magnitude than classically understood growth and shortening phases. These intermediate phase are called stutters. Applying this new method to both 13- and 2-PF MT model simulated data revealed that these stutters not only occur over time durations comparable to the ones for shortening phases, but they also are present during catastrophe events and provide transitions between growth and shortening. The bi-phase assumptions of previous methods that only considered growth and shortening phases made them overlook these stutters. This provides new insights to catastrophe events, which are now understood to occur less suddenly than previously thought, and possibly require some changes to the MT structure prior to beginning a rapid shortening

period. Additionally, stutters observed as part of catastrophes are analogous to some "slow-down" periods observed to occur before the shortening period starts as part of catastrophe events in recent experimental data [21]. However, these findings have only been part of a discussion without explicitly quantifying their rates or frequency of occurrence, since the older methods for extrapolating DI parameters overlooked a third possible phase. The new method for classifying and analyzing DI phases presented in this dissertation does not make any assumptions about the number of phases present in a length history data, and thus is appropriate to make accurate calculations describing DI behavior in high frequency length history data. Also, this method is applicable to any DI data set, sourced either from laboratories or from simulations. The conclusions on stutters presented here were limited to simulations from the 13- and 2-PF MT models. Experimental data is currently being collected so that the presented DI phase analysis method can be applied to *in vitro* data using pure tubulin experiments, and thus adding to the biological relevance of stutter phases results found in this dissertation.

The DI phases identified in the 10 hour long 2-PF MT model simulations provided macro-level DI phase classes for each of the tip structures in the 4.6 million micro-level observations in the simulated data. To test the relationship between the micro- and macro-level data, a supervised machine learning approach was used. First, the micro-level tip structure features were treated as predictor variables, and the macro-level phase classes were treated as response variables. The results indicated that the 2-PF MT tip structure features were indeed capable of predicting the corresponding DI phase during which they occurred, as displayed the confusion matrices in Table 5.1. However, these results were significantly improved when a moving average window of the tip feature data was used as the predictor variables, resulting in the misclassification rates being below 20%. It's worth noting that the shortening phase observations were far more successfully predicted ($\leq 5\%$). The reduced success of

the detection of growth and stutter phases indicated some overlap in the distribution of tip structure features, which adds to the difficulty of distinguishing a separation between the two phases. Nevertheless, those tip features involving the GTP-cap size estimates were particularly capable of predicting DI phases.

Previous studies indicated that the cracked tip region and the tubulin subunits near the bottom of the crack are important factors in determining DI phase transitions like catastrophes [47]. The tip features calculated for the data set used here was inspired by those results. The variable importance results for the tip-to-phase predictive models in Figure 5.26 showed that the estimated size of the GTP-cap plays an overwhelming role in determining DI phases when compared to the other tip structure features. More specifically, the mean decrease in Gini index values drop off after the top five most important tip structural features, which are listed in Section 5.4.1. Future work should consider portions of the MT tip structure that go farther below the cracked region so that more details of the GTP-rich portions of the MT structure is collected. The GTP-cap information used in this study were just estimates, and a better understanding can benefit from additional information. This is admittedly a difficult task, especially considering the lack of a clear boundary between the GTP-cap and the rest of the MT structure. Rather, information like the location for the lowest GTP-bound subunit, local GTP-bound subunit density, and general information of how the GTP-subunits are dispersed would be some ways of calculating this information from simulated data.

Finally, the transitions between DI phases were studied. The last few observations at the end of each phase period were labeled considering the future phase into which the transition was to occur. The observations that were not in this transition range had unchanged labels representing their corresponding DI phase. These phase transition labels were now used as the response variables for conducting predictive test for the three DI phases separately. The overall results showed an improvement in mis-

classification rates when using the moving average data as before. In some cases, an interesting improvement was observed when the transition range was reduced. This improvement could indicate that certain transitions between phases may be taking places more rapidly than others, which require fewer structural modifications to effectively alter the macro-level rates of change to the MT length. Additionally, overall results for transitions out of stutter phases were the most successfully predicted. The specific phase transitions that had the lowest misclassification rates were "Stutter-to-Growth" and "Growth-to-Shortening" classes. Detecting transitions successfully between shortening and the other two phases was expected, since the tip-to-phase predictive models indicated a good separation between shortening tip structures and the rest of the data. However, the best results coming from transitions between growth and stutters was quite unexpected due to the large overlap in the MT tip features for those two phases.

In order to improve the limited success presented in the results of the forecasting methods, a different approach can be taken for detecting upcoming phase transitions. The approach used here considered a set number of observations at the end of a phase as a pre-transition region, which may have inherited some dependencies between the response variable observations. Instead, a forecasting model regressing upon each observation individually to predict the future phase class of the very next observation in a sequence can serve to be more appropriate. These newer forecasting models can be constructed by including a varying number of observations for each sequence of predictor variables. This would part of a first step required to reduce the number of predictor variables by seeking the significance level of coefficients, and cutting off the sequence at that point. Additionally, confidence intervals can be used to describe the likelihood for predicting the phase classes for a number of subsequential states, and not just one observation in the future. After the necessary number of past observations has been determined, then variable reduction methods can be used to

determine which of the MT tip structure features are most successful in dictating the future DI phase. Similar to the conclusions made here, these important tip features will provide insight to the mechanisms involved with different types of DI phase transitions.

In the future, numerical experiments can be conducted to verify these phase transition results. By identifying specific structures associate with individual phase transitions and using them as initial conditions, multiple simulations can be run to calculate the likelihood of phase transition that are anticipated for each MT tip structure. Also, phase transition studies can be expanded to include more details in the forecasting models. Instead of using the trailing moving average data, all of the tip features from every past observations can be combined to create the predictor variables, and the DI phase in the very next time step can be the response variable. This would create a truer forecasting model that would test the prediction of future phase classes by relying on past MT tip structure features. However, the difficulty in this task would be to determine an appropriate time window for including past tip data. The insight from this study would suggest that the past data window would be relatively short, but using decaying weighting coefficients may also serve as a beneficial compromise in using observations from past time steps.

Other future work could include repeating this study for MT tip features of 13-PF MTs. In fact, many of the tip feature formulations used for the 2-PF MT case were treated with care and knowledge so that they may be used to characterize 13-PF MT tip structures as well. Some tip features only apply to individual PFs, some apply to the crack region between PFs, and some are for the entire MT tip. To avoid complications in cases where the lateral bond height is asymmetrical around a given PF tip, structural features may be determined for neighboring pairs of PFs separately, hence extending the 2-PF concepts directly. The 13-PF MT tip structures would certainly yield a larger number of variables, however predictive modeling methodology

can be repeated very easily, since the Random Forest approach is well suited for handling large data.

Presence of the stutter phases have certainly exposed some interesting aspects of the DI behavior overlooked in previous studies, such as rates of MT length change being inconsistent with the assumption that only growth and shortening phases exist, that growth and shortening occur with near uniform rates, and that catastrophe events are sudden transitions. The transitional role of stutters during catastrophe demonstrates the impact of the structural changes on the MT without altering it's length. Once a MT is in a stutter phase, the ability to detect an oncoming macro-level change is far easier when observing the changes taking place in the MT tip structure, mostly related to the GTP-cap size, but also to the dispersion of GTP-bound subunits in the cracked tip region. As laboratory conditions improve, and finer resolution images can be captured at specific instances during MT dynamics, future experiments can be conducted to verify the structural characteristics that correspond to different DI phases.

Also, the results on stutter phases can relate to future work dealing with MT binding proteins that can alter MT dynamic behavior. It could be highly beneficial to engineer binding proteins to detect and attach to the structures associated with stutter phases. After all, MT binding proteins are known for their ability to encourage stabilizing bond formation to promote growth, or they can assist in destabilizing bond breakage to promote shortening instead. After attaching to stutter-type MT structures, a binding protein has an increased chance of success in altering MT behavior as it desires, because the results of this study indicate a higher success rate of predicting transitions out of a stutter phase. Hopefully, this approach could inspire treatments for combating diseases that affect MT dynamics, which can be used to regain a healthy regime of DI behavior.

# APPENDIX A

## GLOSSARY OF TERMS

- AG-subunit: the subunit positioned immediately above the G-subunit
- BAG = Bootstrap aggregate: a method for regaining the original training data size by averaging over the variables of the observations that remain after removing a subset to create a testing data set
- Catastrophe: the moment where a microtubule transitions from a growth to a shortening period
- Crack: the space defined by the laterally unbonded section between protofilaments
- Crack-depth: the size of a crack defined by the shorter of the protofilament tips that create the boundary of the crack
- Cracked MT-tip: the combination of protofilament tips
- DI = Dynamic instability: the characteristic behavior of microtubules, where the microtubule length undergoes transitions between sustained periods of growth and rapid shortening
- Gap statistic: a measurement of inter-cluster dispersion used to determine the ideal number of clusters in conjunction with the $k$-means clustering method
- Gate: the interface between the sequence of lateral bonds and the crack
- Gated MT-tip: the cracked MT-tip combined with the G-subunits
- GDP-bound = Guanosine-diphosphate bound: the low energy nucleotide bound state of a tubulin dimer subunit that promotes bond breaking
- GTP-bound = Guanosine-triphosphate bound: the energy carrying nucleotide bound state of a tubulin dimer subunit that promotes stable bonds

- GTP-cap: the region near the tip of a microtubule that has a high concentration of GTP-bound subunits

- G-subunit: the subunit within a protofilament connected by the top-most lateral bond located at the gate

- $k$-means: an unsupervised machine learning approach used clustering data

- Lateral bond height: the length of the consecutive sequence of lateral bonds between protofilaments

- MT = Microtubule: a helical tubelike biopolymer made of protofilaments held together with lateral bonds

- PF = Protofilament: a sequence of tubulin dimer subunits held together by longitudinal bonds

- PF-tip = Protofilament tip: the sequence of laterally unbonded subunits at the tip of a protofilament, adjacent to the crack

- OOB-errors = Out-of-BAG-errors: misclassification errors used to assess a Random Forest model prediction ability on the testing data left out of the BAG training data

- Random Forest: a supervised machine learning method used for classification as part of the predictive modeling

- Rescue: the rare moment where a microtubule transitions from a shortening to a growth period

- Seam: the sequence of lateral bonds between the first and last protofilaments, where the 1.5 dimer shift creates the helical pattern of a microtubule

- Stutter: an intermediate phase of dynamic instability, during which the overall microtubule length changes are small and the rates of length change are smaller in magnitude than those seen in classically observed growth and shortening phases

- Tip feature: a property of a microtubule tip structure used as a predictor variable when constructing predictive models

# BIBLIOGRAPHY


1. G. M. Alushin, G. C. Lander, E. H. Kellogg, R. Zhang, D. Baker, and E. Nogales. High-resolution microtubule structures reveal the structural transitions in $\alpha\beta$-tubulin upon gtp hydrolysis. *Cell*, 157(5):1117–1129, 2014.

2. T. Antal, P. Krapivsky, S. Redner, M. Mailman, and B. Chakraborty. Dynamics of an idealized model of microtubule growth and catastrophe. *Physical Review E*, 76(4):041907, 2007.

3. A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage $\lambda$–infected Escherichia coli cells. *Genetics*, 149:1633–1648, August 1998.

4. P. M. Bayley, M. J. Schilstra, and S. R. Martin. Microtubule dynamic instability: numerical simulation of microtubule transition properties using a lateral cap model. *Journal of cell science*, 95(1):33–48, 1990.

5. H. Bowne-Anderson, M. Zanic, M. Kauer, and J. Howard. Microtubule dynamic instability: a new model with coupled gtp hydrolysis and multistep catastrophe. *Bioessays*, 35(5):452–461, 2013.

6. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

7. J. H. Brown, C. Cohen, and D. A. Parry. Heptad breaks in $\alpha$-helical coiled coils: stutters and stammers. *Proteins: Structure, Function, and Bioinformatics*, 26 (2):134–145, 1996.

8. L. Brun, B. Rupp, J. J. Ward, and F. Nédélec. A theory of microtubule catastrophes and their regulation. *Proceedings of the National Academy of Sciences*, 106(50):21173–21178, 2009.

9. B. P. Brylawski and M. Caplow. Rate for nucleotide release from tubulin. *Journal of Biological Chemistry*, 258(2):760–763, 1983.

10. M. Caplow and J. Shanks. Evidence that a single monolayer tubulin-gtp cap is both necessary and sufficient to stabilize microtubules. *Molecular biology of the cell*, 7(4):663–675, 1996.

11. Y.-d. Chen and T. L. Hill. Use of monte carlo calculations in the study of microtubule subunit kinetics. *Proceedings of the National Academy of Sciences*, 80(24):7520–7523, 1983.

12. Y.-D. Chen and T. L. Hill. Monte carlo study of the gtp cap in a five-start helix model of a microtubule. *Proceedings of the National Academy of Sciences*, 82(4): 1131–1135, 1985.

13. C. Cohen and D. A. Parry. $\alpha$-helical coiled coils and bundles: how to design an $\alpha$-helical protein. *Proteins: Structure, Function, and Bioinformatics*, 7(1):1–15, 1990.

14. C. E. Coombes, A. Yamamoto, M. R. Kenzie, D. J. Odde, and M. K. Gardner. Evolving tip structures can explain age-dependent microtubule catastrophe. *Current Biology*, 23(14):1342–1348, 2013.

15. A. O. Demchouk, M. K. Gardner, and D. J. Odde. Microtubule tip tracking and tip structures at the nanometer scale using digital fluorescence microscopy. *Cellular and molecular bioengineering*, 4(2):192–204, 2011.

16. S. Denis Chr6tien. Structure of growing microtubule ends: two-dimensional sheets close into tubes at variable rates. *The Journal of cell biology*, 129(5): 1311–1328, 1995.

17. A. Desai and T. J. Mitchison. Microtubule polymerization dynamics. *Annual Review of Cell and Developmental Biology*, 13(1):83–117, 1997.

18. A. Dimitrov, M. Quesnoit, S. Moutel, I. Cantaloube, C. Poüs, and F. Perez. Detection of gtp-tubulin conformation in vivo reveals a role for gtp remnants in microtubule rescues. *Science*, 322(5906):1353–1356, 2008.

19. M. Dobrzynski, J. V. Rodriguez, J. A. Kaandorp, and J. G. Blom. Computational methods for diffusion-influenced biochemical reactions. *Bioinformatics*, 23(15): 1969–1977, May 2007.

20. D. N. Drechsel and M. W. Kirschner. The minimum gtp cap required to stabilize microtubules. *Current Biology*, 4(12):1053–1061, 1994.

21. C. Duellberg, N. I. Cade, and T. Surrey. Microtubule aging probed by microfluidics-assisted tubulin washout. *Molecular Biology of the Cell*, 27(22): 3563–3573, 2016.

22. J. Dushoff, J. B. Plotkin, S. A. Levin, and D. J. D. Earn. Dynamical resonance can account for seasonality of influenza epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16915–16916, 2004. doi: 10.1073/pnas.0407293101.

23. J. Feng. Microtubule: a common target for parkin and parkinson's disease toxins. *The Neuroscientist*, 12(6):469–476, 2006.

24. H. Flyvbjerg, T. E. Holy, and S. Leibler. Stochastic dynamics of microtubules: a model for caps and catastrophes. *Physical review letters*, 73(17):2372, 1994.

25. H. Flyvbjerg, T. E. Holy, and S. Leibler. Microtubule dynamics: caps, catastrophes, and coupled hydrolysis. *Physical Review E*, 54(5):5538, 1996.

26. D. B. Forger and C. S. Peskin. Stochastic simulation of the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2):321–324, 2005. doi: 10.1073/pnas.0408465102.

27. M. K. Gardner, B. D. Charlebois, I. M. Jánosi, J. Howard, A. J. Hunt, and D. J. Odde. Rapid microtubule self-assembly kinetics. *Cell*, 159(1):215, 2014.

28. E. C. Garner, C. S. Campbell, and R. D. Mullins. Dynamic instability in a dna-segregating prokaryotic actin homolog. *Science*, 306(5698):1021–1025, 2004.

29. D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4): 403–434, 1976.

30. D. T. Gillespie et al. Exact stochastic simulation of coupled chemical reactions. *J. phys. Chem*, 81(25):2340–2361, 1977.

31. H. V. Goodson and E. M. Jonasson. Microtubules and microtubule-associated proteins. *Cold Spring Harbor Perspectives in Biology*, 2016. doi: 10.1101/cshperspect.a022608.

32. I. V. Gregoretti, G. Margolin, M. S. Alber, and H. V. Goodson. Insights into cytoskeletal behavior from computational modeling of dynamic microtubules in a cell-like environment. *J Cell Sci*, 119(22):4781–4788, 2006.

33. I. Grundke-Iqbal, K. Iqbal, Y.-C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology. *Proceedings of the National Academy of Sciences*, 83(13):4913–4917, 1986.

34. K. K. Gupta, C. Li, A. Duan, E. O. Alberico, O. V. Kim, M. S. Alber, and H. V. Goodson. Mechanism for the catastrophe-promoting activity of the microtubule destabilizer op18/stathmin. *Proceedings of the National Academy of Sciences*, 110(51):20449–20454, 2013.

35. E. Hamel, A. Del Campo, M. Lowe, and C. Lin. Interactions of taxol, microtubule-associated proteins, and guanine nucleotides in tubulin polymerization. *Journal of Biological Chemistry*, 256(22):11887–11894, 1981.

36. J. A. Hartigan and J. Hartigan. *Clustering algorithms*, volume 209. Wiley New York, 1975.

37. J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28 (1):100–108, 1979.

38. T. L. Hill. Introductory analysis of the gtp-cap phase-change kinetics at the end of a microtubule. *Proceedings of the National Academy of Sciences*, 81(21): 6728–6732, 1984.

39. P. Hinow, V. Rezania, and J. A. Tuszyński. Continuous model for microtubule dynamics with catastrophe, rescue, and nucleation processes. *Physical Review E*, 80(3):031904, 2009.

40. A. A. Hyman, S. Salser, D. Drechsel, N. Unwin, and T. J. Mitchison. Role of gtp hydrolysis in microtubule dynamics: information from a slowly hydrolyzable analogue, gmpcpp. *Molecular Biology of the Cell*, 3(10):1155–1167, 1992.

41. I. Jain, M. M. Inamdar, and R. Padinhateeri. Statistical mechanics provides novel insights into microtubule stability and mechanism of shrinkage. *PLoS Comput Biol*, 11(2):e1004099, 2015.

42. Z. John Lu. The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):693–694, 2010.

43. E. M. Jonasson, A. J. Mauro, E. C. Norby, C. Li, S. M. Mahserejian, J. P. Scripture, I. V. Gregoretti, M. S. Alber, and H. V. Goodson. *In preparation*, 2017.

44. R. A. Keates and F. R. Hallett. Dynamic instability of sheared microtubules observed by quasi-elastic light scattering. *Science*, 241(4873):1642–1646, 1988.

45. M. Kirschner and T. Mitchison. Beyond self-assembly: from microtubules to morphogenesis. *Cell*, 45(3):329–342, 1986.

46. C. Li. Computational modeling of microtubule dynamic instability. *ProQuest Dissertations and Theses*, page 173, 2014. URL `http://proxy.library.nd.edu/login?url=http://search.proquest.com.proxy.library.nd.edu/docview/1547351218?accountid=12874`.

47. C. Li, J. Li, H. V. Goodson, and M. S. Alber. Microtubule dynamic instability: the role of cracks between protofilaments. *Soft matter*, 10(12):2069–2080, 2014.

48. A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

49. S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

50. J. MacQueen et al. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14):281–297, 1967.

51. M. Mahrooghy, S. Yarahmadian, V. Menon, V. Rezania, and J. A. Tuszynski. The use of compressive sensing and peak detection in the reconstruction of microtubules length time series in the process of dynamic instability. *Computers in biology and medicine*, 65:25–33, 2015.

52. E.-M. Mandelkow and E. Mandelkow. Microtubule oscillations. *Cell motility and the cytoskeleton*, 22(4):235–244, 1992.

53. E.-M. Mandelkow, E. Mandelkow, and R. A. Milligan. Microtubule dynamics and microtubule caps: a time-resolved cryo-electron microscopy study. *The Journal of cell biology*, 114(5):977–991, 1991.

54. G. Margolin, I. V. Gregoretti, H. V. Goodson, and M. S. Alber. Analysis of a mesoscopic stochastic model of microtubule dynamic instability. *Physical Review E*, 74(4):041920, 2006.

55. G. Margolin, H. V. Goodson, and M. S. Alber. Mean-field study of the role of lateral cracks in microtubule dynamics. *Physical Review E*, 83(4):041905, 2011.

56. G. Margolin, I. V. Gregoretti, T. M. Cickovski, C. Li, W. Shi, M. S. Alber, and H. V. Goodson. The mechanisms of microtubule catastrophe and rescue: implications from analysis of a dimer-scale computational model. *Molecular Biology of the Cell*, 23(4):642–656, 2012.

57. A. Markov. Theory of algorithms [translated by jacques j. schorr-kon and pst staff] imprint moscow, academy of sciences of the ussr, 1954 [jerusalem, israel program for scientific translations, 1961; available from office of technical services, united states department of commerce] added tp in russian translation of works of the mathematical institute, academy of sciences of the ussr, v. 42. *Original title: Teoriya algorifmov.[QA248. M2943 Dartmouth College library. US Dept. of Commerce, Office of Technical Services, number OTS 60-51085]*, 1954.

58. S. Martin, M. Schilstra, and P. Bayley. Dynamic instability of microtubules: Monte carlo simulation and application to different types of microtubule lattice. *Biophysical journal*, 65(2):578–596, 1993.

59. I. Mazilu, G. Zamora, and J. Gonzalez. A stochastic model for microtubule length dynamics. *Physica A: Statistical Mechanics and its Applications*, 389(3): 419–427, 2010.

60. J. McIntosh, E. O'Toole, J. Austin, R. Ding, E. Ulyanov, F. Ataullakhanov, and N. Gudimchuk. Microtubules grow in vivo through pathways that use bent tubulin protofilaments. *Molecular Biology of the Cell*, 27, 2016.

61. L. Michaelis and M. L. Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49 (333-369):352, 1913.

62. T. Mitchison, M. Kirschner, et al. Dynamic instability of microtubule growth. *Nature*, 312(5991):237–242, 1984.

63. M. I. Molodtsov, E. A. Ermakova, E. E. Shnol, E. L. Grishchuk, J. R. McIntosh, and F. I. Ataullakhanov. A molecular-mechanical model of the microtubule. *Biophysical Journal*, 88(5):3167–3179, 2005.

64. E. T. O'Brien, W. A. Voter, and H. P. Erickson. Gtp hydrolysis during microtubule assembly. *Biochemistry*, 26(13):4148–4156, 1987.

65. A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

66. P. Ranjith, D. Lacoste, K. Mallick, and J.-F. Joanny. Nonequilibrium self-assembly of a filament coupled to atp/gtp hydrolysis. *Biophysical journal*, 96 (6):2146–2159, 2009.

67. P. Ranjith, K. Mallick, J.-F. Joanny, and D. Lacoste. Role of atp-hydrolysis in the dynamics of a single actin filament. *Biophysical journal*, 98(8):1418–1427, 2010.

68. P. Ranjith, A. B. Kolomeisky, and D. Lacoste. Random hydrolysis controls the dynamic instability of microtubules. *Biophysical journal*, 102(6):1274–1283, 2012.

69. L. M. Rice, E. A. Montabana, and D. A. Agard. The lattice as allosteric effector: structural studies of $\alpha\beta$-and $\gamma$-tubulin clarify the role of gtp in microtubule assembly. *Proceedings of the National Academy of Sciences*, 105(14):5378–5383, 2008.

70. J. Rickman, C. Duellberg, N. I. Cade, L. D. Griffin, and T. Surrey. Steady-state eb cap size fluctuations are determined by stochastic microtubule growth and maturation. *Proceedings of the National Academy of Sciences*, 114(13):3427–3432, 2017.

71. D. Sept, N. A. Baker, and J. A. McCammon. The physical basis of microtubule structure and stability. *Protein Science*, 12(10):2257–2261, 2003.

72. S. L. Shaw, R. Kamyar, and D. W. Ehrhardt. Sustained microtubule treadmilling in arabidopsis cortical arrays. *Science*, 300(5626):1715–1718, 2003.

73. E. B. Stukalin and A. B. Kolomeisky. Atp hydrolysis stimulates large length fluctuations in single actin filaments. *Biophysical journal*, 90(8):2673–2685, 2006.

74. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

75. P. Tran, P. Joshi, and E. D. Salmon. How tubulin subunits are lost from the shortening ends of microtubules. *Journal of structural biology*, 118(2):107–118, 1997.

76. V. VanBuren, D. J. Odde, and L. Cassimeris. Estimates of lateral and longitudinal bond energies within the microtubule lattice. *Proceedings of the National Academy of Sciences*, 99(9):6035–6040, 2002.

77. V. VanBuren, L. Cassimeris, and D. J. Odde. Mechanochemical model of microtubule structure and self-assembly kinetics. *Biophysical Journal*, 89(5):2911–2926, 2005.

78. A. Vandecandelaere, M. Brune, M. R. Webb, S. R. Martin, and P. M. Bayley. Phosphate release during microtubule assembly: what stabilizes growing microtubules? *Biochemistry*, 38(25):8179–8188, 1999.

79. W. A. Voter, E. T. O'Brien, and H. P. Erickson. Dilution-induced disassembly of microtubules: Relation to dynamic instability and the gtp cap. *Cytoskeleton*, 18(1):55–62, 1991.

80. R. Walker, S. Inoué, and E. Salmon. Asymmetric behavior of severed microtubule ends after ultraviolet-microbeam irradiation of individual microtubules in vitro. *J. Cell Biol*, 108(3):931–937, 1989.

81. R. Walker, N. Pryer, and E. D. Salmon. Dilution of individual microtubules observed in real time in vitro: evidence that cap size is small and independent of elongation rate. *The Journal of cell biology*, 114(1):73–81, 1991.

82. Z. Wu, E. Nogales, and J. Xing. Comparative studies of microtubule mechanics with two competing models suggest functional roles of alternative tubulin lateral interactions. *Biophysical journal*, 102(12):2687–2696, 2012.

83. S. Yarahmadian, V. Menon, M. Mahrooghy, and V. A. Rezania. Wavelet-based compression and peak detection method for the experimentally estimation of microtubules dynamic instability parameters identified in three states. *arXiv preprint arXiv:1510.07290*, 2015.

84. S. Yarahmadian, V. Menon, and V. Rezania. On using compressed sensing and peak detection method for the dynamic instability parameters estimation for microtubules modeled in three states. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 417–420. IEEE, 2015.

85. P. Zakharov, N. Gudimchuk, V. Voevodin, A. Tikhonravov, F. I. Ataullakhanov, and E. L. Grishchuk. Molecular and mechanical causes of microtubule catastrophe and aging. *Biophysical journal*, 109(12):2574–2591, 2015.

*This document was prepared & typeset with* pdfLaTeX*, and formatted with* NDdiss2$_\varepsilon$ *classfile (v3.2013[2013/04/16]) provided by Sameer Vijay and updated by Megan Patnott.*